

Multi-media based web mining for an information resource

H. S. Tan & S. E. George
University of South Australia, Australia

Abstract

In this paper we examine multi-media based web mining with view to establishing a web-based information resource that is able to retrieve documents based upon multi-media features within the document. A multi-media based 'search engine' potentially far exceeds what could be achieved from a purely textual one, but before such a comprehensive engine could be developed, a data mining process is required to extract appropriate features from unstructured web documents.

Firstly, this paper discusses the need for visualisation of search results and describes some user interface work that has been conducted towards improving the linear list presentation that is commonly found for presenting search results. Secondly, the principles of the self-organising feature map (SOM) are described including the method of training before providing examples of how the SOM can be used for visualisation and how it has been used in various search-related applications. The SOM provides a topological ordering of the clustering. Thirdly, we present how the SOM has been used to present the results of data organisation using the clustering ability of the SOM, operating with appropriate features extracted from the web-pages. The extraction of appropriate features from documents that can then be used to 'index' the document within the multi-media search space.

1 Introduction

This paper presents a novel visualisation method for web-based search in the medical domain, and uses an artificial neural network to do so. It builds upon existing work in visualisation for retrieval tasks which has so far been confined to documents of a specific genre. In particular, the genre of research publication has been utilised, where there is a well defined structure to the document (that importantly includes title and abstract). Typically, keywords from the title or

abstract are utilised to form a clustering of the documents. However, given the vast variety of information within web-based sites we cannot restrict visualisation to such a small genre of documents. This paper broadens the clustering to include many other document 'features' and so we do not restrict the document to being of a particular genre.

First we summarise the range and scope of search in text and non-text-based media considering the issues in text-based search, image and video, audio and textual document image. We include a review of the ways that search results have thus far been visually represented. Secondly, we introduce the artificial neural network, self-organising feature map, that is to be used in this visualisation task outlining how it is 'trained' and subsequently used to cluster data. Thirdly, we explain the feature extraction that we have made from web-documents with various multi-media features, text and image at this current research stage, and how we have utilised these features to make a document clustering. Finally, we present the results of the SOM visualisation and preliminary evaluation of the representation.

1.1 Search mechanisms in standard and non-standard media

1.1.1 Text-based search

Text-based search engines are based upon mechanisms that retrieve appropriate indexing information from the pages that are found. These 'spiders' crawl the 'web' and frequently use meta-tag information if it is supplied in the page. There are many different tags. The keywords tag provides keywords for the search engine to associate with a page and is often used in conjunction with the description tag that returns a description of the page in place of the summary the search engine would ordinarily create.

1.1.2 Image and video-based search

Unlike text that can be alphabetised or numbers that can be ordered, image and video has no general language to aid in its organization. Tools which can understand the content of imagery are still in their infancy, but can provide assistance to search through visual media. At the MIT Media lab Rosalind Picard has developed tools to aid the browsing, retrieval and annotation of video and image documents. Semantic approaches have been devised that involve creating semantic representation of documents and searching among these, as well as a system that learns to retrieve vacation images. Portions of some images are labelled and inferences made to label others in the database. Retrieval is then made based upon learned feedback from the user's input.

1.1.3 Audio search

One of the main applications of audio retrieval is to be found in music information retrieval research. A British initiative has looked at retrieval based upon sound files, listings of MIDI performance directives and encoded score-descriptions. Other approaches have considered using standard approaches of textual information retrieval to music retrieval. Melodies are extracted from

MIDI or music scores, segmented into phrases (a term in text retrieval) using an algorithm that operates with lexical units based on user perception of changes in melodic flow.

1.1.4 Document image search

Other search approaches use the actual document image that is found on a web page to categorise it according to format. There are a variety of different document formats that can be identified from their layout ranging from letter, to newspaper, to memo, journal article and other. Web sites can be categorised according to the images within their sites.

1.1.5 Meta-document criteria

The cognitive overload in web navigation (for any domain) has long since been recognised (Witt and Tyerman 2002) and the need for supporting people in retrieving information is well documented.

1.2 Use of Visualisation in search

Search results are typically presented in a textual linear-list. Gershon et al do find a time when words are more effective than pictures (Gershon, Eick et al. 1998) since experiences, beliefs and culture, all influence the interpretation of pictures. However, there are many problems with such a text based linear representation for search results, not least the competition to be listed at the top of the list; some search engines solicit funds so that ranking on the list can be purchased. Also the linear representation gives little sense of connection between possibly related sites or how much each term in a search query contributed to a given site, although users have an awareness that not all terms in a search query are of equal importance. Grewal et al. (Grewal et al. 1999) made observations about the relative contribution from search terms, and introduced the Results-Wheel to illustrate the contribution from each term to the search result. They also used a Tepee representation to show how much each term contributed to a search result.

There is clearly a place for visualising search results, showing among other things : the connections between sites, the relative importance of search keywords to the retrieval of the site, with view to help the user identify something interesting. Han and Kamber make it clear that the purpose of the visualization is to help the user to identify something interesting found and to make it possible for the user to interact with the system to find more (Han and Kamber 2001). Shneiderman finds that the overriding mantra in visualisation for retrieval is overview first, zoom and filter, then details-on-demand (Shneiderman 1996). Often with the linear list presentation of web search results we loose sight of the 'overview', have no option to 'zoom and filter' and instead are forced to go straight into 'details on demand'.

2 The Self-Organising Map

2.1 Description of the SOM

The Self-Organising Map (SOM) is an artificial neural network that is able to perform a data clustering. One of the most important aspects of the SOM is that it is primarily a visualisation method for the clustering. Unlike statistical methods or other ways of grouping data the SOM provides a topological ordering where the relationships between data items are made apparent. In fact, relationships within data of high-dimensionality are reduced to relationships on a 2-dimensional surface.

Kohonen developed the map and was initially interested in simulating a process of self-organisation that he observed in the 'maps' that occur within the brain (eg the tonotopic map of the auditory cortex). Kohonen developed an artificial neural network called the self-organising feature map (SOM) (Kohonen 1982). The SOM has a 2-dimensional matrix of processing elements (the output layer) and a scalar array of processing elements (the input layer). A training process is necessary to change the initially random weights into values that respond to input data with minimum distance. The output maps forms this 2-dimensional surface that effectively orders the input data once the training is complete.

2.2 Background to visualisation with the SOM

Naturally the SOM has been used for visualisation in many different domains. We focus here particularly upon clustering used for retrieval tasks. Lagus et al (Lagus, Honkela et al. 1996) use WEBSOM method for exploration of clustered full-text articles of Usenet newsgroup. Houston et al (Houston, Chen et al. 1999) describe CancerLit collection of articles is clustered. The collection of articles is 'semi-structured' in that they are all in the format of journal papers and contain an abstract. The keywords that appear in the abstract form the basic feature for clustering. Lin et al (Lin, Soergel, Marchionini 1991) use a self-organising map as a retrieval interface for an on-line biographic system where they particularly looked at AI literature. The words appearing in document titles from the LISA database were used, removing 'stop words' that added no real criteria to the search.

Vesanto (Vesanto 1999) describes visualization that can be done with the SOM. There are three different contexts: (i) Visualising the overall shape and cluster structure of the training data set, (ii) Analysing the prototype vectors (that is the weight vectors), characteristics of clusters and correlation between vector components, and (iii) Examining new data samples. When the training data is visualised there are various representations including the (a) U-matrix that represents the distances of each map unit to each of its immediate neighbours with greyscale shading, (b) the Distance matrix showing the size and shape of each map unit and (c) Similarity coding, where similar map nodes get similar shading. Figure 1 illustrates these different methods of representing the training data.

The SOM has a very important advantage over statistical clustering methods (even those that might be used to visually display results). This advantage is that the SOM develops intermediate categories. These categories may not directly respond to any data item in the training set but represent the generalisation ability of the network; even if an item in the original data set never activates a given output node on the trained SOM, unseen data may respond to these nodes. Statistical clustering methods have no such counterpart since the generated classes fit the data directly and there are no ‘extra’ classes as can be found on intermediate SOM nodes.

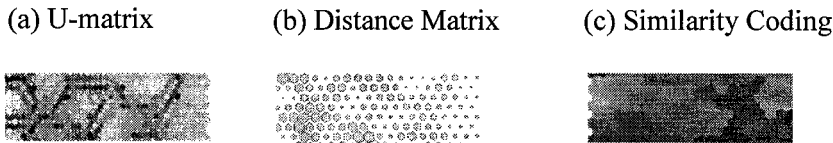


Figure 1. Illustration of different visualisations

Examining the weight vectors of the map can be done with component planes which play a crucial role. Each component plane consists of the values of a single vector component in all map units. Inspection of component planes provides an idea about the spread of values of that component. Finally we can examine new data samples as simply as plotting the bar a distribution of where new examples across the map. These are all general ways of representing the SOM training data, components of the weights and new data.

However, we also identify that where a spatial interpretation can be given to the data we can achieve maps where the nodes represent a particular spatial pattern. Figure 2 (a) illustrates how visual field defects (from the inferior hemi-field) can be plotted on the SOM surface with the blind spot appearing as the dark area on the right hand side of each node (and other visual defects accordingly represented by dark shading). We also have Figure 2 (b) the case where fragments from handwritten digits

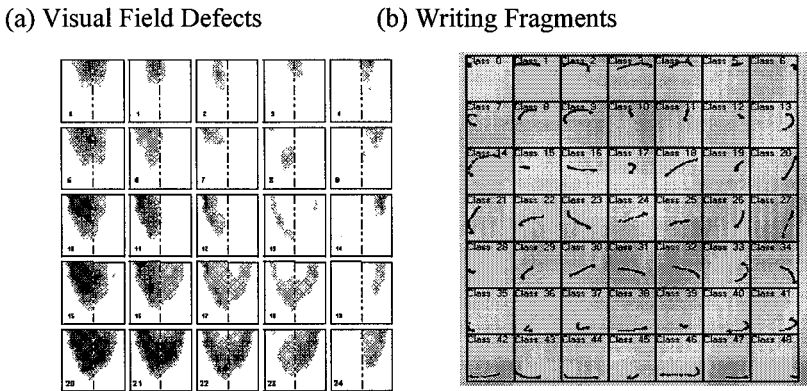
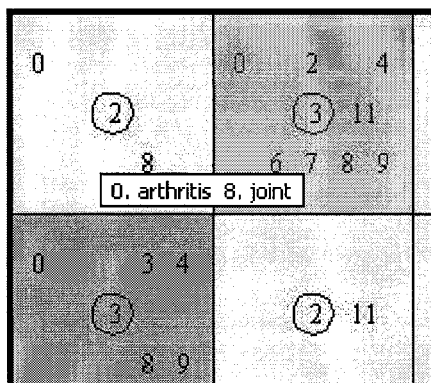


Figure 2. Illustration of different ways in representing the clustering of data with a spatial interpretation.

2.3 Examples of visualisation from document searches

Lin et al. (Lin, Soergel et al, 1991) has used the self-organising feature map (SOM) to present search results among documents in a medical domain. Their representation is similar to that of (Tan, George 2002) that was developed to represent the keywords and document occurrences in a given node. Figure 3 (a) illustrates the top level fragment of a map where the circled number represents document occurrence and the other numbers are keywords, dynamically expanded into text when the cursor is moved over the region. Lagus et al (Lagus, Honkela et al, 1996) use WEBSOM method for exploration of clustered full-text articles of Usenet newsgroup as illustrated in Figure 3 (b). There is a static representation of the words superimposed on the text showing their relative placement over the map.

(a) Keywords from webpage



(b) Text from usegroups

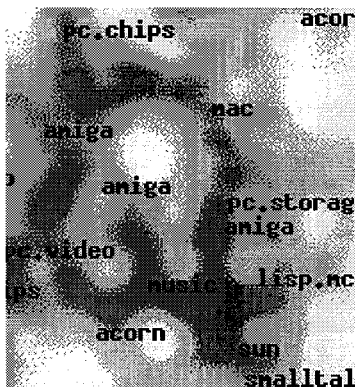


Figure 3. Illustration of clustering documents from a newsgroup and Medical web sites.

3 Feature extraction from web-pages for multi-media based web mining

3.1 Web-page pool

Medical web documents were used fetched from the Internet using the Google Search Engine. The keywords: “arthritis”, “asthma”, “cancer”, and “diabetes” were entered to the search engine and a number of URL’s returned in a linear list. Ten documents for each disease, 40 documents in total, are selected from the results. Using general web page documents extends previous work where the focus of the clustering was confined to either particular documents (with well defined format; abstract, conclusion etc.) or simply the abstracts of documents. It is already a far more ambitious venture to utilise the whole web page. It is important to generate a pool of static web pages since we can control the pool

and make sure only approved medical sites are included, and so not include any random sites from unverified sources where information may be misleading, wrong, incomplete or dangerous. Given the legal implications of medical mispractice, any tool that is to be seriously used by either medical practitioner or other, must contain information from a verified site.

3.2 Extracting text features

The keywords of a document are determined by using the steps that usually occur in automatic indexing: (1) Controlled vocabulary look up, (2) Removing the Stop Word, (3) Word stemming and (4) Term phrase formation.

The vocabulary used for controlled vocabulary lookup is Unified Medical Language System (UMLS 2002) and its NORM lexical tool is used to help do word stemming and term phrase formation. It is also necessary to remove the stop words, that is the words that are considered poor for indexing terms, usually the words that have high frequency of occurrence on the sentence, e.g. 'of', 'the', and 'and'.

The indexing text words of a document are determined using a cooperation between the keywords and the indexing words for each document. Each document is transformed to document vector (the vector represents the indexing words). Again, these features are extracted from the whole document rather than the abstract or the title and so are more general than the other works that have clustered documents.

3.3 Extracting image features

The presence of images within the URL is easy to detect from the HTML tagging. We can extract the keywords from (a) the image description, (b) the surrounding text and (c) from the filename itself. Heuristics are used to obtain potential keywords from areas near the location of the image. It is necessary to remove the logo and other advertising material that may not be relevant to the content of the web page. These spurious images are naturally more restricted on specific medical sites than they may be on sites that contain general advertising.

4 SOM visualisation

4.1 Training a SOM and visualising the result

The document vectors, the results of feature extraction from web-pages, are used as the input of the SOM. For example, documents 0 and 1 are represented as follows: $\text{doc0} = [0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1]$ and $\text{doc1} = [0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1]$. Each vector represents the indexing words [arthritis, asthma, blood, care, cause, control, diabetes, disease, health, link, patient, resource, symptom, time, top, treatment] where the element of the vector in given column has value 1 if the indexing word occurs as the keyword in the document, and 0 otherwise.

A SOM with a 16 by 16 output map is trained using the 40 feature vectors of the documents. The SOM was trained for some 5000 training cycles. This combination of the map size and the number of training cycles was found sufficient to produce a number of cluster classes with all input classes included and each input document is mapped to the right class it belongs. Also, the SOM produces not only all the input classes in the output but also new cluster classes, compare to K-Mean clustering algorithm that only produces exactly the number of input classes or cases if we give K maximum value. The ability to produce the new classes make it possible not to re-train the SOM even though there is new document input, as long as the it has been represented in the new class. For example: If we have 4 documents with keywords as the following: doc0 [0, 0, 0], doc1 [1, 1, 1], doc2 [0, 0, 1] and doc3 [0, 0, 0]. If we use K-mean clustering algorithm with K is 4, that is maximum, the output classes are [0, 0, 0], [1, 1, 1] and [0, 0, 1], but if we use SOM, it is possible to produce the following output classes: [0, 0, 0], [1, 1, 1], [0, 0, 1], [1, 0, 1] and [1, 1, 0]. The two latest classes, that are new classes, will be useful when there are new documents with those keywords. The SOM does not need to be re-train.

After the training process is completed, each document is mapped to the node on the map. Then, the map is labelled by comparing the final weights of each node on the map to the mean of all weights. The node is labelled with the name of the indexing words if the weight for the particular indexing word is greater than a threshold, that is the mean of all weights in this case. The nodes with the same names group to form areas on the map because of the cooperative feature of the neighbouring nodes.

The map below is the lowest level of the map, called level 0 map, the numbers in the circle represent the number of documents mapped to the particular node, while the number surrounding the circle are the coding of the keywords: 0 = arthritis, 1 = asthma, 2 = blood, 3 = care, 4 = cause, 5 = control, 6 = diabetes, 7 = disease, 8 = health, 9 = link, 10 = patient, 11 = resource, 12 = symptom, 13 = time, 14 = top, and 15 = treatment. The colours of the area are used to simply represent the complexity of the keywords.

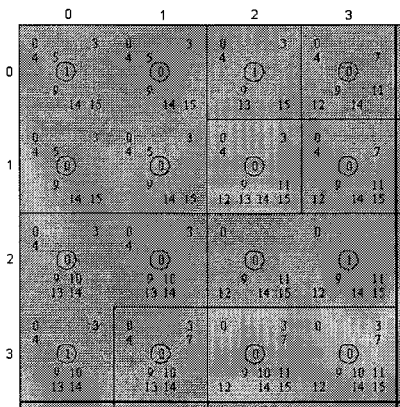


Figure 4. Level 0 Map

4.2 Experiments and evaluation

The result of clustering web documents by involving the image keywords is compared to the result of clustering web documents using only text and title keywords to see whether there is any different between them.

At in previous work, 12 keywords that have most frequent occurrences in all documents are chosen from 308 keywords. These are the indexing word, as shown in table 1a. The number of keywords increases to 621 when involving the images. Again, the keywords that have most frequent occurrences in all documents must be chosen as indexing words.

Table 1. Keywords and their number of occurrences in all documents

(a) Text and Title Keywords

Keyword	Occurrence
Disease	12
Cause	11
Health	11
Treatment	11
Arthritis	10
Diabetes	10
Asthma	9
Cancer	9
Patient	9
Care	8
Joint	8
Pain	8

(b) Involving Image

Keyword	Occurrence
Disease	19
Top	16
Treatment	16
Cause	14
Health	14
Patient	14
Link	12
Symptom	12
Care	11
Resource	11
Arthritis	10
Asthma	10
Blood	10
Control	10
Diabetes	10
Time	10

From table 1, it can be seen that some new words, such as: ‘top’, ‘link’, ‘symptom’, etc, appear as the indexing words when image data is included. This is because these words often occur as the image keywords in the documents. But, there are also some words from table 1 (a) are not chosen as the indexing words involving images and unfortunately, one of these words is ‘cancer’. This is considered important since it is a disease and more significant than other keywords like ‘top’, ‘link’ and ‘time’.

It is possible to include ‘cancer’ as one of the indexing word by reducing the number of minimum occurrences necessary for the keywords to be promoted to indexing words. This has the consequence that there are more indexing words. An alternative approach is to always choose the disease name as the indexing words regardless the number of occurrences.

The change of the indexing words lead to the different mapping of documents, involving the image as keywords. This mapping was richer than the keywords from text alone used to cluster the documents. For example, document 0 that previously mapped to the node labelled by ‘cancer, disease, and treatment’ now is mapped to the node labelled by ‘cause, disease, patient, top, and treatment’ and document 1 previously identified by ‘cause, diabetes, disease, health, and treatment’, now can be identified by ‘blood, cause, diabetes, disease, health, symptom, time, and treatment’.

4.3 Hierarchical visualisation tool

A hierarchical component to the clustering, imposing a hierarchical ‘tree’ structure upon the 2-dimensional space. This is in direct response to Shneiderman’s principle of information visualization which suggests overview first, then zoom and filter, and then details-on-demand (Shneiderman 1996). It is possible that there could be many iterations of the zoom and filter as portions of the space are explored.

The level 0 map, shown in figure 4, can be zoomed out to level 1 map and the level 1 map is zoomed out to the level 2 map, shown in figure 5, by combining the four adjacent grid cells. Then, the node is renamed by choosing the common keywords of the four nodes and responds to the total of four nodes document vectors. For example, node (0,0), (0,1), (1,0), and (1,1) from the level 1 map were combined together to form node (0,0) in level 2 map. The node (0,0) now represents keywords ‘arthritis’ and ‘link’ and responds to 4 document vectors. There are probably no common keywords among the four adjacent nodes, in this case the node will be renamed with all keywords the four adjacent nodes represent, the different colour or style of font can be used to distinct it, meaning that in this level the keywords are unknown but at the next level contains the shown keywords, for example node (1,0) in level 2 map.

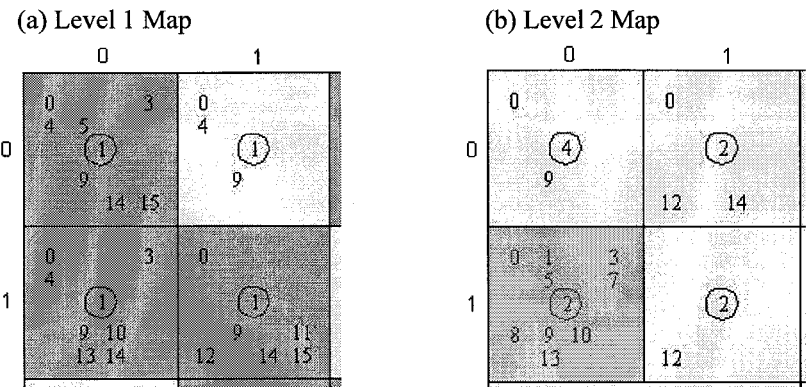


Figure 5. The higher level map

The user then can explore the map from the top/highest level of the map to more detail level of the map at last retrieve the web documents he/she is interested in. For example, given the level 2 map as the top level map, the user that is interested in documents containing 'arthritis' and 'link' can decide either he/she wants to retrieve documents directly or to zoom in to the next level, that will give him/her more document keywords, shown in level 1 map node (0,0), node (0,1), node (1,0) and node (1,1), respectively: (1) arthritis, care, cause, control, link, top and treatment, (2) arthritis, cause, and link, (3) arthritis, care, cause, link, patient, time, and top, and (4) arthritis, link, resource, symptom, top, and treatment. Again, the user can choose to see document contain the keywords or zoom in to the next level, that is level 0 map, to see more document keywords.

5 Conclusions

We believe that it is vital to augment the current linear representation of search results by drawing upon suitable 'features' from the web page. There is particular scope for using the SOM as a visualisation method. Researchers have already performed rudimentary SOM clustering based on text within document abstract or titles. We have extended the clustering to include keywords indexed upon the whole document and so have made the method more generally applicable than those that have relied upon the format of a formal journal paper with title and abstract. This is vital when we consider search within such a general resource as web pages from the Internet.

6 References

- [1] Witt R. and Tyerman S. (2002), Reducing Cognitive Overhead on the World Wide Web, Proceedings of the 25th Australian Computer Science Conference, Melbourne, Australia, Vol 24, No 1. Ed Michael Oudshoorn, Australian Computer Society, 311 – 320, 2002.
- [2] Gershon, N., S. G. Eick, et al. (1998). "Information visualization." *Interactions* 5(2): 9-15.
- [3] Grewal RS, Jackson M, Burden P., Wallis J, A Novel interface for representing search-engine results, (<http://seed.scit.wlv.ac.uk/papers/lostintheweb99.html>, last accessed 15th April 2002).
- [4] Han, J. and M. Kamber (2001). *Data mining: concepts and techniques*. San Francisco, Morgan Kaufmann Publishers.
- [5] Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. Proceedings., IEEE symposium on visual languages, Boulder, CO, USA, IEEE.
- [6] Kohonen, T. (1982). "Self-organized formation of topologically correct feature maps", *Biological Cybernetics* 43, 1982, pp. 59 – 69
- [7] Lagus, K., T. Honkela, et al. (1996). Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration. Proceedings of the

- Second International Conference on Knowledge Discovery and Data Mining. AAAI Press.
- [8] Houston, A. L., H. Chen, et al. (1999). "Medical Data Mining on the Internet: Research on a Cancer Information System." *Artificial Intelligence Review* 13: 347-466.
 - [9] Lin, X., D. Soergel, et al. (1991). A self-organizing semantic map for information retrieval. Proceedings of the fourteenth annual international ACM/SIGIR conference on research and development in information retrieval, ACM Press.
 - [10] Vesanto, J. (1999). SOM-Based Data Visualization Methods. Journal article, in Volume 3(2) of *IDA*. Helsinki University of Technology publications. (<http://www.cis.hut.fi/projects/ide/publications/>, last accessed 24th June 2002)
 - [11] Tan, H. S., S. E. George (2002). Medical Information Resource using a Self-Organising Map for Visualising Web Search Results. 6th World Multiconference on Systemics, Cybernetics, and Informatics (SCI 2002), July 14-18, 2002. Florida, U.S.A.
 - [12] UMLS (2002). Unified Medical Language System. United States National Library of Medicine. (<http://www.nlm.nih.gov/nlmhome.html>), last accessed 24th April 2002).