Contribution

We propose an Adaptive Fuzzy Agent Clustering Algorithm based on the ASM (Ant Sleeping Model) in order to resolve the clustering problem. In the ASM model each data is represented by an agent, its environment being a two dimensional grid. The agents will group themselves into clusters by making simple moves according to some local environment information and the parameters are selected and adjusted adaptively. In order to avoid the agents to be trapped in local minima, they are also able to directly communicate with each other. Furthermore, the agent moves are expressed by fuzzy IF-THEN rules and hence hybridization with a classical clustering algorithm is needless. Also, because no a priori information on the number of clusters is required, this algorithm is a good solution for the web search results clustering problem.

Motivation

The possibility to cluster web search results so that the output would be a list of labeled clusters would be very helpful in our opinion. The best way to explain this is through an example. Suppose a user enters the query “mouse” in a search engine. The result will usually be a list containing sites about “mouse — the animal”, but also sites about “mouse — the device”. We claim that usually a user is not searching for both. So the idea would be to offer the user the possibility to browse through a list of either “mouse” - the animal or “mouse — the device” and hence the importance of finding proper clustering algorithms that cluster web search results.

Token weights

Term frequency inverse document frequency:

\[ tfidf_{a,t} = tf_{a,d} \cdot idf_t \] (1)

\[ idf_t = \log \frac{N}{df_t} \] (2)

N — total number of documents in the collection
df_t — document frequency, the number of documents in the collection containing the term t.

Fuzzy sets

Let X be the universe of discourse. The sets S, D, VD of Similar, Different and VeryDifferent documents respectively are defined in the following way:

\[ S(x) = \begin{cases} 1, & x \in [0, S1D] \\ 0, & \text{otherwise} \end{cases} \] (3)

\[ D(x) = \begin{cases} (x - SD2) + 1, & x \in [SD1, SD2] \\ 1, & \text{otherwise} \end{cases} \] (4)

\[ VD(x) = \begin{cases} (x - VD2) + 1, & x \in [VD1, VD2] \\ 1, & \text{otherwise} \end{cases} \] (5)

Fuzzy variables – Similarity

In the above figure the fuzzy sets S, D and VD corresponding respectively to the linguistic concepts Similar, Different and VeryDifferent are called the states of the fuzzy variable Similarity.

Fuzzy IF-THEN rules

By fuzzy if-then rules we mean conditional statements that comprise fuzzy logic (below, and denote agents):

if similarity(a_i, a_j) is S then move closer
if similarity(a_i, a_j) is D then move further
if similarity(a_i, a_j) is VD then move further.

Case study

We have applied our algorithm [3] for clustering web search results. In order to do this we have made a Java application containing the following components: a web crawler, a weighing component and a clustering component. The crawler receives as input a set of starting pages and it extracts the text and the links. The weighting component has two parts: a MySQL procedure and a Java thread that executes the procedure at a given interval of time. When performing a normal search, the dot product between the query vector and documents from the index is computed and the documents are returned in decreasing order. A matrix of document similarities is given to the clustering component. The clustering component uses the algorithm defined in this paper [3] and outputs the clusters and the document ids from each cluster. For example, suppose we are searching for the word “mouse”. We take only the first 5 search results and send them to the clustering component:

<table>
<thead>
<tr>
<th>Id</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td><a href="http://computer.howstuffworks.com/mouse.htm">http://computer.howstuffworks.com/mouse.htm</a></td>
</tr>
<tr>
<td>1</td>
<td><a href="http://en.wikipedia.org/wiki/Mouse">http://en.wikipedia.org/wiki/Mouse</a></td>
</tr>
<tr>
<td>2</td>
<td><a href="http://www.newsgg.com/">http://www.newsgg.com/</a></td>
</tr>
<tr>
<td>3</td>
<td><a href="http://animal.discovery.com/">http://animal.discovery.com/</a></td>
</tr>
<tr>
<td>4</td>
<td><a href="http://computer.howstuffworks.com/">http://computer.howstuffworks.com/</a> &amp;share-redirect?type=facebook&amp;cid=1106</td>
</tr>
</tbody>
</table>

In order to compare our approach with other clustering methods we have first considered the following data set:

<table>
<thead>
<tr>
<th>test dataset</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.11</td>
<td>0.11</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>1</td>
<td>0.12</td>
<td>0.14</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>3</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>4</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>5</td>
<td>0.81</td>
<td>0.81</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>6</td>
<td>0.81</td>
<td>0.81</td>
<td>0.83</td>
<td>0.81</td>
</tr>
</tbody>
</table>

We have evaluated our algorithm against this dataset and we have obtained the following clusters: Cluster0 (0, 1, 2), Cluster1 (3, 4) and Cluster2 (5, 6). On the same dataset the k-means implementation from Weka reported 3 misclassifications. This suggests that our approach is a promising one.

We also perform tests on bigger datasets like Iris and the results are better compared to k-means: the k-means implementation from Weka has 17 misclassifications while our algorithm has, by now, 10.

Conclusions and future work

Finding proper clustering algorithms for search engines is an important issue. Our algorithm already behaves well on datasets about the size of Iris. In order to perform properly in the area of serach engines an algorithm should also be context aware and perhaps incremental.

References


Acknowledgments

The authors wish to thank for the financial support provided from programs co-financed by The Sectoral Operational Programme Human Resources Development, Contract POSDRU 6/1.5/S/3 — “Doctoral studies: through science towards society”.

Authors:

Radu D. Găceanu

Babes-Bolyai University, Cluj - Napoca, Romania

Eötvös Loránd University, Budapest, Hungary