Improving Flexibility of the Bee Hive Metaphor for Web Search by Adding New Parameters and Modifying Dispatching Strategy

Robert Čapla*
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
capla@fiit.stuba.sk

Abstract

In this paper we introduce improvements of bee hive metaphor model for web search. We identified critical areas in this model and designed several methods to improve behavior in web search. Then we compared experimentally these new methods with existing solutions and original model. Performed experiments demonstrate that both behavior of the bee hive and results of web search have been improved.

Categories and Subject Descriptors

I.2.8 [**Distributed Artificial Intelligence**]: Multiagent systems, Intelligent agents; H.3.3 [**Information Search and Retrieval**]: Internet search

Keywords

information search, Bee Hive At Work, bee hive, foraging algorithm

1. Introduction

Information and number of web pages on the Internet is growing faster every day. It is important for users, that information search returns relevant information, which are up to date.

User's need for actual and relevant information is mostly provided by few popular search engines. All of these search engines do not search directly on the web in real-time, but work with large database, in which all indexed information about individual web pages are stored. This web search, although very popular nowadays, has some minor drawbacks. This includes smaller actuality of indexed web pages or too many irrelevant results. To avoid these problems, we need to make searching more real-time and precise.

By using real-time search engine we should be able to get more recent and actual results. Real-time search engine based on static search engines would be unable to use, because of enormous amount of time spent on downloading web pages. It is important for real-time search engine to use some kind of heuristics that assure to limit downloading web pages as much as possible, but with sufficient results. This can be achieved by using focused web crawler. Focused web crawlers ensure that not all web pages are downloaded, but only on most perspective paths.

In following section we describe related work that deal with different focused web crawlers and online web search engines. Later we describe important features of basic bee hive metaphor model and some existing improvements. In section 4 we introduce new improvements of this model, followed by performed experiments and achieved results. At the end is conclusion, evaluation and future work.

2. Related Work

Main difference between static and online search is that in online search all documents are downloaded during the search, usually after entering the query by user. Crawling through web pages takes most of running time. Therefore online search engines try to download as little web pages as possible. Most common way of limiting crawled pages is using focused web crawler. Focused web crawlers are crawlers that do not download all found web pages, but its target is only relevant pages to defined topic or entered query. They ensure that only perspective sources with high probability to be relevant are downloaded. This behavior shortens time of downloading web pages, which is suitable for real-time web search engines.

Some focused web crawlers are inspired by behavior seen in nature. The idea of designing algorithms inspired by nature is not new. Early attempts include fish search [1] or shark search [3]. In these models individual agents represent fish, which reproduce in area with high density of food. If the area, in which the agent is located, is poor for

*Doctoral degree study programme in field Software Engineering. Supervisor: Professor Pavol Návrat, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, STU in Bratislava. Work described in this paper was presented at the 7th Student Research Conference in Informatics and Information Technologies IIT.SRC 2011.

Čapla, R. Improving Flexibility of the Bee Hive Metaphor for Web Search by Adding New Parameters and Modifying Dispatching Strategy. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies, Vol. 3, No. 2 (2011) 32-36

food, this agent dies. Difference between fish and shark search is in evaluation of page relevance. Fish search uses binary, shark search is using continuous evaluation. More complex models are Focused Ant Crawling Algorithm [2] or bee hive search [6]. Focused Ant Crawling Algorithm is algorithm for hypertext graph crawling, which significantly limits amount of downloaded web pages. Model of bee hive search is described in the next section.

Infospider [5] is multi-agent system used for online search. When agent moves from document to document, it chooses link to follow. They are able to adapt to changing environment using neural network and evolution of agents.

3. Bee hive metaphor and Bee Hive at Work model

This article focuses on improving existing model, called Bee Hive at Work (BH@W), which was presented in [6]. This model was used for web search.

A bee hive in this model consists of three main areas: dispatch room, dancing room and auditorium. Dispatch room serves as repository of all known sources (that have been found and visited by bees) and assign random source to each bee flying out. Dancing room and auditorium are used for communication between bees which communicate via waggle dance.

At the beginning all bees are in the dispatching room. Each bee chooses random source from initial sources and visits this source. After visiting chosen source, each bee evaluates this source and depending on its quality chooses to keep this source or abandon it. When source is abandoned, bee flies to auditorium. If bee chooses to keep the source, it has two choices: (a) return to hive and perform waggle dance for this source or (b) continue exploitation from the source.

This model is used for web search, sources are meant as web pages. Quality of web pages presents relevance to user's query. Quality calculation is based on simple concept: it is divided into three components (distance quality, header quality and count quality). Each component can reach maximum value according to parameter. Sum of these parameters is equal 1, which is maximum overall quality. In web search exploitation of source means following random link from given web page.

3.1 Existing improvements

Several improvements were proposed for model described in section 3. Some improvements were designed to improve results of web search and behavior and others were used to minimize identified problems.

Problem with low promotion of new sources was solved by introducing bee scouts [7]. Foraging bees are split into two categories depending on their success in following dancing bee – scouts and recruits. Because scouts did not find a dancing bee to follow, they search for new sources. On the other hand recruits follow other bees from dancing room. Change is in behavior outside the hive, when scouts automatically dance for found source, unlike recruits, which uses quality of source as probability of dancing and promoting this source to other bees.

Other modification introduced energy concept [4] instead of distance quality component in quality computation. The bee is searching for new source until it has enough energy. When the bee is out of energy it has to return to the hive and follow a bee from the dancing room.

To avoid problem with convergence to local maximum, desirability [8] was added to the model. Desirability is new property of source, similar to quality. It describes how desired is source by bees. The higher is desirability, the more bees will exploit this source.

4. Model improvements

In previous sections we described behavior of basic BH@W model and several improvements and modifications of this model. We created basic model and used this model for several tests, both change quality test, described in [9] and real-time web search on web. During these tests we discovered two important areas of model that are responsible for main functionality of model. If these areas are not designed properly, model is not working as a real bee hive in nature. These two areas are (a) getting trapped in local maximum and (b) dispatch room strategy.

Both areas were studied and several methods to each issue were proposed. We experimentally compared these methods with each other and methods introduced in section 3.1.

4.1 Getting trapped in local maximum

When using bee hive model for web search, most of bees often promote one or only small number of sources. Main reason for this behavior is, that in web search experiment there are too many sources when compared to limited experiment. While in limited experiment we had tens or hundreds of resources, there are thousands or even millions of different sources in web search. Most of these sources have zero quality, because they are not relevant to user's query. Important note is that solution space is unknown, i.e. bees do not have access to all possible sources at the beginning of the experiment. Usually they start with some known sources (explicitly entered by user before start of experiment) from which they continue searching. If bee hive finds a source with very good quality (but not the best one), most of bees will promote this source among other bees and in response more bees will exploit food from this source. The more bees exploit a high quality source, the more bees will promote this source by waggle dance. This will cause very strong promotion of this source, which is local maximum. Result of this promotion may be inability of finding other high-quality sources that are more distant from primary sources or other high-quality sources.

It is necessary to modify this model to be more flexible and more adaptive to possible changes, which can occur during experiments. We proposed several improvements to each mentioned issue and compared them with other methods:

1. Adding inaccuracy to the model 2. Progressive exploitation of sources 3. Adding desirability of sources to the model 4. Quadratic decreasing of desirability of sources

First method was introduced in [5]. It contained two new parameters, information noise and error of quality evalu-

ation. Adding of these parameters has positive effects on dynamics and speed of finding new sources, because bees fly to random sources more often.

Second method is inspired by collecting food in nature. Each source contains only limited amount of food. By each visit of a bee this amount decline. The less food a source contains, the less qualitative is. The food replenishes periodically in time.

Third method was introduced in [8]. Desirability indicates how much the source is demanded for bees. The more bees collect food from a source, the less is this source desirable.

Quadratic decreasing is improved desirability calculation mentioned in previous method. Decreasing of desirability was linear, which was not proper for sources which were promoted by only a few of bees. Quadratic decreasing of desirability is more suitable, where size of the reduction is higher when more bees are promoting one source.

4.2 Dispatch room strategy

Problems associated with dispatch room strategy have various forms. These problems are also associated with source quality calculation. If starting source has zero quality and quality of sources is rated only by occurrences of query in content of web page, then model will find no new sources. The reason is that bees have no motivation to exploit or promote zero quality sources, because bees will never fly to new sources from zero quality sources. Two solutions to this problem are possible: (a) changing source rating strategy or (b) changing dispatch room strategy.

We changed rating of sources compared to original model. New quality calculation contains three quality components: URL quality, headers quality and count quality. Compared to old model, only difference is replacing distance quality for URL quality. Distance quality was not used, because the distance of source from starting page has no impact on actual quality of this source.

Problems caused by adding or removing distance quality are associated with behavior of dispatch room. It is common that solving one problem will cause problem in other part. By behavior of dispatch room is meant strategy of adding sources to dispatch room and strategy of selecting source by a bee. Five methods were proposed solutions to dispatch room strategy problems:

1. Adding all discovered sources to dispatch room
2. Adding only high-quality sources into dispatch room
3. Leaving sources with low quality
4. Dispatch room sources weighting
5. Energy concept

First method changes way of adding of sources to dispatch room. Instead of adding only visited sources, all discovered sources are added to dispatch room. In this case web search continues, even if starting sources have zero quality.

To avoid overload of dispatch room with zero quality sources or poor quality sources, second method was presented. It ensures, that only sources referenced by other high quality sources (selected by threshold quality) are added to dispatch room.

Other method to avoid source overload in dispatch room is leaving sources. It is unlikely, that if several levels of followed sources have poor quality (lower than threshold), they will link to high quality source. Therefore we can stop search on such source and redirect effort of bees to more promising areas.

Fourth method is changing way of selecting source from dispatch room. In basic model was selecting purely random, we introduce weighting sources inside dispatch room. The higher quality has a source, the greater is the chance of selecting it by a bee.c Fifth method was introduced in [4] and is similar to third method. Every bee has initial energy needed for searching of new sources and spends this energy on flying to new sources. If all energy is spent, bee must return to the bee hive (abandon current source).

5. Experiments

We designed two types of experiments to compare newly introduced methods with basic model and other improvements described in section 4: (a) simulated web search and (b) real-time web search. Each experiment was designed for one problem area.

5.1 Simulated web search

First experiment was designed to simulate information search on the Internet. Experiment is inspired by the one performed in [9]. There were only 2 sources in original experiment, with different qualities and in the half of experiment source qualities were switched. Aim of this experiment was to find out, whether the bee hive can determine best source after the change. Compared to original experiment, we added many zero quality sources to simulate irrelevant sources during web search. The change of qualities simulates behavior in web search, when a bee finds better source and other bees should adapt to this change.

The solution space in this experiment contains 100 sources, from which first four sources had following qualities: 80%, 60%, 40%, 20%; and the rest (96) are sources with zero quality. In the half of experiment (500 iterations), quality of one source with zero quality is changed to best – 85%.

We expect that newly introduced methods will be able to recognize highest quality source after the change among all sources. Also, response to change should be more quickly and more dynamic than basic model or models with other improvements. Expected behavior is that in every time the source with highest quality is most promoted in the dancing room. According to the experiment, in the first half of experiment 80% source will be most promoted and after the change[1], most promoted source will be the new best source – 85% source.

We also added model with combination of two methods to the comparison: adding inaccuracy to the model and quadratic decreasing of desirability of sources. These two

[1]Order of downloading the best source

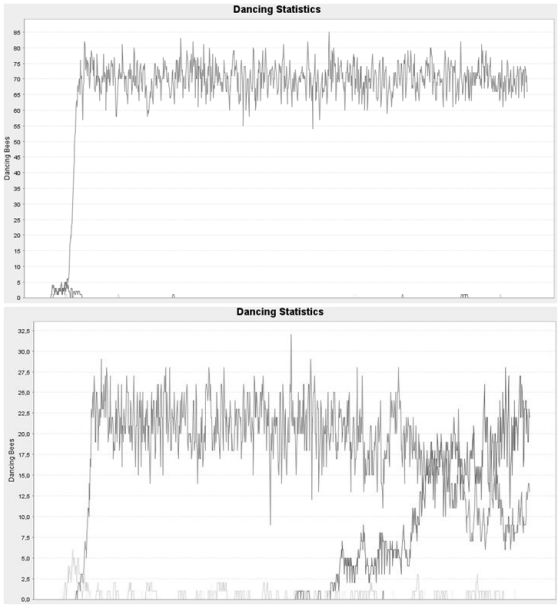

Figure 1: Dancing process of basic model and model with added desirability of sources.

methods do not have opposite effect, so effect to improvements of results should be even better.

5.1.1 Results

In this experiment we observe behavior of bees after change of the best source. Optimal behavior of bees in this experiment is: (a) correct and fast respond to change of best source, (b) sufficient promotion for less qualitative sources and (c) amount of dancing bees is proportional to quality of sources.

In the Figure 1 we can see dancing process of two models: basic model on the left and model with added desirability of sources on the right. Basic model was unable to identify best source after the change. The only source that is promoted in the dancing room by bees is 80% source. Even after the change bees will not promote the best 85% source. Promotion of sources with lower quality was insufficient.

When desirability was added to the source, behavior of bees was more similar to expected behavior. More sources were promoted during whole experiment and the model did respond to change. The speed of response was not as fast as expected and promotion of the 80% source at the end of the experiment is poor. It seems this model can promote only one, although the highest quality source.

In the Figure 2 we can see dancing process of model with two combined methods. Model was able to recognize new best source after change and bees were strongly promoting also second best source. Promotion of all non-zero quality sources was strongest among all models. Every source with non-zero quality was promoted at least by one bee.

According to results of experiments the best behavior is achieved by combination method of adding inaccuracy and quadratic decreasing desirability of sources. All three conditions of optimal behavior mentioned earlier were met by using model with combination method. From individual methods most effective was quadratic decreasing of sources.

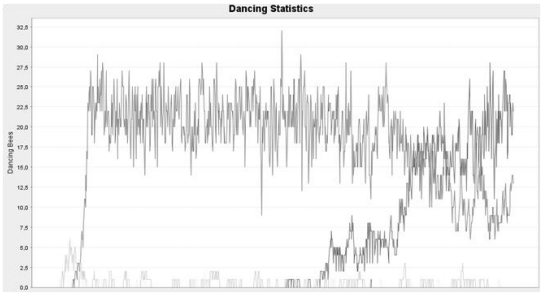

Figure 2: Dancing process of model with combined methods.

5.2 Real-time web search

Second experiment was designed for monitoring behavior during real-time search on the Internet.

Solution space is unknown, because sources are pages on the web and are downloaded during the experiment. For simplification the domain was limited in this experiment to only one domain – http://www.sme.sk (Slovak news portal). Maximum limit of downloaded web pages was 2000. Such limitations are appropriate, because the main aim is to find relevant web pages as soon as possible, with downloading minimum web pages.

Starting web page was http://www.sme.sk; entered query was "Island". The search was made in time of natural disaster on Iceland, when a volcano erupted and whole Europe was affected by this disaster. Starting page contained entered query. Since the topic of search was very popular in the time of experiment, many pages were relevant to query.

We expect that changing dispatch room strategy will cause improvements of the results, both quality of retrieved web pages and behavior of bees in promotion of sources. It is necessary for successful search that the starting source has non zero quality. In other case we need to change behavior of methods, which will stop searching.

For better results we added a combination method to this experiment: adding only high-quality sources into dispatch room and dispatch room sources weighting.

5.2.1 Results

In this experiment we observe dancing for the high quality sources and quality (relevance) of found web pages. Optimal behavior is dancing proportionally to quality of sources and especially ability to adapt to change of source quality. Most important in evaluation is quality of found sources.

Table 1 illustrates time consumption and memory usage during experiments. First two models were both time and memory less demanding, because of less downloaded pages. Most demanding was method adding all sources do dispatch room, mostly because of large number of sources in dispatch room. Method leaving low quality sources did not find the best source.

Table 1: Time consumption and memory usage in real-time web search.

Method	Duration (s)	Downloaded pages	Discovered pages	Iteration [2]	Order[3]	Last source [4]
Basic model	228	328	10418	22	53	1000
Weighting	318	332	10712	51	175	1000
All sources	2038	2000	35027	2	58	101
HQ sources	1509	2000	33165	22	680	128
Leaving LQ	1225	2000	34428	N/A (77)	N/A (1402)	114
Combination	1551	2000	33932	91	1015	216

Table 2: Results of real-time web search.

Method	Retrieved relevant (RI)	Retrieved (I)	Precision	Recall	F1	Average quality [5]
Basic model	11	14	78.6%	15.3%	0.26	62.31%
Weighting	10	12	83.3%	13.9%	0.24	59.77%
All sources	9	13	69.2%	12.5%	0.21	60.24%
HQ sources	26	33	78.8%	36.1%	0.50	71.13%
Leaving LQ	7	10	70.0%	9.7%	0.17	55.52%
Combination	47	51	92.2%	65.3%	0.76	70.53%

In Table 2 is presented comparison of efficiency and success rate of all methods. Total amount of downloaded pages was 6909 (D). All pages with quality more than 50% were identified as relevant (I). Relevance of documents was inspected manually.

Results show that most successful method was combination of methods adding only high quality sources to dispatch room and weighting sources in dispatch room. Most of relevant sources were found and the best web pages were retrieved by using combined method.

6. Conclusions and future work

Main limitation of online search is long response time. It is caused by the time spent on downloading web pages, which takes up to 95% of searching time. Presented model can be used as search engine on fast changing one-domain web portals, such as news portals. Retrieved results are very actual and the response time is not a big issue. The biggest difference between this model and static search engines is when using for full internet search. In this case long response time makes presented model very slow.

In this work we present several improvements of Bee Hive at Work model. We performed and described experiments that show improvements in behavior of bees when dancing and also better results in real-time web search. Model using presented methods achieve better results than basic model or models with existing improvements.

There are many options for future research. One way is to limit downloaded web pages even more and to retrieve relevant documents in such short time that model can compete with static search engines. Other option for research is finding application of this model, where long response time is not critical issue. Different way of further research is improving behavior of bees, especially by dancing.

Acknowledgements.
This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

[1] P. De Bra, G.-J. Houben, Y. Kornatzky, R. Post. Information Retrieval in Distributed Hypertexts. In: *Proceedings of RIAO'94, Intelligent Multimedia, Information Retrieval Systems and Management,*, 1994.

[2] P. Dziviński, D. Rutkowska. Ant Focused Crawling Algorithm. ICAICS 2008, LNAI 5097, 2008, pp. 1018-1028.

[3] A. Hersovici, M. Jacovi, Y. Maarek, D. Pelleg, M. Shtalhaim, S. Ur. The shark-search algorithm – An application: tailored Web site mapping. In: *Proceedings of the 7th international conference on World Wide Web*, 1998.

[4] L. Jastrzembská. Model of information retrieval inspired by social insects. Bc. thesis, Slovak University of Technology in Bratislava, (2007). (in Slovak).

[5] F. Menczer, A. Monge. Scalable Web Search by Adaptive Online Agents: An InfoSpiders Case Study. Intelligent Information Agents. Springer, (1999).

[6] P. Návrat. Bee Hive Metaphor for Web Search. International Conference on Computer Systems and Technologies - CompSysTech' 06, 2006.

[7] P. Návrat, L. Jastrzembská, T. Jelínek. Bee Hive At Work: Story Tracking Case Study. IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society, (2009), pp. 117-120.

[8] Š. Sabo. Odporúčanie informácií vo webovom prostredí. Bc. thesis, Slovak University of Technology in Bratislava, (2007).

[9] T. D. Seeley, S. Camazine, J. Sneyd. Making in Honey Bees: How Colonies Choose Among Nectar Sources. Behavioral Ecology and Sociobiology, (1991), pp. 277–290.