Cluster, clustering & cluster analysis

Some texts in library and information science (LIS) uses the word 'cluster' synonymous with the word 'class', while other texts make a differentiation between those words.

Those who differentiate between 'cluster' and 'class', are directly or indirectly inspired by the work of Ludwig Wittgenstein, who proposed (in his *Philosophical Investigations*, 1953) the idea of cluster, or 'family resemblance', concepts: some terms by their nature do not admit of an essentialist definition, but are rather characterized by a diffuse network of more or less loosely interconnected properties. Any particular instantiation of the concept may draw on a subset of such threads, even though there is a limit to such conceptual 'plasticity'. Wittgenstein's famous example is the idea of a game: the more one thinks about it, he said, the more it is clear that it is difficult to list a set of characteristics that are necessary and sufficient to define what we mean by 'game'.

It is an open question, however, whether Wittgenstein was right. Sutcliffe (1993, p. 42) writes from the point of view of Aristotelian theory:

„2.2.1 Wittgenstein on ‚family' and 'family resemblance'

Wittgenstein (1953), having had difficulty specifying defining conditions for the class *language* (*language games*), gave up the search for necessary and sufficient conditions, and then (without proof) asserted that:

'These phenomena have no one thing in common which makes us use the same word for all . . . You will not see something which is common to all, but similarities, relationships, and a whole series of them at that. . . . We see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail. I can think of no better expression to characterize these similarities than 'family resemblances' for the various resemblances between members of a family (pp. 31-2)'

After a thorough discussion of this problem, Sutcliffe (1993, p. 48) concludes: "The explanation of Wittgenstein's difficulty, then, is that in concentrating on 'family resemblances' he remained in the wrong context A^1, when, to find the needed genus-definition for the monothetic concept *language-games*, he should he should have shifted to context A^0 within which one can state the conditions which set off *language-games* as a class from other things which are not *language-games*".

Sutcliffe thus undermines the argument to regard 'clusters' as being different from 'classes'.

Other definitions of 'cluster' include:

"The term cluster refers to the grouping together of elements within a domain- usually spatial" (Wikipedia, 2005a).

"Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Machine learning typically regards data clustering as a form of unsupervised learning." (Wikipedia, 2005b).

"A cluster, from its astronomical origin, tends to connote a crowd of points close together in space, but a class is a set of objects assembled together for any kind of reason," (Hartigan, 2001, p. 15014).

Informationsordbogen (1991): "Cluster in LIS-connections means a collection of words or expressions, which are associatively connected without necessarily to have mutual semantic relations which can be formalized. The purpose of clusters is to facilitate information retrieval by pointing to search terms which may supplement or substitute a given search term." (Translated from Danish).

The concept "cluster" has also been used in another meaning within LIS. In the Danish union catalog "ALBA/SAMKAT" were "clusters" used about different bibliographical descriptions of the same document. Those descriptions originated from different libraries but were clustered by means of their common ISBN number.

The term “cluster analysis” was first used by Tryon (1939). Cluster analysis encompasses a number of different classification algorithms that aim to organize information into "clusters".

"Document clustering is an information-retrieval approach. Unlike text categorization, it does not involve pre-defined categories or training documents and is thus called unsupervised. The clusters and, to a limited degree, relationships between clusters are derived automatically from the documents, and the documents are subsequently assigned to those clusters" (Golub, 2005, p. 52-53).

Ereshefsky (2000, p. 15) writes about methods for classifying the world's entities: "[T]hree general philosophical schools will be presented: essentialism, cluster analysis, and historical classification.

- Essentialism sorts entities according to their essential natures. (See: Essence in Epistemological Lifeboat)
- Cluster analysis divides entities into groups whose members share a cluster of similar traits, though none of those traits are essential.
- The historical approach classifies entities according to their causal relations rather than their intrinsic qualitative features." (Format with bullets added).

Ereshefsky (2000, p. 24-28) presents an overview of cluster analysis as a method for classification. He finds that "All forms of cluster analysis make two common assumptions: the members of a taxonomic group must share a *cluster* of similar traits, and those traits need not occur in all and only the members of a group. Still, cluster analyses vary: first, on the breath of similarities desired among members of a group, and second, on the relationship between similarity and theory".

Literature:

Cooper, Rachel (2005). Classifying Madness: A Philosophical Examination of the Diagnostic and Statistical Manual of Mental Disorders. Berlin: Springer. (Chapter 4.1.1 Cluster analysis within psychiatry, pp. 95-103).

Ereshefsky, M. (2000). *The Poverty of the Linnaean Hierarchy : A Philosophical Study of Biological Taxonomy*. Cambridge: Cambridge University Press.

Golub, K. (2005). *Automated subject classification of textual web pages, for browsing*. Lund: Lund University, Department of Information Technology.

Hartigan, J. A. (2001). Statistical Clustering. IN: Smelser, N. J. & Baltes, P. B. (eds.) International Encyclopedia of the Social and Behavioral Sciences. Oxford. (Pp.15014-15019).

Hearst, M. A. (2006). Clustering versus Faceted Categories for Information Exploration. *COMMUNICATIONS OF THE ACM*, 49(4), 59-61.
http://web.archive.org/web/20070318090819/http://flamenco.berke

Informationsordbogen. Ordbog for informationshåndtering, bog og bibliotek. 2. udg. Udarbejdet af J. B. Friis-Hansen, Torben Høst, Poul Steen Larsen & Henning Spang-Hanssen. [Hellerup]: Dansk Standardiseringsråd, 1991. (DS/INF 27).

Kowalski, G. J. & Maybury, M. T. (2000). *Information storage and retrieval systems: Theory and implementation. 2nd ed*. Norvel, Mass.: Kluwer Academic Publishers. Chapter 6: Document and term clustering, pp. 139-163.

Sutcliffe, J. P. (1993). Concept, class, and category in the tradition of Aristotle. IN: van Mechelen, I.; Hampton, J.; Michalski, R. S. & Theuns, P. (eds.). *Categories and Concepts*. London: Academic Press, pp. 35-65.

Tryon, R. C. (1939). *Cluster analysis.* New York: McGraw-Hill.

Wikipedia. The free encyclopedia. (2005A). Cluster.
http://en.wikipedia.org/wiki/Cluster

Wikipedia. The free encyclopedia. (2005B). Data clustering.
http://en.wikipedia.org/wiki/Data_clustering

Willett, P. (1988). Trends in Hierarchic Document Clustering: A Critical Review. *Information Processing and Management, 24*(5), 577-597.

Wittgenstein, L. (1953/1958). *Philosophical Investigations*. Third edition. Translated by G.E.M. Anscombe. Englewood Cliffs, NJ: Prentice Hal. (1st edition 1953).

See also: Text categorization

Birger Hjørland

Last edited: 07-10-2009

Home