Mining associative relations from website logs and their application to context-dependent retrieval using spreading activation.

Johan Bollen, Herbert Vandesompel and Luis M. Rocha
Los Alamos National Laboratory, NM
jbollen@lanl.gov, herbert.vandesompel@rug.ac.be, rocha@lanl.gov
http://lib-www.lanl.gov/ jbollen

ABSTRACT
We have devized a methodology that mines sequential navigation patterns from a website's logs to enable us to identify the most significant associative links in the networks. Spreading activation can then be applied to the generated network of weighted hyperlinks enabling the content-dependent, semantic retrieval of nodes in the network. This approach to information retrieval avoids many of the existing problems in IR associated to synonymy and polysemy in lexically matching query approaches.

KEYWORDS: Adaptive hypertext, associative networks, spreading activation, information retrieval.

1 Introduction.

1.1 Lost in Hyperspace, again.

The problems users of hypertext networks experience in browsing the networks' hyperlink structure for the retrieval of information are well-documented. Users report difficulties in finding the information they need, assessing the general structure of the network they are browsing, the feeling of getting lost and not knowing where to go [10]. Several authors have identified hyperlink structure as one of the important factors in determining the efficiency of retrieval and identification of information from hypertext systems since the early stages of their growing popularity [22] [23] [28]. For the most part, the work concerned with improving human information retrieval from hypertext systems, however, has since then mainly been involved with other issues such as search engines using lexical matching (node-keyword matching), user modeling for adaptive information access and presentation [6][3], visualization techniques, automated browsing agents, browsing assistance [18], etc... Although certain attempts have been made to measure degrees of user link satisfaction and apply this data to a hypertext's network structure improvements[15][20], hyperlink structure has mostly been taken as a given. Hypertext network structure has been considered a factor that eluded quantitative analysis due to its subjective nature.

1.2 IR and hypertext.

Research on automated retrieval, for the same reason, has also avoided dealing with hypertext hyperlink structure and has focussed on applying traditional IR methods. Use of the semantics expressed in hypertext network structure has mainly been left to human navigators and designers. Automated information retrieval research in some respect disregarded hypertext network structure and tried to operate on the more tangible aspects of hypertext content like e.g. text, keywords, etc... The IR approach to retrieval from hypertext networks and the WWW has led to the implementation of keyword-based search engines for automated retrieval. Most search engines for the WWW today are based on the lexical match principle in which documents are retrieved by lexically matching a set of provided keywords. This approach has a number of fundamental problems with polysemy and synonemy in which a same word can have different meanings and the same meaning can be expressed by different keywords. An interesting approach to this problem is the use of Latent Semantic Indexing (LSI) which resolves many of the problems associated with keyword-based IR, but it does not address better or more efficient use of the information stored in the hypertext network's structure [2][12].

1.3 Hyperlink structure as a resource for IR and human retrieval.

Due to the problems inherent in applying traditional IR retrieval methods to a corpus the size and variety of the WWW, the recent years have seen a shift towards methods of analyzing, improving and using hyperlink structure for human as well as automated retrieval. The study of human factors in hypertext browsing[19][27][7], ways to adaptively improve hypertext network structure

[6][11] and automated methods of retrieval that rely on hyperlink structure [16] are all examples of this trend. The view that hyperlink structure is an important resource, not only to improve human navigation but also for the automated retrieval of information, has gained ground and has led to interesting applications like e.g. Chakrabarti and Kleinberg's system[8] to uncover authorative pages by iteratively analyzing hubs and authorities in the web's hyperlink structure.

We believe an integration of the methods put forward in the domain of data mining web user data [25][29], the field of adaptive hypertext and the use of hyperlink structure for automated retrieval is most promising. First, the analysis of human navigation can aid strongly in uncovering relations among hypertext pages not expressed in the pre-designed pattern of linkage. Second, hypertext network structure can be adapted to these findings to improve hyperlink structure and consequently enable more efficient human and automated retrieval. Third, these measurements will aid automated retrieval methods in that they provide additional information on the nature and weight of the existing links and pages in the network. In this paper we therefor propose a methodology that uses an adaptive hypertext system to derive associative relations between the pages in a hypertext network and then applies spreading activation retrieval to the generated associative networks. We present some promising prelimenary results with a prototype of this system that will eventually lead to the implementation of a service that combines automated reshaping of a hypertext network's linkage patterns according to user preferences with a highly efficient recommendation system.

2 Methodology.

2.1 Design structure vs Use structure.

As mentioned in the introduction, hypertext network structure has generally been treated as a difficult to handle resource due to the subjective nature of both its design and use. An assessment of any hypertext network's link structure would involve the analysis of the designer's semantic motivations and intentions, which would prove be a practical and quantitative impossibility. From the viewpoint of optimal web or hypertext esign, however, the author's intentions and motivations are not necessarily the most relevant factors. Rather than relying on measurements of the designers preferred structure, an analysis directed towards optimal hypertext design would have to focus on user preferences and these can more easily be subjected to quantitative analysis.

2.2 Navigation patterns and associative relations.

The basic assumption of our research is that the usage of connections in a hypertext network or web can be used as a measure of the "goodness" or "relevancy" of hyperlinks. The amount of use for a given hyperlink can be said to be an expression of the strength of the associative relation among the webpages in the users' mind. Measuring traversal frequencies for hyperlinks [25] in a hypertext network has been proven to be a quite succesful technique to not only predict future user link preferences [14] but also to interactively shape the structure of existing hypertext networks [5].

The information required to measure user link traversals can be derived from on-line user activity measurements as well as a web site's logs. A sufficient number of link usage measurements will express the linkage structure for a hypertext network as it is desired by its users and will allow us to empirically inform and even automate the design and improvement of hypertext networks.

2.3 Adaptive hypertext: user preferences as a design principle.

How measurements of user link preferences can be used as a design principle for hypertext networks is illustrated by the Adaptive Hypertext experiments we conducted late 1995 and 1998[4]. The experiments showed how a randomly structured network of 150 hypertext pages (representing the 150 most frequent words in the English) can evolve to a well structured hypertext network by adapting link weights in the network to user traversal frequencies. The networks consisted of a set of randomly connected hypertext pages. A set of three learning rules continuously adjusted hyperlink weights according to certain patterns and frequency of link traversal thereby reshaping the network's structure according to user navigation patterns:

1. frequency: reinforced the hyperlink between nodes a and b with a certain amount anytime it has been traversed
2. transitivity: reinforced the connection between nodes a and c with a certain amount when the connection $a \rightarrow b \rightarrow c$ has been traversed
3. symmetry: reinforced the hyperlink between a node a and a node c with a certain amount when the hyperlink $c \rightarrow a$ has been traversed

Changes in link weight were fed back to the users of the system by a rank ordering of the hyperlinks from a given hypertext page.

The experimental networks of 150 nodes (150^2 possible connections) were shown to be able to, only after a relatively small number of visits (4000 selections), converge to a stable and well-structured state representing the users preferred pattern of linkage [4]. The associative data of this weighted hyperlinks network could then in its turn be used to analyse hyperlink structure by

spreading activation. A first experiment with spreading activation on the structure of a trained adaptive hypertext network was carried out in 1995 and is still available online (http://pespmc1.vub.ac.be/spreadact.html).

2.4 Spreading Activation as retrieval and analysis mechanism.

The technique of spreading activation is based on a model of facilitated retrieval [21] from human memory [1][9] and has at least once been implemented for the analysis of hypertext networks structure by [24]. The model assumes that the coding format of human memory is an associative network in which the most similar memory items have strongest connections[13][17]. Retrieval by spreading activation is initiated by activating a set of cue nodes which associatively express the meaning of the to be retrieved nodes. Activation energy spreads out from the cue nodes to all other related nodes modulated by the network connection weights as shown in fig. 1.

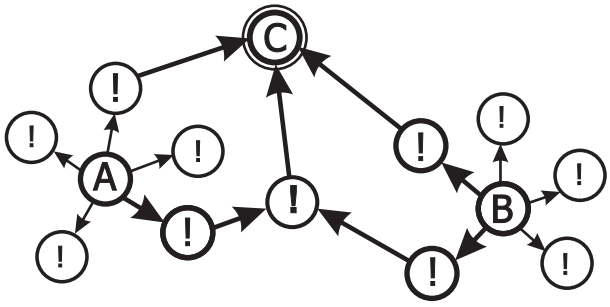

Figure 1: Activity spreads from activated nodes A and B and accumulates in retrieval result C.

The retrieved nodes are the ones that directly or indirectly accumulate most (or above a certain threshold) activation energy. The retrieval results can be used to predict how starting from a certain set of nodes of interests, users will hone in on certain parts of the network. Due to this property, spreading activation can aid designers in the improvement of hyperlink structure to better fit the users associative preferences. For the same reason spreading activation also has interesting applications in the implementation of search engines that are not dependent on explicit word matches, but rather use associative descriptions, i.e. they enable content-dependent retrieval.

Spreading activation for hypertext networks can be implemented by the following iterative algorithm.
Let V be the set of all n_i hypertext nodes,
and M be a matrix whose elements w_{ij} are the normalized strength of association between two nodes n_i and n_j so that for every n_i, $\sum_{j=1}^{n} = 1$.
A vector A_0 represents the set of initially activated cue nodes in the network; its values a_i indicate the level of activation for the nodes n_i

The spreading of activation in the network can then be iteratively modelled over discrete time steps
$t = 1, 2, \ldots, n$
as:
$A_{(t+1)} = A_0 + M.A_{(t)}$ or
for a give number k iterations:
$A_t = A_0 + \sum_{t=1}^{k}(A_0.M^t)$
Applications such as those of [24] implement specific relaxation and decay functions for the activation of nodes in the network using a modified matrix M. We have chosen for an implementation in which the activation value of any node in the network decays with a factor 0.2 for every iteration.

3 A case-study: the Principia Cybernetica Project.

3.1 A systems science hypertext encyclopedia.

In this paper we will present a case-study in which the web logs for the Principia Cybernetica Project's (PCP) website (see fig. 2) (http://pespmc1.vub.ac.be/)
have been analyzed and mined for association strengths according to the Adaptive Hypertext method. The gathered data has then been used for a spreading activation application.

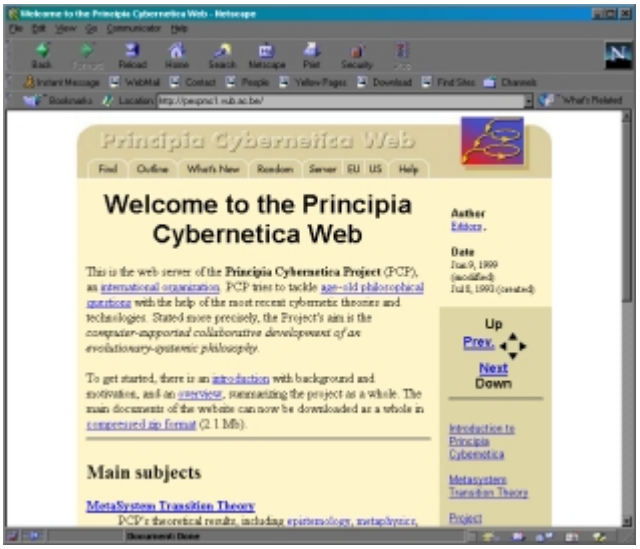

Figure 2: The Principia Cybernetica Website

The Principia Cybernetica website is devoted to the foundational principles of cybernetics and systems science. Established in 1993 as one of the very first web sites in the world, it has since then been the largest and most complete on-line hypertext encyclopedia on the philosophical principles underlying systems science. It also hosts a range of organizations, mailing-lists and other information sources for the scientific community. The website was a prime subject for the application of our methodology. Its structure and content have been the work of a small design team who were all strongly committed to an on-going and conscious effort to design the network in the most logical and well-structured

way. Its content is furthermore a heterogeneous mixture of highly abstract concepts (e.g. life, space, time, etc...) and less abstract and more general pages such as an introduction to the project, its participant, a mailing list and general scientific concepts and principles. PCP web contains about 900 general webpages including a linked dictionary of cybernetics and a library of digitalized books. The website attracts a wide audience of scientists and laypersons. It receives up to 5000 requests daily and keeps extensive logs of its usage.

3.2 Web site's existing structure.

The website designers team issued a list of the 423 most relevant web pages on the site. An agent scanned the website for these pages and identified the embedded hyperlinks. Based on this data, the agent generated an adjacency matrix representing the network's actual, hardwired structure that will later be compared to the structure of users' associations.

3.3 Web Logs.

The PCP website's server records one log entry per HTTP request. Each of these records contains a request IP (originator of the request), request URL (page requested), number of bytes transferred, status of the request (whether the request was successfully completed or not) and the time and date of the request. Extensive logs have been kept for all years since the creation of the website and have been preserved for later analysis.

4 A spreading activation retrieval engine.

4.1 Adaptive hypertext using Log analysis.

We used the PCP web logs that were kept for June 98 for a first analysis. This log contained 98654 requests. A filter removed all HTTP requests that had either failed, had exactly the same timestamp (robots) or had not been refered by the PCP website. Using the IP and time stamp on each request, a navigation pattern for each request originator (denoted by IP number) was composed according to the adaptive hypertext learning rules.

A linear series of minimal 3 HTTP requests was derived for each same IP number and taken to denote an individual navigation path. Using this path an adaptive hypertext network of the 423 chosen web pages was trained according to the methodology discussed in 2.3, with the only difference that web restructuring wasn't taking place on-line by the users' navigation patterns but rather as an after the fact reconstruction of user activity. All log lines were read, filtered and used to train the adaptive hypertext network (see fig. 3).

A 423 x 423 matrix representing the associative structure of the website's hyperlink structure was generated from the final network's structure. The resulting associative matrix's values were normalized and transformed into association strengths.

4.2 Implementation.

The implemented spreading activation algorithm consisted of three java packages that will shortly be combined to a public internet service (see fig. 5) but is now available as an off-line package for experimental purposes (see fig. 4)

First, a keyword matching module receives a set of keywords and consequently generates an activation vector representing the initial activation state of the nodes in the hypertext network (cue node activation). This module links keywords to webpages by matching the provided keywords to a stored list of webpage titles and URLs. The number of matches for a specific webpage title indicates that page's activation strength. Second, a spreading activation module implements the actual spreading activation algorithm that is highly similar to the one implemented and proposed by Pirolli et al (1996). The algorithm carries out an iterative multiplication of the (initial or resulting) activation vector and the normalized association matrix. For each webpage a maximum activation value over all iterations is calculated. The 15 top ranking pages are selected as the retrieval results. Third, an interface module was implemented that received user requests and generates the result presentation format. This interface module will shortly be converted to a servlet interface, so that the retrieval algorithm can be made available to all web site's users.

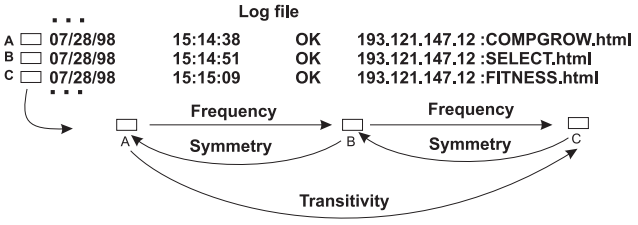

Figure 3: Analysing logs by adaptive hypertext training.

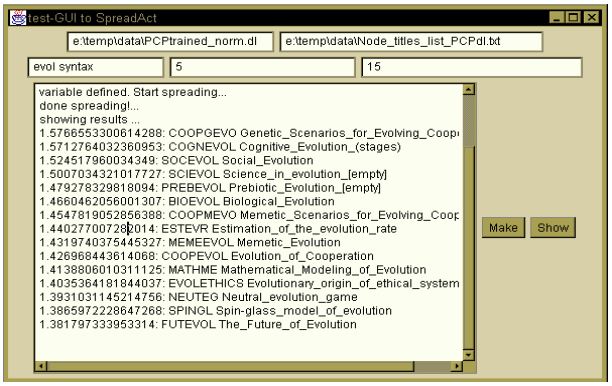

Figure 4: Off-line spreading activation package

4.3 Preliminary results.

Table 1 provides a number of retrieval examples using the spreading activation system. Activation keywords were chosen specifically to investigate how well the algorithm could approximate the meaning represented by the activation keywords. The search engine did indeed repeatedly seem to catch the central meaning of what the activation keywords were trying to express. For instance, when cued with the activation keywords "web link semant", the system generated a list of retrieval results containing: "Adaptive hypertext networks", "Consensus Building", "Hypertext web as a semantic network" and "Links on Cognitive Science and AI". Even more remarkable is that when the system was cued with the keywords "begin life" it generated a list of retrieval results containing among other things a page on "Arguments for and against the Existence of God", indicating that the system instead of focussing on the individual keywords alone in some way grasped the combined meaning expressed by the cued nodes.

Another example of how the spreading activation managed to query the hypertext network based on the meaning expressed in the set of activated keywords can be found in table 2. This table shows the retrieval results for both "people syntax" and "people PCP". For the first combination the retrieval engine found words such as "semantics", "message and "meaning", indicating the combined meaning of both "people" and "syntax" refered to language and semantics. For the second combination however the results contained nodes that refered to the social aspects of the Principia Cybernetica project, like e.g. mailing lists, symposia, etc...

These spreading activation results can be said to reflect the relations and structure the users had expressed in their use of the hypertext network and as such are an indication of whether the network has been organized adequately or not. The results were convincing enough to start using the spreading activation system as a full blown retrieval engine for the network in the very near future. Users will however have to get used to a different query format: rather than expressing word matches and logical queries, the spreading activation queries should express the meaning and associations of the sought after information.

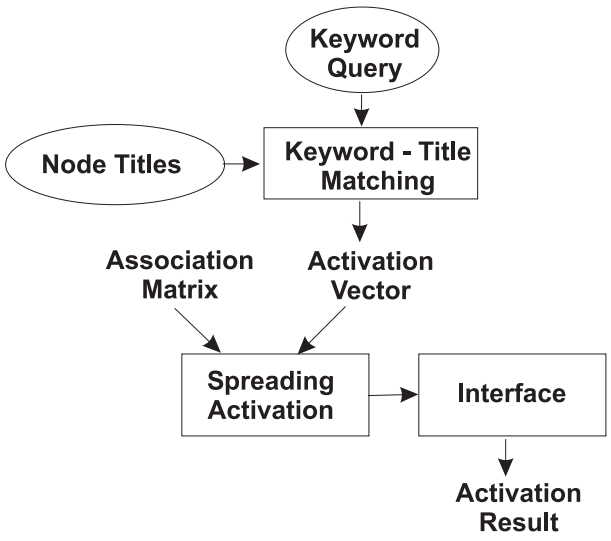

Figure 5: Spreading activation modules

5 Conclusion.

The spreading activation technique combined with the Adaptive Hypertext methodology applied to the associative analysis of web logs has shown to be a promising method for improving the structure of hypertext networks, informing web designers on the underlying structure of user preferences and as an powerful retrieval application. Our past experiments with adaptive hypertext networks [5] had already provided some insight into how spreading activation and association data (http://pespmc1.vub.ac.be/spreadact.html) could improve the link structure of entire hypertext networks and even automate their design. Unfortunately, the earlier results had limited face value due to the experimental reduction of hypertext networks to word networks. We see these results as a natural extension and a prime application of that work. Further experiments with the system will concentrate on the further analysis of the gathered data, fine-tuning of the spreading activation parameters and better visualization techniques for its results. Recent research in the more controlled setting of citation databases has also shown considerable promise for this technique in terms of automated resource linking and user recommendations [30]. Finally, we are also pursuing a structural analysis and user adaptation of the same data within the Adaptive Recommendation Project [26] (http://www.c3.lanl.gov/ rocha/lww).

REFERENCES

1. John R. Anderson. A spreading activation theory of memory. *Journal of Verbal Leaning and Verbal Behaviour*, 22:261–295, 1983.

2. M.W. Berry, S.T. Dumais, and G.W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595, 1995.

3. Michael Bieber, Roberto Galnares, and Qiang Lu. Web engineering and flexible hypermedia. In Peter Brusilovksy and Paul de Bra, editors, *Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia (Hypertext 98)*, Pittsburgh, USA, June 1998.

4. Johan Bollen and Francis Heylighen. Algorithms for the self-organization of distributed, multi-user

networks. In R. Trappl, editor, *Proceedings of the 13th European Meeting on Cybernetics and Systems Research*, pages 911–917, Vienna, Austria, 1996. Austrian Society for Cybernetic Studies.

5. Johan Bollen and Francis Heylighen. A system to restructure hypertext networks into valid user models. *The new review of Hypermedia and Multimedia*, 4, 1998.

6. Peter Brusilovsky. Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6(2–3):87–129, 1996.

7. Lara D. Catledge and James E. Pitkow. Characterizing browsing strategies in the world wide web. *Computer Networks and ISDN Systems*, 27:1065–1073, 1995.

8. Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopallan, David Gibson, and Jon Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. 7th International World Wide Web Conference*, Brisbane, Australia, April 1998.

9. A.M. Collins and E.F. Loftus. A spreading activation theory of semantic processing. *Psychological Review*, 82:407–428, 1975.

10. D.M. Edwards and L. Hardman. Lost in cyberspace: Cognitive mapping and navigation in a hypertext environment. In R. McAleese, editor, *Hypertext: Theory into practice*, chapter 7. Ablex Publishing Corporation, New Jersey, 1989.

11. John Eklund. The value of adaptivity in hypermedia learning environments: a short review of empirical evidence. In Peter Brusilovksy and Paul de Bra, editors, *Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia (Hypertext 98)*, Pittsburgh, USA, June 1998.

12. Peter W. Foltz. Using latent semantic indexing for information filtering. In R. B. Allen, editor, *Proceedings of the Conference on Office Information Systems*, pages 40–47, Cambridge, MA, 1990.

13. G. Hinton and J.R. Anderson. *Parallel Models of Associative Memory*. Hillsdale Publishers, New Jersey, 1981.

14. T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In *Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97)*, 1997.

15. C. Kaplan, J. Fenwick, and J. Chen. Adaptive hypertext navigation based on user goals and context. *User Models and User Adapted Interaction*, 3(2), 1993.

16. J.M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, 1998.

17. W. Klimesch. *The Structure of Long Term Memory: A connectivity Model of Semantic Processing*. Lawrence Erlbaum and Associates, Hillsdale, 1994.

18. H. Lieberman. Letizia: An agent that assists web browsing. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Montreal, August 1995.

19. Gary Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, Cambridge, UK, 1995.

20. N. Mathe and J.R. Chen. User-centered indexing for adaptive information access. *International Journal of User Modeling and User Adapted Interaction*, 6(2–3):225–261, 1996.

21. D.E. Meyer and R.W. Schvaneveldt. Facilitation in recognition pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90:227–234, 1971.

22. J. Nielsen. The art of navigating through hypertext. *Communications of the ACM*, 33(3):298–310, 1990.

23. H. Van Dyke Parunak. Ordering the information graph. In E. Berk and J. Devlin, editors, *Hypertext and Hypermedia Handbook*, pages 299–325. McGraw-Hill, 1991.

24. Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sow's ear: Extracting usable structure from the web. In *Proceedings of CHI'96 (ACM), Human Factors in Computing Systems*, Vancouver, Canada, April 1996. ACM.

25. James Pitkow. In search of reliable usage data on the www. In *Proceedings of the Sixth International WWW Conference*, Santa Clara, California, April 7-11 1997.

26. Luis Mateus Rocha. Talkmine and the adaptive recommendation project. In *Proceedings of ACM Digital Libraries 99*, Berkeley, California, August 1999.

27. J.F. Rouet, editor. *Hypertext and Cognition*. Lawrence Erlbaum Publishers, New Jersey, 1996.

28. P.L. Schoon. World wide web hypertext linkage patterns. In *Proceedings of the International Meeting of the World Conference of the WWW, Internet and Intranet*, Toronto, Canada, October 1997.

29. Myra Spiliopoulou. The laborious way from data mining to web mining. *Int. Journal of Comp. Sys., Sci. & Eng., Special Issue on "Semantics of the Web"*, Mar. 1999.

30. Herbert Vandesompel. Reference linking in a hybrid library environment (ii). *D-Lib Magazine*, 5(4), 1999.

6 Acknowledgements: This research has been made possible by the disposition of user data and logs by the PCP team at the Vrije Universiteit Brussel and the Los Alamos National Laboratory. Special thanks are extended to Francis Heylighen at the Leo Apostel Center (Vrije Universiteit Brussel) for his kind collaboration.

"people syntax"	"people PCP"
Semantics	Reactions, discussions, comments
Message	The Global Brain Group
Meaning	Symposium on Memetics
Syntax	Subscription to the Global Brain mailing list
Command	Symposium: Theories and Metaphors of Cyberspace
Johan Bollen	Criticisms of Principia Cybernetica
Historic record	Subscription to PRNCYB-L
Symbol	Management of Principia Cybernetica
Semiotic Terms	About Jean-Marc Dewaele
Abstraction	Subscription to PCP-news
Intuition	Sample Issue of 2-monthly PCP-news
Sentence	Problem-solving
Language	Gathering a variety of contributions
Doubt	Principia Cybernetica Web and the "Best of the Web" awards
Truth	Submitting Nodes for Inclusion in Principia Cybernetica Web

Table 1: Ordered results for spreading activation retrieval for "people syntax" and "people PCP" keywords.

"web link semant"	"begin life"
Adaptive hypertext networks	Semantic Analysis
Consensus Building	Links on Cognitive Science and AI
Semantic Analysis	Multiple axiomatization sets, a metaphor for
Links on Cognitive Science and AI	Arguments for and against the Existence of God
Fuzzy logic and sets [empty]	Methodology for the Development of MSTT
Principia Cybernetica in "Wired" magazine	Links on Future Development
Semiotic Terms	Meaning Goes First [empty]
Links on Computer Interfaces and the Web	Multicellular organisms
Direct Interfaces into the Global Brain	MST as the quantum of evolution [empty]
Principia Cybernetica Copyright Statement	Consensus Building
collaborative granularity [empty]	The Origins of Life
Hypertext web as a semantic network	Sexuality as a Metasystem Transition
Structure of Principia Cybernetica Web	Biological Evolution
Links on Future Development	History of the Principia Cybernetica Project
The Social Superorganism and its Global Br	Foundational Concep

Table 2: Ordered results for spreading activation retrieval for "web link semant" and "begin life" keywords.