Information Processing and Management 53 (2017) 608–623

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

An ant-colony based approach for real-time implicit collaborative information seeking

Alessio Malizia[a,*], Kai A. Olsen[b,c], Tommaso Turchi[a], Pierluigi Crescenzi[d]

[a] Human Centred Design Institute, Brunel University London, London, UK
[b] Faculty of Logistics, Molde University College, Molde, Norway
[c] Department of Informatics, University of Bergen, Bergen, Norway
[d] Department of Information Engineering, University of Florence, Florence, Italy

ARTICLE INFO

Article history:
Received 17 July 2015
Revised 15 December 2016
Accepted 23 December 2016

Keywords:
Ant Colony Optimization
Cooperative systems
Evolutionary computation
Information filtering
Information retrieval
Recommender systems
World wide web

ABSTRACT

We propose an approach based on Swarm Intelligence — more specifically on Ant Colony Optimization (ACO) — to improve search engines' performance and reduce information overload by exploiting collective users' behavior. We designed and developed three different algorithms that employ an ACO-inspired strategy to provide implicit collaborative-seeking features in real time to search engines. The three different algorithms — NaïveRank, RandomRank, and SessionRank — leverage on different principles of ACO in order to exploit users' interactions and provide them with more relevant results. We designed an evaluation experiment employing two widely used standard datasets of query-click logs issued to two major Web search engines. The results demonstrated how each algorithm is suitable to be employed in ranking results of different types of queries depending on users' intent.

1. Introduction

Traditionally text retrieval was based on keywords. However, not all documents had been adequately tagged, neither could the keywords describe all aspects of a document. With faster computers, it became possible to perform full-text searches. Then we got the problem of too many hits, i.e. the supplied keywords were found in too many documents. One tried to cope with this by determining relevance as the number of occurrences of each search term in the document, in relation to document size. The first search engines on the Web used this approach.

There were several disadvantages to this approach. Looking for information on a given car, using maker and model as keywords, the search engine did not direct you to any official site. Instead, one was overloaded with car for sale advertisements, as these had a good occurrence to size ratio of the keywords. It was also quite easy to fool the search engines, for example by adding long list of repeated keywords to a Web page, often using a small white font so that it did not clutter the page.

Google's PageRank algorithm saved the day. By letting relevance be determined by the number of links to a page, adding up the score if the links also came from pages that had many links to them, Google had captured a semantic understanding of the relevance concept. For example, many Web pages may say something about The White House, and there may be

* Corresponding author.
E-mail addresses: alessio.malizia@brunel.ac.uk (A. Malizia), kai.olsen@himolde.no (K.A. Olsen), tommaso.turchi@brunel.ac.uk (T. Turchi), pierluigi.crescenzi@unifi.it (P. Crescenzi).

http://dx.doi.org/10.1016/j.ipm.2016.12.005
0306-4573/

many white houses, but Google puts the official site on top, most probably the page that the user wants. And every time someone makes a link to this page, they increase its relevance.

The disadvantage of this approach is that it is static. Pages found important by the PageRank algorithm will probably get more important as they are found by the search engine. That is, important pages will get more important. New pages on similar topics will be hard to find, i.e. placed further down on the search engine result page, and will thus be considered less important. Over time, an algorithm used to determine relevance might be self-fulfilling.

Ideally, we would need a search algorithm that were more dynamic, but still gave a good idea of relevance. Our idea is to use data from the actual searches–what we call dynamic trail information. While PageRank uses static information as link structure, we want to collect data from the actual searches performed by other users. For example, you may be interested in renting a boat to go deep-sea fishing outside the Lofoten Islands in Northern Norway. Your keywords may be rent, boat, fishing, Lofoten. The search engine will then return a standard list of relevant pages; however, in addition you will find a list that says: "other users found these pages". That is, the system have collected data on what other users with similar query terms did. They may have started with the same keywords as you, but may also tried other searches, ending up with a few interesting pages. That is, the effort that other users have put in finding relevant pages can be important information to you.

The data needed to offer an "other users found these pages" list can be collected quite easily, but one will need access on the server level, i.e. to collect data from many users. One could strengthen the trail if the user performed some action at the end. This could either be implicit, such as noting that the users stayed on the site for some time, typed in data, printed from the page, bought or booked, etc. Alternatively, it could be explicit, where the users use a "like" button to tell that the page is interesting, e.g. the Google+1 service[1].

Such an approach falls into the implicit collaborative information-seeking area in which developing new collaborative search interfaces is still needed, as recently suggested by Hearst (2014).

According to Golovchinsky, Pickens, and Back (2009), a collaborative information search system can be either implicit or explicit, meaning that users can explicitly collaborate on query formulations and review search results or can implicitly take advantage of other users' search intents. Normally, implicit collaboration systems provide a recommendation and filter the results already explored by previous users, making them available to others with similar information needs.

The majority of studies in the implicit area are based on collaborative querying techniques that upgrade information systems with data on past query preferences related to other users. As recently demonstrated (Yue, Han, He, & Jiang, 2014), such studies primarily tested implicit collaborative information-seeking systems using simulated query formulation instead of employing user analysis involving human participants. In our research, we employed a classic approach by using two existing datasets to simulate queries to evaluate our system in a real setting.

Hence, we deal with the problem of improving search engines' performance by exploiting the actions performed by users. In fact, search engines are tools designed to help people solve their own informational needs and significant room exists for improvements. Queries submitted to search engines can be clustered into three main categories on the basis of users' aim (Broder, 2002):

Informational queries are issued by users willing to acquire information that they assume is present on one or more Web pages;
Navigational queries are being used to get to a particular Web page belonging to an organization or an individual; and,
Transactional queries are issued to perform activities using the Web, such as booking a trip or downloading a file.

Ranking results produced through navigational queries can be effectively addressed using existing Link Analysis Ranking (LAR) algorithms, such as PageRank, Hyperlink-Induced Topic Search (HITS), or Stochastic Approach for Link-Structure Analysis (SALSA): a higher number of hyperlinks pointing toward one particular page results in a higher page relevance (in other words, algorithms assume that this page is the one that the user was looking for when she issued the query). Ranking results of informational and transactional queries is another matter: given the high frequency of Web pages' updates and the ever-increasing need to obtain answers in real time, the World Wide Web hyperlinks' configuration is no longer the only effective relevance measure that users assign to Web pages. Thus, devising new relevance indicators — to be placed alongside the existing ones (in other words, those based on Link Analysis Ranking) — with the goal of further improving the ranking by considering other relevance measures valued by users is necessary (Zareh Bidoki, Ghodsnia, Yazdani, & Oroumchian, 2010).

In this paper, we propose to employ the concepts of Swarm Intelligence (SI) in relation to the Ant Colony Optimization(ACO) meta-heuristic to improve search engine performance and to reduce the information overload[2] by exploiting collective users' behavior in their usage of search engines.

[1] https://developers.google.com/+/web/+1button/.

[2] The inability to make a decision because of the huge quantity of information obtained by the users.