Concept Maps: Making Learning Meaningful
Proc. of Fourth Int. Conference on Concept Mapping
Viña del Mar, Chile, 2010

# FINDING THE NUMBER OF CONCEPTS FOR MAPPING FRENCH CANADIAN HEALTH NETWORKS

*Louise Bouchard1, Marcelo Albertini2, Aubain Hilaire Nzokem1, Isabelle Gagnon-Arpin1*
*1. Population Health Institute, University of Ottawa, Canada; 2. Universidade de São Paulo, Brazil.*
*louise.bouchard@uottawa.ca*

**Abstract.** The concept mapping method was used to investigate the issues faced by stakeholders of the francophone minority health movement in Canada. Data gathered from the meeting at the Ontario regions were analyzed in order to determine the most satisfactory method of identifying the appropriate number of concepts. After exploring various statistical methods, the cluster procedure within the SAS software, with the Ward method as a parameter, proved to be the most satisfactory method to produce the appropriate number of concepts. By analyzing the graphical tool computed from the SAS cluster procedure output; we were able to choose 8 and 9 clusters as the appropriate cut-off number of clusters. The content of every grouping was closely examined by the research team and deemed qualitatively valid. Therefore, we conclude that in addition to being statistically consistent, the Ward method within SAS has a strong qualitative coherence and is therefore the most satisfactory method to identify the appropriate number of concepts for our study.

## 1 Introduction

In the cont ext of Canada's official languages (English and French), where the francophone majority lives in the province of Quebec and the Anglophone majority in all other provinces, we investigated the situation of the health services offered to linguistic minorities. We used the method of concept mapping to investigate the reality of Francophones living in minority in the province of Ontario, and constituting 4.5% of the total population of that province.

The method of concept mapping involves bringing together a number of participants concerned with a specific problem and identifying all possible statements related to that problem. In our case, it consisted of the issues faced by stakeholders of the francophone minority health movement in Canada. The gathering of information takes place in different stages (Dagenais, 2009). The first stage is the definition of a "focus question" that guides participants during the elaboration of statements. Similar statements will form concepts surrounding the focus question. In the next stage, the participants individually make groups of statements, attaching a label or "concept" to it, from the statements generated according to their background knowledge. Also, each participant will attribute a rating from 1 to 5 to each statement. The last step is the generation of a concept map based on the agreement among opinions of participants. This stage is based on three computing procedures. The first procedure is to convert the grouped organization of statements made by the participants into a matrix of n x n dimensions, with n being the number of statements. Every pair i and j of statements organized in the same group adds a unit to the position (i,j) of the matrix. The next procedure consists of applying a dimensionality reduction method to represent the distance between each statement in two dimensions. The last procedure is to execute a hierarchical clustering algorithm that helps the decision of which statements represent elements of a concept. The hierarchical cluster algorithm outputs a dendrogram (or tree) that may help in choosing the number of appropriate concepts.

There are several criteria that allows us to determine the optimal number of clusters for general clustering algorithms and abstract applications (Brock, 2008), such as internal (connectivity, sillhouette, Dunn index) and stability measures (modification of intra-cluster variance, rate of non-changing of cluster, among others). A rule oriented to concept mapping is to pick the average number attributed by the participants as the desired number of clusters. An ad hoc rule called the rule of thumb is also available, which defines the number of clusters considering only the number of elements to be clustered.

However, according to our experiments with data collected on issues faced by stakeholders of the francophone minority health movement in Canada, none of the previous criteria or rules is satisfactory to find the number of clusters for concept mapping. The criterion based on stability does not work mostly because of the relatively low amount of information. The internal criterion cannot be blindly trusted as it mainly provides extreme suggestions: too high or low numbers. Finally, although the rule oriented to the concept mapping provides reasonable suggestions, it does not take into account the clustering process.

Considering these drawbacks, we propose a dendogram-oriented method that uses the variance of distance among clusters in the process of agglomeration to decide the cut-off point on the hierarchical clustering tree to determine the number of clusters. Our experimental results suggest that this method provides a good orientation for the choice of the number of clusters for concept mapping.

## 2 Design of concept maps

There are two ways to approach the design of conceptual maps (Trochim, 1989, Novak & Gowin, 1984). In the first, specialists manually analyze data collected about the subject. Such data is often obtained from interviews and it is not structured. In this case, the design of a map relies on guidelines that will help the specialists organize data and extract general, broad and relevant concepts. The second way is driven by methods that use structured data and automatic analysis (Trochim, 1989). In such methods, the structured data provides support to the generation of tools to automatically create maps, visualize the possible concepts and quantify their intersection. The method employed in this work is based on basic units of knowledge called “Statements”, that are the most fine-grained concept possible to have. From the clustering of many fine-grained concepts, we visualize the emergence of general, broad and relevant concepts.

The clustering of “Statements” uses a notion of proximity defined by consensus among the individuals who took part in their creation: when most of the participants individually say that two “Statements” belong to the same group of concept, then they are very similar. The search of general, broad and relevant concepts and the definition of similarity of the concepts lead us to observe the generation of groups by a diagram named dendrogram, which is in the format of a tree and contains similar concepts in close nodes. The first node in a dendrogram represents the extreme case where all “Statements” form only one concept group. More groups are visualized according to the increasing demands of similarity of “Statements” within the same group. As the demand for similarity increases, the number of fine-grained concepts within a concept in a dendrogram decreases. The groups in the last level are called leaves and are composed of only one “Statement”. Then, naturally raises the question of how close “Statements” should be to the constitution of a concept. Also, this question can be translated into how many concepts there will be in the concept map.

From the active research field of clustering algorithms, there are ad-hoc heuristics to answer questions of how similar elements and how many concepts should be chosen, such as (Xu & Wunsch, 2008): to look for groups with uniformly distributed elements, to find a balance of minimum distance among elements in the same group and maximum distance among elements of different groups, and to minimize the amount of elements to define a meaningful group. However, we believe that from the point of view of specialists, the most satisfactory method is the one that finds the most fine-grained concepts, instead of looking for only statistically/mathematically sounding group configurations. The fine-grained, or minimal concepts, are those that characterize a minimum line of thoughts or subjects, yet still constituent and consistent to an idea. From fine-grained and consistent concepts, broader and more general concepts can be organized by the specialists, when needed.

Our position can be translated in terms of a dendrogram with the following sentence: “to stop the sequence of agglomerations in order to avoid the increasing of variance of dissimilarities among groups”. The main reasoning is that when the agglomerations start to increase their disparities (in terms of variance) they become broader concepts. The detailed proposal of such method is described in the following section.

## 3 Experiments with the French Canadian Health Networks Dataset

The evaluation of the proposed method was carried out using a dataset collected in 2009. The dataset is meant to provide a view on the current state of policies over the French Canadian Health Networks (FCHN). The FCHN dataset was collected from eight meetings, with a total of 80 participants. All eight meetings considered the focus question “When you think about the future of French Canadian Health Networks you think of...”.

The meetings generated 117.35 statements in average. The eight meetings included participants from provinces: North-West Provinces (NWP), Atlantic provinces (ATP), Western provinces (WP), Ontario province (OP), and from Ontario regions: Ottawa, Sudbury, Timmins, and Toronto. The results emerging from the Ontario meeting are the ones presented below. The generation and evaluation of concepts for the FCHN dataset of experiments relied on a comparison with state-of-art criteria and was followed by a step of human inspection.

## 4 Method to find the number of concepts using agglomerative hierarchical clustering

Before performing the hierarchical cluster analysis, we used the Concept System software to explore the distribution of every statement that emerged from the meeting with the Ontario regions. This visual representation is presented in the form of a Scatter plot, shown at Figure 1 (below). Subsequently, the standard hierarchical agglomerative clustering algorithm starts by taking each input data as a sub-group. Then, iteratively, dissimilarities for every pair of groups are computed and the most similar ones are joined. This procedure is repeated until all sub-groups form only one. The algorithm outputs the sequence, with respective dissimilarities, at which sub-groups are agglomerated.

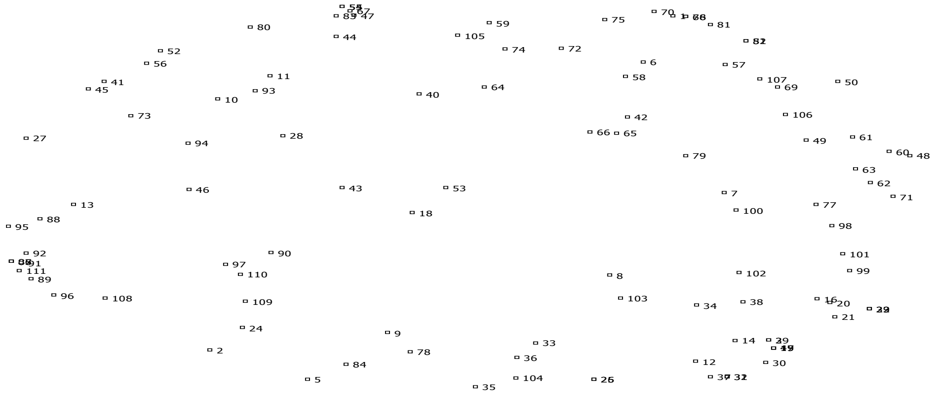

**Figure 1.** Scatter plot of statements from the meeting with the Ontario regions

### *4.1 Cluster analysis using SAS*

The SAS software was used to overcome the weaknesses of the concept system software in performing the cluster analysis. The data, similarity matrix, was imported from the Concept System software and a transformation was made in order to have the dissimilarity matrix, also called the distance matrix. The matrix was then stored in a SAS data file. In order to perform the hierarchical cluster analysis, the SAS Cluster procedure was invoked. The previous data set is used as an input. In addition, the Ward method was chosen as the criterion of clustering the statement. The following operations are implemented by the cluster procedure: 1) Initialisation: at the beginning, each statement is a cluster. 2) Iteration: the following two steps are computed to obtain only one cluster, 2.1) Grouping together the two closest clusters using the chosen Ward distance, 2.2 Updating the previous distance matrix by overwriting the two closest clusters by a new cluster and computing their distance with other old clusters.

A graphical tool helps to choose the number of clusters. Based on the statistics from the cluster procedure, we graph a decreasing curve that takes on the horizontal axis the different number of clusters and on the vertical axis the inter-class variance. As a criterion, the curve should be read from the right to left and we should stop at the first important skip on the graphic, and from then select the corresponding number of clusters. The output of this procedure is the tree data set, containing all the elements needed for choosing the number of clusters and also building up the dendrogram. After choosing the number of clusters, we are in good position to graph the dendrogram. We make use of the tree procedure in SAS in order to draw the tree diagram.

## 5 Application of the method on the ONTARIO case: presentation of the main results

### *5.1 Selection of the number of clusters*

As showed in figure 1 below, from the right to the left, we notice an important skip between 9 and 8 clusters in red colour, and between 8 and 7cluster in green colour. The analysis shows the relative importance of the inter-class variance if we choose to group at 9 or 8 clusters.

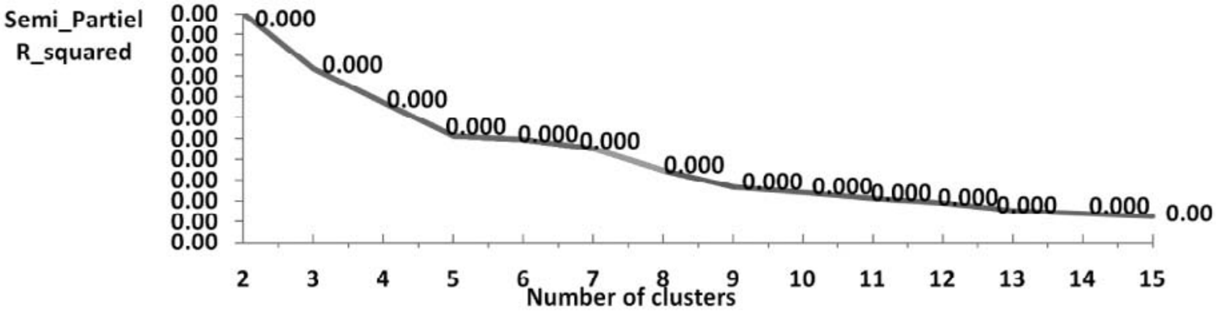

**Figure 2.** Decreasing in the inter-class variance during clustering process

Based on 8 and 9 clusters, a qualitative analysis was done on each number of clusters (8 and 9 clusters) in order to select the most logic and coherent cluster. As a result of the complementary analysis, 9 was chosen as the number of clusters for the study.

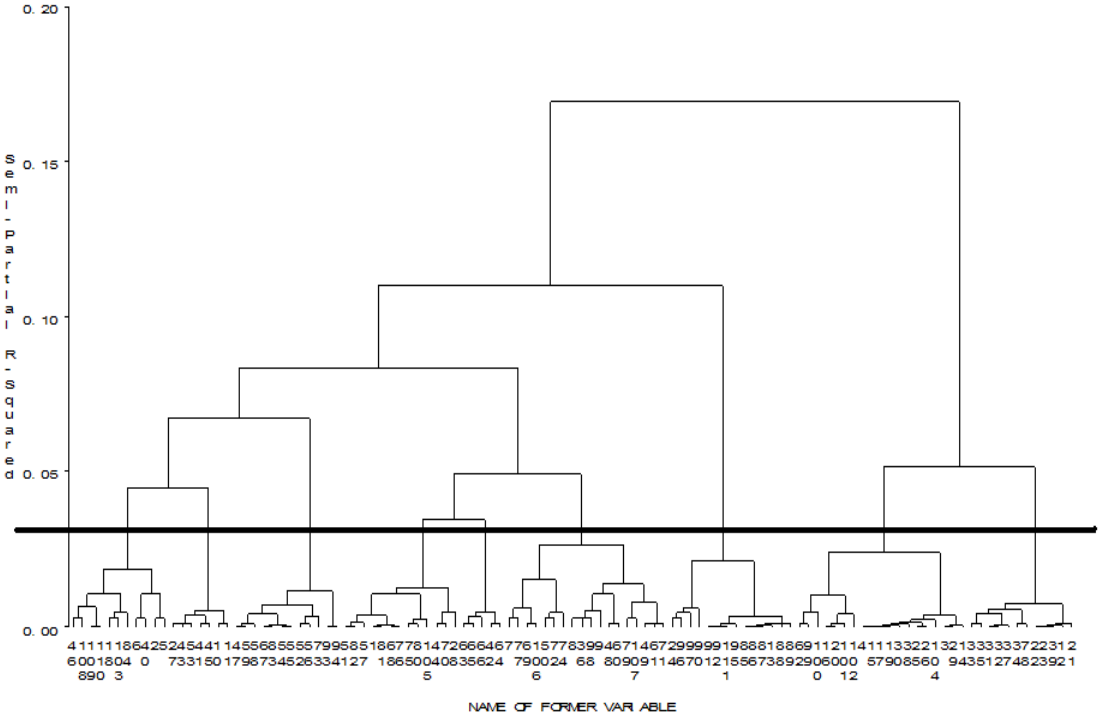

**Figure 3.** Ontario tree diagram with 9 clusters

## 6 Summary

The Concept System software did not provide us with a robust technique to help determine the appropriate number of concepts that would have been used to group the generated statements. In the initial absence of an objective method, the research team had to rely on visual judgment based on the general dispersion of dots on the scatter plot to establish the number of clusters. After exploring various statistical methods, the cluster procedure within the SAS software proved to be the most satisfactory method to determine the number of clusters. By analyzing the graphical tool computed using the SAS cluster procedure with the Ward method as a parameter, we were able to choose 8 and 9 clusters as the appropriate cut-off number of clusters. The content of every grouping was closely examined by the research team and deemed qualitatively valid by the participants. Therefore, we conclude that in addition to being statistically consistent, the Ward method using SAS has a strong qualitative coherence and is therefore the most satisfactory method to identify the appropriate number of concepts for our study.

## 7 Acknowledgments

Grant awarded by the Social Sciences Humanities Research Council of Canada for the study Canadian linguistic minority and interventions in the field of health

## References

Rui Xu and Donald C. Wunsch II, “Clustering”, IEEE Press / Wiley, 2008.

Brock, G.; Pihur, V.; Datta, S. & Datta, S., “clValid: An R Package for Cluster Validation”, Journal of Statistical Software, 2008, 25, 1-22

Trochim, W. M. K., An introduction to concept mapping for planning and evaluation, Evaluation and Program Planning, 1989, 12, 1 - 16

Novak, J. & Gowin, D., Learning how to learn, Cambridge University Press, 1984