How the Vivísimo Clustering Engine Works

Document clustering refers to the grouping of a document collection based on similarity. Document clustering is one type of the more general problem of cluster analysis, for which many algorithms have been published, and even textbooks written.

Unlike many other uses of cluster analysis, in document clustering for the display of documents to a user, the quality of the resulting cluster descriptions (or annotations) is of highest importance. The problem of good-quality descriptions has been the main research challenge in document clustering over the years.

To add significant value over just seeing a ranked list of documents, the cluster descriptions should be:

1. Concise – not occupy much screen space or much user attention
2. Understandable – look as human-like as possible
3. Accurate – fairly reflect what is inside the cluster
4. Distinctive – make clear at a glance how a given cluster is distinct from neighboring clusters

The Vivísimo Clustering Engine, when used to cluster search results, uses only the returned title and abstract for each result. The similarity between documents is based only on this raw material (the visible text of the search result, not the entire article) and nothing else.

The proprietary Vivísimo algorithm then puts documents together (clusters them) based on textual similarity. However, this raw similarity is augmented with heuristics that address each of the four points above, so textual similarity is only one of the two factors that determine the eventual clusters. The other factor summarizes human knowledge – coded by Vivísimo's programmers and partly invented by them – of what users wish to see when they examine clustered documents.

Vivísimo does not use a pre-defined taxonomy or controlled vocabulary, so every cluster description is taken from the search results within the cluster. Thus, do not expect the clusters to correspond to your pre-conceived, expert's internal model of how the material that matches your query should be arranged. One reason for this is that the top actually-returned search results may not be what you expect. Or, the prominent similarities that the algorithm finds may not exactly match your

taste, which in many cases may reflect your pre-conceptions and are not even based on the returned textual descriptions.

The Vivísimo Clustering Engine will not force each document into only a single place in the cluster hierarchy. Documents can be about multiple themes, so it is best to place each document where it seems to fit. Of course, a poor algorithm would place every document just about everywhere; little value would be added to just the raw, ranked list. So the trick is to make good placement judgments, similarly to how a human classifier would operate if clustering an unfamiliar domain by hand. Depending on the browser interface used by the licensee of Vivísimo's clustering engine, you will see highlighted the clusters into which a selected document has been placed.

In normal use, the raw search results will be placed unchanged on half the screen. The folders that represent the clusters are sorted according to the number of search results in each, and according to the overall rank, in the search engine's output, of the individual search results.

Vivísimo is not perfect. In fact, it's hard to know what perfection would even mean. Thus, sometimes an article will be placed in a cluster which it doesn't really match, in the eyes of an expert. Or, two different meanings of a word or phrase may be conflated The overall claim is that using Vivísimo's clustered search results, a user will be able to see further, with less effort, into a given body of literature.

info@vivisimo.com

Vivísimo