Towards a Decentralized Search Engine for P2P-Network Communities

Herwig Unger Markus Wulff
Fachbereich Informatik
Universität Rostock
18059 Rostock, Germany
{hunger|mwulff}@informatik.uni-rostock.de

Abstract

The paper considers the data search problem in distributed P2P networks. Since no central search engines are available, other effective methods must be developed to avoid a complete search of all nodes each time. We investigate mechanisms and introduce the main structure and functionality of such a decentralized cooperative search engine. It is working on the basis of so called ants. This search engine reduces the search time for information, which are needed by more than one node in a peer-to-peer network community. Therefore, the ants are able to switch in a special manner between different behavior strategies to search, concentrate and return the respective information. The various strategies are derived from ant colonies and were simulated in a small world community environment.

1. Introduction

Availability, coverage, and timeliness are the most problems with todays search engines, which are the only possibility to locate any contents in the Internet.

Peer-to-peer (P2P) systems like Gnutella [9] and Freenet [4] consisting of a number of nodes, which are built by computers running a certain software. These nodes are able to communicate with each other and hold all available information of the community in a decentralized manner—each one a piece. This information includes the structural information as well as the information about available services and the services itself. The structural information contains the knowledge of the addresses of several other machines in the P2P-network community. These address information connects the nodes with each other and thereby a logical network structure is created on top of the physical one.

Appropriate protocols allow to search and to access all available information in the whole community network. The collaboration of all nodes, which can be clients as well as servers, is a basic precondition for the work within such a system. Therefore, besides processing its own information management and retrieval, every node also offers the execution of basic services for incoming requests from other community members.

Work division and collaboration are the key to further improvements in community based system. The current paper intends to show how the search of information could be done in the given context of a P2P-community more efficiently.

Therefore the paper is structured as follow. The below section 2 discusses information management problems in the todays Internet as well as in P2P-networks with their advantages and disadvantages. Section 3 introduces our new ideas basing on the use of modified ant-like mechanisms for P2P-networks and section 4 shows the respective simulation results. These results are the basis for the introduction of the concept of a collaborative and decentral working search engine in section 5. Section 6 concludes the paper and gives an outlook on further research topics.

2. Related, Internet oriented Search Mechanisms

Nowadays, the "intelligence" of the Internet is built by the information kept on different machines and the links between them. The huge amount of data makes the Internet to the most important medium for the exchange of (HTML-) documents in different languages.

The structure of the WWW is mostly used for navigation proposes, i. e. to "surf" from one document to the other one following the links and maybe recovering the history of a session from a list. In particular, searching the web for a special information becomes a tedious task, since there is no connection between the place, where an information is stored and the information itself. Centralized search engines are the only sufficiently working tool filtering out web-content related to a certain topic. These engines search the web and return the documents that contain certain key-

Proceedings of the Eleventh Euromicro Conference on Parallel,Distributed and Network-Based Processing (Euro-PDP'03)
0-7695-1875-3/03 $17.00 © 2003 IEEE

words provided by the user. However, the working principle of such a search causes the above cited well known problems, which are more detailed shown below.

- **Availability**
  Due to network and server overload and system breakdowns the centralized search engine servers may be not available or may have long response times. Different attacks (e. g. denial-of service) may cause additional security problems for any kind of centralized, but necessary, network instances.

- **Coverage**
  Studies have shown that no search engine covers more than a small part of the web, and the union of 11 major search engines covers less than 50% of the web [12].

- **Timeliness**
  Search engines are often out-of-date partially due to limited crawling speeds and the low average life-span of a web page.

- **Correctness**
  Most search engines return a huge quantity of web pages for a single request at the expense of precision. In addition to that, there is no automatic support to restrict or extend a search procedure by new keywords from the desired target subject.

- **Trust**
  There is no chance to evaluate the trustworthiness or correctness of any information obtained from a web-document and search engines do not contribute to solve this problem.

To find a way to solve some of these problems, a number of new, less centralized algorithms have been developed in the past. These algorithms try to analyze locally the link topology of the web in order to derive a logical, content-oriented structure to the WWW on top of the physical network connections. The basis for doing so is the following. Author of a web document normally include only those links to other documents that are relevant to the general subject of the page. Thus, locating one interesting document may be sufficient to find further information on that issue. High quality documents, that contain clear and useful information, are likely to have many links pointing to them while low quality documents probably have few or no incoming links. Thereby, an implicit preference function is given for the pages. The following algorithms take advantage from this fact: PageRank [2] and HITS [8]. They determine the quality or "authority" of a web page on the basis of the number and quality of the pages that link to it. HITS produces two types of pages: *authorities* (highly referenced pages), that are pointed to by many good "hubs" (lists of web pages) and *hubs*, that point to many good authorities to pull them together. Another approach tries to overcome the described problems of the search engines by using a different way to identify communities which may enable web crawlers to effectively focus on narrow but topically related subsets of the web. Here, a community is defined as a set of web pages that link, in either direction, to more web pages in the community than to web pages outside the community [7].

The disadvantage of these methods is that they are static. They merely use the linking pattern that already exists and do not allow the web to adapt to the way it is used. To achieve this, other implicit information can be used [10]. People who browse the web by navigating from page to page express their preferences by the links they choose. In other words, the frequency of their link selection provides information about their preferences. A number of algorithms have been developed to use this implicit information by "collaborative filtering" [13].

Peer-to-peer (P2P) computing has recently gained an increasing popularity due to the emergence of file sharing systems like Gnutella [9], and Freenet [4], but may also used to make the information management more flexible. The most advantage of P2P-systems is that peers may come and go. Nevertheless, all available peers may be used to manage the information processing in the whole system at every moment. This fact rapidly increase the capabilities of the overall system concerning its flexibility, its reconfigurability and adaptability when user requirements are changing or mistakes appear. In addition, the privacy and anonymity of the users will be kept and a significant growth of the number of machines and users does not automatically result in overloaded networks and servers. In such a manner, the peer-to-peer computing paradigm overcomes a lot of disadvantages known from classical client-server systems and builds a promising computing scheme corresponding to the properties of the todays Internet and the needs for mobile computing environments.

Most P2P-networks deploy in an evolutionary process starting with a small group of at least two individuals and then growing up by adding new members to this initial group. In such a manner, the structure of P2P-networks reflects similar interests of groups users and their relations to each other. Like in the human society P2P-communities emerge and are implicitly defined by the neighborhood relations between all the individuals, which is generally known as *social network* of the community. Even though the building process may seem anarchic, most of those networks are so called *s*mall-world networks. This term traces back to the work of Stanley Milgram, who, performing a number of experiments as early as in 1967, concluded that any two people in the United States could be connected by only a short chain of about six people [14]. This fact is today well

known as "six degrees of separation" and was generalized to a lot of social networks.

Most systems use message chain mechanisms [15, 4, 16] for communication. Hereby, an information will be transferred from a source node to other nodes by sending the data successively from one node to one of its (direct) neighbors until a suitable or the specified target node is reached.

The message contains, besides the data or the requested operation itself, the address of the origin node and a hop counter. The hop counter is decreased every time the message is forwarded to the next node to ensure that this operation is finite. In accordance to its contents, a message can cause any kind of action at every node on its way as well as at the destination node and will be forwarded to a chosen neighbor in each case, unless the hop counter has reached zero. Thus, every node of the network is able to find information at previously unknown nodes and can learn about any other node in the community. Since the forwarding procedure of the message chain is normally a random procedure, every node of the community can be reached with a certain probability. Therefore, a probabilistic search over the whole network system can be realized. Merging of different message chains can be used to increase cooperation effects and to keep the network load low [15].

Although the network load is almost equally distributed over the whole system, response times maybe high through longer and repeated message chain operations. Furthermore, the larger the respective P2P-network communities are the longer it may take to find the respective information. Therefore a few works like [11] trying to identify communities from the WWW or to divide large communities into smaller sub-communities.

Summarizing can be said that there is a need for research in the area of information management in decentralized P2P-communities.

3. Collective Network Discovering

3.1. General Remarks on Community Information Management

Whenever a set of autonomous individuals acts in a given, shared environment, mutual influences and interactions emerge, if a number of (limited) resources is used by the individuals. It is a well studied effect that these dependences may result in manifold relations between two or more individuals and may also yield to a communication and/or a hierarchical structuring within the groups. Various sciences have studied the social behavior of biological individuals, like ants, bees and even people [1]. Thus concluding that although the intelligence of a single individual may vary in different environments and systems [10], some equal characteristics can be found in all those groups of individuals acting in a shared environment after some time.

1. Every single individual can personally identify a few members of the whole group and can exchange information with these members.

2. No single individual has the capability to know or to control all the activities or structures of the whole group.

3. The group of individuals can be smart in a way that none of its members could.

4. Most groups are structured in a hierarchical manner, with one individual at the top directing the activities of the other individuals.

Using this and the above said facts about the Web, the key to solve a lot of information management problems seems to be in the following statements.

- Every node in the network can store a given amount of data, only. In addition, it can only know a few other computers.

- An increased cooperation between different machines (single individual) and a respective work division must be organized between the machines through respective algorithms. New information, for example, might be shared with other community members.

- Idle network and computing facilities must be used to organize the structure and functionality of the machines and to built the *social memory* of the community. For instance, new information can be propagated in the whole community or in parts of it.

It is known that an ant collective behaves similar to the self-organizing systems studied in physics and chemistry [1]. Very large numbers of simple components interact locally and may produce a global organization and adaptation to its environment. The same effect might also appear in computer systems, in case the respective mechanisms are installed in the right manner there.

From the above said it becomes clear that information management means to find a balance between the propagation and the search of information. In such a manner, a decentralized search engine could be understood as a global system containing the following three components.

1. A parallel working search mechanism exploring the whole community space (e. g. a set of decentral controlled agents).

2. A decentral concentration process for information, which seems to be interesting for more than one machine. Hereby, the information shall be stored on a set of machines which are distributed over the whole community.

3. A global known mechanism (e. g. a number of recognizable paths) directing any user to the points, where information is concentrated.

The following parts of the paper intend to describe the necessary basic mechanisms to solve these three problems in the context of the WWW.

3.2. Message Chains, Wanderer and Ants

In several articles the behavior of ants was described and later used to solve computer science related problems [6, 3]. We also believe that similarities to the ant behavior may help us to solve information management and search problems in P2P-network communities.

It was found that the medium used to propagate information among individuals consists of pheromone trails. A moving ant leaves a fixed amount of pheromone on its path. Each pheromone marking decays exponentially with time.

An ant is able to detect existing trails and follows the trail having the highest pheromone concentration with a very high probability. In case no trail is detected, the ant moves randomly. Through this mechanism existing trails will be reinforced, but an exploration of new trails is also possible.

The most similar correspondence to the movement of ant along a trail is given in P2P-network communities by the message chains [15, 4]. These chains follow a route given through the links in neighborhood relation of every node of the community. If the hop-counter of such a message is set to ∞ (realized by a value ≤ -1), a message chain is transferred into an endless walking *wanderer* through the community. Since the origin of the chain is stored within its data, optionally a periodic or random return to this node may be selected.

Pheromone trails can be easily implemented by adding additional attributes to each node and each entry stored in the neighborhood warehouse. Each node can simulate the decay of the pheromones and update the value in case an arriving message chain is processed and forwarded. Since one individual does not have a significant influence on the behavior of the whole community, manipulations of these values forced from the (egoistic) interests of a single node do not disturb the system too much.

In such a manner, a wanderer is a pretty good simulation of the behavior of an (biological) ant within the context of our P2P-network community. The following subsection describes the respective simulation experiments more detailed.

4. Experiments with Modified Ants

4.1. Simulation Environment

In order to obtain reproducible results the developed algorithms were considered within a simulation. The goal of the simulation was it to show that algorithms emulating the behavior of ants meet the requirements of an information management system introduced in Sec. 3.1.

Following the survey given in [5], the P2P-network community was modeled as a graph. Since we are not interested in growth processes of such communities yet, the edge-reassignment method was used. This method generates a *small world* network with a fixed number of nodes. The small world term denotes a set of clusters which are highly intraconnected whereas the connection between the clusters is comparatively weak.

In order to have a good compromise between a possible real community sizes and reasonable simulation times, a community size of 2048 members was selected. From experience can be said that in networks with more than 2048 nodes the intended effects appear even better.

Every neighborhood warehouse of a node may obtain (suitable to the number of nodes) 16 entries, i. e. 16 possible ways (trails) exist to leave any node. In addition the probability for the edge reassignment was set to 0.4.

The generated graph was kept and used for all (below described) experiments. For the decay of pheromones a time constant of 0.02 gave the best results.

Finally it was decided to eliminate the influence of the existing network bandwidth from our experiments at this moment. Therefore, it was assumed that the transfer of any wanderer between any two node takes one time unit. Due to the high robustness of the described methods all mechanisms also work in a network with limited bandwidth. A computation of the available bandwidth within the pheromone attributes of every node may even improve the efficiency of the developed mechanisms.

All simulations were run over 10,000 time steps in order to ensure stabilized conditions inside the simulation environment.

4.2. The MinimumAnt

If the information search is considered at first, a maximal number of ants is needed, exploring all possible nodes of the community in a parallel manner. In order to ensure that

- all nodes are visited,
- no node will be seen twice or oftener and
- network load shall be equally distributed,

the first modified ant shall follow the link with the lowest pheromone concentration. This ant shall be called the *MinimumAnt*. Figure 1 shows that such an ant really visit all nodes of a given network with almost the same probability. The figure plots the record of the path of a single ant through the community network.

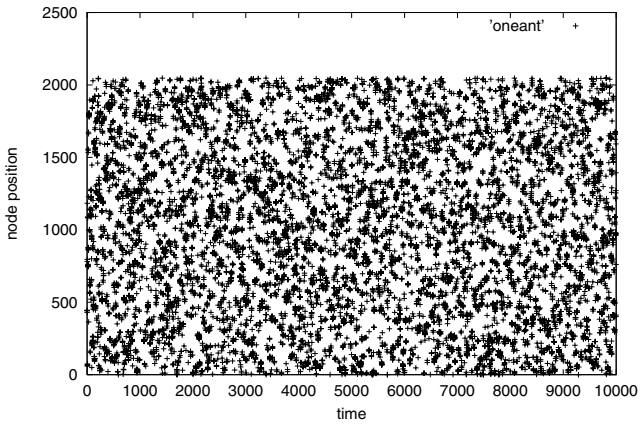

Figure 1. The behavior of a MinimumAnt.

Note that there is an indirect communication/coordination between all ants through the commonly used pheromone system.

4.3. MaximumAnts

In a next basic task, ants should put the collected information on a set of machines. These machines should be determined by the ants themselves. Following the example from the nature, the ants may put their information along a strong trail, i. e. a path which exceed a given pheromone concentration. On such a trail, the selection of the concrete machine is done randomly. While doing so, a lot of ants are able to find this data later.

A strong trail will easily be found from any point of the network by a so called *MaximumAnt* following the highest pheromone concentration. Furthermore, this trail will also be reinforced by walking on it (see Fig. 2). Because of the small world structure of our P2P-network community, the ant remains with a high probability in the stronger connected environment of a local sub-area.

Differing to completely centralized search engines, the loss or breakdown of a few machines along a strong trail of course do not result in problems for the whole P2P-network.

Although the MaximumAnt may concentrate information on a self-selected number of nodes, problems can arise from overloaded trails and respectively overloaded network paths in the P2P-communities. Therefore, it makes sense to modify the MaximumAnt. For this purpose a value *limit*

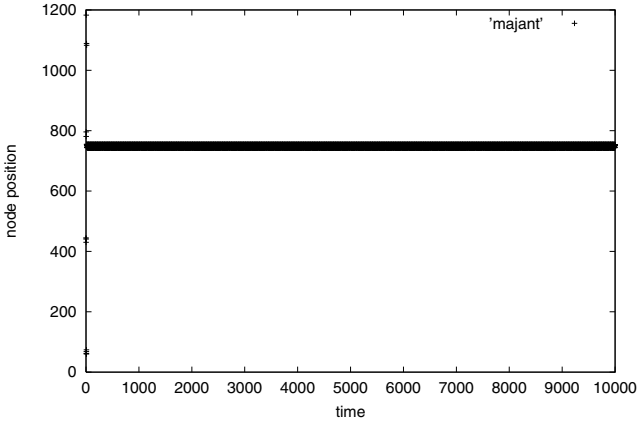

Figure 2. Behavior of a typical MaximumAnt.

is introduced describing the maximal pheromone level for a path. An ant must not follow this trail if the pheromone concentration exceeds this limit. This ants should be called *Maximum-2-Ants*.

Because a high pheromone concentration emerge when walking in a local sub-community, now the strong trail of the Maximum-2-Ant is normally distributed over a set of local sub-communities (Fig. 3).

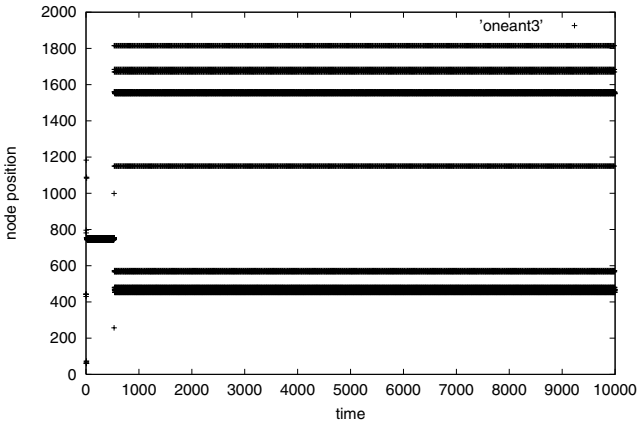

Figure 3. The behavior of a Maximum-2-Ant.

Since the pheromone decays with the time, all the trails become cycles with a high probability. The ants will follow a strong trail again, if the concentration of the pheromone is fallen below the given *limit*. The length of this cycle depends for the Maximum-2-Ant directly from the selected *limit*. Figure 4 shows that a smaller *limit* will cause a longer cycle. The reason is that an ant with a small *limit* avoids the return to a strong 'smelling' part of the path for a longer time than an ant with a higher value of *limit*.

Proceedings of the Eleventh Euromicro Conference on Parallel,Distributed and Network-Based Processing (Euro-PDP'03)
0-7695-1875-3/03 $17.00 © 2003 IEEE

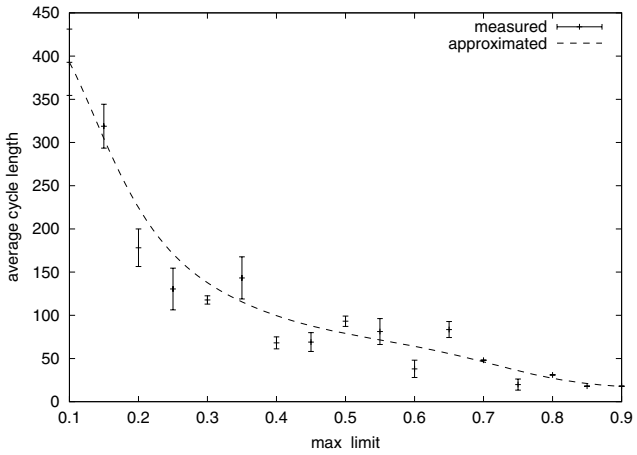

Figure 4. Length of the cycle of a Maximum-2-Ant depending on the value of 'limit'

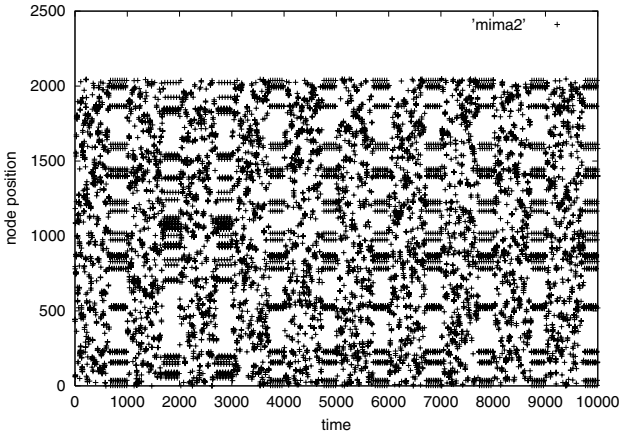

Figure 5. The MinMaxAnt (700 steps working as MinimumAnt followed by 300 steps MaximumAnt)

4.4. MinMaxAnts

The outcome of the above is that the combination of the behavior of the MinimumAnt and the Maximum-2-Ant leads to a mechanism to for an efficient information search in the P2P community.

In a first period an ant must collect information until it found a certain amount of data. In this time the ant can also distribute the collected information over the visited nodes. Then, each ant must concentrate the gathered data for collaboration proposes on a set of machines which is often visited by other ants.

To fulfill this task, the *MinMaxAnt* was created. Figure 5 show its behavior. After a (fixed) period of data collection and distribution the ant concentrate them on selected machines along a trail with a high pheromone concentration. It is to be seen that this trail is selected slowly and changes often in the beginning of the simulation. After a few periods the selected trail is very stable and can be recognized by other ants too.

The positive effect in a community of ants (see Fig. 6) is that this self-stabilization effect works much faster the more ants exist in the system. In addition, the number of visits of every node will also increase with a higher number of ants.

5. Concept of a Decentralized Search Engine

In the above section 3.1 the principal components of a decentralized search engine have been given, which can be realized in the context of the web.

Wanderer are the main instances in the data search and concentration process. These wanderer may behave like the different types of ants depending on the respective situation. In addition, every machine of the community is able to store and process information corresponding to the pheromone concentrations of the different paths as well as a limited number of (external) information brought by the wanderers.

The data search procedure works now as shown in Fig. 7. If a special information is searched by a node, it generates a new wanderer. This wanderer follows at first the Maximum-2-Ant strategy. Therefore, the wanderer will find the nodes of an already established trail with a significantly high pheromone concentration or take part in the building process of such a trail. The nodes of this trail are acting like an information cache. If the search on these nodes is already successful, the wanderer directly return the searched information to its origin.

Otherwise, the wanderer changes to the minimum ant strategy after some time. Now it explores the whole community and tries to collect as much as possible pointers to nodes with the respective information. Due to the indirect communication over the pheromones the minimum strategy ensure that parallel working wanderer are equally distributed over the whole community. The duration of the search period in general is limited by the capacity of the wanderer and a search time limitation (timeout).

When this search period is over—due to a time out or the limited storage capacity of the wanderer—the wanderer switch back to the Maximum-2-Ant behavior and returns to the trail with the high pheromone concentration. Now, the wanderer looks for a fixed time for a free place where the found information can be put. In addition, the wanderer stores the location of the oldest unused information during this time. In case no free place could be located, the wanderer stores the found information on the place with the oldest unused information (realizing a Last-Recently-Use-

Proceedings of the Eleventh Euromicro Conference on Parallel,Distributed and Network-Based Processing (Euro-PDP'03)
0-7695-1875-3/03 $17.00 © 2003 IEEE

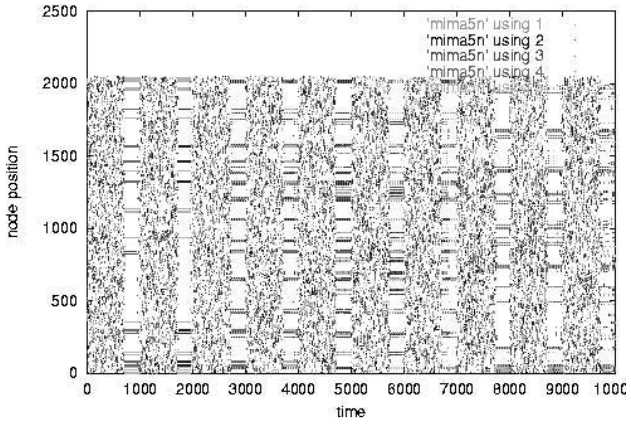

Figure 6. The paths of 5 MinMaxAnts in a common environment

(LRU) replacement strategy).

Last but not least it returns the results to its origin, where a decision can be made whether the search shall be continued/repeated or terminated.

It should be mentioned, that the trail with the high pheromone concentration acts like a cooperative, social memory of the whole community and the decentralized search engine. Information, searched by different machines or with a higher frequency, can be found along the trail with a higher probability. The less machines are looking for a special information the higher the probability is that the respective search must be performed as a search over the whole distributed community. This is an analogue to the social behavior, which can be found in the human society: a community keeps mostly the information a broad majority is interested in, while outsider have to invest more to obtain their needed information.

The building procedure of the high concentration pheromone trail must be discussed more detailed in the future. It is known from our simulations that the trail is built and its size will be adapted to the parameters of the community. In addition, any faults of machines along the trail will be automatically tolerated.

However, in our experiments only one trail emerges at every time. Nevertheless, special (irregular) community structures may cause in an emergence of several, non-connected trails. This phenomenon may be used to identify and to separate sub-communities.

6. Conclusion and Future Work

Ant-like mechanisms can be used for the information management in P2P-network communities. Especially the data search can be supported by a de-central search engine,

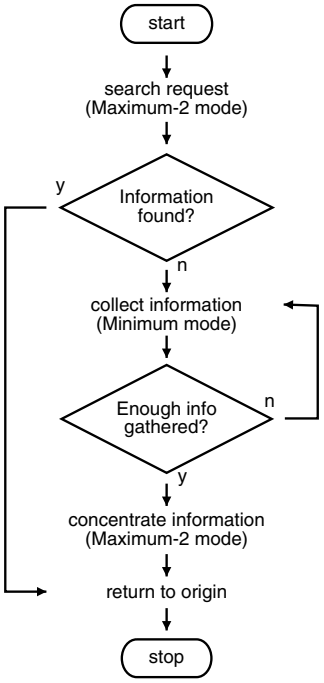

Figure 7. Flow chart of a search process.

which can offer a higher availability, coverage and actuality than known centralized approaches.

Therefore, the behavior of an ant was modified and considered in several experiments. It become clear that a combination of different strategies within the wanderer (i. e. the realization of a ant in a P2P-network community) can be used for a two period data search. In a first period information will be collected. In a second phase the information will be concentrated for cooperation proposes along a trail over a well-defined set of machines. The built trail is organized and maintained by the community itself as well as the storage of data on the contributing machines. So, the emerging pheromone trails act like a social memory and are the basis for some cooperation effects within the population. The developed methods are very robust and fault tolerant.

In the present contribution we have considered the basic mechanisms for a cooperative, completely decentralized search engine only. To show the effects more clearly, the time behavior of the ants/wanderers was fixed and no bandwidth problems were considered.

In the future, improvements probably could be achieved by an increased intelligence of the ant/wanderer. Therefore, more suitable environment (network) parameters must be investigated and the respective learning and adaptation mechanisms and strategies must be developed.

References

[1] E. Bonabeau, G. Theraulaz, J. Deneubourg, S. Aron, and S. Camazine. Selforganization in social insects. *Trends in Ecol. Evol. 188–193*, 1997.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[3] G. D. Caro and M. Dorigo. AntNet: A Mobile Agents Approach to Adaptive Routing, 1997.

[4] I. Clarke, O. Sandberg, B. Wiley, and T. Hong. Freenet: A Distributed Anonymous Information Storage and Retrieval System. In *ICSI Workshop on Design Issues in Anonymity and Unobservability*, Berkeley, CA, 2000.

[5] N. Deo and P. Gupta. World Wide Web: A Graph Theoretic Approach. CS TR-01-001, University of Central Florida, 2001.

[6] M. Dorigo, V. Maniezzo, and A. Colorni. Ant System: Optimization by a Colony of Cooperating Agents. *IEEE Trans. on Systems, Man, and Cybernetics–Part B*, 26(1):29–41, 1996.

[7] G. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August 20–23 2000.

[8] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

[9] Gnutella. www.gnutellanews.com, 2001.

[10] F. Heylighen. Collective Intelligence and its Implementation on the Web: Algorithms to Develop a Collective Mental Map. *Computational & Mathematical Organization Theory*, 5(3):253–280, 1999.

[11] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *WWW8 / Computer Networks*, 31(11-16):1481–1493, 1999.

[12] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400:107–109, 1999.

[13] V. Menko, D. J. Neu, and Q. Shi. AntWorld: A Collaborative Web Search Tool. In K. et al., editor, *Distributed Communities on the Web (DCW)*. Springer Verlag Berlin, 2000.

[14] S. Milgram. The small-world problem. *Psychology Today*, 1967.

[15] H. Unger and T. Böhme. Distribution of information in decentralized computer communitites. In A. Tentner, editor, *ASTC High Performance Computing*, Seattle, Washington, 2001.

[16] H. Unger, P. Kropf, G. Babin, and T. Böhme. Simulation of search and distribution methods for jobs in a Operating System (WOS). In S. A. Tentner, editor, *ASTC High Performance Computing*, pages 253–259, Boston, 1998.

[17] M. Wulff. A decentral library for scientific articles. In H. Unger, T. Boehme, and A. Mikler, editors, *2. Conference on Innovative Internet Computing Systems*, pages 153–162, Kuehlungsborn, 2002. Springer Verlag Berlin.

Proceedings of the Eleventh Euromicro Conference on Parallel,Distributed and Network-Based Processing (Euro-PDP'03)
0-7695-1875-3/03 $17.00 © 2003 IEEE