Neurocognitive Approach to Clustering of PubMed Query Results

Paweł Matykiewicz[1,2], Włodzisław Duch[1],
Paul M. Zender[3], Keith A. Crutcher[3], and John P. Pestian[2]

[1] Nicolaus Copernicus University, Torun, Poland
[2] Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[3] University of Cincinnati, Cincinnati, Ohio, USA

Abstract. Internet literature queries return a long lists of citations, ordered according to their relevance or date. Query results may also be represented using Visual Language that takes as input a small set of semantically related concepts present in the citations. First experiments with such visualization have been done using PubMed neuronal plasticity citations with manually created semantic graphs. Here neurocognitive inspirations are used to create similar semantic graphs in an automated fashion. This way a long list of citations is changed to small semantic graphs that allow semi-automated query refinement and literature based discovery.

1 Introduction

Structured electronic databases and free text information accessible via Internet completely changed the way information is searched, maintained and acquired. Most popular search engines index now well over 10 billion pages, but the quality of this information is low. Instead of boosting productivity increasingly large proportion of time is spent on searching and evaluating results. An ideal search and presentation system should be matched to the neurocognitive mechanisms responsible for understanding information [1]. Visualization of clusters of documents that are semantically related should reflect relations between configurations of neural activations in the brain of an expert. Many books have been written on various *concept mapping* or *mind mapping* techniques [2] that essentially recommend non-linear notes in form of graphs containing interrelated concepts. These techniques are also supported by a large number of software packages, known as *mind mapping* software (see the Wikipedia entry on *mind map*). There are some indications that these techniques indeed help to learn and remember written material in a better way [3]. However, creation of mind maps has so far been manual, and there is a clear need to introduce these techniques in query refinement [4] and literature based discovery [5].

We are especially interested in search and visualization of information in the life sciences domain, therefore the experiments reported below have been done on a PubMed, a collection of over 18 million citations. In [6,7] a prototype of a Visual Language (VL) system has been used on manually create semantic graphs that represent semantically related key biological concepts manually extracted from findings reported in the literature from the PubMed database. The purpose of this paper is to show that similar results can be obtained using computational methods.

M. Köppen et al. (Eds.): ICONIP 2008, Part II, LNCS 5507, pp. 70–79, 2009.

There are several ways in which information retrieval can be improved. A novel approach introduced recently is based on asking the user a minimum number of relevant questions to disambiguate the topic in a precise way [8]. An interesting research direction is based on knowledge-based clustering techniques [9]. In this paper neurocognitive inspirations [1] are used to create semantic graphs that represent PubMed search results. In the next section the outline of this approach is presented, followed by the description of methods and experiments on a restricted PubMed domain. The neurocognitive approach increases quality of query result clusters enabling to create useful input for the Visual Language system that changes textual representation to icon-based graphical representation. The end-user is presented iconic graphs instead of a long list of citations enabling faster query refinement and literature based discovery.

2 Neurocognitive Inspirations in Information Retrieval

The purpose of the **VL project** is to test and report the effectiveness of the icon-based visualization in comparison with an existing text-based display approach to internet or database queries. The goal is a display that will enable scientists to identify biological concepts and their relationships more quickly, leading to insight and discovery. In order to make the project precise yet extensible the initial prototype is based on the Unified Medical Language System (UMLS) [10], a well-developed biomedical scientific ontology. Moreover the domain for developing the VL was narrowed to neuronal plasticity in Alzheimer disease (a field of expertise of one of the authors). Hence, to make the project rather general from a design perspective, graphic design and information visualization principles will be limited to provide a syntax of visual form for systems biology. However, systematic visualization techniques for the representation of biological concepts and their relations are applicable to the visualization of concepts within any field of biology.

For the first VL project experiment papers were reviewed manually based upon a random selection of 40 citations from PubMed resulting from search of the terms *Alzheimer Disease* and the protein *ApoE*. From these papers 20 terms were extracted that express important key concepts [7]. These concepts were represented in a semantic graph shown in the left side of fig. 1 (edges of the graph represent semantic relations between concepts manually extracted from the 40 citations).

Next, graphical designers converted each conceptual object into a visual icon, adding specific modification of shapes as one means to systematically depict basic categories of things, processes and actions. These modifiers should perform functions similar to the role adjectives or adverbs serve in natural language [11]. One visual icon systems (designed by students Sean Gresens, David Kroner, Nolan Stover and Luke Woods at the University of Cincinnati) was used to change graph shown in the left fig. 1 to an icon representation shown in the right fig. 1. Visual elements are quickly associated with the desired information that is being searched for. The final step of this project is to measure accuracy and time for completing a list of cognitive tasks that are selected and sequenced to represent the workflow of a scientist conducting a search. These measures will both demonstrate whether the effects of the display are significant, and also provide feedback for future improvements for the VL prototype.

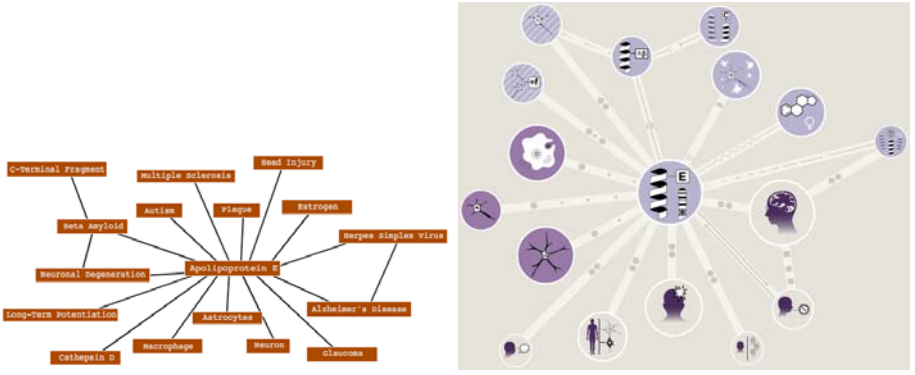

Fig. 1. Manually created textual and iconic representation of a semantic graph based on random sample of 40 publications about the Alzheimer disease and the protein ApoE

An intermediate step for the VL project is to create graphs similar to the one in the left fig. 1 in an automated fashion. This is achieved by clustering query results (similar to *Teoma, Ask, Vivisimo* or *Clusty* search engines), enhancing the initial concept space with semantically related terms and representing each cluster with by a semantic graph. Since each citation is represented by a set of biological concepts enhancing the initial set concepts with semantically related concepts is similar to automated priming effect in a human brains [12]. This means that whenever someone sees word *dog* most likely he will also think about concept *cat*. Semantic priming was studied for over 30 years but never incorporated *per se* into practical computational algorithms. Multiple evidence for priming effects comes from psychology and neuroimaging. For that reason presented here technique for information retrieval is **neurocognitively inspired** [1,13,14].

3 Methods

Large number of PubMed citations have been annotated using Medical Subject Headings (MeSH), providing keywords that characterize the content of an article. MeSH is a hierarchical controlled vocabulary created by National Library of Medicine, used also for indexing books and other documents. Moreover, MeSH is a part UMLS [10], much larger vocabulary that combines over 140 biomedical ontologies and enhance MeSH terms with additional semantic relations that come from other sources. The use of MeSH terms allows for some standardization of searches.

In our experiment a search query *"Alzheimer disease"[MeSH Terms] AND "apolipoproteins e"[MeSH Terms] AND "humans"[MeSH Terms]* was submitted to the PubMed server. It returned 2899 citations along with 1924 MeSH terms. MeSH terms are organized in 16 hierarchical trees. Concepts from only four trees were selected: *Anatomy; Diseases; Chemicals and Drugs; Analytical, Diagnostic and Therapeutic Techniques and Equipment*. This narrowed down the number of concepts to 1190. A binary document/concept matrix was created with information whether a citation had a given MeSH term assigned to it or not. Creating graphs similar to one in the left fig. 1 involves iterative

use of **cluster analysis** and **automated priming** technique. Once a desired clustering quality is achieved a **semantic graphs** as an input for the VL prototype can be computed.

Cluster analysis proposed in this paper is composed of three steps: calculating a distance matrix, organizing documents into hierarchical clusters, and choosing a number of clusters based on quality measures. Since the data is binary various similarity measures from the R language statistical package called *simba* can be used [15]. Let's assume that M is a matrix with rows corresponding to documents and columns to MeSH concepts. Then $M_i^j = 1$ means that a document i has MeSH concept j assigned to it. For this experiment Legendre measure is chosen as a distance function (among fifty other measures it was found to give good results, publication in preparation), defined by [16]:

$$\mathrm{d}(\mathbf{x},\mathbf{y}) = 1 - \frac{3\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{x},\mathbf{y}\rangle + \langle \mathbf{x},\mathbf{x}\rangle + \langle \mathbf{y},\mathbf{y}\rangle} \in [0,1]. \tag{1}$$

Clustering is done using Ward's minimum variance rule [17]. This method is known to produce clusters of almost equal size, but is sensitive to outliers [18]. This is not the optimal clustering algorithm but it suffice for presented here experiments (relation between clustering algorithm and automated priming will be published elsewhere).

In order to choose the optimal number of clusters a combination of four quality measures from the *clusterSim* R language statistical package is used [19]. This includes Davies-Bouldin's index which needs to be minimized [20], Calinski-Harabasz pseudo F-Statistic which needs to be maximized [21], Hubert-Levine internal cluster quality index which needs to be minimized [22] and Rousseeuw Silhouette internal cluster quality index which needs to be maximized [23]. A combination of all four indexes can be normalized between 0 and 4 so that 4 means perfect agreement between all the measures as for the number of clusters. Prior work that used automated priming approach showed usefulness of combined indices in discovery of interesting clusters in patient discharge summaries [1,13,14].

Automated priming can be explained by a spreading activation theory [12]. It gives mathematical simplicity therefore it was used here for modeling the priming effect. At the initial step $t = 0$ of the spreading activation only a subset columns in the M matrix have non-zero elements. As activation spreads from initial concepts to semantically related concepts the number of non-zero columns M increases. The initial concept space representation is symbolized here by ${}_0M$. Let's assume that R is a symmetric matrix with all possible semantic relations between MeSH terms, derived from the UMLS. Then $R_i^j = R_j^i = 1$ means that there is a semantic relation (e.g. *is_a, is_associated_with, may_be_treated_by*, etc.) between concept i and a concept j (e.g. *cerebral structure is_finding_site_of alzheimer's disease*). The simplest mathematical model of spreading activation scheme can be written as:

$${}_{t+1}M = \mathrm{f}({}_tM + \mathrm{f}({}_tMR)). \tag{2}$$

This process enhances the previous feature space ${}_tM$ with new, semantically related concepts, defined by the R matrix. Simple spreading activation algorithms have been already investigated by [24,25,26,27,28]. These approaches did not exploit the full potential of semantic networks (only parent/child relationships were used).

On the other hand if the full potential of semantic relations from UMLS are used for spreading activation it may result in associations similar to that in a schizoid brain (almost everything is associated with everything) [29]. One solution to this problem is to inhibit certain relations by removing those concepts that do not help in assigning documents into right clusters. They may be roughly identified using feature ranking techniques. Ranking is done using Legendre distance (Eq. 1) between columns in the ${}_{t+1}M$ matrix and the cluster labels. The main difference of this technique from a standard feature filtering algorithm is that the ranking is done for each of the cluster separately. So if there are three clusters identified then there are three sets of ranking. The best concepts are taken from each ranking and are put in a set that will excite only those concepts that passed ranking test. This set is called here ${}_{t+1}\mathcal{E}$ and can be roughly defined by:

$$ {}_{t+1}\mathcal{E} = \bigcup_i \text{ the best concepts according to eq. 1 that represent } i\text{th cluster.} \tag{3} $$

The original adjacency matrix R was inhibited to a ${}_tR'$ matrix according to a rule:

$$ {}_tR' = \begin{cases} 0 & \text{if } j \notin {}_{t+1}\mathcal{E} \\ R_{ij} & \text{otherwise} \end{cases}. \tag{4} $$

Enhanced with additional semantic knowledge matrix ${}_{t+1}M$ was calculated using inhibited ${}_tR'$ matrix and the simple neuron threshold output function:

$$ {}_{t+1}M = \mathrm{f}({}_tM + \mathrm{f}({}_tM\,{}_tR')). \tag{5} $$

Semantic graphs are computed using high quality PubMed query results clusters. These graphs are called *graphs of consistent concepts* (GCC) [30]. The idea of GCC that represents a PubMed query results is to show an optimal number of concepts that represent each query cluster. There should be maximum number of connection between concepts that belong to the same cluster and minimum number of connection between concepts that represent different clusters. Increasing the number of concepts that represent each cluster increases also chance that the lower rank concept will have semantic connections with concepts that represent other clusters.

First step is to rank concepts for each cluster separately using Legendre distance from eq. 1. Next a function is defined:

$$ \mathrm{Concepts}(i, n) = \text{ a set of } n \text{ best concepts using eq. 1 that represent } i\text{th cluster.} \tag{6} $$

In order to make the GCC optimization process simple n was varied the same way for each cluster. This gives a set of of concepts that will be used to create adjacency matrix:

$$ \mathcal{E}^n = \bigcup_i \mathrm{Concepts}(i, n) \tag{7} $$

After $iter$ number of iterations of clustering and spreading activation the R matrix can be modified according to following rule to create an adjacency matrix for the semantic graphs:

$$ {}^{iter}_{n}R = \begin{cases} R_{ij} & \text{if } i \in \mathcal{E}^n \text{ and } j \in \mathcal{E}^n \\ 0 & \text{otherwise} \end{cases} . \tag{8} $$

At the beginning n, the best n representants for each cluster, was unknown. It was computed maximizing following function:

$$ \begin{aligned} \mathrm{gcc}({}^{iter}_{n}R) = \sum_{k} \left(\sum_{\{i|\mathrm{Concepts}(i,n)\}} \sum_{\{j|\mathrm{Concepts}(j,n)\}} {}^{iter}_{n}R^{j}_{i} \right) /2o \\ - \sum_{\substack{l,m \\ l \neq m}} \left(\sum_{\{i|\mathrm{Concepts}(i,n)\}} \sum_{\{j|\mathrm{Concepts}(j,n)\}} {}^{iter}_{n}R^{j}_{i} \right) /2o + p/o \end{aligned} \tag{9} $$

where p is the number of clusters, o is the number of active concepts ($o = |\mathcal{E}^n|$). The first term of this equation sums relations between concepts that represent the same cluster, second term sums relations between concepts representing different clusters, while the last term adds the number of clusters. All terms are divided by the total number of concepts o to assure that $\mathrm{gcc}({}_{n}R) = 1$ when all concepts representing the same clusters are connected by a minimum spanning trees and there are no connections between concepts that represent different clusters.

A summary of the algorithm for creating semantic graphs that can be used as input to the VL prototype can be written as a pseudo-code (algorithm 1). The algorithm takes as an *input* data matrix, semantic relations, a number of iterations and *outputs* optimal GCC.

Algorithm 1. GCC algorithm

```
function GCC(R, 0M, iterations)
    for t in 0 to iter do
        compute distance matrix based on tM matrix
        compute tree based on distance matrix and Ward's clustering algorithm
        compute number of clusters based on combined and normalized cluster quality index
        compute the best ranked features separately for each cluster that will be activated
        compute tR' by inhibiting original tR matrix using the best ranked features
        compute t+1M based on inhibited tR', tM and neuron output function
    end for
    compute ranking of features for each cluster separately
    compute n best ranked features that maximize graph consistency index
    compute iter nR adjacency matrix with optimal number n best ranked features
    return graph of consistent concepts
end function
```

4 Results

Concepts that occurred in less than three documents, and documents that had less then three concepts were removed. This created a matrix with 2575 documents and 416 concepts. Two constrains were applied to the system: concepts that receive activation from at least two other concepts are activated (threshold of the $\mathrm{f}(\cdot)$ function is set to 2), and semantic relations are used if they are mentioned in at least two UMLS sources (e.g. MeSH and SNOMED CT).

Six steps of spreading activation were applied to the initial data matrix. The feature space increased from 416 to only 423 concepts. Every step of spreading activation increased quality of clusters for all measures: Davies-Bouldin's index was reduced from 2.2844 to 1.3849, pseudo F-Statistic was increased from 112. 96 to 940.15, Hubert-Levine index was reduced 0.4574 to 0.4414, and Silhouette cluster quality was increased from 0.1515 to 0.3867.

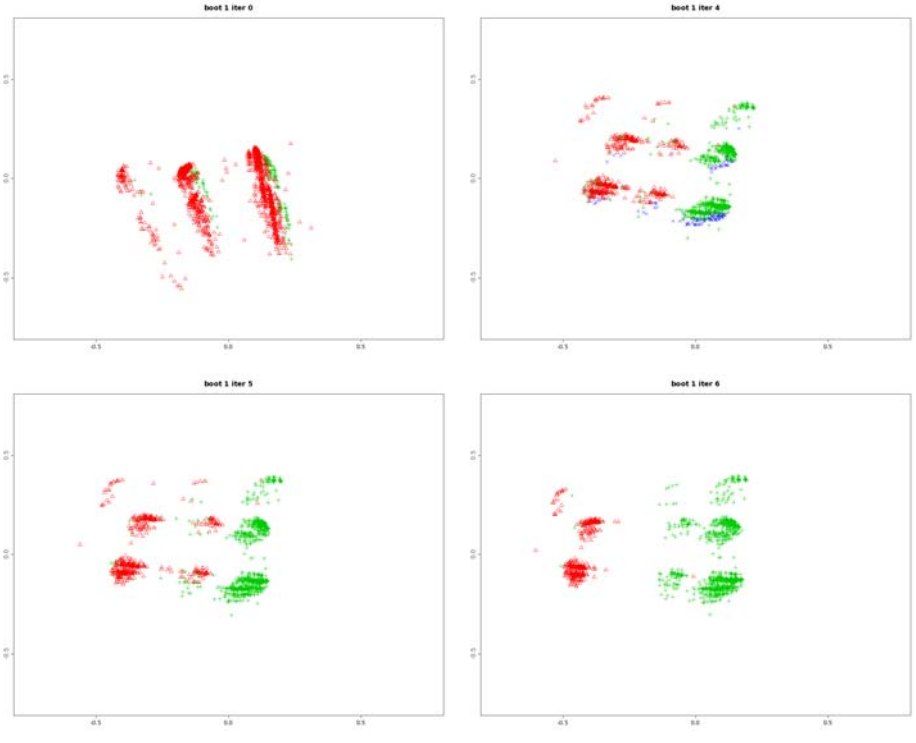

Fig. 2. Multidimensional scaling for *"Alzheimer disease"[MeSH Terms] AND "apolipoproteins e"[MeSH Terms] AND "humans"[MeSH Terms]* PubMed query at the initial step ($t = 0$), after fourth ($t = 4$), fifth ($t = 5$) and sixth step ($t = 6$) of spreading activation. Red triangle sign (first cluster), green plus sign (second cluster) and blue x sign (third cluster) show to which cluster a document concept belongs to. Each step of feature space enhancement creates clearer clusters.

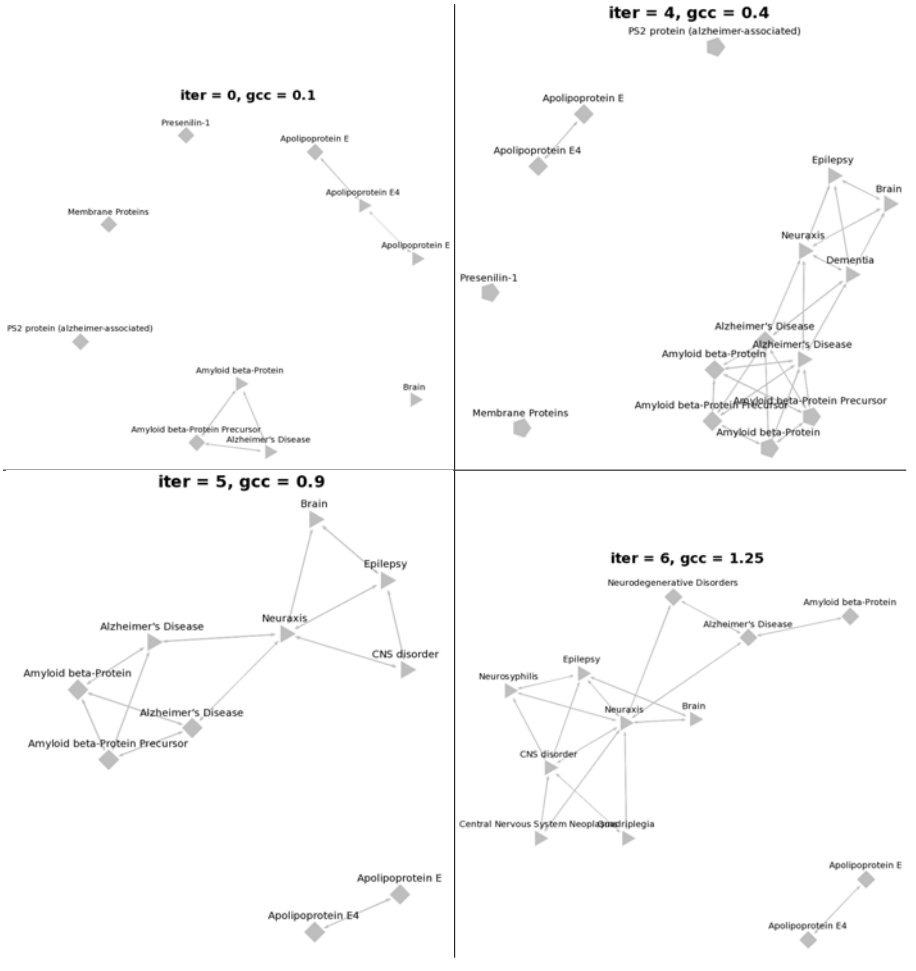

Fig. 3. Semantic graphs for *"Alzheimer disease"[MeSH Terms] AND "apolipoproteins e"[MeSH Terms] AND "humans"[MeSH Terms]* PubMed query at the initial step ($t = 0$), after fourth ($t = 4$), fifth ($t = 5$) and sixth step ($t = 6$) of spreading activation. Triangle (first cluster), rectangle (second cluster) and pentagon (third cluster) nodes show to which cluster a concept belongs to (see eq. 6). Each image shows the optimal value of gcc function (see eq. 9).

The most important finding is that the initial graph that represents query clusters has low maximum consistency measure $\mathrm{gcc}({}^{0}_{n}R) = 0.1$. Adding new information that simulates associative processes in the expert's brains using spreading activation networks increases not only the cluster quality but also consistency of semantic graph that represents same query results clusters $\mathrm{gcc}({}^{6}_{n}R = 1.25)$. Fig. 3 shows the increase of the consistency measure calculated using eq. 9 while fig. 2 shows changes in the 2D projections of the high dimensional concept space using classic multidimensional scaling.

5 Conclusion

In computational intelligence community WebSOM [31] has been the most well-known attempt to visualize information using clustering techniques. However, in many respects 2-dimensional visual representation of hierarchical visualization of clusters offered by SOM is not sufficient: relations between documents are much more complicated and may only be shown in a non-planar graphs. Instead presentation of query results that match human cognitive abilities in a best way may be done using neurocognitive inspirations.

The algorithm presented here tries to re-create pathways of neural activation in the expert's brain, enhancing the initial concepts found in the text using relations between the concepts. This may be done in the medical domain because specific relations between MeSH terms present in a citation may be extracted from the huge UMLS resources [10]. Background knowledge is added using adjacency matrix describing semantic relations. The whole algorithm is presented in the matrix form, making it suitable for efficient large-scale retrieval systems. The goal here is to automatically create graphs of consistent concepts using documents that result from specific queries, and present these concepts using Visual Language iconic system [7]. This paper has demonstrated the usefulness of neurocognitive inspirations approximating brain processes with a combination of clustering, feature selection and neural spreading activation techniques. While there is an ample room for improvement using better clustering techniques and feature selection methods a significant increase of the quality of the initial GCC graph has been obtained, giving much better representation of the information in a visual form.

References

1. Duch, W., Matykiewicz, P., Pestian, J.: Neurolinguistic approach to natural language processing with applications to medical text analysis. Neural Networks (in press, 2008), doi:10.1016/j.neunet.2008.05.008
2. Buzan, T.: The Mind Map Book. Penguin Books (2000)
3. Farrand, P., Hussain, F., Hennessy, E.: The efficacy of the mind map study technique. Medical Education 36(5), 426–431 (2002)
4. Kraft, R., Zien, J.: Mining anchor text for query refinement. In: Proceedings of the 13th international conference on World Wide Web, pp. 666–674. ACM, New York (2004)
5. Gordon, M.D., Lindsay, R.K.: Toward discovery support systems: a replication, re-examination, and extension of swanson's work on literature-based discovery of a connection between raynaud's and fish oil. J. Am. Soc. Inf. Sci. 47(2), 116–128 (1996)
6. Zender, P.M., Crutcher, K.A.: Visualizing alzheimer's disease research: a classroom collaboration of design and science. In: SIGGRAPH 2004: ACM SIGGRAPH 2004 Educators program, p. 24. ACM, New York (2004)
7. Zender, M., Crutcher, K.A.: Visual language for the expression of scientific concepts. Visible Language 41, 23–49 (2007)
8. Duch, W., Szymaski, J.: Semantic web: Asking the right questions. In: Gen, M., Zhao, X., Gao, J. (eds.) Series of Information and Management Sciences, California Polytechnic State University, CA, USA, pp. 456–463 (2008)
9. Pedrycz, W.: Knowledge-Based Clustering: From Data to Information Granules. Wiley Interscience, Hoboken (2005)

10. U.S. National Library of Medicine, National Institutes of Health: Unified medical language system (January 2007), http://www.nlm.nih.gov/research/umls/
11. Zender, M.: Advancing icon design for global non verbal communication: Or what does the word bow mean? Visible Language 40, 177–206 (2006)
12. McNamara, T.P.: Semantic Priming: Perspectives From Memory and Word Recognition. Psychology Press, Taylor & Francis Group (2005)
13. Duch, W., Matykiewicz, P., Pestian, J.: Towards Understanding of Natural Language: Neurocognitive Inspirations. In: de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D.P. (eds.) ICANN 2007. LNCS, vol. 4669, pp. 953–962. Springer, Heidelberg (2007)
14. Duch, W., Matykiewicz, P., Pestian, J.: Neurolinguistic approach to vector representation of medical concepts. In: Press, I. (ed.) Proc. of the 20th Int. Joint Conference on Neural Networks (IJCNN), August 2007, p. 1808 (2007)
15. Jurasinski, G.: simba: A Collection of functions for similarity calculation of binary data, R package version 0.2-5 (2007)
16. Legendre, P., Legendre, L.: Numerical Ecology. Elsevier, Amsterdam (1998)
17. Anderberg, M.R.: Cluster Analysis for Applications. Academic Press, New York (1973)
18. Milligan, G.W.: An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45, 325–342 (1980)
19. Walesiak, M., Dudek, A.: clusterSim: Searching for optimal clustering procedure for a data set, R package version 0.36-1 (2008)
20. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1, 224–227 (1979)
21. Calinski, R.B., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics 3, 1–27 (1974)
22. Milligan, G.W., Cooper, M.C.: An examination of procedures of determining the number of cluster in a data set. Psychometrika 50, 159–179 (1985)
23. Kaufman, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis. Wiley, New York (1990)
24. Hotho, A., Staab, S., Stumme, G.: Wordnet improves text document clustering. In: Proc. of the Semantic Web Workshop at SIGIR-2003, 26th Annual International ACM SIGIR Conference (2003)
25. Sedding, J., Kazakov, D.: Wordnet-based text document clustering. In: Pallotta, V., Todirascu, A. (eds.) COLING 2004 3rd Workshop on Robust Methods in Analysis of Natural Language Data, Geneva, Switzerland, August 2004, pp. 104–113 (2004)
26. Struble, C.A., Dharmanolla, C.: Clustering mesh representations of biomedical literature. In: Lynette, H., James, P. (eds.) HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases, May 2004, pp. 41–48. Association for Computational Linguistics, Boston (2004)
27. Yoo, I., Hu, X., Song, I.-Y.: A coherent biomedical literature clustering and summarization approach through ontology-enriched graphical representations. In: Tjoa, A.M., Trujillo, J. (eds.) DaWaK 2006. LNCS, vol. 4081, pp. 374–383. Springer, Heidelberg (2006)
28. Jing, L., Zhou, L., Ng, M.K., Huang, J.Z.: Ontology-based distance measure for text clustering. In: Proceedings of the IV Workshop on Text Mining; VI SIAM International Conference on Data Mining (April 2006)
29. Kreher, D., Holocomb, P., Goff, D., Kuperberg, G.: Neural evidence for faster and further automatic spreading activation in schizophrenic thought disorder. Schizophrenia bulletin 34, 473–482 (2008)
30. Matykiewicz, P., Duch, W., Pestian, J.: Nonambiguous concept mapping in medical domain. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 941–950. Springer, Heidelberg (2006)
31. Kohonen, T., Kaski, S., Lagus, K., Salojrvi, J., Paatero, V., Saarela, A.: Organization of a massive document collection. IEEE Transactions on Neural Networks 11(3), 574–585 (2000)