Chapter 25
Visualization and Storage of Big Data for Linguistic Applications

Dan Ophir
Ariel University, Israel

ABSTRACT

The following two main tendencies occur: 1) increase in the amount of the computational power around the world and increasing its sensitivity and functionality (communication, imaging, voice recording, position retrieval, etc.) causing growth of data; 2) decrease in the qualifications which are required to operate that computational power are two phenomena feeding themselves mutually: increasing the amount of PC's, Laptops, iPads, iPhones available with simpler and more intuitive operating instructions. This situation requires simplifying the data perception by the more developed human sense – the vision by untrained person even by a kid who doesn't know how to read and write. Such approach may easily make the media accessible to more and more people. Therefore, so called human interfaces are mainly supported by the vision, the most accurate human sense which demands developing the present methodology of visualizing huge data.

INTRODUCTION

The universe may be considered as a super object containing a system of objects with mutual properties and relations. The objects, properties and the relations are dynamic and thus generate new information freely. Humans collect some of these signals of information and this phenomenon occurs *exponentially*. This information is transformed mainly to at least one of the following forms: images, sounds, physical data and linguistic interpretations.

Here a philosophical question arises regarding the context of the mathematical data and what implication it may have on understanding the matter: "Is the mathematical theorem an invention or a discovery?" I will let the reader be puzzled. The above issue is comparable to the fact that the treatment of data in the context of Big Data is relevant only to the interpreted data or to candidate data for such an interpretation and not to "undiscovered existing data".

DOI: 10.4018/978-1-5225-3142-5.ch025

Verbal data are an example of such a transformation of the existing data to relevant data which is in our scope, namely: AOI – Area of Interest. Verbal data refers to data represented in a comprehensible manner i.e. in a language invented or developed by human beings such as natural language, programming language, symbolic language for mathematical expressions, *body language* or any combination of them. Such combinations would approximately describe the physical phenomena. Note that verbal language would never entirely represent the physical world; it serves only as its model. Verbal language is discrete and the real world is assumed continuous (excluding quantum theory).

This situation has strongly motivated searching to upgrade the existing methodology used for treating data. The classic means of manipulating data appears to be insufficient. The current chapter focuses on treating various aspects of the verbal information, especially its *visualization***,** which enables its quick comprehension. The status of the present ways of coping with the enormous amount of data will be summarized.

The treatment of data is classified into two main categories: morphological or significant. A morphological category of data is a category whose members have a different shape but it can have the same meaning, for example, synonyms in a natural language. A significant category is a category in which the members have a different significance but it may have the same configuration, for example, the Adenine, which is a nucleotide in a genome (DNA sequence) that has the same structure independently of its location but its functional significance is different.

Another classification is according to the type of data carrier, which may be digitized characters, sound signals, images or movies. The purpose of the data is another property; the data can be classified as professional, colloquial or serving the entertainment. Another way of classifying data is by differentiating the origin of the data, whether it is a human being or some sensors such as climate indicators (direction, temperature and velocity) or images made by a camera in a closed loop.

The present chapter will emphasize the visualization of data related to the processing of various types of languages. The languages considered were chosen from various fields of science and life: natural languages (text and speech), programming language, body languages, mathematical and formal languages, and the language of DNA. There are several degrees of visualization. The same information might be transmitted in several channels. For example, some story may be represented as text in a book or as a movie which is in some ways higher order visualization.

MORPHOLOGY VERSUS SIGNIFICANCE

Morphology

Textual and phonetic morphology is quite developed. There are operations such as morphological searching, compression (Salomon, D., 2008), encryption / decryption (Goldreich, 2004) or recognition – optical (OCR Optical Character Recognition) (Schantz, 1982), which is alphabet oriented or phonetic (Yu, D. & Deng, L., 2014). Phonetic morphological recognition can be distinguished by several types: identification of phonemes, identification of the speaker, identification of the speaker's gender and age estimation.