Social Insect Inspired Algorithm to Detect and Track Topics in Dynamic Documents

Štefan Sabo*
Institute of Informatics, Information Systems and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia
stefan.sabo@stuba.sk

Abstract

In our work we present a novel approach to identification and tracking of news stories on the web. We utilize a set of social insect inspired agents to acquire news articles and subsequently analyse relationships between articles based on story words. Story words represent our concept for modelling terms relevant to news stories as a whole, instead of using keywords relevant only to a single document. We leverage behavioural patterns inspired by honey bees when foraging for food in order to design a self adjusting and self prioritizing mechanism that allows for dynamic response to changing news story landscape. Due to the independent nature of agents, the resulting system offers flexibility, scalability and distributivity while maintaining high level of cooperation during identification and tracking of currently unfolding news stories.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models, search process*; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*multiagent systems*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

beehive metaphor, kernel methods, keyword extraction, multi-agent systems, online learning, topic detection and tracking, web crawling

1. Introduction

When visiting a web based news portal, human visitor finds himself browsing a wide range of articles that cover the current news stories, events and developments. Based on their content and also context the user can piece together a picture of what is currently going on and make decisions upon whether he wants to pursue further reading, or make some research of his own. Our goal is to be able to provide support this process by analysing the available web based articles and providing an overview of current events that are unfolding around us.

This synthesis of larger picture is however not a simple process from the information point of view, due to sheer complexity of additional contextual information and knowledge that a human user is able to leverage when analysing content of a given news article. In our work we take a different approach and instead of trying to analyse implications of information contained within individual news articles, we propose a social insect inspired approach that utilizes a set of agents to determine relationships between articles using syntactic analysis. Based on spacial distribution of these relationships within the article space, we are then able to establish which articles comprise distinct news stories in a dynamic and flexible fashion.

2. Article Acquisition and Analysis

Articles that we analyse in order to determine current news stories are extracted from the web. Therefore the first step is to acquire and process raw article documents. When designing our article acquisition mechanism, we consider following aspects of the news domain:

Dynamics. One of the most important aspects of the news domain is the dynamics of the presented data. News stories are added on a daily, even hourly basis as new events unfold and it is imperative that these changes are accounted for. Therefore current articles need to be updated to reflect the most recent changes and new articles need to be retrieved as they become available.

Lack of structure. A news article generally represents an unstructured textual information. Certain parts of an article, such as its title or individual paragraphs may be identified through parsing of the document structure. However the topic of a news article, key figures, characters and events described within the article are not readily available and need to be identified through deeper analysis of plain text.

*Recommended by thesis supervisor: Prof. Pavol Návrat
To be defended at Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.

Succinct language. A web based source of news articles needs to convey information about the content of an article succinctly and unambiguously. This way a potential reader may quickly decide whether to visit the given article or not. Due to the limited amount of space available in the title of an article, the information about the article content needs to be compressed into few words, at most few short sentences. Therefore article titles will often contain named entities or unique terms that identify the news story being covered unambiguously enough so that it may be easily recognized by an average reader.

2.1 Crawling Policies and Techniques

In order to acquire articles from the web we utilize a set of independent agents. Each agent acts as an advanced web crawler and is autonomously capable of traversing the web and acquiring articles, which are further processed and evaluated. In order to facilitate the interaction of agents with the web environment, each agent adopts the following web crawling policies and techniques.

Revisiting policy. Revisiting is a crucial policy identified by Castillo [4] as one of main cornerstones of effective crawling. Revisiting policy determines how long an agent needs to wait before revisiting an article that had already been visited in the past. Our revisiting policy sets the minimum delay between successive visits to 30 minutes, as we generally do not assume that an article will be updated at a more frequent rate.

Parsing. Parsing of a retrieved document is done through automated parsing tools and aims to extract article content along with article title, paragraphs, hypertext links and optional time and location stamps.

Courtesy delays. Short delays between requests are implemented in order to alleviate load of multiple successive requests on a single given web host. Our agents are set to observe 5 second minimum courtesy delays. In addition to saving resources this also helps to control the pace of article acquisition during periods of high latency.

2.2 Story Word Extraction

During the article acquisition process, the articles are analysed by agents in order to identify relationships between articles that share topical similarities. For this purpose we have proposed a concept of *story words* which represent specific terms linked to news stories. Story words may be either existing terms that are linked to a particular story such as *Watergate*, or even completely new terms such as more recent *Brexit.* The only requirement is that each story word represents a given news story at a given time.

Using the concept of story words we are able to break down news stories into specific aspects which we can track using our set of agents. Each agent is assigned a specific *story word candidate* and proceeds to evaluate the content of articles in order to determine the popularity and relevance of its current story word candidate. The aim of this process is to determine the most prominent story words and map out relationships between individual articles based on these story words. However, the dimensionality of the article—story word space is generally too high to evaluate all possible combinations of articles and story word candidates, therefore we utilize a bio inspired mechanism to focus the evaluation effort onto the most promising story word candidates and articles.

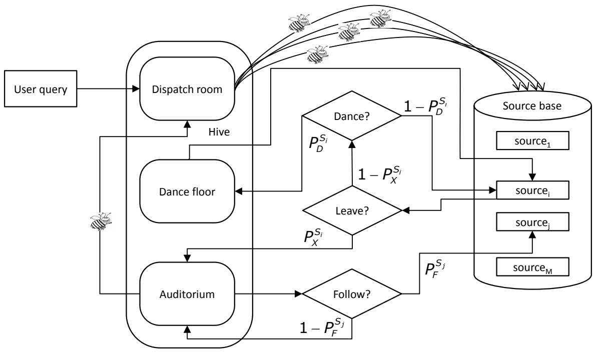

Figure 1: Decision making process of an agent as described in Beehive Metaphor [6].

3. Honey Bee Inspired Mechanism of Agent Coordination

In order to provide a coordination mechanism that would allow the agents to prioritize combinations of articles and story words we look for inspiration in nature. A similar problem of prioritizing sources in a decentralized environment is observed by honey bees (*apis mellifera*) when deciding on which food sources to visit while foraging for food. Honey bees have therefore developed a communication mechanism that allows them to share information about the most suitable food sources through their typical *waggle dance.*

Waggle dance is a specific movement pattern of a bee that encodes information about the direction of and distance to a food source. If a bee finds a suitable source of food it may choose to promote it by sharing information through engaging in a waggle dance. If a bee returns from a foraging trip and has not found a suitable source, it may adopt a different source through observing of dancing bees. Thus good sources of food are promoted and foraged from while unsuitable sources are abandoned.

3.1 Allocation of Agents

The idea of modelling agent coordination mechanism after waggle dance based coordination of honey bees was first utilized in Artificial Beehive Colony (ABC) model proposed by Karaboga [5] for numeric optimization. This model was further enhanced by Návrat [6] by adjusting it for web based document retrieval. This enhanced version of ABC model is known as Beehive Metaphor and it is also the model upon which our work is based. The general overview of Beehive Metaphor is given in Figure 1.

Based on the quality of its current source, each agent is able to perform foraging, dancing or observing task at a time. Unlike the original Beehive Metaphor model, our agents do not carry document, but rather each agent is assigned a *story word candidate* to evaluate. A story word candidate is a term which potentially represents a certain news story. The task of an agent is to determine, whether a story word candidate represents a viable story word.

In order to determine viability of a story word candidate, agents visit multiple articles and evaluate relevance of the currently carried story word to the visited articles. Depending on the relevance of the current story word each agent may select either to adopt foraging, dancing or observing task in accordance to the Beehive Metaphor model. Foraging is the basic task in which agent evaluates its current story word candidate. During the dancing an agent propagates its source so that other agents may adopt it. Observing task is adopted if an agent deems its current story word candidate as unsuitable and proceeds to select a new candidate from the pool of candidates currently being propagated by dancing agents.

3.2 Identification of Relationships

The main goal of article analysis is to determine which story words are relevant to the currently ongoing news stories. In order for a story word to be considered suitable, it needs not only to be relevant to a single article, but to connect multiple articles forming a given story line. Therefore our mechanism of story word identification is aimed to find relationships between articles based on the evaluated story words.

If an agent discovers that its current story word is relevant to multiple articles, it will evaluate the given story word positively and may establish a relationship between the relevant articles based on its current story word. By recording relationships between articles we are able to map out the explored parts of the article space and determine the closeness of given articles, which can be in turn used to determine the shape of currently ongoing news stories.

3.3 Advantages of Social Insect Inspired Coordination

Utilizing a set of agents inspired by honey bee inspired agents in order to determine the story words most relevant to the currently ongoing stories is a novel approach that has few specific advantages over traditional topic detection and tracking methods.

Dynamic story tracking capability. Ability to dynamically track news stories represents the most important characteristic of our news story tracking system, as it allows us to track news stories in real time as they develop. The key feature of our system that enables us to track stories in real time is the social insect inspired inspired behaviour of the agent swarm. Instead of treating the web as a static repository of articles, we are able to maintain a persistent presence on the web and acquire articles as they emerge. Although it is not feasible to track changes in every single article, the self prioritizing capability of our agent swarm allows us to focus the exploration effort around the most relevant articles where the actual development takes place.

Iterative article evaluation. Iterative evaluation of articles is closely related to the dynamic story tracking capability of our approach. It refers to the ability of our approach to process new articles iteratively as they emerge. When extracting a set of story words relevant to the set of articles there are two general possibilities how to do it. Batch algorithms such as Probabilistic latent semantic analysis [3] or Latent dirichlet allocation [2] treat the set of articles as closed, with classification always analysing the whole set in each run. This process is costly therefore the rate at which new articles may be introduced and evaluated is limited. Through decomposition of individual news stories into a set of story words that may be tracked independently and individually for each article, our approach offers iterative processing capability. Therefore we are able to process each article as it emerges and incorporate it into the whole news story landscape without having to reassess the whole article set.

No learning required. With our approach no learning or supervision is needed in order to classify articles according to news stories they cover. The reason lies in the way how news stories are decomposed into sets of related story words. If we had represented news stories as latent distributions over terms from a vocabulary, we would need to provide a set of training examples in order to determine the structure of each news story. However our representation of news stories through a set of story words enables us to move away from complex analysis of the whole story and instead focus on tracking of individual story words. This is a more straightforward task as we only need to evaluate the relevance of a story word candidate to an article with can be done by analysing the content of the given article without previous training.

Scalability and distributivity. Final advantage of our approach lies in its scalability and distributivity, both of which are supported by the fact that the crucial step of article acquisition and evaluation may be performed in a decentralized manner by a set of independent agents. There are no global decisions or dependencies that would serve as bottlenecks in the communication scheme of the proposed system. Each agent is capable to perform its tasks individually and independently, thus no system overhead is added even after introducing new agents into the system. The decisions of an agent are based on its local attributes such as the story word it carries or content of the article it visits. This allows to deploy our proposed approach in a flexible system with loosely coupled individual modules. Addition or subtraction of resources may be dynamically performed through introduction of new agents or removal of existing ones without compromising the integrity of the article evaluation process.

4. Article Space Model

The result of article acquisition and evaluation process is a model of the article space that includes articles, hyperlinks, identified story words and their respective relationships to related articles. In order to capture relationships between individual acquired entities we have elected to model the article space as a single large graph structure.

4.1 Graph Representation of Article Relationships

There are two types of nodes in the article graph, *articles* and *story words*. Nodes are interconnected by two types of edges, *hyperlinks* and *relationships*. A hyperlink edge connects two articles that are also interconnected by a hyperlink on the web. The relationship type edges connect either an article to a story word node if the particular story word is related to the article, or two articles if a connection between two articles has been established

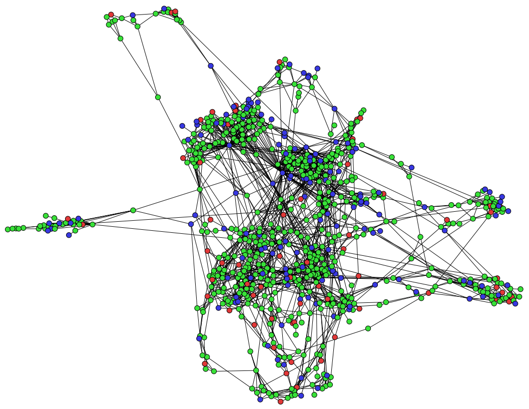

Figure 2: Graph representation of the article space. Nodes are coloured according to their type. Green colour represents visited article nodes, red colour represents unvisited article nodes and blue colour represents story word nodes.

on a basis of certain story word. Each relationship edge is weighted according to the confidence of the given relationship. An example of an article graph is given in Figure 2.

4.2 Louvain Algorithm Based News Story Extraction

Although the article graph structure provides us with an overview of articles and their respective relationships, it is by itself not sufficient to provide insight into the ongoing news stories. In order to extract news stories information we need to transform identified relationships into a complex news story structure.

For this purpose we utilize Louvain graph algorithm [1] to detect modules within the article graph. By using Louvain algorithm we partition the graph into subgraphs called *modules* is such a was as to maximize the number of connections between nodes within individual modules and to minimize the number of connection between nodes in different subgraphs.

In general we assume that if two articles cover the same news story their probability of sharing relevance to a common story word is higher than with random two articles that cover different news stories. Therefore if we partition a graph into modules in such a way as to minimize the number of connections of nodes across different modules we are maximizing the probability of articles within the same module to cover the same news story. Resulting modules represent news stories identified by our approach. Each story therefore consists of a set of articles along with a set of related story words and corresponding relationships amongst them. An example of a resulting graph with nodes color coded according to their affiliated news story is given in Figure 3.

5. Contributions and Research Goals

There are two main contributions of our approach in relation to the current state of the art, which map onto our research goals of topic representation suitable for swarm based extraction and scalability of an insect inspired topic detection and tracking approach.

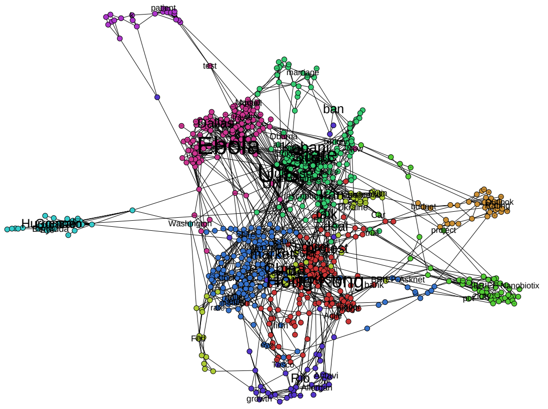

Figure 3: Article graph with identified stories. Nodes are colour coded according to their news story affiliation and story words are labelled.

5.1 Topic Representation Suitable for Swarm Based Extraction

The first contribution is the representation of a news story suitable for evaluation by a set of independent agents. In topic detection and tracking a topic is generally viewed as a probabilistic distribution over a set of terms, which can be sampled and evaluated as a single abstract concept. We present a different view with topics of news stories decomposed into multiple story related terms which we call story words. This provides us with a model of story decomposed into multiple aspects that can be tracked and evaluated independently, while maintaining accuracy of story detection comparable to batch processing algorithms, as shown by our experiments [7][8][9][10]. This model also provides a richer representation of the news story semantics, as the underlying relationships between individual story words allow us to explore the various aspects of the story on a more tangible level than with traditional latent stories represented by probabilistic models.

5.2 Scalability of Insect Inspired TDT Approach

The second contribution of our approach is the proposal of a decentralized self prioritizing algorithm for classification task within non parametric or semi parametric spaces. Standard online methods for classification in non parametric or semi parametric spaces share a common feature of being computationally dependent on the number of individual data points in the given space. The general notion of our approach is similar to kernel based methods in that we survey the space by performing one to one comparisons of individual data points in order to establish their relative distance.

However our social insect inspired approach introduces a self prioritizing capability that allows to determine the most relevant data points and prioritize their comparisons over the others, while at the same time rejecting comparisons of data points that are deemed irrelevant. In addition to prioritizing the search of document space we are also able to control the rate at which the search occurs by changing the number of involved agents. This provides for a scalable solution which addresses the common issue of kernel methods with sensitivity to the number of available data points.

6. Conclusion

In our work we have developed a novel approach to news story identification and tracking using a set of independent agents inspired by social insect. The social insect inspired model of the agent behaviour allows for a dynamic self prioritizing system that identifies and tracks the most relevant parts of news stories in time. The advantages of such approach lie in its ability to dynamically identify relationships between news articles through evaluation of relevance to common story related terms called story words.

By utilizing a honey bee inspired mechanism of coordination, our agents are able to determine the most promising story words to evaluate and thus focus on the most relevant part of article space. This allows for a dynamic and flexible approach that detects news stories as they evolve. Furthermore the story word based representation of topics allow for fine grained tracking of individual aspects of each news story. Through the decomposition of topics into interconnected terms we are able to deploy a method similar to traditional kernel based approaches that circumvents the high dimensionality of news story space. Scalability drawbacks of kernel methods when processing high number of inputs are addressed through social insect inspired coordination mechanism that allows to prioritize which articles and which story words to analyse first. The underlying scheme is general and thus it is our hope that it may find further applications in analysis of high-dimensional or non-parametric spaces beyond news story identification.

References

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[2] L. Bolelli, e. Ertekin, and C. Giles. Topic and trend detection in text collections using latent dirichlet allocation. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 776–780. Springer Berlin / Heidelberg, 2009.

[3] T. Brants, F. Chen, and I. Tsochantaridis. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 211–218, New York, NY, USA, 2002. ACM.

[4] C. Castillo. Effective web crawling. *SIGIR Forum*, 39(1):55–56, June 2005.

[5] D. Karaboga. An idea based on Honey Bee Swarm for Numerical Optimization. Technical Report TR06, Erciyes University, Oct. 2005.

[6] P. Navrat. Bee hive metaphor for web search. *Communication and Cognition-Artificial Intelligence*, 23(1-4):15–20, 2006.

[7] P. Navrat and S. Sabo. What's going on out there right now? a beehive based machine to give snapshot of the ongoing stories on the web. In *Nature and Biologically Inspired Computing (NaBIC), 2012 Fourth World Congress on*, pages 168 –174, nov. 2012.

[8] P. Navrat and S. Sabo. Beehive based machine to give snapshot of the ongoing stories on the web. *Transactions on Computational Science XXI, Special Issue on Innovations in Nature-Inspired Computing and Applications*, pages 296–314, 2013.

[9] S. Sabo, A. Kovarova, and P. Navrat. Multiple developing news stories identified and tracked by social insects and visualized using the new galactic streams and concurrent streams metaphors. *International Journal of Hybrid Intelligent Systems*, 12:27–39, 2015.

[10] S. Sabo and P. Navrat. Social insect inspired approach for identification and dynamic tracking of news stories on the web. In *Nature and Biologically Inspired Computing (NaBIC), 2013 World Congress on*, pages 226–231. IEEE, 2013.

Selected Papers by the Author

S. Sabo, P. Navrat. Bee Inspired Detecting and Tracking of Currently Developing News Stories From the Web. [submitted to] In: *International Journal of Bio-Inspired Computation, Inderscience Publishers, 2016.*

S. Sabo, P. Navrat. Social insect inspired approach for identification and dynamic tracking of news stories on the Web. In: *Nature and Biologically Inspired Computing (NaBIC), 2013, World Congress on*, IEEE, 2013, pp. 226–231.

P. Navrat, S. Sabo. What's going on out there right now? A beehive based machine to give snapshot of the ongoing stories on the Web. In: *Nature and Biologically Inspired Computing (NaBIC), 2012 Fourth World Congress on*, 2012.

P. Navrat, S. Sabo. Beehive Based Machine to Give Snapshot of the Ongoing Stories on the Web. *Transactions on Computational Science XXI, Special Issue on Innovations in Nature-Inspired Computing and Applications*, 2013, pp. 296–314.

S. Sabo, A. Kovarova, P. Navrat. Multiple developing news stories identified and tracked by social insects and visualized using the new galactic streams and concurrent streams metaphors. *International Journal of Hybrid Intelligent Systems*, 2015, vol. 12, pp. 27–39.

Š. Sabo. Dynamic Detection and Tracking of Stories in News Articles from the Web [in Slovak]. In: *WIKT 2013, 8th Workshop on Intelligent and Knowledge oriented Technologies*, Centre for Information Technologies, 2013, pp. 167–171.

A. Kovárová, Š. Sabo. Visualization of News Articles Identified by Bees [in Slovak]. In: *WIKT 2013, 8th Workshop on Intelligent and Knowledge oriented Technologies*, Centre for Information Technologies, 2013, pp. 35–40.

S. Sabo. Beehive Metaphor Inspired Web Crawler. In: *8th Student Research Conference in Informatics and Information Technologies Bratislava*, Nakladateľstvo STU, 2012, pp. 249–254.

P. Návrat, Š. Sabo. Determining Keywords for Unfolding Stories Using Swarm of Social Agents [in Slovak]. In: *WIKT 2012, 7th Workshop on Intelligent and Knowledge oriented Technologies*, Nakladateľstvo STU, 2012, pp. 37–40.

Š. Sabo. Tracking of a Story on the Web Using Multi-Agent System Inspired by Social Behaviour of Bees [in Slovak]. In: *WIKT 2011, Proceedings 6th Workshop on Intelligent and Knowledge oriented Technologies*, Košice, Technická Univerzita, 2011, pp. 167–171.