Chapter 20

Application of the Cluster Analysis in Computational Paleography

Loránd Lehé Tóth
Budapest University of Technology and Economics, Hungary

György András Jeney
Budapest University of Technology and Economics, Hungary

Raymond Eliza Ivan Pardede
Budapest University of Technology and Economics, Hungary

Ferenc Kovács
Budapest University of Technology and Economics, Hungary

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

ABSTRACT

This chapter presents a method to determine the actual version of a script used in constructing of a script relic from unknown origin. The glyphs belong to graphemes as models are realized in the relics as symbols. Some group of glyphs may transform their shape (shapeshifting) through time which produces various versions of scripts that use different glyphs to express the same grapheme. These glyph variants can be identified from extant relics, mainly from historical abecedaries that are used as references. Our algorithm can determine whether or not an abecedary is related to the symbols of a relic from unknown origin by means of the canonical decomposition of the glyphs and symbols. From there an aggregated value called fingerprint is created and it is unique for each relic. The fingerprints then are evaluated by clustering technique using various metrics. As the result of performing comparative evaluations the Minkowski metric provides the most interpretable clustering structure. The results of the evaluations, conclusions, and future work are also presented.

1. INTRODUCTION

The writing system is a symbolic representation of a language described in terms of linguistic units (Malatesha & Aaron 2006). The script is the graphic format in which a writing system is represented. Some examples of scripts are the following: Aramaic, Arab, Batak, Brahmi, Carian, Carpathian Basin

DOI: 10.4018/978-1-4666-9479-8.ch020
A character is an extension of grapheme with a computer code assigned to it. The orthography is a representation system for using a particular script to write a certain language. A script can have various versions depending on the age and place; each of them has its own collection of glyphs of graphemes called an alphabet. The inscription of unknown origin is an extant inscription whose origin, date, and author are unknown. The inscription under test is the examined inscription and, usually, has unknown origin.

The explored inscriptions that were written in ancient times provide challenges for researchers to identify them. The reason for this apart from missing the writing support materials (wood, stone, brick, paper, etc.) is that the glyphs of the graphemes used in these scripts might have changed numerous times over the ages. Moreover, it is also possible that different ancient script relics were written in different handwritings or written by people with altering writing skills. As a result, the topological properties and styles of glyphs used in script relics may vary during history.

This chapter introduces a new method for identifying the version of a certain script (an actual alphabet) used for making an examined inscription of unknown origin. We suppose that the script used for writing the examined inscription is already identified; however, the actual version of the script used for making that particular inscription is still undetermined. The basis of the method is comparing the symbols of the examined inscription to the glyphs of historical abecedaria and other deciphered script relics of a certain script. The method was verified by applying it to the Szekely-Hungarian Rovash (other spelling: Rovas, pronounced “rove-ash”) script that is used for representing the Hungarian language (Hosszú 2013b). This script was selected for verification because

1. In the Szekely-Hungarian Rovash script each grapheme generally corresponds to one phoneme in the Hungarian language; in such a way, the symbols of an examined rovash inscriptions of unknown origin can easily be associated to rovash graphemes and to appropriate Hungarian phonemes.
2. The Szekely-Hungarian Rovash script has several different versions (alphabets), which are more-or-less known by extant abecedaria and deciphered relics of known origin; therefore, these alphabets are usable to verify the developed identification method.
Related Content

Participatory Management of Protected Areas for Biodiversity Conservation and Social Inclusion: Experience of the SimParc Multi-Agent-Based Serious Game
Jean-Pierre Briot, Marta de Azevedo Irving, José Eurico Vasconcelos Filho, Gustavo Mendes de Melo, Isabelle Alvarez, Alessandro Sordoni and Carlos José Pereira de Lucena (2017). Multi-Agent-Based Simulations Applied to Biological and Environmental Systems (pp. 295-332).
www.igi-global.com/chapter/participatory-management-of-protected-areas-for-biodiversity-conservation-and-social-inclusion/173224?camid=4v1a

Making Memories of a Lifetime
www.igi-global.com/chapter/making-memories-lifetime/10171?camid=4v1a

Architecture with Multi-Agent for Environmental Risk Assessment by Chemical Contamination
www.igi-global.com/chapter/architecture-with-multi-agent-for-environmental-risk-assessment-by-chemical-contamination/173219?camid=4v1a

An Exposition of Uncertain Reasoning Based Analysis: An Investigation of Total Hip Arthroplasty
www.igi-global.com/chapter/exposition-uncertain-reasoning-based-analysis/53306?camid=4v1a