Clustering Web Search Results Using Fuzzy Ants

Steven Schockaert,* Martine De Cock,† Chris Cornelis,‡ Etienne E. Kerre§
Department of Applied Mathematics and Computer Science,
Ghent University, Fuzziness and Uncertainty Modelling Research Unit,
Krijgslaan 281 (S9), B-9000 Gent, Belgium

Algorithms for clustering Web search results have to be efficient and robust. Furthermore they must be able to cluster a data set without using any kind of a priori information, such as the required number of clusters. Clustering algorithms inspired by the behavior of real ants generally meet these requirements. In this article we propose a novel approach to ant-based clustering, based on fuzzy logic. We show that it improves existing approaches and illustrates how our algorithm can be applied to the problem of Web search results clustering. © 2007 Wiley Periodicals, Inc.

1. INTRODUCTION

Most existing Web search engines respond to a user's query by returning an ordered list of links to Web pages that are considered relevant. The majority of these queries consist of only a few keywords, which are often ambiguous or too general for accurately expressing the user's information need. As a consequence, typically only a small fraction of the search engine results are really relevant. In this way, users are often forced to sift through a long list of search results to find the information they are looking for. Algorithms for clustering Web search results try to overcome this problem by converting the output of an existing search engine to a list of labeled clusters. Well-known clustering algorithms such as k-means and the more general fuzzy c-means depend on an initial estimation of the number of clusters (i.e., k or c). Hence they are not very suitable in the context of search results clustering where such a priori information is not at our disposal. Other clustering algorithms such as agglomerative hierarchical clustering (AHC) are too slow for online clustering of Web search results.

*Author to whom all correspondence should be addressed: e-mail: Steven.Schockaert@UGent.be.
†e-mail: Martine.DeCock@UGent.be.
‡e-mail: Chris.Cornelis@UGent.be.
§e-mail: Etienne.Kerre@UGent.be.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 22, 455–474 (2007)
 Published online in Wiley InterScience (www.interscience.wiley.com). • DOI 10.1002/int.20209

Ant-based clustering algorithms are usually inspired by the clustering of dead nestmates, as observed with several ant species under laboratory conditions.[1] Without negotiating about where to gather the corpses, ants manage to cluster all corpses into one or two piles. The conceptual simplicity of this phenomenon together with the lack of centralized control and a priori information are the main motivations for designing a clustering algorithm inspired by this behavior. However, most ant-based clustering algorithms are only suitable for a visual representation of the data on a two-dimensional grid (e.g., Ref. 2) or require a hybridization with a classical clustering algorithm such as *k*-means (e.g., Ref. 3) or AHC (e.g., Ref. 4).

Real ants are, because of their very limited brain capacity, often assumed to reason only by means of rules of thumb.[5] Inspired by this observation, we propose a clustering method in which the desired behavior of artificial ants (and, more precisely, their stimuli for picking up and dropping items) is expressed flexibly by fuzzy IF–THEN rules. In this way, we obtain a genuine clustering algorithm in which hybridization with a classical clustering algorithm becomes superfluous. Moreover, because no a priori information on the number of clusters is needed, our algorithm is a very suitable candidate for the task of clustering the results of a search engine. Note that although we use fuzzy IF–THEN rules, the result of our algorithm is a crisp clustering. This article is an extended version of Refs. 6 and 7. As an extension of our previous work, in this article we compare results of our algorithm with other clustering techniques and, most importantly, we show its applicability for the clustering of Web documents.

The article is organized as follows: In Section 2, we review existing approaches to ant-based clustering, in particular the algorithm of Monmarché, which served as our main source of inspiration. In Section 3 we outline the structure of our clustering algorithm and motivate its key design principles. A comparison of our algorithm with other clustering methods is presented in Section 4. In Section 5, we apply our algorithm to the clustering of Web search results. Finally, Section 6 offers some concluding remarks.

2. RELATED WORK ON ANT-BASED CLUSTERING

Deneubourg et al.[1] proposed an agent-based model to explain the clustering behavior of real ants. In this model, artificial ants (or agents) are moving randomly on a square grid of cells on which some items are scattered. Each cell can only contain a single item, and each ant can move the items on the grid by picking up and dropping these items with a certain probability that depends on an estimation of the density of items of the same type in the neighborhood. Lumer and Faieta[8] extended the model of Deneubourg et al., using a dissimilarity-based evaluation of the local density to make it suitable for data clustering. Unfortunately, the resulting number of clusters is often too high and convergence is slow. Therefore, a number of modifications were proposed, by Lumer and Faieta themselves as well as by others (e.g., Refs. 2 and 9).

Monmarché[3] proposed an algorithm called AntClass in which several items are allowed to be on the same cell. Each cell with a nonzero number of items