Categorization Process and Data Mining

Maria Suzana Marc Amoretti
Federal University of Rio Grande do Sul (UFRGS), Brazil

INTRODUCTION

For some time, the fields of computer science and cognition have diverged. Researchers in these two areas know ever less about each other's work, and their important discoveries have had diminishing influence on each other. In many universities, researchers in these two areas are in different laboratories and programs, and sometimes in different buildings. One might conclude from this lack of contact that computer science and semiotics functions, such as perception, language, memory, representation, and categorization, reflect our independent systems. But for the last several decades, the divergence between cognition and computer science tends to disappear. These areas need to be studied together, and the cognitive science approach can afford this interdisciplinary view.

This article refers to the possibility of circulation between the self-organization of the concepts and the relevance of each conceptual property of the data-mining process and especially discusses categorization in terms of a prototypical theory, based on the notion of prototype and basic level categories. Categorization is a basic means of organizing the world around us and offers a simple way to process the mass of stimuli that one perceives every day. The ability to categorize appears early in infancy and has an important role for the acquisition of concepts in a prototypical approach. Prototype structures have cognitive representations as representations of real-world categories.

The senses of the English words *cat* or *table* are involved in a conceptual inclusion in which the extension of the superordinated (animal/furniture) concept includes the extension of the subordinated (Persian cat/dining room table) concept, while the intention of the more general concept is included by the intention of the more specific concept. This study is included in the categorization process. Categorization is a fundamental process of mental representation used daily for any person or for any science, and it is also a central problem in semiotics, linguistics, and data mining.

Data mining also has been defined as a cognitive strategy for searching automatically new information from large datasets or selecting a document, which is possible with computer science and semiotics tools. Data mining is an analytic process to explore data in order to find interesting pattern motifs and/or variables in the great quantity of data; it depends mostly on the categorization process. The computational techniques from statistics and pattern recognition are used to do this data-mining practice.

BACKGROUND

Semiotics is a theory of the signification (representations, symbols, categories) and meaning extraction. It is a strongly multi-disciplinary field of study, and mathematical tools of semiotics include those used in pattern recognition. Semiotics is also an inclusive methodology that incorporates all aspects of dealing with symbolic systems of signs. Signification join all the concepts in an elementary structure of signification. This structure is a related net that allows the construction of a stock of formal definitions such as semantic category. Hjelmeslev considers the category as a paradigm, where elements can be introduced only in some positions.

Text categorization process, also known as text classification or topic spotting, is the task of automatically classifying a set of documents into categories from a predefined set. This task has several applications, including automated indexing of scientific articles according to predefined thesauri of technical terms, filing patents into patent directories, and selective dissemination of information to information users. There are many new categorization methods to realize the categorization task, including, among others, (1) the *language model based classification*; the maximum entropy classification, which is a probability distribution estimation technique used for a variety of natural language tasks, such as language modeling, part-of-speech tagging, and text segmentation (the theory underlying maximum entropy is that without external knowledge, one should prefer distributions that are uniform); (2) the *Naïve Bayes classification*; (3) the *Nearest Neighbor* (the approach clusters words into groups based on the distribution of class labels associated with each word); (4) *distributional clustering of words to document classification*; (5) the *Latent Semantic Indexing (LSI)*, in which we are able to compress the feature space more aggressively, while still maintaining high document classification accuracy (this information retrieval method improves the user's ability to find relevant information, the text categorization method based on a combination of distributional features with a Support

Vector Machine (SVM) classifier, and the feature selection approach uses distributional clustering of words via the recently introduced information bottleneck method, which generates a more efficient representation of the documents); (6) the *taxonomy* method, based on hierarchical text categorization that documents are assigned to leaf-level categories of a category tree (the taxonomy is a recently emerged subfield of the semantic networks and conceptual maps). After the previous work in hierarchical classification focused on documents category, the tree classify method internal categories with a top down level-based classification that can classify concepts in the document.

The networks of semantic values thus created and stabilized constitute the cultural-metaphorical 'worlds' which are discursively real for the speakers of particular languages. The elements of these networks, though ultimately rooted in the physical-biological realm can and do operate independently of the latter, and form the stuff of our everyday discourses (Manjali, 1997, p. 1).

The prototype has given way to a true revolution (the Roschian revolution) regarding classic lexical semantics. If we observe the conceptual map for *chair*, for instance, we will realize that the choice of most representative chair types; that is, our prototype of chair, supposes a double adequacy: referential because the sign (concept of chair) must integrate the features retained from the real or imaginary world, and structural, because the sign must be pertinent (ideological criterion) and distinctive concerning the other neighbor concepts of chair. When I say that *this object is a chair*, it is supposed that I have an idea of the *chair sign*, forming the use of a lexical or visual image competence coming from my referential experience, and that my prototypical concept of chair is more adequate than its neighbors *bench* or *couch*, because I perceive that there is a back part and there are no arms. Then, it is useless to try to explain the creation of a prototype inside a language, because it is formed from context interactions. The double origin of a prototype is bound, then, to shared knowledge relation between the subjects and their communities (Amoretti, 2003).

MAIN THRUST

Hypertext poses new challenges for a data-mining process, especially for text categorization research, because metadata extracted from Web sites provide rich information for classifying hypertext documents, and it is a new kind of problem to solve, to know how to appropriately represent that information and automatically learn statistical patterns for hypertext categorization. The use of technologies in the categorization process through the making of conceptual maps, especially the possibility of creating a collaborative map made by different users, points out the cultural aspects of the concept representation in terms of existing coincidences as to the choice of the prototypical element by the same cultural group. Thus, the technologies of information, focused on the study of individual maps, demand revisited discussions on the popular perceptions concerning concepts used daily (folk psychology). It aims to identify ideological similarity and cognitive deviation, both based on the prototypes and on the levels of categorization developed in the maps, with an emphasis on the cultural and semiotic aspects of the investigated groups.

It attempted to show how the semiotic and linguistic analysis of the categorization process can help in the identification of the ideological similarity and cognitive deviations, favoring the involvement of subjects in the map production, exploring and valuing the relation between the categorization process and the cultural experience of the subject in the world, both parts of the cognitive process of conceptual map construction.

The concept maps, or the semantic nets, are space graphic representations of the concepts and their relationships. The concept maps represent, simultaneously, the organization process of the knowledge, by the relationships (links) and the final product, through the concepts (nodes). This way, besides the relationship between linguistic and visual factors is the interaction among their objects and their codes (Amoretti, 2001, p. 49).

The building of a map involves collaboration when the subjects/students/users share information, still without modifying the data, and involves cooperation, when users not only share their knowledge but also may interfere and modify the information received from the other users, acting in a asynchronized way to build a collective map. Both cooperation and collaboration attest the autonomy of the ongoing cognitive process, the direction given by the users themselves when trying to adequate their knowledge.

When people do a conceptual map, they usually privilege the level where the prototype is. The basic concept map starts with a general concept at the top of the map and then works its way down through a hierarchical structure to more specific concepts. The empirical concept (Kant) of *cat* and *chair* has been studied by users with map software. They make an initial map at the beginning of the semester and another about the same subject at the end of the semester. I first discussed how cats and chairs appear, what could be called the structure of cat and chair appearance. Second, I discussed how cat and chair are perceived and which attributes make a cat a cat and a chair