In: Series of Information and Management Sciences, California Polytechnic State University, Proc. of the 9th International Conference on Information and Management Sciences (IMS 2010), Urumchi, China, pp. 130-138

Dynamic Semantic Visual Information Management

Julian Szymański
Department of Electronic, Telecommunication and Informatics
Gdańsk University of Technology, Gdańsk, Poland
Email: julian.szymanski@eti.pg.gda.pl

Włodzisław Duch
Department of Informatics, Nicolaus Copernicus University, Toruń, Poland.
Google: W. Duch

Abstract: Dominant Internet search engines use keywords and therefore are not suited for exploration of new domains of knowledge, when the user does not know specific vocabulary. Browsing through articles in a large encyclopedia, each presenting a small fragment of knowledge, it is hard to map the whole domain, see relevant concepts and their relations. In Wikipedia for example some highly relevant articles are not linked with each other. Static links do not express the need to show dynamic subsets of interconnected concepts that reflect particular interest of the user. Using neurocognitive inspirations to understand information representation based on natural language we have developed various ways for dynamic representation of relevant information. Preliminary experiments with Wikipedia articles are used to demonstrate this approach.

Keywords: Semantic Web, semantic memory, episodic memory, information retrieval, knowledge acquisition, NLP, Wikipedia, text clustering, SOM, PCA.

I. Introduction

Rapid growth of information resources in the Internet makes the extraction and management of useful information increasingly difficult. On the one hand specific information based on specialized keywords unique for a given field makes it relatively easy to find specialized articles on that subject, provided that the user knows these keywords. Biomedical applications of natural language processing (NLP) are based on such specialized vocabulary stored in standardized ontologies, and therefore are relatively easy to implement. The Unified Medical Language System (UMLS) is a huge collection of medical ontologies [1] and indexing terms selected for Medical Subject Headings (MeSH) ontology are at the foundation of information retrieval in Medline, the biggest repository of medical articles. Commonsense knowledge is much more difficult to verbalize; human ability to describe well-known objects using words is rather poor. For example, there are more than 400 breeds of dogs that we can distinguish visually, but it is almost impossible to describe such images using keywords in sufficient details to identify a particular breed.

Despite continuous improvements of search engines, retrieval and management of textual information is getting increasingly difficult. The quality of information in the Internet is degraded by a large number of spam pages with advertisement, poor quality, copies of the same information, making the search for relevant information increasingly difficult. Semantic Internet [2][1][3] is still far from realization and will not help to avoid information clutter, it may actually make the situation even worse. Definitions of word meaning added to HTML text using XML (eXtended Markup Language) or RDF (Resource Description Framework) markup tags for describing information and resources on the web can easily be copied also on spam pages. Searching for more information about the breaking news returns hundreds of essentially identical copies provided by numerous news services, and adding semantic description will not change it.

We have already commented on the problems of semantic web [4] and proposed to use a query-based approach based on neurocognitive inspirations to create semantic memory model [5]-[11], asking minimum number of questions to define the subject of the query more precisely. Precisiation of natural language concepts has also been addressed from the fuzzy logic perspective [12], but with the focus on understanding quantitative concepts only, like “most”, or “usually”. This is interesting but rather limited approach that is important only in relatively rare situations.

In this paper another aspect of information management, not covered by the semantic web, is addressed: browsing for information in new domains, where the user does not know precise keywords or specific vocabulary. In such cases a good place to start is from systematic knowledge sources, such as articles in specialized journals, or articles in a large encyclopedia, each presenting a small fragment of knowledge. Connections of relevant concepts or papers for further reading are captured by links that are manually added either by the authors of the information or the editors of journals or databases. For example, PubMed library provides many kinds of links for each article, and Wikipedia, Scholarpedia and many other encyclopedic and library resources available in the Internet include internal links to other articles within each resource, and additional links to the external sources. Whole domains of knowledge may be visually mapped in this way, showing relevant concepts and their relations.

The link-based approach to information management faces three problems:

- First, links are never complete and not always stable. In Wikipedia and similar resources some highly relevant articles miss direct links to each other.
- Second, some pages have high number of links, making visualization difficult.
- Third, static links point to all kinds of documents, while the user may be interested in a particular aspect that requires dynamic selection of subsets of links and visualization of interconnected concepts that reflect particular interest of the user.

These problems are addressed below. Using neurocognitive inspirations we test different text representations based on words and links which allow to process information according to its semantics. Using this representation we have developed various ways for dynamic retrieval of relevant information. A few experiments with Wikipedia articles are used to demonstrate this approach.

II. Visualization and Clustering of Text-based Information.

Visualization of information, also known as InfoVis or in scientific domains SciVis, is quite popular subject, with a number of books [13]-[16] and software tools [17]-[19] for visualization. In particular graphical representation of linked articles, co-citations, documents or web sites is rather straightforward. Our Gossamer component[1] has also been used as a graphical tool for visualization and editing[2] [20] of Wordnet [21] concepts, and visual browsing Wikipedia articles[3].

Searching for information on some topic one creates a set of expectations, priming the brain, facilitating disambiguation of concepts, and increasing attention to relevant information. The effect of priming is to change perceived similarity between documents. Traditional approach to information search treats documents that contain a set of given keywords as relevant. Evaluation of semantic similarity tries to cluster whole documents using some measures of overall similarity. The goal of dynamic visualization of semantic relations is to model these priming effects, select documents that are similar from a particular point of view, and present them in visual form.

The algorithm for dynamical semantic clustering includes the following steps:

1. Define the search domain and cluster all documents D_i in this domain using the vector space model, i.e. representing documents by vectors $V(D_i)$.
2. Provide a set of keywords $\{Key_j\}$ for priming, or alternatively a reference document from which such keywords will be extracted.
3. Determine additional concepts $\{K_k\}$ that are correlated with keywords $\{Key_j\}$, including their synsets, and calculate their linear correlation coefficients π_k.
4. Rescale components of $V(D_i)$ vector corresponding to expanded keywords using correlation coefficients $\pi_k > 0.6$.
5. Perform new clustering based on modified vectors.
6. Visualize clusters, providing new view on relations among documents in the search space.

This general algorithm may be implemented in many ways. In the examples shown below, cosine similarity measure between vectors representing texts has been selected, proportional to normal Euclidean distance if vectors are normalized to have a unit length $|V(D_i)|=1$. WordNet [21] is used to determine expanded set of keywords. Correlations between concepts are calculated using **L**atent **S**emantic **A**nalysis [22], which essentially is the **P**rincipal **C**omponent **A**nalysis [23],[24] on the matrix of all $V(D_i)$ vectors.

Clustering of documents is used for visualization of texts, showing maps with peaks corresponding to similar documents that may be labeled by the relative frequency of keywords. So far larger scale experiments of this sort have been done only with Kohonen's Self-Organizing Maps (SOM) [25]. The WebSOM approach [26] has been used to analyze in this way one million documents taken from the Internet discussion groups. Below we have also used SOM for clustering and initial visualization, but connections between clusters may frequently be easier to follow using graph-based techniques. Our main purpose will be to re-define clusters in such a way that they will match user needs, thus reducing irrelevant information.

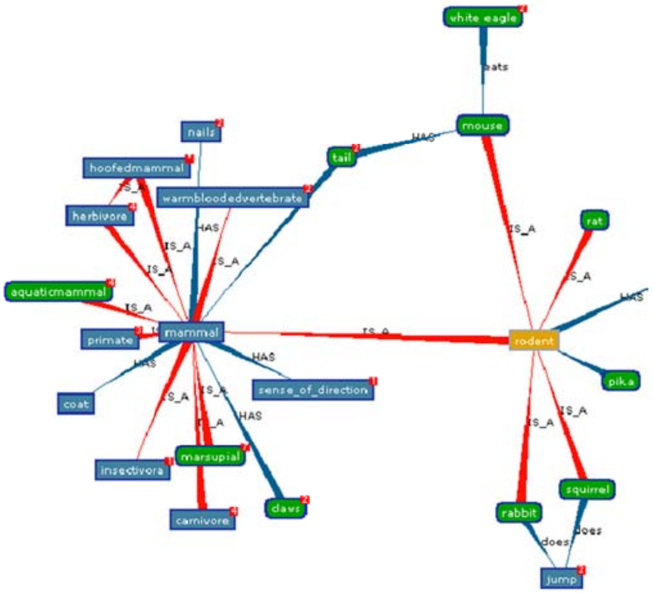


Fig. 1. Visualization of Semantic Space for animal kingdom.

[1] http://gossamer.eti.pg.gda.pl
[2] http://wordventure.eti.pg.gda.pl
[3] http://swn.eti.pg.gda.pl

III. Illustrative examples

Dynamic semantic clustering may be done for any set of concepts, articles, documents or other entities for which vector-based representation can be obtained.

Animal database. In the first example relations between concepts developed in our 20 question game based on semantic memory [8] are presented. The game has been restricted to 50 animals and the questions involved 270 of their features. In Fig. 1 graph illustrating visualization of semantic network for these animals is presented. Overall similarity of animal properties may also be presented using Principal Component Analysis. Analysis of eigenvalues (Fig. 2) shows that the first 3 PCA components may already be useful to display similarity of animals, and Fig. 3 shows a projection of this 270-dimensional data on the first 3 most important principal components. Clustering of similar animals is quite clear, showing natural taxonomy into reptiles, birds, mammals, but also a few untypical animals that form their own subclusters, like bats or marine mammals (here represented by a single species). This type of visualization represents typical view based on all properties. Application of **S**elf-**O**rganized **M**apping with 5x5 network shows similar clustering (Fig. 4).

To browse such information from a particular perspective keywords are given to prime the network and words correlated with these keywords are also used to weight similarity. For the word *beak* 3 strongly correlated words have been found: *bird, fly, eggs*, with linear correlation coefficients C = [0.8948, 0.7362 and 0.6426].

The W weights of the relations between these 4 features and objects they are related to have been enhanced. The enhanced W' value has been produced accordingly to feature coefficient values as:

$$W' = W + W^*(C)$$

In effect cluster with animals that have beaks and lay eggs has been created (Fig. 5). Few birds are in a separate cluster and platypus, quite untypical animal with a beak that lays eggs, is nearby. Other clusters have also changed, but they are rather far and their content may be removed from the dynamic query graph.

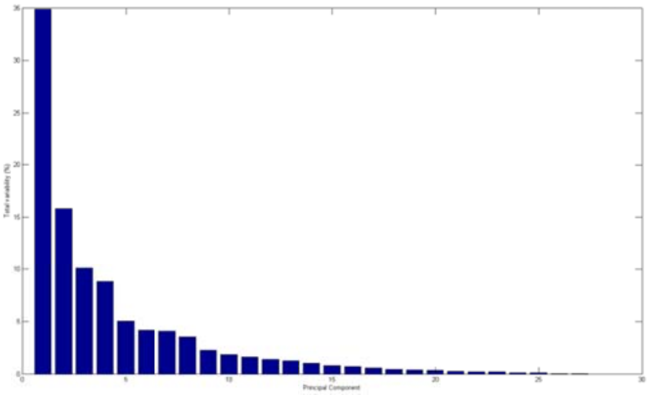
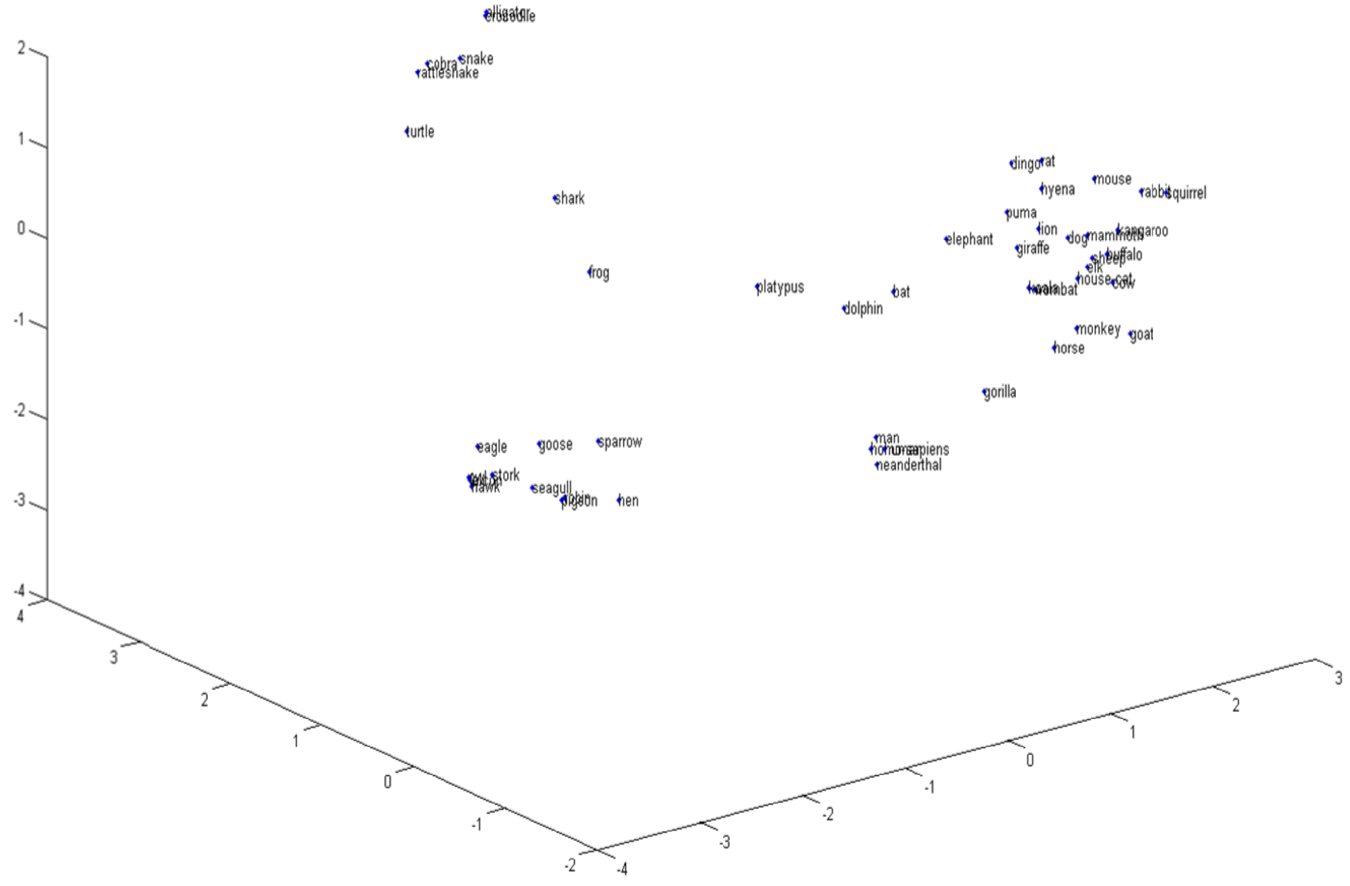


Fig. 2. Percentage of variance for the PCA components for the 270-dimensional space with 50 animals.



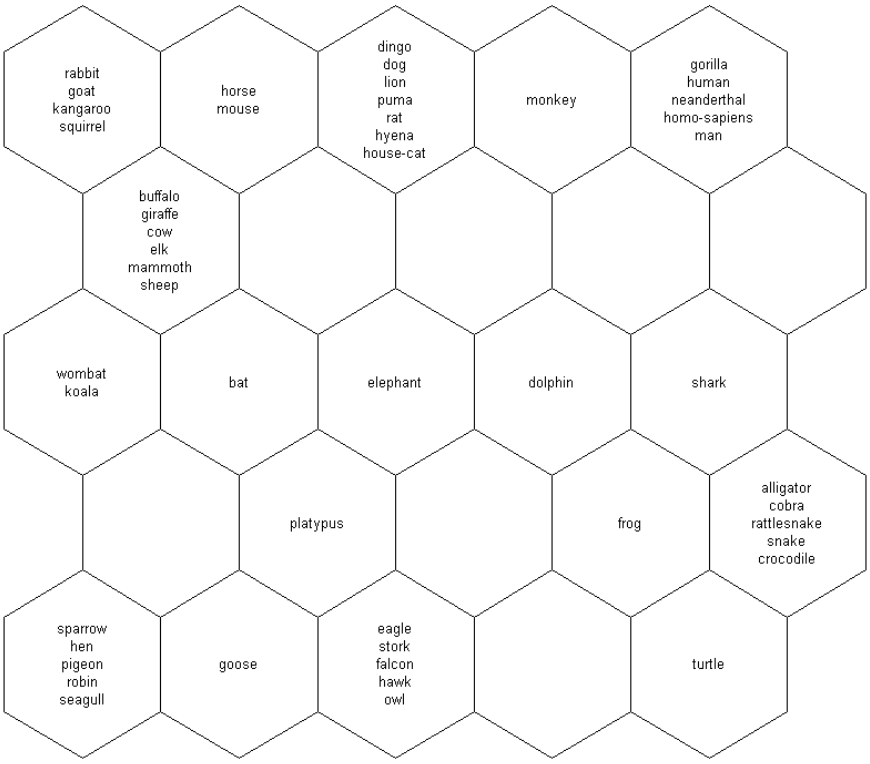


Fig. 4. SOM map generated for semantic network of 50 animals

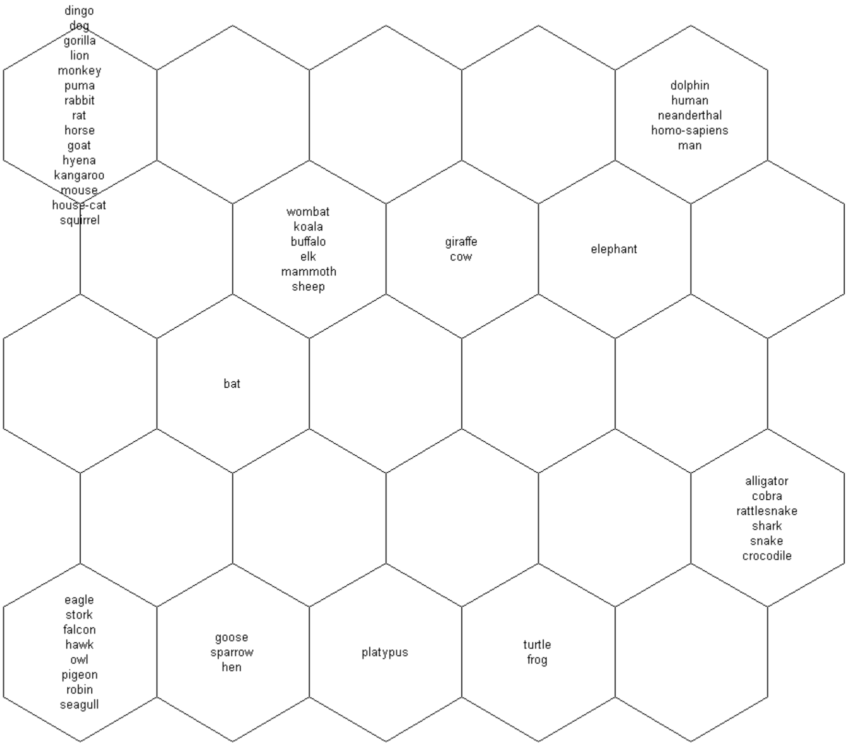


Fig. 5. SOM map generated for 50 animals, weighted by keyword "beak" and words correlated with this keyword: bird, fly and eggs.

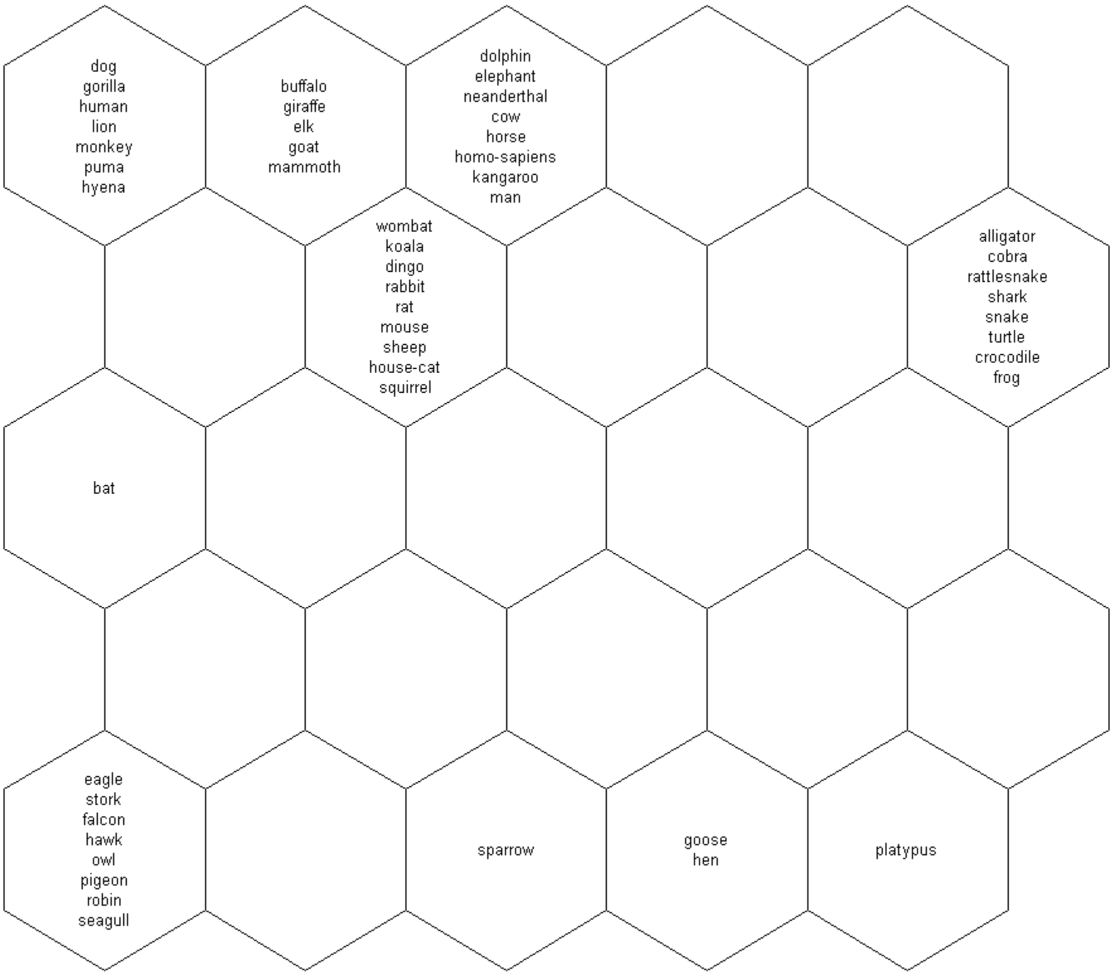
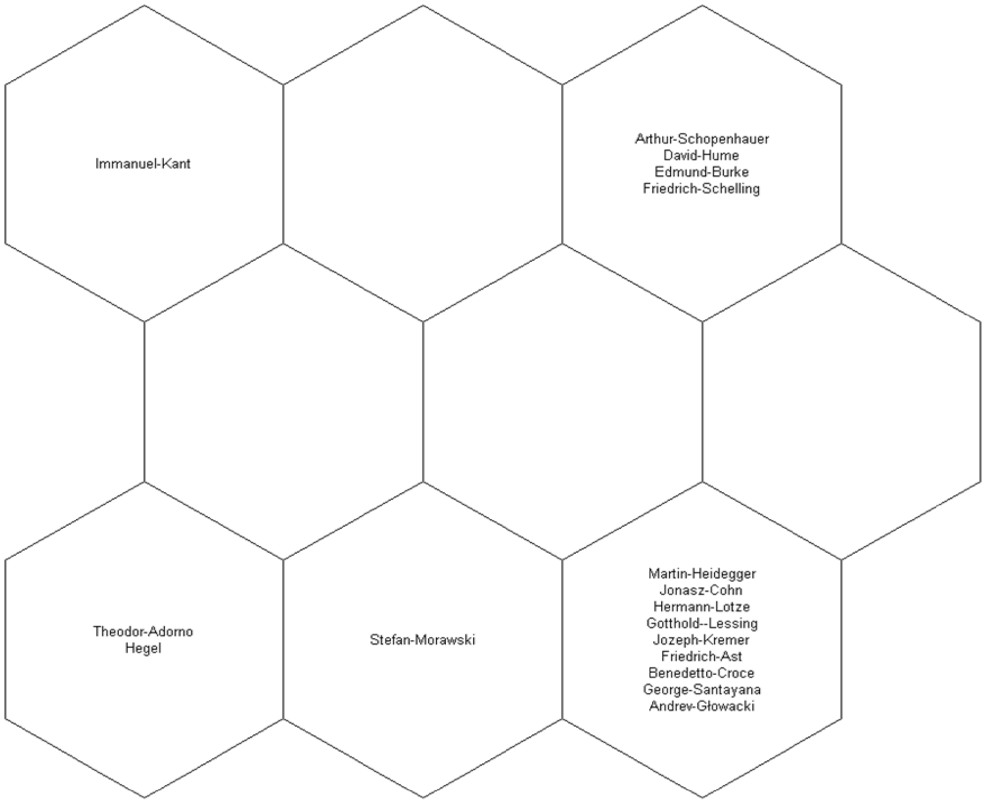


Fig. 6. Adding features correlated with "fly" separates birds from other animals.

Further separation depicted in Fig. 6 is achieved if „fly” is used as a keyword. This feature is correlated with *eggs*, *beak*, *bird* with coefficients 0.621, 0.7362, 0.8466. Most birds are now in one cluster, with domestic birds in another and sparrow (much smaller than all other birds) in a separate cluster, while turtle and frog are shifted to a big cluster.

Wikipedia. The same approach may be used to search and organize larger database of articles. English Wikipedia has now (May 2010) over 3.2 million articles (in 2006 it had only 1.4 million articles), and several times more in a large number of other languages; 5 largest collections are in English, German, French, Polish, and Japanese. While there are many attempts to create peer-review reliable encyclopedic resources none is at such large scale as Wikipedia. Semantic map of the whole 2005 collection of articles has been created [27] showing clusters corresponding to manually designed categories. These categories have been designed manually, in a hierarchical and logical way, but at quite different ontological level, form rather large branches to individual and specialized topics. Fixed categories will never be sufficiently flexible to express all points of view of interest to the users. For example, selecting neuroscience articles that are relevant to human, related to computational models, or to molecular neuroscience, or to combination of such categories, cannot be done by browsing through categories.

In this series of experiments we have used different representation of similarity between articles, based on co-citations. This type of representation has been used before with good results for analysis of the whole branches of science [28], but here the purpose is quite different, searching for specific visual representation (as in Fig. 1) of articles that are related to each other from a specific point of view. Articles about 17 philosophers interested in the philosophy of art have been selected. There were 660 citations in all these articles, so similarity between these philosophers may be evaluated in this space. It is well known that reduction of dimensionality by the **P**rincipal **C**omponent **A**nalysis (called in natural language processing “**L**atent **S**emantic **A**nalysis”) leads to more



rticles.

robust similarity evaluations for documents [22]. It was found that first PCA component has been dominating, capturing 85% of all variance, with 15 other components accounting for the remaining 15% (Fig. 7).

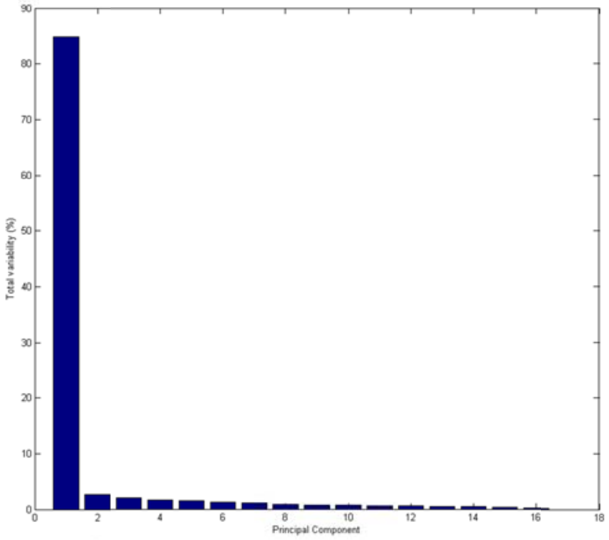


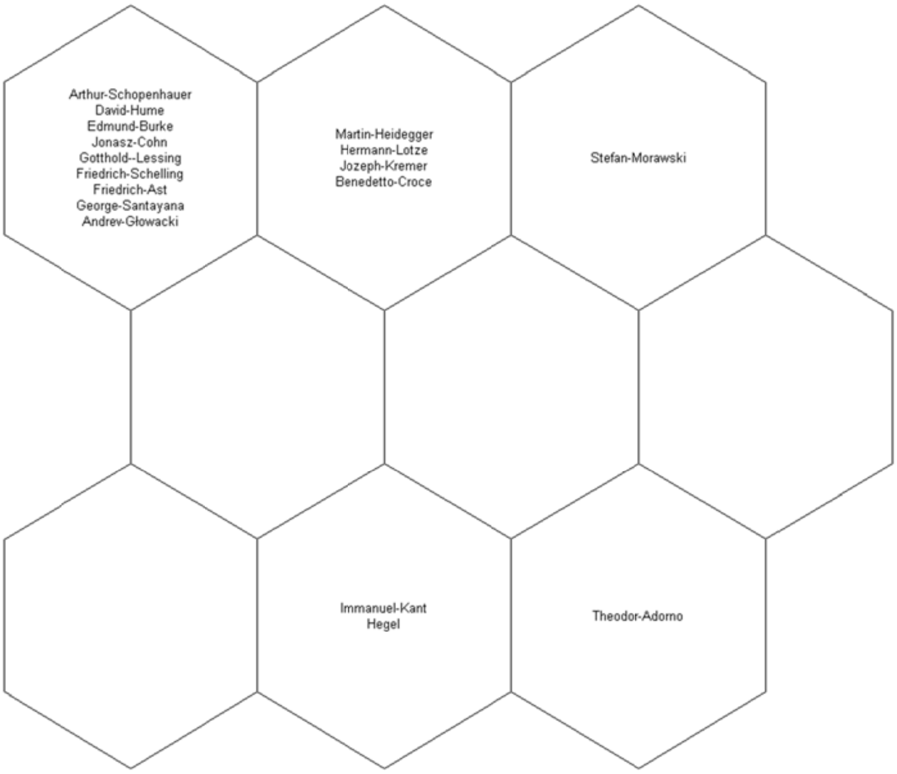
Fig. 7. PCA contributions to variance for 17 philosophers based on co-citations found in Wikipedia articles.

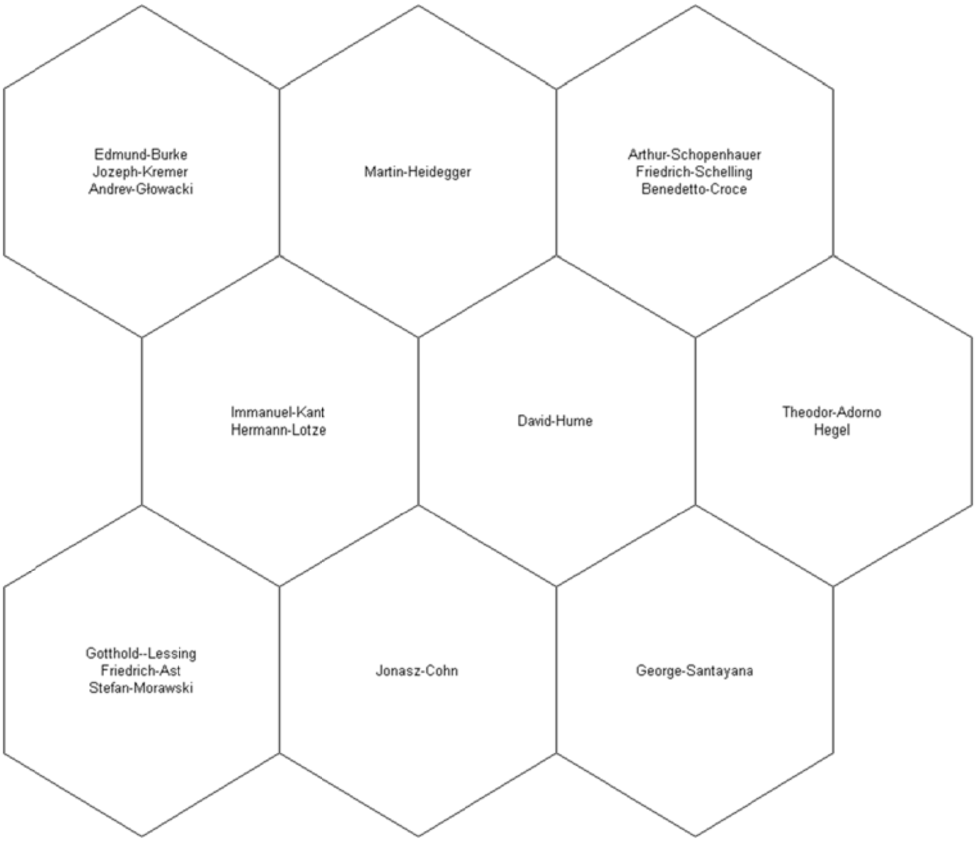
SOM maps based on original 660-dimensional binary vectors and 16 PCA components look identical (Fig. 8), displaying general similarity of articles.

Since this map is based on binary co-citations rather than concepts, it is not possible to modify it by adding keywords and correlated concepts to general similarity calculated in the original or in the PCA space. Instead one may add weights to the PCA components that form new features used for similarity evaluation. This leads to different views on the documents, as shown in SOM maps in Fig. 9 and 10. The first of these maps (Fig. 9) is based on rescaled second component which has been multiplied by 10 to increase its importance to about half of the first component. In Fig. 10 SOM map for similarly rescaled principal component no.12 is shown.

As can be seen in Fig. 9 and 10 the influence of such rescaling is quite strong. However, interpretation of these differences is not as straightforward as in the case of word-based semantic representation. Increasing the weight of one of the component amounts to more confidence or importance given to those links and citations that have relatively large coefficients in the corresponding principal vector. Although evaluation of confidence in particular sources is possible it may be rather difficult to quantify.

The co-citation based estimation of similarity captures some important information that is rather difficult to capture in a concept-based vector model. Two articles that quote the same papers or link to identical articles are similar in some aspects. Therefore it may be better to combine co-citation with the term-based representation, facilitating interpretation and keywords to define points of view.


Arthur-Schopenhauer
David-Hume
Edmund-Burke
Jonasz-Cohn
Gotthold--Lessing
Friedrich-Schelling
Friedrich-Ast
George-Santayana
Andrev-Głowacki
Martin-Heidegger
Hermann-Lotze
Jozeph-Kremer
Benedetto-Croce
Stefan-Morawski
Immanuel-Kant
Hegel
Theodor-Adorno


Edmund-Burke
Jozeph-Kremer
Andrev-Głowacki
Martin-Heidegger
Arthur-Schopenhauer
Friedrich-Schelling
Benedetto-Croce
Immanuel-Kant
Hermann-Lotze
David-Hume
Theodor-Adorno
Hegel
Gotthold--Lessing
Friedrich-Ast
Stefan-Morawski
Jonasz-Cohn
George-Santayana

IV. Discussion and future developments

We have identified several problems that have not been addressed by the semantic web approach and it does not seem likely that these problems will be addresses in the near future. Creating dynamical maps that show information maps restricted to what is interesting at a given moment is a novel concept that should help to avoid information clutter created by current visualization methods. Semantic similarity should help to avoid problems with missing links and incomplete relational information. The algorithm presented here may use similarity evaluation based on concepts found in the documents, existing links between documents, and common citations.

There are many sophisticated methods for evaluation of similarity and clustering of documents. Although SOM has been used here the final results are frequently better represented by maps linking similar documents that are clustered in a few adjacent SOM nodes. One way to represent such concepts is by using minimum spanning trees [29]. Another way is to present keywords from major PCA components on a radar plot, showing interesting directions (such visualization has been used to show major components in clusters [30]). A combination of semantic similarity with search directions represented in radar plots may also be an interesting way to define suitable perspective for visualization. We have used linear correlation between concepts to rescale a group of related features in vector representation of documents, but in the Interspace project [31] a sophisticated system for guided navigation in the space of concepts has been described. Several ways of initial query expansion have been described in the literature [32]-[34].

Combining overall similarity with filtering based on keywords simplifies browsing through new domains. Visual representation of linked documents or web pages has been used for some times but did not gain large popularity, despite such interfaces as TheBrain[4] that should linked notes and webpages in form of a big graph. Adding the ability to create semantic links and filtering such links through specific lens defined by keywords should make them much more useful.

Such approach will be useful in many information management platforms. For example, *Neurocommons*[5] is an open source knowledge management platform for neuroscience research that "seeks to make all scientific research materials - research articles, knowledge bases, research data, physical materials - as available and as usable as they can be. [...] We want knowledge sources to combine easily and meaningfully, enabling semantically precise queries that span multiple information sources."

International Neuroinformatics Coordinating Facility (INCF)[6] (with headquarters based in Sweden) has similar goals and tries to rely on the semantic web techniques, integration and management of text data and other symbolic information. Such national neuroinformatics platforms are being developed in collaboration with INCF in many countries. Some projects are focused on specific subfields, for example the Visome project is concerned integration of experimental databases, software for data analysis, models of neural processes and relevant literature for color vision [35]. Educational resources sharing educational material, such as Connections[7] contain thousands of modules that are quite hard to browse, as the amount of information is overwhelming and there are no good tools for viewing it. The need for better search and browsing tools is indeed great and therefore development of good tools along the lines sketched in this paper is highly desirable.

References

[1] Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research (32) 267 – 270

[2] Berners-Lee T, Hendler J. and Lassila O, The Semantic Web. Scientific American, May 17, 2001.

[3] Davies J, Semantic Web Technologies: Trends and Research in Ontology-based Systems. J. Wiley 2006.

[4] Szymanski J, Duch W, Semantic Memory Architecture for Knowledge Acquisition and Management. 6th International Conference on Information and Management Sciences (IMS2007), July 1-6, 2007, California Polytechnic State University, CA, pp. 342-348.

[5] Szymanski J, Duch W, Semantic Memory Knowledge Acquisition Through Active Dialogues. 20th Int. Joint Conference on Neural Networks (IJCNN), Orlando, IEEE Press 2007, pp. 536-541

[6] Szymanski J, Duch W, Semantic Memory Architecture for Knowledge Acquisition and Management. 6th International Conference on Information and Management Sciences (IMS2007), July 1-6, 2007, California Polytechnic State University, CA, pp. 342-348

[7] Szymanski J, Sarnatowicz T, Duch W, Towards Avatars with Artificial Minds: Role of Semantic Memory. Journal of Ubiquitous Computing and Intelligence, American Scientific Publishers, 2, 1-11, 2008.

[8] Duch W, Szymański J, Semantic Web: Asking the Right Questions. In: Series of Information and Management Sciences, M. Gen, X. Zhao and J. Gao, Eds, California Polytechnic State University, CA, USA, pp. 456-463, 2008.

[9] Szymanski J, Duch W, Knowledge representation and acquisition for large-scale semantic memory. World Congress on Computational Intelligence (WCCI'08), Hong Kong, 1-6 June 2008, IEEE Press, pp. 3117-3124

[10] Duch W, Neurocognitive Informatics Manifesto. In: Series of Information and Management Sciences, Cali-

[4] http://www.thebrain.com

[5] http://neurocommons.org

[6] http://incf.org

[7] http://cnx.org

fornia Polytechnic State University, 8th Int Conf on Information and Management Sciences (IMS 2009), Kunming-Banna, Yunan, China, pp. 264-282.

[11] Duch W, Matykiewicz P, and Pestian J, Neurolinguistic Approach to Natural Language Processing with Applications to Medical Text Analysis. *Neural Networks* 21(10), 1500-1510, 2008

[12] Zadeh L.A, Precisiated natural language (PNL), AI *Magazine* 25, 74–91, 2004.

[13] C. Ware. Information Visualization. Elsevier. 2004

[14] C. Chen (2004) Information Visualization: Beyond the Horizon. Springer.

[15] A.C. Telea, Data Visualization, A K Peters, Ltd. 2007

[16] R. Mazza, Introduction to Information Visualization. Springer 2009

[17] Touchgraph, see: http://www.touchgraph.com/

[18] AiSee, http://www.aisee.com/

[19] yEd, http://www.yworks.com/en/products_yed_about.htm

[20] J. Szymański, Developing WordNet in Wikipedia-like style, Proc. of the 5th International Conf. of the Global WordNet Association, Mumbai, India 2010.

[21] G. Miller, R. Beckitch, C. Fellbaum, D. Gross, K. Miller, Introduction to WordNet: An On-line Lexical Database. Cognitive Science Laboratory, Princeton University Press, 1993.

[22] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis. Journal of the American society for information science, Vol. 41, 1990, pp. 391 – 407.

[23] I.T. Jolliffe, Principal component analysis. Springer Verlag, 2002.

[24] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008. pp. 109-133, 253-419.

[25] T. Kohonen, Self-Organizing Maps. Springer Series in Information Sciences, 3rd ed, 2000

[26] Lagus K, Kaski S, and Kohonen T. Mining massive document collections by the WEBSOM method. Information Sciences, 163, 135-156, 2004.

[27] T. Holloway, M. Bozicevic, K. Börner, Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. Complexity 12(3), 30-40, 2007.

[28] Moya-Anegón, F., Vargas-Quesada, B., Victor Herrero-Solana, Chinchilla-Rodríguez, Z., Elena Corera-Álvarez & Munoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes & categories. Scientometrics, 61(1): 129-145.

[29] Duch W, Matykiewicz, P. Minimum Spanning Trees Displaying Semantic Similarity. Intelligent Information Processing and Web Mining, Advances in Soft Computing, Springer, pp. 31-40, 2005.

[30] A. L. Kaczmarek, "Clustering by Directions algorithm to narrow search queries," in Proc. of Human System Interaction Conference, Krakow, Poland, pp. 689-694, 2008.

[31] Schatz B, The Interspace: Concept Navigation across Distributed Communities, IEEE Computer, 35(1): 54-62, 2002.

[32] H. Chim and X. Deng, "Efficient Phrase-Based Document Similarity for Clustering," IEEE Transaction on Knowledge and Data Engineering 20: 1217-1229, 2008.

[33] B. M. Fonseca, P. Golgher, B. Pôssas, B. Ribeiro-Neto and N. Ziviani, "Concept-based interactive query expansion," in Proc. of the 14th ACM Conf. on Information and Knowledge Management, New York 2005, pp. 696-703.

[34] B. Zhang, Y. Du, H. Li and Y. Wang, "Query expansion based on topics," in Proc. of 5th Conf. on Fuzzy Systems and Knowl. Discovery, Vol. 2, IEEE Press 2008, pp. 610-614.

[35] Usui S, Visiome: neuroinformatics research in vision project. Neural Networks 16(9), 2003 , pp. 1293-1300

Julian Szymański received the BEng, MSc and PhD degrees from Gdańsk University of Technology in computer science and MSc degree in philosophy from the Nicolaus Copernicus University; he works currently as the Assistant Professor at Gdańsk University of Technology.

Włodzisław Duch received the MSc, PhD and DSc degree from the Nicolaus Copernicus University, and is the Head of Department of Informatics at this university; recently (2003-2007) he has worked as a Visiting Professor at Nanyang Technological University, Singapore. Currently he serves as the President of the European Neural Networks Society (www.e-nns.org).

For more information Google: W. Duch.