20. BioQSpace: An interactive visualisation tool for clustering MEDLINE abstracts.

Anna Divoli 1,2, Rasmus Winter 2, Steve Pettifer 2 & Terri Attwood 1,2

1 Faculty of Life Sciences & 2 School of Computer Science, The University of Manchester, UK

BioQSpace combines a set of algorithms and an architecture for interactive clustering and browsing MEDLINE abstracts n a 3D virtual environment. The clustering is based on document similarity measures calculated on user-specified eighting of certain attributes (MeSH-terms, word usage, specialised dictionaries, etc). Several selection, navigation and reconfiguration options are provided.

We have developed BioQSpace, a highly interactive visualisation tool for clustering MEDLINE abstracts in 3D space. BioQSpace has been built on an existing application named Q-SPACE[1], a conceptual environment consisting of objects positioned within 3D space, and uses MAVERIK[2], a publicly available virtual reality system.
BioQSpace is an environment presented as a window in a desktop graphical user interface (GUI). The users can query abstracts from PubMed, using an embedded search facility. BioQSpace performs pairwise similarity calculations between all the abstracts based on a set of individual attributes such as structure, function, disease and therapeutic compounds word lists (used by BioIE[3]), MeSH-terms, word usage, PubMed related articles, publication date, and so on. These attribute measures are given more or less importance according to user-manipulated sliders specifying the weight attached to them. Users can easily change these weightings at any time to regenerate the BioQSpace. The collection of weighted measures are summed and scaled to produce a value between 0 (not similar at all) and 1 (identical), giving a final similarity measure between abstracts. A similarity matrix for all abstracts is then produced. This yields a strict triangular matrix with overall similarity values for each pair of abstracts. An ordered list is then created from the matrix, positioning the most similar abstracts at its head and the least similar at its tail, which in turn, generates a Minimal Spanning Tree (MST), using Kruskal's algorithm. The tree is then traversed, 'colouring' each of the nodes (representing the MEDLINE abstracts) to indicate which grouping that node belongs to. A node is considered as belonging to a different group to its parent in the MST if its similarity value to that parent is less than a given threshold. Once the groups of similar abstracts have been identified, a force placement algorithm is used to generate a 3D configuration of the nodes and their links. The algorithm works by calculating repelling forces for all participating nodes, and then calculating attractive forces for only those nodes connected by edges. It considers the dominant nodes of each group first, and then the rest within a given group. Simultaneously, an algorithm for determining and rendering the Minimal Convex Hulls to contain the groups/clusters is executed. Abstracts that are very similar and grouped together, are being represented by the same colour nodes and encapsulated in a semitransparent hull. The hulls can be reshaped dynamically as the nodes move in
space. Besides the weight-based fine-tuned similarity calculation, the users can control navigation through the abstracts (named with their PMID by default) with the mouse or with buttons on the GUI; select abstracts to view their details; highlight abstracts with certain keywords, terms or phrases; or delete irrelevant abstracts. The abstract details include information derived from PubMed as well as additional BioQSpace generated information such as title and abstract top keywords (in format of stems produced using the Porter algorithm) with their scores, based on IDF calculations.
This application may be used by biologists and bioinformaticians, to navigate through the full of complex concepts, biomedical literature, by providing custom requirements. The clustering algorithm

and the user-adjustable weights provide a number of alternative ways to select related articles and the 3D virtual environment allows easy navigation while exploring for associations between various biomedical concepts (entities, diseases, drugs and so on). Cluster-formed comparisons can be made for, for instance, semantic clustering (using only MeSH-term similarity) vs. syntactic clustering (using word usage) vs. specialised interest clustering (using the predefined and user defined word lists and their combinations); whereas clustering based on the publication date can provide hints on the evolution of biological knowledge. More importantly, users' selection of the weights given to a combination of these attributes can provide domain specialised cluster formations, where in turn, the users can create their own literature network by the (optional) trail-lines created during the custom navigation.

References:

[1] Pettifer, S., Cook, J. and Mariani, J. (2001) Towards Real-Time Interactive Visualisation in Virtual Environments: A Case Study of Q-SPACE,
Proceedings International Conference on Virtual Reality 2001, Laval, France, pp.121-129, May.
[2] Hubbold, R., Cook, J., Keates, M., Gibson, S., Howard, T., Murta, A., West, A. and Pettifer, S. GNU/MAVERIK: A micro-kernel for large-scale virtual
environments, Presence, Teleoperators and Virtual Environments, pp.22-34, I SSN 1054-7460, Vol.10 (1), February, MIT Press.
[3] Divoli, A. and Attwood, T.K. (2005) BioIE: extracting informative sentences from the biomedical literature Bioinformatics 21: 2138-2139