

• REPORT OF AN EXPERIMENT IN CLUSTERING

◦ *hans varghese mathews, The Centre for Internet and Society, Bangalore*

0 The object of the exercise described in what follows was to see if the various editors of a frequently edited Wikipedia document could be clustered in different groups, each distinguished from the others by some particular interest, through some *quick* machine process requiring minimal human intervention. The experiment was not entirely unsuccessful; but the supervision of human judgement upon machine decision was more haphazard than one could wish.

1 The Wikipedia page named **Evolution** was chosen as the primary document; and for our collection of editors we chose, out of all the editors that this page had had from November 2008 through January 2009, those who had in that same period edited at least one other Wikipedia page. The rationale for proceeding so was to compute in some very quick and convenient way a *measure of similarity* between each pair of the editors to be clustered: without *at all* considering, at this stage, the textual character of their individual interventions. The data from which these similarities were computed constituted a matrix, with a row for each such ancillary document and a column for each editor; and the entry in the i^{th} row and the j^{th} column was 1 if the i^{th} page had been edited by the j^{th} editor, and 0 otherwise. The similarity between any pair of editors was then assessed by counting the pages they had both edited against the counts, for each, of those pages that one had edited but the other not: but in such a way that, among the pages both had edited, those that were *on the whole* less edited received more weight; while among the pages only one or other had edited, those that were on the whole less edited received less weight.

2 The similarities thus obtained were collected in a square and *symmetric* matrix, having as many rows and columns as there were editors: with the similarity between the i^{th} and the j^{th} editor being the entry in the i^{th} row and the j^{th} column, and the entry in the j^{th} row and the i^{th} column as well. This matrix of similarities was then used to cluster the editors in a variety of ways, using various techniques. Three among these clusterings were retained, on considerations of variation in size between, and of *coherence* within, their constitutive clusters: with the coherence of a given cluster of individuals being measured by $\lambda_1/(\lambda_1 + \lambda_2 + \dots + \lambda_k)$, where λ_1 is the largest among the positive eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_k\}$ of the submatrix recording the similarities between the individuals in that cluster.

A square matrix A of numbers, with m rows and columns say, may be regarded as device for *rotating the direction* and *rescaling the length* of any vector \mathbf{v} having m numerical entries or components: through the usual multiplication, with the i^{th} component of the resultant vector $A \cdot \mathbf{v}$ being $(A_{i1}v_1 + A_{i2}v_2 + \dots + A_{im}v_m)$, where A_{ij} is the entry in the i^{th} row and j^{th} column of A , while v_j is the j^{th} component of \mathbf{v} . An *eigenvector* \mathbf{u} of A is a vector of *unit length* whose direction is only reversed, if it is at all rotated; and one has $A \cdot \mathbf{u} = \lambda \mathbf{u}$ for some *eigenvalue* λ now, with a reversal of direction if $\lambda < 0$, and with the size or 'absolute value' $|\lambda|$ being the rescaling of length.

A vector \mathbf{v} has the usual *Euclidean* length $\|\mathbf{v}\| = (v_1^2 + v_2^2 \dots + v_m^2)^{\frac{1}{2}}$ here. To specify the direction of \mathbf{v} we must consider the m *standard basis vectors* we obtain by taking the $\mathbf{0}$ vector here, whose m components are all zeroes, and replacing any one of these with 1; and \mathbf{e}_k usually denotes the vector which has 1 for its k^{th} component and zeroes everywhere else.

Now for each $j \in \{1, 2, \dots, m\}$ the *direction-number* $v_j/\|\mathbf{v}\|$ is the cosine of the angle between \mathbf{v} and \mathbf{e}_j in the plane this pair of vectors would determine, whenever \mathbf{v} is not a multiple of \mathbf{e}_j : whenever \mathbf{v} cannot be obtained from \mathbf{e}_j by multiplying each component of the latter with some number. Note that when \mathbf{v} is so obtained the direction-number $v_j/\|\mathbf{v}\|$ would be either $1 = \cosine(0^\circ)$ or $-1 = \cosine(180^\circ)$: which is appropriate, since \mathbf{v} would lie either along \mathbf{e}_j itself or along its reverse $-\mathbf{e}_j$ in any plane that \mathbf{e}_j might help determine. The word “plane” has its common geometrical meaning now.

We note that such a criterion of coherence would favour those clusters in which the similarities between members were more uniform than elsewhere: were the similarities between them used to locate all the editors in some Euclidean space, using some technique like multidimensional scaling for instance, our criterion would favour *globular* clusters over *flat* or *chainlike* ones. Regarding variation between the sizes of constitutive clusters in a clustering, the desideratum was that the largest such cluster should not exceed too much the mean size of the remainder, when their difference is scaled by the variation in size there.

The specifics of the clustering techniques that were employed, and the statistical tests which decided the retention of clusterings, are set out in the technical supplement to this report; and the computation of similarities described in **1** may be found there as well.

3 Each cluster within a retained clustering was next used to induce a weighting on the collection of edited pages, in the evident way, with the weight of a document relative to a cluster being the proportion of the individuals there who had edited it; and the names of its most frequently edited pages were taken as indicative of the interest of that group or ‘pack’ of editors. A choice between the three retained clusterings might now be made by considering how *distinctively* the separate interests of their several constitutive packs could be characterised.

Such judgements are apt to vary considerably between individuals, of course: but we note that there were appreciable congruences between the three clusterings that were retained. Each clustering yielded one notably coherent pack, consisting of the same individuals almost in each clustering, set apart by their attention to *creationism* and *intelligent design*. Each retained clustering also yielded a pack whose members seemed to have a special interest in *evolution as theory and fact*; and these cognate groups almost overlapped as well. Another group identifiable in all three clusterings had its members linked by their common attention to the Wikipedia *Sandbox*. These results are hardly surprising. But we note that all three clusterings did indicate a small group, somewhat more coherent in one than in the other two, of individuals linked by an interest in the *gastrointestinal tract*.

4 Given the ‘virtual’ character of our packs or groups we should expect that some individuals may sit well in more than one of them; and we should find a way to allot such an individual to each among the groups with which he might have comparable affinities. A convenient way to do so would be to locate our editors in some Euclidean space, again, so that each group may be identified with the *centroid* of the points to which its members are assigned. An individual who is markedly further from the centroid of his own group, than his fellows there, might now be allotted to some other groups: if, for instance, his distances from their centroids, compared to his distances from the rest, are markedly closer to his distance from

his assigned centroid. Proceeding so would work best with globular spatial clusters, again, and we should be cautious about how similarities between individuals will be used to spatially locate them: especially when some very few of the clusters in a clustering are markedly more coherent than the others.