Organix:
Creating Organic Objects from Document Feature Vectors

Robert J. Hendley, University of Birmingham, UK

Barry Wilkins, University of Birmingham, UK

Russell Beale, University of Birmingham, UK

ABSTRACT

This article presents a mechanism for generating visually appealing but also effective representations for document visualisation. The mechanism is based on an organic growth model which is driven by features of the object to be visualised. In the examples used, the authors focus on the visualisation of text documents, but the methods are readily transferable to other domains. They are also scaleable to documents of any size.The objective of this research is to build visual representations that enable the human visual system to efficiently and effectively recognise documents without the need for higher level cognitive processing. In particular, the authors want the user to be able to recognise similarities within sets of documents and to be able to easily discriminate between dissimilar objects.

Keywords: Document Analysis, Feature Vector, Graphics, Visual Interpretation, Visualisation

1.0 INTRODUCTION

Information visualisation has been defined as “The use of computer-supported, interactive, visual representations of abstract data to amplify cognition” (Card et al., 1999). This can involve many processes, such as filtering, abstracting or re-organising the data. Critically, though, it requires the generation of a visual representation which makes best use of the human visual system to allow efficient interpretation of the information.

DOI: 10.4018/jcicg.2010010104

It is important to note the difference between physical and abstract data. Physical data has some physical form on which the visualisation can be based, for example archaeological data, human body (medical) data, earth data, etc., and this type of data is represented by scientific visualisations. In contrast abstract data has no such basis, examples include World Wide Web data, software modification logs, etc., and is represented by information visualisations. Effective information visualisation methods produce cognitive amplification, in which visualisation methods help to shift the work load from the cognitive to the perceptual system, expand the working memory and allow a high level of interaction. The user is thus aided in their goals of the confirmation and discovery of knowledge.

As the use of the internet increases the amount of information becoming accessible to users grows rapidly. A large percentage of this data is in text form. Often it is unstructured which makes it difficult for users to find specific information in a single document let alone in collections containing hundreds of documents. A common task faced by users is to identify documents similar in content to a particular document they already know is relevant. The most common tool for undertaking this task on the Web is the search engine. Unfortunately, having submitted a query the user is often faced with a high recall to precision ratio. Results tend to be formatted as lists of text 'snippets' which the user then has to scan through in the hope of finding something useful. We propose a novel system that attempts to visualise documents within a collection as organic shapes. It is hoped that by producing visual representations of documents users can identify similar documents more easily. Within the context of search engine results this system could be used on the 'snippets' themselves, or on the entire document. Alternatively the system could be regarded on a more artistic level as simply producing visually interesting shapes.

The following sections describe related work, the model used in the system, results obtained, future work and conclusion.

2.0 RELATED WORK

A wide variety of document collection visualisations have been developed. Bead (Chalmers & Chiston, 1992) uses physically based modelling techniques to produce document clusters. This approach can be computationally complex, an alternative and more efficient algorithm has been developed (Chalmers, 1996). WEBSOM as described by Lagus et al. (1996) uses a self organising map (SOM) algorithm to produce a map of documents with similar documents located in closely related regions of the map. Themescapes (Wise et al., 1995) visualises the thematic content of a document collection as a 3D landscape, stronger themes are give a higher elevation. A network is used by Singhal and Salton (1995), Salton (1995), the resulting structure of the network and the number of incident lines (or degree) at a particular node can give insights into the core documents or paragraphs within a particular article. The research and approaches used for text visualisation are extensive. Card et al. (1999, p409-461) contains a selection of papers discussing 1D, 2D and 3D text visualisation. A comprehensive review of document visualisation has been written by Morse (1998).

A novel system is described by Roher et al. (1998). This approach generates a document feature vector, maps the weights for each feature to distances along each axis and the eight bisecting quadrants, place spheres at the end points and finally produce a 3D amorphous shape. This allows up to 14 dimensions of the document to be viewed as a single shape. Documents can then be compared, with similar documents having similar shapes. It was this idea that inspired the current work.

Chernov faces are very simple 2D line drawings of faces where the features of the face are determined by the data. Since humans are good at recognising faces and facial expressions, it was reasoned that this may be an appropriate strategy for representing data in a form other than a face. The system presented here is based on the same principles. Morris et al. (1999) discuss the merits of such a strategy, and indicate that in theory objects other than faces could be used to represent data in this way.

3.0 MODEL

3.1 Document Feature Vector Extraction

It should be noted that the current system considers individual words within the document to be the base unit. However, it has been written so that two or three consecutive words, sentences

Digital Images: Interaction and Production

Linda Matthews and Gavin Perin (2011). *International Journal of Creative Interfaces and Computer Graphics (pp. 27-41).*

www.igi-global.com/article/digital-images-interaction-production/54332?camid=4v1a