www.ichim.org

Les institutions culturelles et le numérique

Cultural institutions and digital technology

École du Louvre
8 - 12 septembre 2003

SOFT ONTOLOGIES AND SIMILARITY CLUSTER TOOLS TO FACILITATE EXPLORATION AND DISCOVERY OF CULTURAL HERITAGE RESOURCES

Aviles Collao, J., Diaz-Kommonen, L., Kaipainen, M., Pietarila, J., University of Art and Design Helsinki

« Acte publié avec le soutien de la Mission de la Recherche et de la Technologie du Ministère de la Culture et de la Communication »

Abstract

In this essay, we describe a system and tools that we are creating and that allows us to produce similarity (SC) cluster representations of the contents of our Culture Heritage (CH) Forum. The motivation for introducing these types of technologies in the context of cultural heritage materials is to allow the use of spatial vector-based computational techniques that result in similarity mapping. We propose that similarity mappings can be used to build navigation tools that enable users to explore these types of materials without knowing the search parameters in advance. Also, this method of access to cultural heritage materials might be less dependent of a priori determined ontologies than conventional curatorial practices. This is on line with our final objective of developing representation methods and visualization tools that describe cultural heritage material while at the same time allowing the user freedom to interpret the material, i.e. the open interpretation approach.

Introduction

One of the aims of the CIPHER project is to develop innovative technologies and methodologies that enable the exploration of large information repositories containing cultural heritage knowledge on a global scale. We want to use these tools to empower visitors to investigate, navigate, and research the contents of the CH Forums created. In addition, users will be encouraged to produce their own personal spaces, as well as shared spaces owned by emerging communities of interest.

Cultural heritage artifacts

The system is being tested with two cultural heritage artifacts: A *Description of the Northern People, 1555*, and the *Carta Marina of 1539*. Olaus Magnus, the last Catholic bishop of Uppsala, Sweden is the author of both of these items. Considered by many to be

one of the great achievements of European cartography, *Carta Marina* provided the first comprehensive description of the landscape, the people, and the customs of the Nordic region. In addition to documenting many of the ethnographic aspects, the pictorial elements also elaborate the mythical narratives of the region. This is particularly true of the myriad representations of monstrous figures included in the map. The map consists of nine separate wood-cut sheets put together into a map, the total size of which is 1,25 m x 1,70 m.

It can be argued that the *Description of the Northern People, 1555* is the written counterpart of *Carta Marina*. It is a chronicle written in Latin containing 22 Books. Each Book is further subdivided into chapters for a total of 778 chapters. The work examines the history, landscape, beliefs and customs of the people in the Nordic countries.

Similarity cluster maps computed by the Self-Organizing Map (SOM)

In our study, similarity cluster (SC) maps are computed using the Self-Organizing Map algorithm, which has the advantages of being more flexible with respect to data representation and updating than e.g. multidimensional scaling techniques. Generally, artificial neural networks (ANNs) can be thought about as non-linear, multilayered, parallel regression techniques. There are two classes: supervised and unsupervised ANNs. Supervised ANNs are techniques for extracting from data input-output relationships and for storing those relationships into mathematical equations that can be used for forecasting or decisions making. Unsupervised ANNs can be used as techniques for classifying, organizing and visualizing large data sets.

The Self-Organizing Map (SOM) [1] is an example of an unsupervised artificial neural network approach. This approach—SOM—has been around since the early 1980's and has widely been applied in engineering and many other fields. References to applications of unsupervised neural networks and SOM can be found in [2]. Through a process called self-organization, SOM configures the output nodes into a topological representation of the original data and organizes similarity clusters that can be seen as soft edged classes or

fuzzy sets emerging from statistical correlations. [3] There is also a reduction of high-dimensional data so it can be projected into a 2D space. The SOM can thus serve as a clustering tool as well as a tool for visualizing high-dimensional data.

A self-organizing map consists of two layers of processing nodes: The first is an input layer containing processing nodes for each element in the input vector; the second is an output layer or grid of processing nodes that is fully connected with those at the input layer.

Visualization and architecture of Self-Organizing Maps

The task of visualization is to deal with the design of the visual representation of data objects and their relationships. Aside from being effective sources of communication, good visualizations can provide the ability to comprehend huge amounts of data, reduce the visual time needed to comprehend the information presented, reveal relations otherwise not noticed, as well as facilitate hypothesis formulation. From this point of view, SOMs have been described as a promising algorithm for organizing large volumes of information. [4]

SOM can provide easy visualization and is able to detect isolated patterns and structures from large data set. SOM can also allow the user to browse a graphic display of the data based on similarities. In contrast to conventional data retrieval techniques that assume that the user knows what to search for, this approach might better support an exploratory approach to data. However, many issues remain to be solved about interface solutions to such techniques.

General description of system

The system we have designed is available to participants in the CIPHER project via the World Wide Web (WWW). The purpose of the system is to integrate the methods and tools that allow us to create similarity cluster maps of her data. Overall the work of the

system is currently divided into two major areas. There are the tasks related to analytical processing of data so as to prepare it for encoding and subsequent display as a SC map. These tasks can be divided into: Input data, analysis, algorithm and semantic space construction, and visualization. Then there is the area that pertains tasks related to the control of the flow of work between the different components of the system.

Additionally there are the tools that allow for automatic input of data into the system. Among the tools we have created and are testing is an *Automatic Description Engine* (ADE). With this tool textual corpora can be processed into a document vector that allows us to generate a SC map. There is also a tool that permits the domain expert to manually enter a *Soft Ontology Layer* (SOL) of non-hierarchical properties and feature descriptions of a heritage object that are then clustered according to similarity relations and displayed in a SC map. The tools differ from each other primarily in that the methodologies used in encoding the data are distinct. Both tools, however, make use of the SOM algorithm for clustering of the data according to relations of similarity.

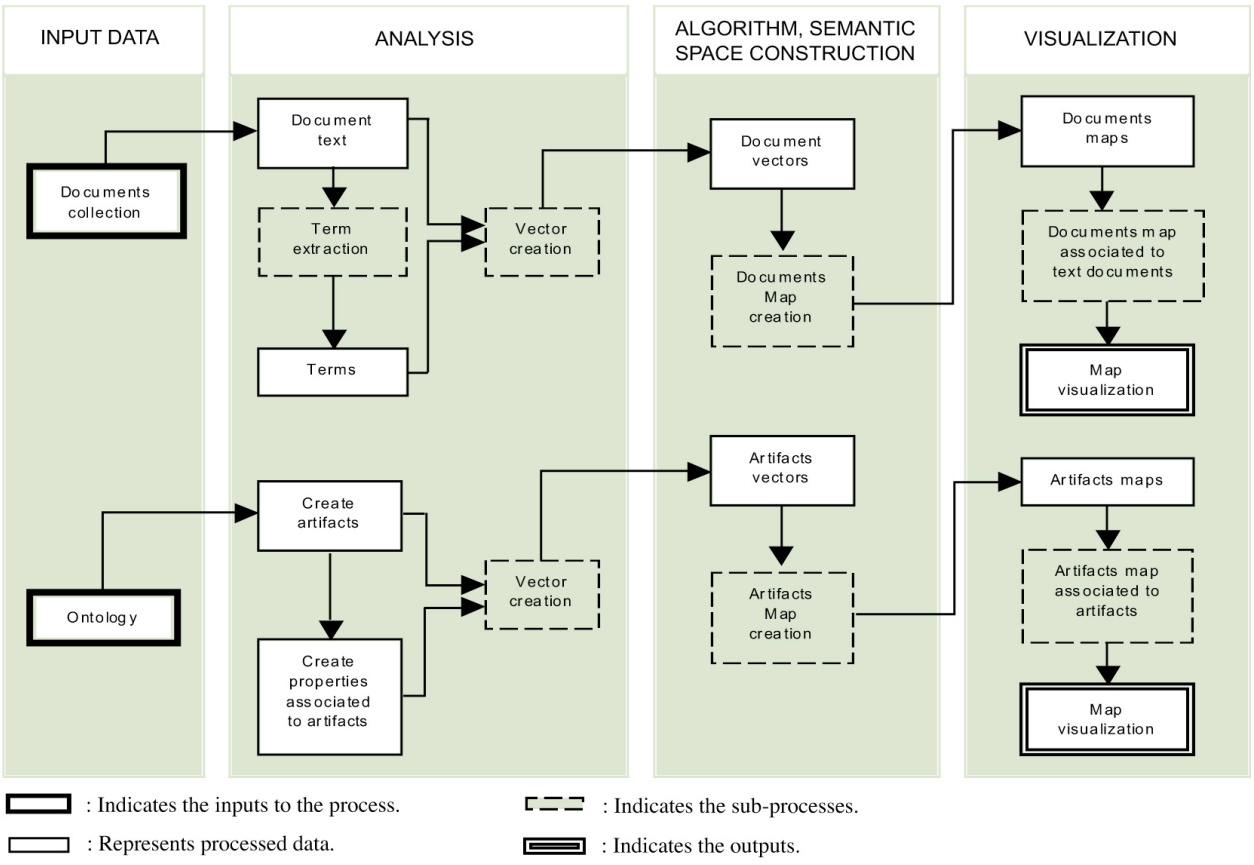

Fig. 1 : Analytical tasks of the system.

Automatic Description Engine (ADE) and semantic space construction with Latent Semantic Indexing (LSI)

A *Document Collection* is a collection that refers to a set of documents. A *document* refers to a piece of textual information about one item of the information. A *term* can be simply a *content word*, or it can also be construed as a *multi-word concept*, further representative of a concept in the document domain. A *semantic space* can be derived on the basis of analyzing relationships among terms and documents in the entire collection.

The initial step of converting the textual information of *A Description of the Northern People* into a *Documents Collection* of digital format occurs through a scanning process. This process is also used for capturing the illustrations, titles and notes contained in each chapter. Though the title and images that accompany each chapter are entered into the database, they are not factored in the semantic analysis process but rather, reserved for later use in the visualization tasks. The program we use in the scanning of the textual components is Optical Character Recognition OmniPage Pro X, Scansoft. The data obtained via this process is divided into documents and stored in a database.

The formatted file of the database is processed so as to extract a semantic space. The semantic space is derived on the basis of analysing relationships among linguistic object-such terms, sentences, or documents-in, the entire *Documents Collection*.

The method we are currently using to create the semantic space is Latent Semantic Indexing (LSI). Latent Semantic Analysis or Latent Semantic Indexing, LSA/LSI is a mathematical learning method based on statistics, developed in the late 1980's. [5][6]

LSI takes advantage of the implicit higher-order structure of the association of terms with documents to create a multi-dimensional semantic structure of the information. [7] Through the pattern of co-occurrences of words, LSI is able to infer the structure of relationships between documents and terms. Singular-value decomposition (SVD) of the term by article association matrix is computed producing a reduced dimensionality matrix containing the best K orthogonal factors to approximate the original matrix as the model

of *semantic space* for the collection. This semantic space reflects the major associative patterns in the data while ignoring some of smaller variations that may be due to idiosyncrasies in the term usage of individual documents.
LSI uses no linguistic knowledge such as morphology or grammar to calculate this similarity between elements in the text collection. LSI also does not take word order into consideration but treats the text collection like a "bag of words." In addition, since LSI does not use any linguistic knowledge but only statistics of co-occurrences, the language or languages of the text collection does not matter.

Pre-processing of the text using LSI includes the following steps: Input data, term extraction, vector creation, document vector. We are now working with diverse applications of the LSI algorithm including the Telcordia Latent Semantic Indexing software, by Telcordia Technologies, Inc., and the General Text Parser (GTP) developed by S. Howard, H. Tang, M. Berry, and D. Martin at the University of Tennessee (Department of Computer Science). Both applications of the LSI algorithm can be licensed for non-commercial purposes.

Soft Ontology Layer (SOL) and Random Vector Encoding (RVE)

The method used for encoding data in SOL was originally developed [8] as an automatic method for encoding large text corpora, in order to be further processed into similarity clusters by the self-organizing map. The great advantage of this method is the flexibility with respect to the number of attributes. Unlike with LSI, these may change as needed during the encoding work.
Each property that the user identifies as a defining attribute of one or more artefacts of the domain is assigned an attribute vector. It is a high dimensional vector of which the components are randomly chosen (Random Vector Encoding, or RVE). The dimension is globally fixed to a large number, e.g. 90, to ensure that each attribute vector is dissimilar with respect to other attribute vectors.

The attribute vectors altogether form a global set of attributes that is used to describe every object of the domain. The number of attribute vectors can grow infinitely as the

description of the domain requires. That is, the ontology is flexible, not assuming any fixed number of attributes.
For every object there is a weight vector with the globally shared dimensionality that specifies the degree to which each of the attributes characterizes it. For every attribute, the (expert) estimates a weight that may correspond to her judgment of the relevance, membership, probability of occurrence, propiscuousness, or size of the property the attribute corresponds to. The object-specific weight vector consists of the weights for each of the attributes.

The unique description of an object using the RVE is computed as the object-specifically weighted sum of all of the attribute vectors, or the object vector. For most computational purposes, this will usually need to be normalized.

We have used the method to create an ontology that features descriptions of the monster figures from the *Carta Marina*. With the exception of the real versus mythical category, the encoded descriptions of the monsters focus on the pictographic aspects of the creature represented. These pictographic aspects are elaborated within a framework having to do with prototypical, or family resemblance, categories. The initial attribute vectors represented are:

The position of the figure with respect to the physical characteristic of the map. That is the section or plate in which it is inscribed, with every alternative section as an independent attribute vector.
Two coordinate vectors for relative latitude and longitude.
The type of figure in terms of real or mythical animal, expressed as separate attribute vectors. For this latter mythical aspect, additional descriptive properties are assigned. The properties are constructed according to degrees of similarity. For example: Is this sea monster serpent-like?
The basic physiognomy of the creature as it appears in the map. This splits into a multitude of properties including basic characteristics such as physical disposition of the head, trunk, extensions, and appendages, represented by an attribute vector for each.

The location of the figure in relation to its geographic surroundings. For example, is it a land or sea creature? This is represented as an attribute vector assigned to every alternative.

The data has been entered manually one by one, determining the weight for each attribute. We are recording typical questions and problems encountered in this attempt, as well as ideas and implications that can be used for further development of the representation and the interface.

Clustering and representing the data with SOM_PAK

In both ADE and SOL, the SOM neural network model is used to cluster and represent the data. There are many implementations of the SOM algorithm. We are currently using SOM_PAK, and SOM Tools. We seek to extend the scope of these tools, especially with regards to issues pertaining interface design and visualization.

SOM_PAK is a software tool that contains all the programs for the correct application of the self-organizing algorithm. [2]. SOM_PAK runs under UNIX and MS_DOS, and it requires only a (ANSI-C) C compiler. The SOM_PAK's user interface may be used only from the shell command line. The package and the documentation of SOM_PAK can be downloaded from the Internet. [9] The "SOM Toolbox" is a proper SOM-library created for the Matlab computing environment. [10]

Relevance of Soft Ontologies in the cultural heritage domain

The relevance of the similarity cluster maps depends solely on the representation. Meaningful similarity appears on the SC map level only if the representation includes attributes with respect to meaningful difference and similarity that occur in the data, supporting exploratory search, and discovery.

The soft-ontology approach allows for the expert to create rich descriptions of a domain that are based on a natural language. In addition, it supports the open interpretation approach allowing the user to view the data from her/his perspective without being fixed to the curator's point of view.

Conclusion

Similarity clustering representations of objects allow a spatially defined ontology, and a range of mathematical operations, in particular those that make it possible to compare the vector in terms of proximity and distance, corresponding to holistic similarity vs. difference. SC maps based on the spatially defined ontology of objects, can be used as the basis for similarity-based exploratory search.

The similarity clustering (SC) representation, or soft ontology, is inherently flat, not offering any means of explicit expression of hierarchy. On the other hand, similarity maps computed from data that have intrinsic hierarchical properties will display hierarchical order in terms of embeddings of map responses. It serves the goal of open interpretation not to include the hierarchic structure of the researcher to the representation.

Bibliography

[1] Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. Biological Cybernetics 43:59-69.

[2] Kohonen, T. (1997, 1995) Self-organizing Maps, Berlin:Springer-Verlag.

[3] Zadeh, L.A. (1965) Fuzzy sets. Information and Control, 8, 338-353.

[4] Börner, K., Chaomei, C. and K. Boyack. (2003) "Visualizing Knowledge Domains," Annual Review of Information Science and Technology, Vol. 37. (In press.)

[5] Deerwester, S., Dumais, S., Landauer, T. Furnas, G., and Harshman, R., (1990) "Indexing by latent semantic analysis", Journal of the Society for Information Science, 41(6), pp. 391-407.

[6] Landauer, T., Foltz, P. and D.Laham. (1998) "Introduction to Latent Semantic Analysis" Discourse Processes 25, pp. 259-284.

[7] Dumais, S., Furnas, G., Landauer, T., and Deerwester, S., (1988) "Using latent semantic analysis to improve information retrieval," Proceeding of CHI'88 Conference on Human Factors in Computing, New York, ACM, pp. 281-285.

[8] Honkela, T., Kaski, S., Lagus, K., Kohonen, T., (1996) Newsgroup Exploration with WEBSOM Method and Browsing Interface. Helsinki University of Technology, Lab. Of Computer and Information Science. Report A 32.

[9] (http://www.cis.hut.fi/research/som_pak/)

[10] (http://www.cis.hut.fi/projects/somtoolbox)