Traversing News with Ant Colony Optimisation and Negative Pheromones

David M.S. Rodrigues and Vitorino Ramos [david.rodrigues@open.ac.uk, vitorino.ramos@ist.utl.pt]

The past decade has seen the rapid development of the online newsroom. News published online are the main outlet of news surpassing traditional printed newspapers. This poses challenges to the production and to the consumption of those news. With those many sources of information available it is important to find ways to cluster and organise the documents if one wants to understand this new system.

Traditional approaches to the problem of clustering documents usually embed the documents in a suitable *similarity space*. Previous studies have reported on the impact of the similarity measures used for clustering of textual corpora [1]. These similarity measures usually are calculated for bag of words representations of the documents. This makes the final document-word matrix high dimensional. Feature vectors with more than 10,000 dimensions are common and algorithms have severe problems with the high dimensionality of the data.

A novel bio inspired approach to the problem of traversing the news is presented. It finds Hamiltonian cycles over documents published by the newspaper *The Guardian*. A *Second Order Swarm Intelligence* algorithm based on Ant Colony Optimisation was developed [2, 3] that uses a negative pheromone to mark unrewarding paths with a "no-entry" signal. This approach follows recent findings of negative pheromone usage in real ants [4].

In this case study the corpus of data is represented as a bipartite relation between documents and keywords entered by the journalists to characterise the news. A new similarity measure between documents is presented based on the Q-analysis description [5, 6, 7] of the simplicial complex formed between documents and keywords. The eccentricity between documents (two simplicies) is then used as a novel measure of similarity between documents.

The results prove that the *Second Order Swarm Intelligence* algorithm performs better in benchmark problems of the travelling salesman problem, with faster convergence and optimal results. The addition of the negative pheromone as a non-entry signal clearly improved the quality of the solutions. The application of the algorithm to the corpus of news of *The Guardian* creates a coherent navigation system among the news. This allows the users to navigate the news published during a certain period of time in a semantic sequence instead of a time sequence.

This work as broader application as it can be applied to many cases where the data is mapped to bipartite relations (e.g. protein expressions in cells, sentiment analysis, brand awareness in social media, routing problems), as it highlights the connectivity of the underlying complex system.

References

[1] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64, 2000.

[2] David M. S. Rodrigues, Jorge Louçã, and Vitorino Ramos. From standard to second-order swarm intelligence phase-space maps. In Stefan Thurner, editor, *8th European Conference on Complex Systems*, Vienna, Austria, 9 2011.

[3] Vitorino Ramos, David M. S. Rodrigues, and Jorge Louçã. Second order swarm intelligence. In Jeng-Shyang Pan, Marios Polycarpou, Michał Woźniak, André C.P.L.F. Carvalho, Héctor Quintián, and Emilio Corchado, editors, *HAIS'13. 8th International Conference on Hybrid Artificial Intelligence Systems*, volume 8073 of *Lecture Notes in Computer Science*, pages 411–420. Springer Berlin Heidelberg, Salamanca, Spain, 9 2013.

[4] Elva J.H. Robinson, Duncan Jackson, Mike Holcombe, and Francis L.W. Ratnieks. No entry signal in ant foraging (hymenoptera: Formicidae): new insights from an agent-based model. *Myrmecological News*, 10(120), 2007.

[5] Ronald Harry Atkin. *Mathematical Structure in Human Affairs*. Heinemann Educational Publishers, 48 Charles Street, London, 1 edition, 1974.

[6] J. H. Johnson. A survey of q-analysis, part 1: The past and present. In *Proceedings of the Seminar on Q-analysis and the Social Sciences, Universty of Leeds*, 9 1983.

[7] David M. S. Rodrigues. Identifying news clusters using q-analysis and modularity. In Albert Diaz-Guilera, Alex Arenas, and Álvaro Corral, editors, *Proceedings of the European Conference on Complex Systems 2013*, Barcelona, 9 2013.