BEE INSPIRED DETECTING AND TRACKING OF CURRENTLY DEVELOPING NEWS STORIES FROM THE WEB

[1]ŠTEFAN SABO, [2]PAVOL NAVRAT

[1]Faculty of Informatics and Information Technologies, Slovak University of Technology, Bratislava, Slovakia

[2]Faculty of Informatics and Information Technologies, Slovak University of Technology, Bratislava, Slovakia

E-mail: [1] stefan.sabo@stuba.sk, [2] pavol.navrat@stuba.sk

ABSTRACT.

In this paper we present an approach to detection and tracking of currently unveiling news stories through analysis of news articles obtained from the Web. Articles are obtained using algorithm inspired by honey bees foraging for food. Due to the distributed dynamic nature of communication protocol utilized among bee workers when foraging for food, we are able to detect, evaluate and retrieve articles related to current news events from the web. During this process key terms related to news stories are identified and relationships between articles are established based on these terms. This results in a graph of nodes modelling the article space. The graph is updated and appended in a flexible way based on the real time changes in the articles. News stories are subsequently extracted from the graph using modularity optimizing graph algorithms.

Keywords: *Dynamic Story Tracking, Detecting A Developing News Story, Beehive Metaphor, Nature Inspired Algorithm, Story Word*

1. INTRODUCTION

Web based news articles are nowadays a major source of news related coverage. When a newsworthy event occurs a plethora of articles is published, covering various aspects of a news story from different angles. Although a specific news story is not precisely defined by its content a human reader has little problem to identify the story covered in an article by leveraging contextual information such as knowledge of overall situation, previous experience with news stories and general judgement about which information comprises a news story. However, such an information is not easily available to an automated system, therefore in our work we take a slightly different approach to news story identification. Our goal is to develop a system that would be able to autonomously detect news stories based solely on the content of the articles, without using additional knowledge or previous training.

Whenever a news story is being referred to, certain terms will become representative, be it actual words of the language that acquire new connotation, such as *Watergate*, or even completely new terms, such as a more recent *Grexit*. In order to effectively measure representativeness of a given story related term we need to track it across multiple articles as the news story develops. A dynamic approach is required that would be able to evaluate terms based on the actual set of articles and subsequently update the evaluation as new articles surface.

When looking for such a dynamic approach we find inspiration in the world of nature. Members of honey bee species *Apis Mellifera* utilize waggle dance to relay information about suitable sources of food. This mechanism enables bees to focus on the most suitable sources of food at a given time. Bees react dynamically to the quality of the source and if the source starts to lose its appeal, the number of bees dancing for the source will decrease. Experiments performed by Seeley et al. [1] have shown that bees can react to changes in food sources quite dynamically, reflecting the changes in source quality in a matter of minutes. Based on this mechanism we present a dynamic approach to news story identification and tracking.

2. ARTICLE ACQUISITION

The first step necessary in order to obtain information about current events from the Web is obtaining articles discussing the news stories. When designing approach to acquire articles within our system, we consider three main aspects of the news domain:

Dynamics. New news stories are added on a daily, even hourly basis and existing articles may be further edited. Therefore it is necessary for our system to identify new articles, while also regularly revisiting current articles and updating the set of current news stories accordingly.

Lack of structure. Although certain specific parts of an article, such as title or links may be automatically identified, an article generally contains plain unstructured text. All information about the topics covered needs to be extracted from the content of the article.

Succinct language. Titles and links are the first points at which the reader decides whether he is interested in reading an article. In order to convey the information about the topic of the article, concise information about its content needs to be compressed into just few words or at most a single sentence. Therefore titles and links often contain entities or phrases that are either unique to the covered news story or at least representative enough to distinguish the story from other stories covered at a given time.

Considering these three aspects of news domain we have designed an approach to news story acquisition and tracking based on the natural behaviour of honey bees foraging for food. Our system consists of a set of independent agents each carrying a certain story related term. The task of an agent is to navigate the article space and to evaluate relevance of the carried term to the visited articles and establish links between related articles. This results in a graph of news articles along with their connections from which the current topics are later extracted through graph analysis.

2.1 Coordination of Agents

When foraging for food honey bees use a specific coordination model based around the waggle dance, first decoded by von Frisch [2], to determine the best sources on which to focus their foraging effort. If a foraging bee considers its last visited source to be suitable, it may engage in a waggle dance which encodes the information about location and viability of a source, recruiting other bees to visit the source. The higher the quality and safety of the source, the more often and more vigorously do foragers engage in the waggle dance, thus engaging more workers to follow them. This model has been adopted by Karaboga [3] for numeric optimization and further refined into *Beehive Metaphor* model for use in web search by Navrat [4].

In our work we use a modified version of Beehive Metaphor to coordinate agents evaluating news articles. Instead of selecting web pages, our agents select topic related words and evaluate them according to their relevance to currently ongoing stories. Most popular words are propagated and the number of agents evaluating them increases, while words not connected with stories are gradually discarded. The details of story evaluating mechanism are given in Section 3.

The main advantage of this approach lies in its flexibility. Due to the decentralized nature of the system, agents are working in real time with articles directly from the Web as they are published. There are no preprocessing steps and knowledge of the entire dataset is not required as new articles are incorporated into the set incrementally. In addition, because of no central point of processing the system scales well by adjusting the number of utilized agents when necessary.

2.2 Crawling

When retrieving articles from the Web our agents act as regular web crawlers. In addition to the Beehive Metaphor based model of coordination, additional policies are adopted in order to best suit the crawling process to our goal of topic detection and tracking.

Revisiting policy states that each article will be re-retrieved if at least 30 minutes have passed since its last retrieval. Within the 30 minute period articles are cached, indexed within our storage and may be still visited and evaluated by individual agents.

Parsing of the retrieved web page is performed using Jericho HTML parser [5]. From each web page the content of the article is extracted along with the title of the article, first paragraph, links to other articles and optional time stamp of the article publication and location to which the article is relevant.

Courtesy delays are implemented in order to alleviate the stress of a large number of requests in a quick succession to a single web host. In our work we have been using a courtesy delay of 5 seconds between two requests to the same web host.

3. STORY IDENTIFICATION AND TRACKING

The primary function of the agents described in the Section 2 is to map relationships between articles. A relationship signifies that the two articles share a common relevance to a given term. This section outlines how a relationship between two articles is established and how the relationship graph is subsequently used to determine the stories discussed in given article set.

3.1 Story Words

Rationale behind our story identification mechanism is based on the concept of story words. A story word is a term that closely identifies with a given news story. When a news story starts to unravel and news articles about it emerge, certain terms will come to identify the news story in news coverage. These terms may either be unique and represent the story unambiguously, e.g. "*Charlie Hebdo shooting*", or they may refer to a certain news story in the context of current events, such as "*US presidential elections*". Each story may be connected to multiple such story related terms and our goal is to map the connections between articles and story words.

3.2 Article Relationships

In order to map out the relationships between story words and articles we utilize a set of agents described in Section 2. There are two types of relationships that we track.

- Links between articles facilitate navigation of agents in the article space.
- Shared story words denote relevance of both articles to a common story word.

Links are the byproduct of article evaluation process and their collecting is trivial. In this subsection we focus on the method of Identifying the shared story words. Shared story words are identified by honey bee inspired agents described in Section 2. In accordance to Beehive Metaphor model each agent is attributed with a single source at a given time. We refer to the words carried by agents as *story word candidates*, or SWCs. The purpose of agents carrying SWCs is to establish the relevance of articles to the SWC and in the process evaluate the popularity of the SWC.

Actions of an agent are governed by the popularity of the SWC. We use the popularity of a SWC in place of the *quality of source* attribute used in the original Beehive Metaphor. A scheme the of Beehive Metaphor model is given in Figure 1. The time of an agent is partitioned into turns. Depending on the popularity of the carried SWC an agent selects one of three tasks each turn.

Foraging represents an activity in which the actual evaluation of article occurs. During foraging an agent visits articles and evaluates their relevance to the currently carried SWC. Each time an article is visited, the relevance of the SWC to the article is calculated as

$$R_s^a = \frac{w_t o_s^t + w_1 o_s^1 + \sum_{i=2}^{n} w_p o_s^i}{w_t + w_f + n w_p}$$

where w_t, w_1 and w_p denote the weight of SWC occurrence in the title, the first paragraph and in other paragraphs, respectively while o_s^t and o_s^1 assume value of 1 if the SWC occurs in the title and the paragraph i of the article or 0 otherwise. The relevance R_s^a of SWC s to the article a is subsequently used to establish the relationships between the articles. If relevance $R_s^a > \alpha$, relationship based on *s* is established between article *a* and the last visited article *b* for which $R_s^b > \alpha$. Relevance threshold value α is a parameter of the model. In our experiments we used value $\alpha = 0.8$. The new relationship is assigned a confidence factor of

$$w = \min\left(R_s^a, R_s^b\right)$$

and R_s^a is used to adjust the overall popularity R_s of SWC *s*. R_s is calculated as the average of R_s^a over all articles visited so far during the current foraging session. Depending on the popularity of the SWC, an agent carrying the SWC decides whether to continue pursuing the current SWC with probability of $P_s^c = R_s$, or abandoning the SWC otherwise. If an agent decides to abandon the SWC, the agent adopts the *observing* task during which the agent attempts to select a new SWC. If an agent instead decides to keep the current SWC, it decide further on whether to propagate the SWC with probability of $P_s^d = R_s$ or continue foraging with the probability of $P_s^f = 1 - R_s$, in accordance to diagram given in Figure 1.

Dancing represents the means of recruiting agents to propagate popular SWCs. The higher the popularity of a SWC, the greater the chance of an agent dancing for it. Therefore the most popular SWCs will be evaluated more thoroughly and the chance of identifying a relationship between two articles will be higher. When in dancing state, an agent is idle for a set number of turns calculated as $t_d = \lceil \beta R_s \rceil$. Here R_s is popularity of currently carried SWC and β denotes *maximum dancing time* in turns, which is a parameter of Beehive Metaphor model. During the dance an agent retains its SWC so that it is clear what SWC is being propagated by the dancing agent. After t_d turns of dancing have passed an agent resumes the foraging activity.

Observing enables agents that have abandoned their SWC to adopt a new one from the pool of SWCs propagated by dancing agents. If an agent decides to abandon its SWC because of its poor popularity the agent switches to observing state for a set number of turns denoted γ, which is a parameter of Beehive Model called *observation time*. In each round of observation an agent decides whether to adopt a new SWC using probability of

$$P_a = \frac{N_d}{N}$$

where N_d is the number of currently dancing agents and N is the number of all agents. If an agent decides not to adopt a new SWC, it skips the current turn and proceeds to decide again, if there are turns left. If an agent instead decides to adopt a SWC, it select one from the pool of propagated SWCs with probability of selecting a SWC i being

$$P_s^i = \frac{N_d^i}{N_d}$$

where N_d^i is the number of agents propagating SWC *i*. As the probability of adopting any SWC is equivalent to the proportion of dancing agents and the probability of selecting a given SWC is equivalent to the proportion of agents propagating the SWC, the observing process of an agent *a* may be implemented by sampling a single agent *b* from the total pool of agents. If agent *b* propagates a SWC, agent *a* will adopt the same SWC as well, otherwise agent *a* skips a turn. Using this technique the numbers N_d and N_d^i need not to be actually calculated, therefore each agent is able to accomplish its observation turn in constant time.

If an agent selects a SWC during observation phase, it starts a new foraging session using the new SWC for analysing article relationships. However if an agent fails to select a new SWC during allotted time, it needs to select a new SWC at random. This is an important point, as this selection mechanism allows for agents to find new story words that may have just emerged, potentially exposing emerging news stories.

3.3 Semantics

The ability of our system to correctly identify news stories depends on the ability to evaluate the relevance of a given term to a given news article. The relevance is used in order to direct agents towards appropriate tasks in a stochastic way. A precise value of the relevance is not necessary, however we still need to be able to at least broadly assess whether an article is relevant in respect to a given term, or not.

The relevance given in Subsection 3.2 requires only the ability to identify whether a given text contains the assessed SWC. In case of proper noun SWCs it is a relatively easy task, however in case of common nouns we need to make sure that we also take into account occurrences of synonyms and other related terms. In order to be able to generalize the SWCs we utilize *Wordnet* by Miller [6]. This allows us to broaden the set of terms that we look for in an article. In addition to the original term (such as *plane*) we also look for synonyms (*aircraft*), hypernyms (*vehicle*) and hyponyms (*jet*). To factor in the occurrence of related forms of a SWC we adjust the calculation of occurrence factors o_s^t and o_s^1 from Subsection 3.2 so that instead of boolean value of 0 or 1, both factors can attain value within interval of $[0,1]$ using the formula

$$o_{adj} = o_s + \sigma_{syn} n_{syn} + \sigma_{hr} n_{hr} + \sigma_{ho} h_{ho}$$

where o_s is the original value based on the presence or absence of SWC in a text segment, n_{syn}, n_{hr} and n_{ho} are numbers of synonyms, hypernyms and hyponyms of the original SWC present in a text segment and σ_{syn}, σ_{hr} and σ_{ho} represent weighing factors for the respective groups. In our work we have used weighing factors of $\sigma_{syn} = 0.8$, $\sigma_{hr} = 0.6$ and $\sigma_{ho} = 0.4$. Other values may be used as well, however we recommend that $1 > \sigma_{syn} > \sigma_{hr} \geq \sigma_{ho}$. We prefer hypernyms over hyponyms as the semantic field of hypernym encompasses the semantic field of original SWC, which is however not the case with hyponyms. If value of $o_{adj} > 1$ we assign $o_{adj} = 1$.

The outlined adjustment allows us to factor in the presence of alternative forms of SWC that may be used interchangeably with the original term. If the original term is present in a text segment, the occurrence factor o_{adj} is equal to 1 as $o_s = 1$. Otherwise alternative forms of SWC are considered in order to measure the relationship between an article and evaluated SWC.

3.4 Extraction of Stories

In order to represent the part of article space already explored by the agents we utilize a graph containing nodes both for articles and story words. Nodes are interconnected by edges representing both relevance relationships and hyperlinks. Semantics of the resulting graph structure is given in Table 1.

.

Table 1: Semantics Of Article Graph

	article a_1	story word s_1
article a_2	hyperlink between a_1 and a_2	s_1 relevant to a_2
story word s_2	s_2 relevant to a_1	-

This representation gives us an overview of relationships between articles and story words. However in order to extract information about current events we need to forge the abstract concept of mutually relevant articles into more substantial news stories. For this purpose we use Louvain graph algorithm by Blondel et al. [7] to identify modules within the article graph. We assume that if two articles both cover the same news story there is a higher chance of them to share relevance to common story words when compared to two articles covering different stories. Therefore we partition the graph in such a way as to maximize the number of connection within subsets and minimize number of connections between subsets. Resulting partitions represent individual stories identified by our approach.

A visualization of our graph representation is given in Figure 2. Subfigure 2a shows unprocessed article graph with visited articles, unvisited articles and story word nodes coloured green, red and blue, respectively. Subfigure 2b shows article nodes from the same graph coloured according to their recentness with blue being the most recent and red being the least recent. Story word nodes are omitted in this figure as they carry no time information. Subfigure 2c shows article graph partitioned according to identified stories with story words being labelled. Each label is scaled according to the story word popularity. Finally subfigure 2d represents the same partitioned graph, however within this graph hyperlink edges are omitted and only relationships identified by our approach are shown.

4. EVALUATION

In this section we outline the experiments performed in order to validate and quantify performance of our approach. All experiments described in this section have been performed online on Reuters website.

4.1 Parameter Setup

Throughout all of our experiments we have used the following values of model parameters:

- MDT = 4,
- OT = 4,
- BISB = 100.

All of these are standard parameters of Beehive Metaphor model. *MDT* stands for *maximum dancing time* and represents a number of turns an agent would propagate an ideal story word candidate. *OT* denotes observation time and represents a maximum number of turns an agent may continue observation task. Parameter *BISB* stands for *bees in source base*. It determines the initial distribution of agents between tasks of foraging and observing by setting the percentage of agents initially allocated to foraging. For more detailed discussion of Beehive Metaphor parameters please see work of Navrat and Kovacik [8].

The measurement for evaluation was performed in the period from 5th to 31nd October 2014 on Reuters website, during which 14 separate runs were performed using 50 agents for 500 iterations with default parameters. Each attempt was performed independently on a different day in order to assemble multiple separate datasets on which we could perform both evaluation of results and analysis as a basis for future work.

4.2 Precision and recall

The aim of first experiment was to evaluate the ability of our approach to correctly determine the currently unfolding stories. In order to evaluate the ability of our system to identify stories correctly we have decided to measure the *precision* and *recall* of classification of articles into stories. Due to the fact that assignment of articles into stories does not constitute a binary classification, the standard *true / false / positive / negative* decomposition is not applicable. Therefore we split the results along two axes of *correctness* and *certainty*.

With our approach every article is assigned a story, therefore there are no true and false positives or negatives. Instead the correctness axis divides the whole result set into correctly and incorrectly identified results. In order to establish the correctness of identified results we have manually classified all the retrieved articles by grouping them together according to mutual relevance to the same news story. This has provided a baseline ground truth to which the results obtained by story extraction algorithm have been compared. According to the baseline we have been able to determine for each article whether its related news story was identified correctly. Thus we have divided all the obtained results into two sets of correctly and incorrectly classified.

The certainty axis divides the results into two categories of certain and uncertain. This division is achieved by introducing a membership function $f^M(R) = m^R$ that assigns each result R a membership factor of m^R. A result R is considered certain if and only if $f^M(R) \geq \sigma$ where σ denotes a membership threshold value parameter of the experiment. The membership function allows us to refine the results by disregarding the classification results with arbitrarily low confidence. For the purpose of calculating precision and recall our scheme maps to the traditional true – false, positive – negative scheme in a following way:

- correct certain (*CC*) – true positive
- correct uncertain (*CU*) – false negative
- incorrect certain (*IC*) – false positive
- incorrect uncertain (*IU*) – true negative

For each value of σ we are able to calculate precision P_σ as

$$P_\sigma = \frac{CC_\sigma}{CC_\sigma + IC_\sigma}$$

and recall R_σ as

$$R_\sigma = \frac{CC_\sigma}{CC_\sigma + CU_\sigma}$$

It is important to add that we have measured only the precision of the article classification. The relevance of individual story words was not evaluated as identifying a news story for individual story word may become subjective in case of general story words, such as *United States*, *diplomacy* or *crisis*. The identified stories along with the number of related articles are summarized in Table 2.

Table 2: Overview Of Identified News Stories

	CC_0 + IC_0	CC_0	CC_0[%]
Ebola outbreak	91	88	96.70
Asian economy	70	34	48.57
Business news	66	52	78.79
Islamic State fighting	64	57	89.06
Stock trading news	63	58	92.06
Brazil corruption	62	40	64.52
Hong Kong protests	55	41	74.55
Russia events	42	25	59.52

Recent court rulings	33	15	45.45
US tax legislature	32	11	34.38
China economics	31	19	61.29
Atlantic city casino	26	8	30.77
Pharmaceutical news	14	14	100.00
Scotland vote	6	6	100.00
Total	655	468	71.45

The zero index used in the table header signifies that no membership function was used in order to refine results. All results, both correct and incorrect are considered certain. In other words, $CU_o = IU_0 = 0$, therefore $CC_0 = CC_0 + CU_0$ and $IC_0 = IC_0 + IU_0$. The overall attained precision of story identification was 71.45%. Highest precision was attained for pharmaceutical news and Scotland independence vote, which both represented minor stories, followed by the major story of Ebola outbreak, which was at the time of the measurement focused heavily around potential of carrying Ebola virus into mainland US. Worst precision was attained for less notable stories about Atlantic city casino, followed by stories about US tax legislature changes and recent court rulings. All of these article sets contained significant portion of articles either incorrectly identified as relevant to the other stories, or articles only remotely related to the given story. The relevance of articles to such vaguely defined news stories is rather difficult to assess both for an automated system and for human reader as well.

After introducing membership functions, we have been able to influence the precision and recall measures by adjusting the membership threshold, effectively affecting the sensitivity of the underlying model. In general this has allowed us to trade off precision for recall and vice versa, however there are significant differences in the outcomes depending on which membership function was used. The impact of membership functions is compiled in Figure 3.

The first membership function f_C^M is based on local clustering coefficient of an article node. Clustering coefficient c^R gives us a measure of interconnection between neighboring nodes of R. We assume that we are able to use coefficient local to an article node as a membership coefficient of article within a story. Due to the fact that clustering coefficient attains values from interval $[0,1]$ there is no need for normalization and we can use clustering coefficient c^R of an article node R directly as a value of membership function $f_C^M(R) = c^R$. The graph of precision and recall as functions of membership threshold $\sigma\,\sigma$ with f_C^M as membership function is given in Figure 3a.

The second membership function f_E^M is based on eigenvector centrality measure. Higher eigenvector value of a node presumes greater centrality of a node, thus we have used eigenvector centrality measure as a second candidate membership function. As with clustering coefficients, eigenvector values also attain values only from interval $[0,1]$, therefore we use eigenvector value e^R of an article node R directly as a value of membership function, $f_E^M(R) = e^R$. The graph of precision and recall as functions of membership threshold σ with f_R^M as membership function is given in Figure 3b.

The third membership function f_D^M is based on the degree of a node. If a node is interconnected with more nodes, we can generally assume that it will be more central to a news story, therefore we may assign higher membership value to it. Outlying articles are assumed to have lower degree than central ones. Membership function based on node degree $f_D^M(R)$ is then calculated as normalized node degree by calculating

$$f_D^M(R_i) = \frac{\deg(R_i) - \min(\deg(R))}{\max(\deg(R)) - \min(\deg(R))}$$

The graph of precision and recall as functions of membership threshold σ with f_R^M as membership function is given in Figure 3c.

To summarize the results, in this experiment we have measured the precision and recall of news story classification by first manually classifying articles into news stories and then comparing the results of automatic classification with the manually obtained ground truth. The precision of story classification ranged from 30.77% achieved for *Atlantic city casino* up to 100.00% achieved for *Scotland vote* and *Pharmaceutical news* followed by *Ebola outbreak* story with precision of 96.70%. Overall achieved precision of story identification was 71.45%.

This precision may be further enhanced by refining through filtering of results based on a membership function. We have evaluated three membership function candidates based on *clustering coefficients*, *eigenvector values* and normalized *node degrees*. Function based on clustering coefficients failed to yield significant improvement of precision even for extreme values. We comment this as an evidence that our assumption on relationship between clustering coefficient and article relevance was wrong. Functions based on eigenvectors and node degrees both allowed us to improve precision of the story identification by gradually filtering out remote nodes on the outskirts of story clusters at the cost of recall. It is important to note that although introduction of membership function failed to provide improvement of *F1 measure*, we do not consider this to be a drawback due to the fact that recall in our experiment represents only a synthetic measure presented in order to compare the impact of various membership functions on the number of articles that we are able to correctly classify. However our approach does not specifically require precise classification of individual articles. The aim is rather to outline current news stories and provide certain degree of background of the story, high recall statistic would translate simply into high detail in which the story is covered.

4.3 Task popularity

The aim of the second experiment was to obtain insight into the dynamics of agent swarm behaviour over the course of a single run. The idea of utilizing swarm inspired behaviour of agents presumes that depending on current needs of the system, agents will be able to dynamically adjust the number of agents distributed over foraging, dancing and observing tasks. Therefore we are interested in obtaining data on real behaviour of our system. We performed five independent measurements during five consecutive days. Each measurement used 50 agents and consisted of a 3000 iterations long run which translates approximately to 5 hour long runs. During each iteration we have recorded the number of agents currently performing each tasks. The patterns for each run were rather similar. Regardless of the initial allocation the proportion of dancing agents stabilized around 60%, foraging was performed by around 30% and observing was generally attended by just a small portion of beehive population.

The last attribute that varies between individual runs is the fluctuation rate of agent numbers during stable phases. In order to quantify the rate of fluctuation, for each run *R* we have calculated the running standard deviation

$$\sigma_R = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_i)^2}$$

where $\mu_i = \frac{1}{10}\sum_{i}^{i+9} x_i$

is a running average for the floating window of size 10. The calculation was performed non incrementally over two passes in order to minimize the floating point error.

In all runs the observation task exhibited notably lowest standard deviation, ranging from 0.41 to 2.45, even after attributing for lower mean number of agents performing observation task. On the other hand, both dancing and foraging exhibit similar rates of variation within each run, ranging from 2.26 and 2.24 for foraging and dancing respectively up to 6.13 and 6,01. We conclude that for each run the fluctuation rates for dancing and foraging tasks are related. This is consistent with our predictions based on the nature of Beehive Metaphor Model, as agents usually switch back and forth between dancing and foraging, depending on the desirability of their respective candidates. The rate of fluctuation is inversely related to the mean number of observing agents as high number of observing agents leaves less agents to switch back and forth between foraging and dancing tasks. This is consistent with our observations. As to the fluctuation rate of observing agents, we have been unable to connect it to any other internal variable of the system. At this time we predict the fluctuation rate of observing agents to be solely dependent on data being processed.

Two most distinct of the task allocation patterns are given in Figure 4 for the reference. Figure 4a illustrates a run with three distinct distribution patterns of agents. One iteration lasts approximately 6 seconds, therefore the first pattern lasted for approximately 25 minutes since the start after which it was followed by second pattern that lasted for 50 minutes and finally transitioning into third pattern that lasted slightly under 4 hours. Figure 4b shows different behaviour with one pattern stabilizing shortly after the start and lasting to the end of the run.

5. DISCUSSION

In this section we would like to discuss advantages and disadvantages of the proposed approach both generally and with focus on specific design decisions. We also outline proposed scenarios of use along with examples from real life use. Finally we conclude the section by discussing alternative approaches to the news story tracking task and compare them to our approach.

5.1 Advantages

The main aim of utilizing a story tracking mechanism inspired by honey bees is to provide a dynamic system that would be able to respond to changes in news story landscape as new articles are published. The most notable advantages of our approach from our point of view are the following:

- Dynamic story tracking
- Iterative evaluation
- No learning necessary
- Scalability

Dynamic story tracking is the most important characteristic of our approach. It gives us the ability to track news stories in time as they unfold which was our primary goal when designing the approach. Due to dynamic story tracking capability we are able to analyse new articles on the fly and identify new stories but also to take note of new developments in stories that are already unfolding as new articles are acquired.

Incremental evaluation of new articles is closely connected with dynamic story tracking capability. Due to the incremental evaluation capability of our approach it is possible to add new articles into the set by performing evaluation only on a limited subset of the articles, regardless of the article set size.

No learning or supervision is necessary due to the way how news stories are decomposed into individual story words. Instead of learning what attributes or characteristics of an article need to be extracted in order to track a given news story we focus only on tracking of story words. Instead of training agents to extract the most appropriate terms from an article, we evaluate individual story words through comparing of story word occurrences in multiple articles.

Scalability and distributivity of our approach are supported by the decentralized nature of our approach. In order for each agent to operate it needs to communicate only with a single other agent at a time. Thus assuming a suitable implementation it is able to attain a sufficiently scalable and distributive system.

5.2 Disadvantages and Limitations

The use of beehive inspired mechanism for news story tracking, along with story word representation of a news stories brings certain disadvantages as well. The most notable disadvantages of our approach in our view are:

- Reliance on data organization
- Cold start
- Domain restrictions

Reliance on data organization is related to the way how agents acquire articles. It is necessary for articles to be interconnected by hyperlinks or some other means, so that agents are able to move from one article to another while exploring the article space. A connection between two articles presumes some level of mutual relevance, which can be exploited to direct agents towards groups of interesting articles. Otherwise the analysis of relationships would be reduced to random article matching with no additional information.

Cold start applies to our approach as agents need some time to start assembling the story graph. Even with a small number of initial articles, agents are able to find new articles through hyperlinks and find connections between articles. However, the low number of articles during the initial phase of the run affects our ability to identify news stories. With few articles it is difficult to integrate the connections between articles and identify news stories. Story extraction is possible only after the number of articles has reached at least few dozen, which may take up to few minutes, depending on the settings.

Domain restrictions are in place because of the assumption that each news story may be represented by story words that are specific enough to unambiguously represent a news story. Thus our approach is not well suited to domains such as celebrity news or comic articles that often do not provide representative article titles. Best results are achieved in domains that support succinct and comprehensive information forms, such as current events coverage, economics and technology.

5.3 Alternative Approaches

We conclude this section by discussing alternative approaches to news story detection and tracking. The two most important features that we consider when designing a news story detection system is online versus batch analysis and supervision requirements. Our approach utilizes online analysis that allows us to process new articles as they emerge in an iterative manner. The other option would be to instead adopt a batch analysis based topic model that would first accumulate articles and subsequently partition individual articles into news stories such as *Probabilistic latent semantic indexing* (PLSI) or *Latent dirichlet allocation* (LDA). In general, utilization of such models would translate into better precision of topic detection as they capture the underlying relationships between individual terms in documents better than a kernel based approach. The obvious downside of a batch approach is the necessity to periodically reprocess the whole batch of articles as they emerge. However the more important shortcoming when processing articles from the Web is the inability to focus the article acquisition so that popular, relevant or otherwise desirable articles are given preference. With the social insect inspired algorithm we are able to continuously direct the article acquisition by sharing the information about most notable story words between agents, which is a behaviour pattern which cannot be replicated with batch algorithms, as batch algorithms inherently treat data acquisition and processing as separate steps.

Another alternative is to determine news stories through an approach utilizing supervised learning. Each news story can be represented by a feature vector and a supervised classifier may then be trained to classify articles into these stories. However a supervised approach will not be able to identify a new story which was not included in its training set. Thus supervised approach would be suitable to track long term stories, possibly with greater precision that achieved by our approach. However autonomous identification of new stories that have just emerged would be difficult due to lack of training examples.

6. RELATED WORK

In this section we provide overview of related works that have influenced the design of our approach. We divide it into two main areas on which we focus, namely *Topic Detection and Tracking* and *Bio Inspired Algorithms*.

6.1 Topic Detection and Tracking

Topic detection and tracking has been a focus of research for some time and there is a wide range of various techniques how to approach it. For us the most relevant are the ones that are applicable in the field of news story identification. A standard method to topic detection and tracking is through extraction of keywords for individual documents. Bohne et al. [9] proposes a keyword extraction mechanism relying on Helmholtz principle. By utilizing principle that treats keywords of a document as statistical deviations from random background noise the given approach extracts keywords of a document on a language and grammar independent basis. Another approach for keyword extraction through multiple modifications of TF-IDF measure is proposed by Lee and Kim [10].

Another approach to news story identification is to determine relationships between articles through clustering algorithms. Cheng et al. [11] proposed an algorithm based on agglomerative clustering that groups the articles according to similarity measures. Clustering is performed in multiple rounds and parameters of clustering are optimized through simulated annealing. An approach proposed by Vadrevu et al. [12] clusters news search results using an incremental version of k-means algorithm. The approach is designed for dynamic sets of documents as results are periodically reprocessed after the number of new articles exceeds a given threshold.

A different approach to clustering algorithms is topic detection by probabilistic methods. One of the most notable examples is Latent dirichlet allocation by Blei et al. [13]. LDA is a topic model that extends Probabilistic latent semantic analysis proposed by Hofmann [14]. The basic notion of LDA is that it treats documents as a collection of topics from which individual words are generated. By learning the distribution of topics over documents it is possible to model real topics and to classify documents, both of which are widely used in news article processing.

In addition to content analysis it is also possible to group documents based on non textual information. An approach proposed by Toda et al. [15] analyses temporal data in documents to identify documents co occurring in time and thus determine both the topical structure of the documents and to some degree the meaning behind documents grouped together. A more recent work of Li et al. [16] proposes an algorithm to detect and track topics in news broadcast data containing both textual and visual information through hierarchical And-Or graphs. Instead of tracking purely textual topics, images and text are treated as a joint multi-modal channel that carries the information about topic discussed in a given document.

6.2 Bio Inspired Algorithms

Bio inspired algorithms represent a broad class of algorithms with wide applications in many areas of research. Our research focuses on algorithms that utilize self organizing capabilities of social insect for optimization and search tasks. A classic example of an algorithm based on self organizing social insect colony is the Ant colony optimization algorithm by Dorigo et al. [17]. Beehive metaphor algorithm used in our work was proposed by Navrat [4] as a platform for decentralized decision making of a swarm of independent agents suited for tasks of web crawling and web search. Beehive metaphor is based upon previous idea of Artificial bee colony proposed by Karaboga [3] as a social insect inspired model for numeric optimization.

7. CONCLUSION

In this work we describe our approach to news story detection and tracking based on behaviour of honey bees foraging for food. Our algorithm leverages the communication pattern within a bee swarm that allows to disseminate information about the best food sources. Each news story is represented by a set of relevant terms called *story words*. In order to identify the story words from a set of articles, candidates are selected by individual agents and each agent then proceeds to look for articles that share relevance to its current story word. Based on the viability of individual story word candidates, agents decide whether to analyse the candidates, propagate them further or discard candidates of low relevance.

The result is a model of news stories in which we account only for story word candidates that are relevant to currently ongoing stories. Relevance of story words to articles is determined through analysis prioritised based on current popularity of individual story word candidates and the layout of the article graph so that the most prominent articles and most centralized articles are given the highest priority and subsequent comparisons spread further.

Main advantages of such approach are its dynamics and flexibility. Due to the semi parametric nature of the underlying model, the number of relevance checks is not determined either by the number of articles or number of story word candidates. Rather it depends on a number of agents. This allows for a scalable system that prioritizes its resource allocation based on its evaluation of story word significance. Furthermore, the evaluation of articles is performed in real time on articles acquired from the web as they emerge, which is also facilitated by the self organizing nature of the swarm of agents that process the acquisition and evaluation tasks.

Although the implementation of our approach for news story identification and tracking as outlined in this work is constrained by assumptions about article organization that stem from news story domain, it is our belief that the core idea of a self prioritizing approach operating over semi parametric spaces would find potential applications beyond the scope of news story tracking.

ACKNOWLEDGEMENTS

This work was partially supported by the Scientific Grant Agency of Slovak Republic, grants No. VG 1/0646/15 and 1/0667/18 and by the Research and Development Support Agency, grant No. APVV-15-0508.

REFERENCES:

[1] Seeley, T. D., Camazine, S., Sneyd, J., “Collective decision-making in honey bees: how colonies choose among nectar sources.”, *Behavioral Ecology and Sociobiology*, Vol. 28, No. 4, 1991, pp. 277-290.

[2] von Frisch, K., “The dance language and orientation of bees.”,Belknap Press of Harvard University Press, 1967.

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

[4] Navrat, P, “Bee hive metaphor for web search.”, *Communication and Cognition-Articial Intelligence*, Vol. 23, 2006, No. 1- 4, pp. 15–20.

[5] http://jericho.htmlparser.net/

[6] Miller, G A, Wordnet, “A lexical database for English. “, *Commun. ACM,* Vol. 38 No. 11, 1995, pp. 39 -4.1

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

[8] Navrat, P., Kovacik, M., “Web search engine as a bee hive.”, *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence,* IEEE Computer Society (Washington, DC, USA), 2006, pp. 694-701

[9] Bohne, T., Rönnau, S., Borghoff, U. M., “Efficient keyword extraction for meaningful document perception”, *Proceedings of the 11th ACM symposium on Document engineering,* ACM (New York, NY, USA), 2011, pp. 185–194.

[10] Lee, S. and Kim, H.–J., “News keyword extraction for topic tracking.”, *Networked Computing and Advanced Information Management, 2008. NCM '08. Fourth International Conference on,* Vol, 2 , 2008, pp. 554–559.

[11] Cheng, J. , Zhou, J. , Qiu, S., “Fine-grained topic detection in news search results”, *Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC '12*, ACM (New York, NY, USA), 2012, pp. 912–917.

[12] Vadrevu, S. , Teo, C. H. , Rajan, S. , Punera, K. , Dom, B., Smola, A. J., Chang, Y., and Zheng, Z., “Scalable clustering of news search results”, *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM (New York, NY, USA), 2011, pp. 675–684.

[13] Blei, D M, Ng, A Y, and Jordan, M I, “Latent dirichlet allocation”, *J. Mach. Learn. Res.* Vol. 3, 2003, pp. 993–1022.

[14] Hofmann, T., “Probabilistic latent semantic indexing”, *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '99,* ACM (New York, NY, USA), 1999, pp. 50–57.

[15] Toda, H., Kitagawa, H., Fujimura, K., Kataoka, R., “Topic structure mining using temporal co-occurrence”, *Proceedings of the 2nd international conference on Ubiquitous information management and communication,* (New York, NY, USA), 2008, pp. 236–241.

[16] Li, W., Joo, J., Qi, H., Zhu, S., “Joint image-text news topic detection and tracking with and-or graph representation”, CoRR, 2015, abs/1512.04701.

[17] Dorigo, M., Birattari, M., and Stutzle, T., “Ant colony optimization.”, *Computational Intelligence Magazine, IEEE*, Vol. 1, No. 4, 2006, pp. 28 –39.

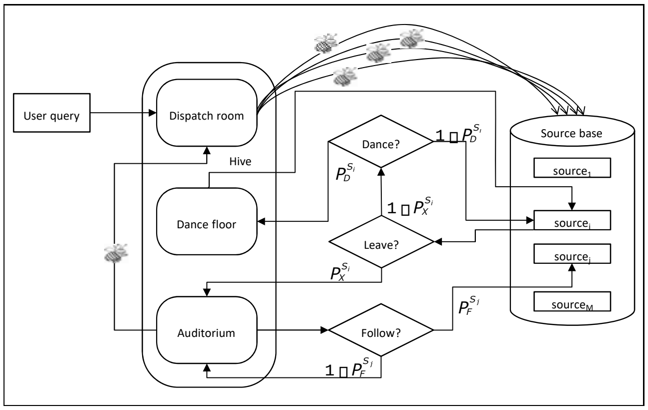
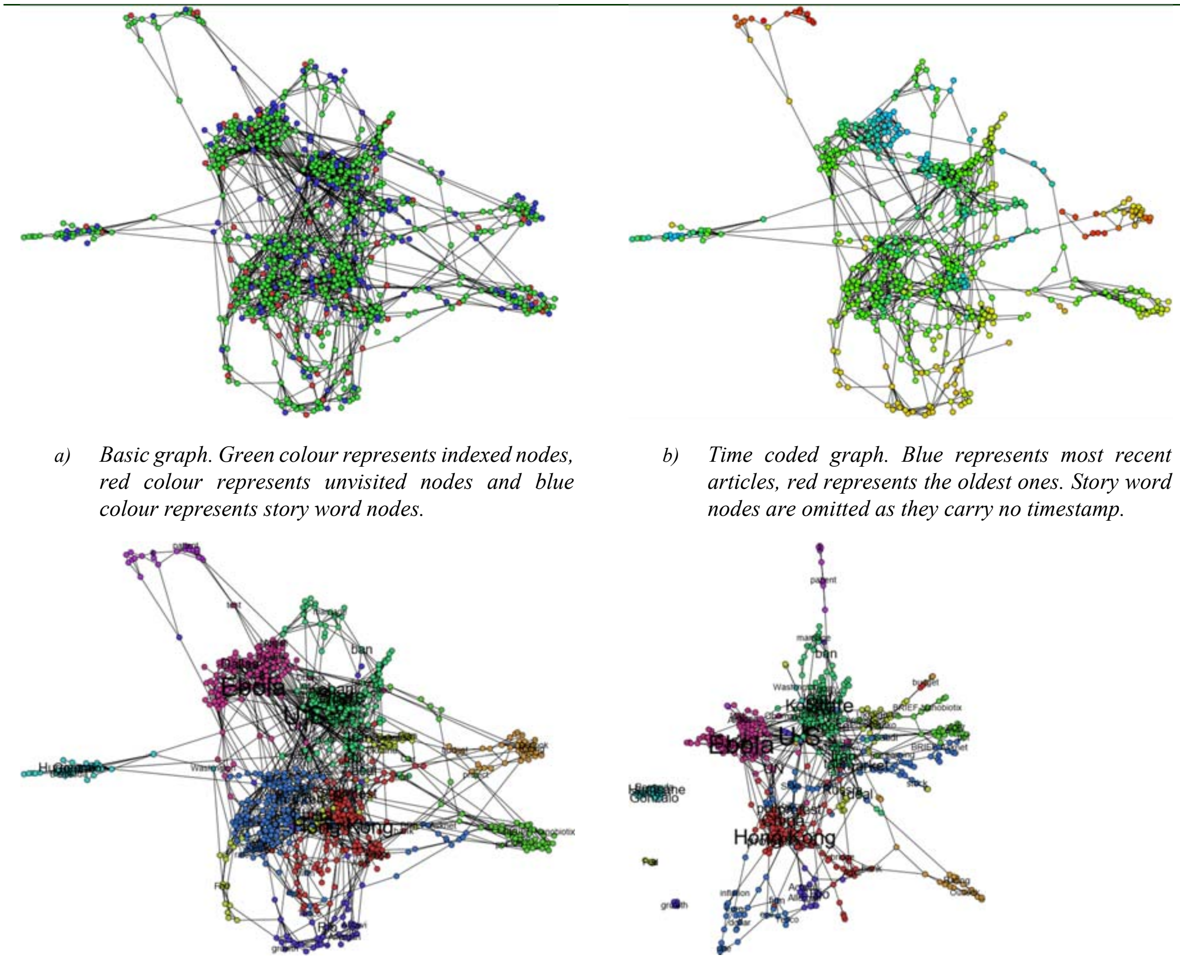


Figure 1: Decision Making Process Of An Agent As Described In Beehive Metaphor By Navrat[4].



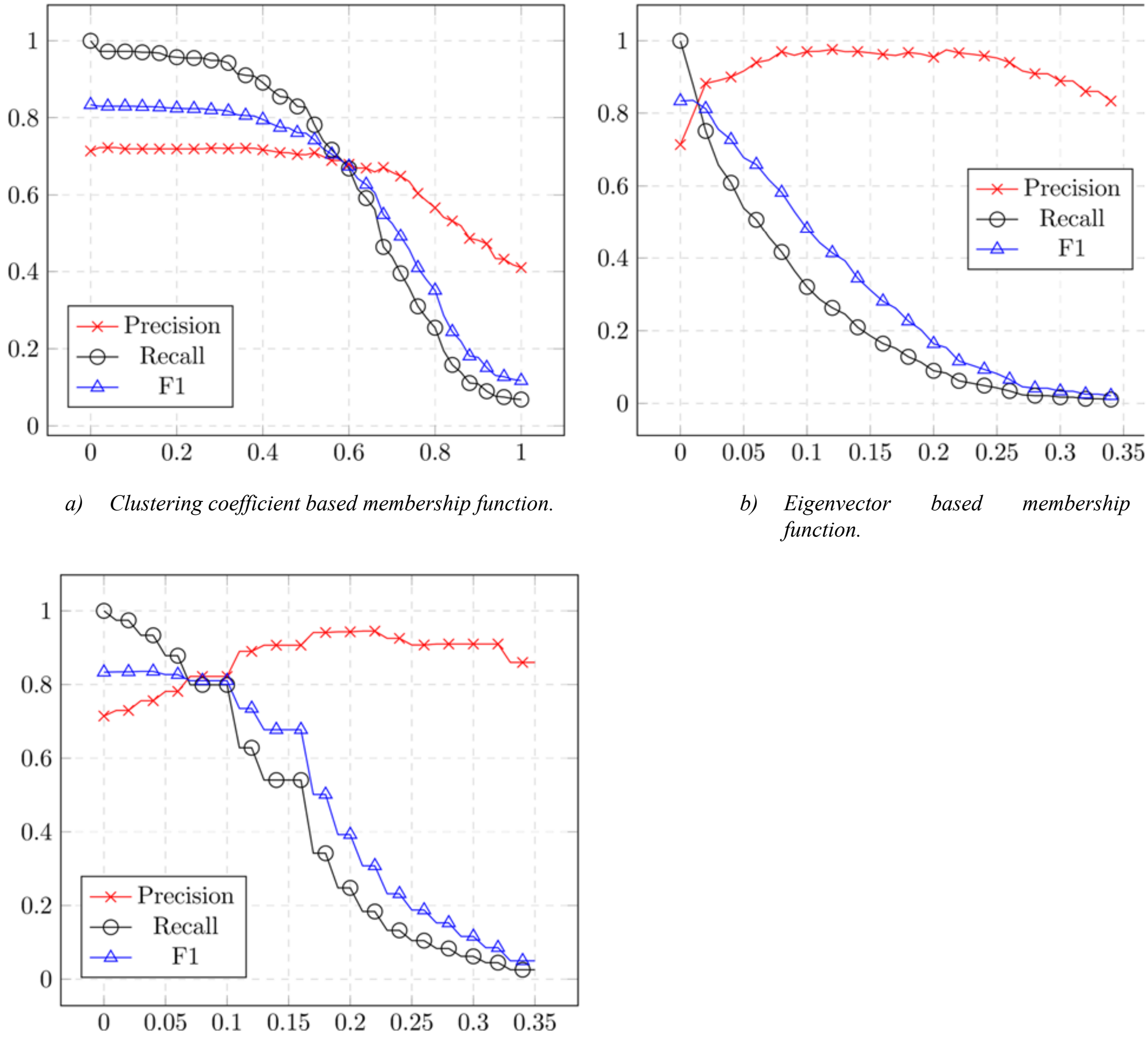
a) Basic graph. Green colour represents indexed nodes, red colour represents unvisited nodes and blue colour represents story word nodes.

b) Time coded graph. Blue represents most recent articles, red represents the oldest ones. Story word nodes are omitted as they carry no timestamp.

c) Story graph. Coloured regions represent identified news stories. Story words show captions scaled according to their popularity

d) Filtered story graph. Hyperlink edges are omitted, all edgesdenote relevance relationship.

Figure 2: Graph Representation Of Article Space.

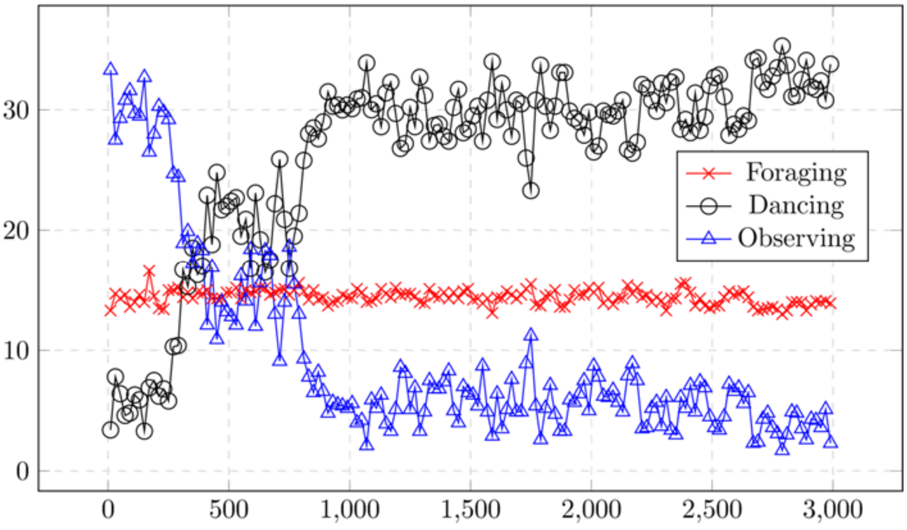


a) Clustering coefficient based membership function.

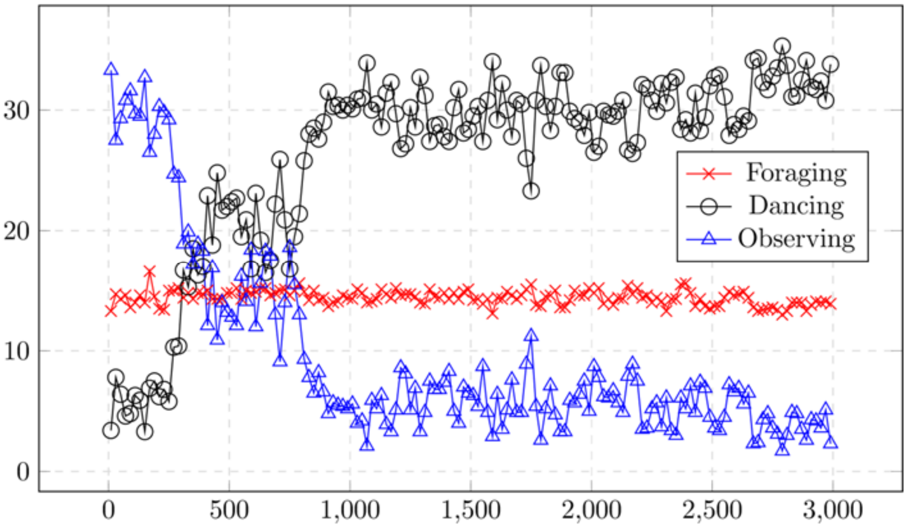
b) Eigenvector based membership function.

c) Node degree based membership function.

Figure 3: Impact Of Various Membership Functions.



a) Three level distribution of agents.



b) Single level distribution of agents.

Figure 4: Allocation Of Agents Among Tasks Over Time.