**OntoForce Challenge – Linking authors to works**

James Malone, malone@ebi.ac.uk

There are three components to linking authors on PubMed to their published works:

1. Identify important features which make authors identifiable
2. Use data mining algorithms to automatically cluster authors based on these feaures
3. Publish these clusters as linked data.

**What makes an author identifiable?**

Without a unique identifier (e.g. Orcid or an email address) there are several features which me may consider as important when attempting to identify an author. These include:

1. Institutes an author is affiliated with
2. Co-authors
3. Journal names an author published in
4. ‘Topic map’ describing files of study an author publishes in
5. Dates which the author is active

By computing a matrix of these as input features which describe an author we can end up with an amount of information for each ‘name’ whereby a name is the textual string found in the author part of a publication. For instance, “J Malone”.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| S\_Jupp | H\_Parkinson | D\_Welter | T\_Burdett | D\_Vasant | E\_Birney | D\_Hawkins |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Here you see a snippet of correlations to the author string “J Malone” searched for in PubMed. Each row is one publication and a 1 indicates another author (string) was a co-author on a particular publication.

We may also consider adding information such as topic maps of the journal articles as a column or affiliation as a discretised value. Here we would use an ontology to control these such as EDAM <http://bioportal.bioontology.org/ontologies/EDAM>

For example:

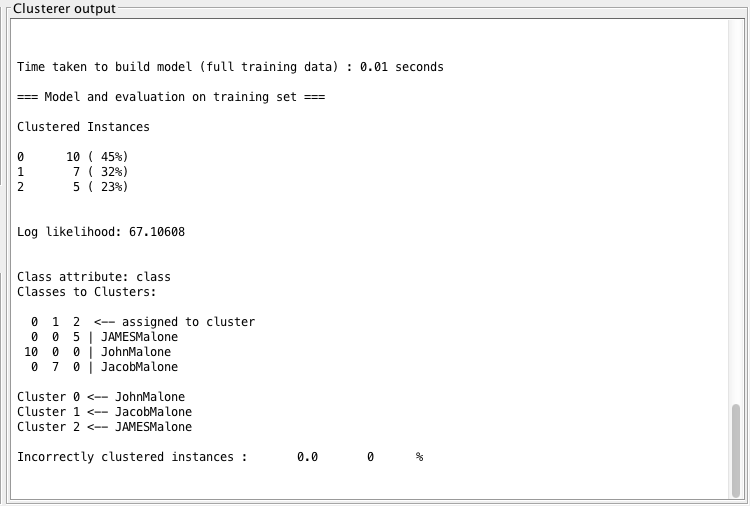
<http://bioportal.bioontology.org/ontologies/EDAM/?p=classes&conceptid=http%3A%2F%2Fedamontology.org%2Ftopic_0091&jump_to_nav=true>

**Identifying an author from a string**

Given the above matrix, we can now consider using machine learning/data mining techniques to identify authors. We consider an author a single cluster, extracted from all of the information for a given search string. For example, J Malone is the string for which several authors may be identified (different individuals).

If we use algorithms such as Expectation Maximisation we are able to automatically identify clusters around these matrices, although ideally we would want to identify clusters by using approaches such as the Elbow Method.

Given an examples (J Malone) for which I have completed three (subsets) of matrices for three known different authors, we can extract the three clusters perfectly when the cluster number is known. This a screenshot from from Weka[[1]](#footnote-1):



However, since in reality the real number of authors for a given string is not known, we would ideally want the machine learning or statistical algorithm to identify this for us. This is more difficult and indeed as we add more features we also find that the accuracy can drop. For instance, adding in the ‘topic map’ reduces accuracy from 100% to 69% suggesting this feature may not be a useful signal as the cross-cluster signal is stronger than the intra-cluster signal, i.e. that J Malone that is me shares more in common with another J Malone publishing in bioinformatics than I do with the sparse matrix of co-authors (for instance that I have published once with Ewan Birney). Two methods would help here; elbow method which can be used to identify when variance between clusters has begun to tail off, suggesting an optimal number of clusters (good article on this here: <https://datasciencelab.wordpress.com/2013/12/27/finding-the-k-in-k-means-clustering/>). We would also ideally employ a weighting system to demote topic to a less strong signal, thereby reducing this potential source of error.

Ideally the following method is used:

1. Identify important features using a training set.
2. Once selected, perform cluster analysis for each set of strings (here one stirng set is one name “J Malone”)
3. Repeat for every string, producing one matrix for each and perform clustering for each, thereby identifying an individual, their papers and their co-authors (and potentially more information such as topic under which they publish).

**Linking Data**

Once we have perform the above steps we have a set of ‘authors’ identified for a given string as well as information on that author such as their co-authors, PubMed IDs and potentially topic maps.

Each cluster should be assigned a URI – the string J Malone has a URI. Note all we are saying here is that there is a shared concept of “J Malone” the string and not that this is an individual. For this cluster we link new clusters for the authors which dropped out and attach information to these individual authors, for instance, as follows:

“J Malone” information

pubmedID

Has\_orcid

Orcid

S Jupp

Has\_co\_author

Publishes\_in\_  
topic

Genetics

Bioinformatics

has\_related\_string

J Malone Author 3

J Malone Author 1

J Malone Author 2

1. http://www.cs.waikato.ac.nz/ml/weka/ [↑](#footnote-ref-1)