

Categorization of web pages – Performance Enhancement to Search Engine

S. Lakshminarayana
National Institute of Oceanography (CSIR)
Regional Centre, 176 Lawsons Bay
Visakhapatnam – 530 017, India
sln@nio.org

ABSTRACT

With the advent of technology man is endeavoring for relevant and optimal results from the web through search engines. Retrieval performance can often be improved using several algorithms and methods. Abundance in web has impelled to exert better search systems. Categorization of the web pages abet fairly in addressing this issue. The anatomy of the web pages, links, categorization of text and their relations are empathized with time. Search engines perform critical analysis using several inputs for a keyword(s) to obtain quality results in shortest possible time. Categorization is mostly done with separating the content using the web link structure. We estimated two different page weights (a) Page Retaining Weight (PRW) and (b) Page Forwarding Weight (PFW) for a web page and grouped for categorization. Using these experimental results we classified the web pages into four different groups i.e. (A) Simple type (B) Axis shifted (c) Fluctuated and (d) Oscillating types. Implication in development of such categorization alleviates the performance of search engines and also delves into study of web modeling studies.

Keywords: Categorization, web pages, link analysis, information retrieval.

INTRODUCTION

The rapid growing web drew in many researchers to study the Information Retrieval (IR) studies for delivery of relevant output. Better results are obtained through the technological improvement and handling huge data bases. Machine dependence is grown considerably for the results in search techniques. Anatomy of the web pages, links and their relations were understood at large during the development of various methods and algorithm complexity. Search engines contribute significantly through a well defined process with inherent ranking methods in time. Keyword(s) play a larger role in this process. Many researchers worked through content categorization to address this issue and we assorted these web pages into four different types with new page parameters i.e. (a) Page Retaining Weight (PRW) and (b) Page Forwarding Weight (PFW). We also provide the method of such categorization and the advantages. Optimal path to the spider/worm can be recommended through clustering of pages and the location of the link. In the absence such works, the spider/worm either moves to the next page available at the least time or by network selection. This classification serves in judgment of traversal of web spider/worm and minimization. Such processes are the major areas of research in IR and strive to improve the effectiveness of interactive IR and can be used as performance evaluation tool.

The classification studies at early stages were with strong human interaction than machine learning. The term reference and feedback processes are used for classification by Amanda Spink¹ (1995), Xiaofeng He et.al² (2002), Agosti et.al³ (1991), Lifantsev⁴ (2001), Schettini et.al⁵ (2002) and many other researchers. Kraft and Lee⁶ (1979) and Lakshminarayana⁷ (2003) described few stopping rules and the effect on expected search length which is useful for spider/worm traversal. Classifications of web pages were also discussed on the basis of content link structure by Kleinberg⁸ (HITS Algorithm), Chakrabarti et.al⁹ (2002) and Borodin et. al¹⁰ through a process of iterative calculation of Hub and Authority weights. Many user queries are limited to one or two words (Ross and Wolfram¹¹ 2000) and hence these context free key words are more used in web page dynamics and categorization.

Reflexive, symmetric, transitive and anti-symmetric properties of graph theory are well practiced by Kleinberg⁸, George Meghabghab¹² and Lakshminarayana et.al¹³. Patten reorganization techniques, Artificial Intelligence methods (Karen Sparck Jones¹⁴), Fuzzy ranking techniques (Nikraveshe Masoud¹⁵, Tran and Duckstein¹⁶, 2002, Cohen and Feigenbaum¹⁷, 1982) are also contributed in ranking methods. Brin and Page¹⁸ studied the hypertext anatomy and

developed a new ranking method (Google). Latent Semantic Indexing is also used for information retrieval methods. (Frei and Stiger¹⁹). Storage and query processing methods improved the efficiency in getting output from the large databases. The time for decomposition is a factor in all above methods. As decomposition is one-time expense, vector space methods could be used in information retrieval (Lakshminarayana²⁰). Linear algebraic methods were used for better indexing the content (Ramara G Kolda and Dianne P O'Leary²¹). Application of Neural Networks to Information Retrieval was also discussed by Kwok²². The TREC Methods by McBrayn²³ (1994), Hub and Authority methods of Kleinberg⁸ (1999) and Chakrabarti et. al⁹ (1998) and Statistical methods by Gupta and Kapoor²⁴ (1983) contributed to categorize the web pages through an inherited weighted factor. Data modeling for retrieving with structured and unstructured data and issues in usage of the architecture, design and development of web were addressed by Agosti et. al²⁵ (1991) and Croft & Thompson²⁶ (1989). Weaver et. al²⁷ (1989) suggested a frame based language for classification of data in information retrieval. Categorization of web pages in respect of a keyword will play a predominant role in such search process optimization.

Experimental results of TREC methods, Hub and Authority method and Statistical methods were applied to a set of odd 8000 pages. We observed that there is about 70% of pages are non-relevant in case of TREC method, 65-68% pages does not weigh a reasonable factor in the Hub and statistical methods, they are 26% related and 74% non-relevant in case of Statistical methods (Lakshminarayana⁷, 2003). Several authors reported that instances were observed that information is available but not relevant to the key word(s).

METHODS

Search methods discussed earlier are from negotiating with the produced output. The processing and ranking depends on the search engine algorithm and traversal methods in the web. Optimized traversal of a worm/spider collects the best information from the web. A set of given hyper linked web pages will have several paths that a spider or a worm could walk to get the information. The next page to be visited is vital in many situations. However, due to advantage of computing machines, the spiders are able to browse more than 400 documents/second (Risvik and Michelsen²⁸, 2002).

A network of 8000 pages is taken for our study. The weight of the each page was computed initially using Kleinberg⁸ (1999) hub and authorities method. Leaf node page weights

(end pages) are taken as unity. A user has three choices after browsing a web page (a) to go back-ward i.e. to visit to the earlier page (b) to visit a page/link which is linked to the existing page or (c) to leave and go to new process. We considered that the page should have some useful information which is interesting to remain. This is assigned as Page Retaining Weight (PRW). The user will have a tendency to move forward to the pages/links that are connected to the existing page. This is rated as Page Forwarding Weight (PFW). This study excogitates the concept of Klienberg⁸ and Chakrabarti et.al⁹. Initially the PFW is taken as zero for all the pages and PRW is the page-weight itself.

The following new formulae were used to computer PFW and PRW.

(1) Page Forwarding Weight (PFW) = $\frac{\sum X}{n}$ where n is the total number of hyper links exists

on that page and $\sum X$ is the sum of the weight of all such links appear in that page and

(2) Page Retaining Weight (PRW) as $A = \frac{AXB}{A^2 + B^2 - AXB} - \frac{\sum X}{n}$ where A is the weight of

page A; B is the weight of the page B where the page A is connected. $\frac{\sum X}{n}$ is the PFW of

the page A.

Results after 300 iterations for all the pages in the network succumbed to categorize these pages into four different types.

RESULTS

Classification of web pages was discussed by klienberg⁸, Chakravarthi et.al⁹, Borodin et.al¹⁰ and many others for better understanding of the anatomy of web structure and to get relevant information. We have classified the pages in to four types with the values of page weight graph depicted after 300 iterations (Fig 1). The page diagrams indicate that they follow a pattern. It is also tenable that these could be categorized into four different categories (A) Simple type (B) Axis shifted (c) Fluctuated and (d) Oscillating type.

(A) **Simple type:** In this category the PRW became zero after a finite number of iterations.

This is because of continuous deduction of PFW from PRW. A decline was observed with PRW.

(B) **Axis shifted:** In this category the PRW does not become zero but reduces to a constant and thereafter no change in this PRW. (PFW tends to zero after few iterations)

(C) **Fluctuated:** In this type of pages the PRW will be fluctuating within a range and does not converge at all. The value can not be predictable but falls with in a range. Thus these pages will have different in nature.

(D) **Oscillating:** There are some pages which fall in this category in which the page forwarding weight will be oscillating between specific ranges.

Such classification facilitates the search engine spider/worm to get the optimal traverse so that extraction of information, transform its methods and re-load of the pages for page ranking becomes simpler.

The diagrams of first 23 pages (1-23) are given at Fig 1. It is observed that the pages 1,2,5 and 6 are of type (A); 10,11,12,13,14,15,17,20,21 and 22 are of type (B); 4,7,8,9,18,19 and 23 are of type (C) and 3 and 16 are of type (D). From the 8000 pages it is observed that 27-30% of pages fall in ‘A’ category 41-45% fall in ‘B’ category, 28-32% fall in ‘C’ category and the remaining in ‘D’ category. (The range in classification is due to fuzziness)

We have analyzed these groups with five odd key words for relevance separately and observed that more number of better ranked pages is within B-group (Table 1).

Group	No. of pages	No. of pages with Page Weight > 0.5 (normalized 0-1)	Approx % of pages	Average page Weight
A	1564	94 – 140	6 - 9	0.62 - 0.65
B	3360	1460 – 1612	44 - 48	0.78 – 0.86
C	2486	472 – 596	19 - 24	0.69 – 0.73
D	590	160 - 210	31 - 36	0.67 – 0.71

Table 1: Category wise distribution of web pages and their weights for five keywords

CONCLUSIONS

Classification of web-pages in this method accomplished for pigeonholing the high weighed pages together and facilitates a direction to the web crawler/robot. The Spider/worm could be designed to traverse towards a better clustered ‘B’ group. A better verge for spider can be reached at shorter number of node (page) traversals. Besides relevance optimality and precision can be arrived at in less number of iterations. Further such classification helps out in designing a performance enhancement tool for a search engine.

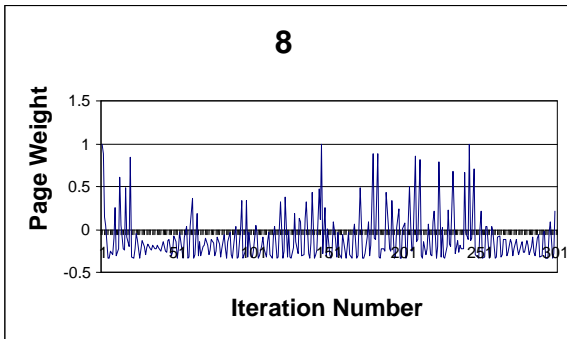
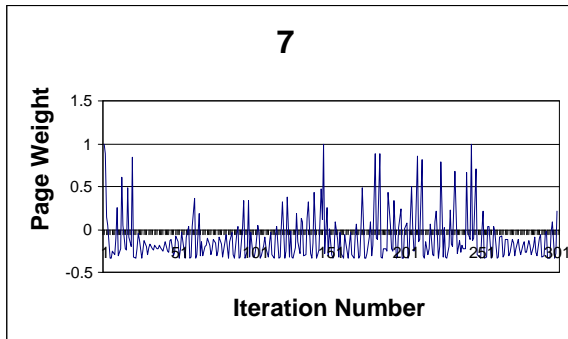
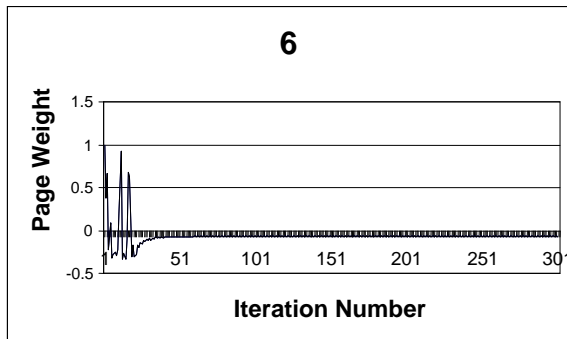
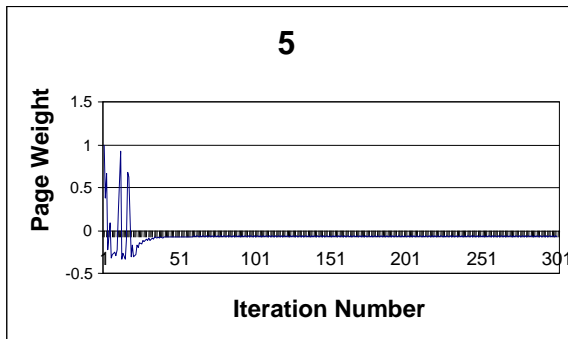
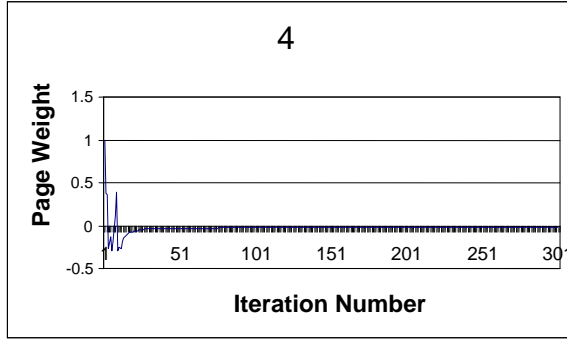
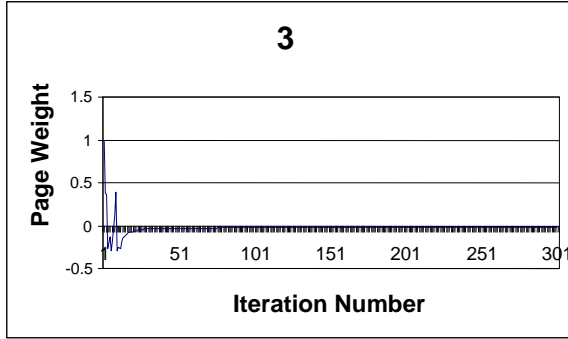
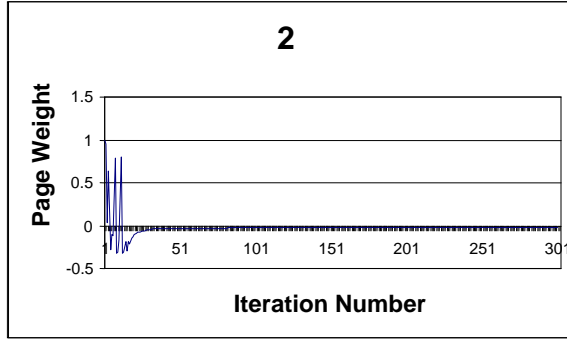
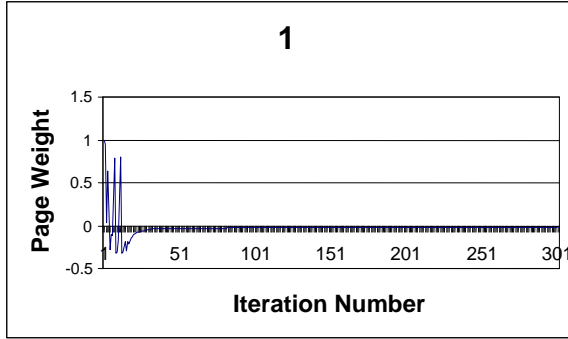
ACKNOWLEDGMENTS

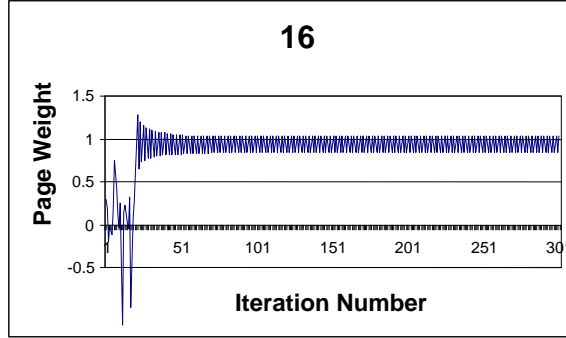
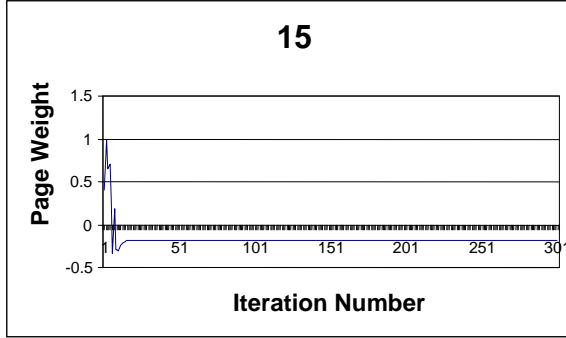
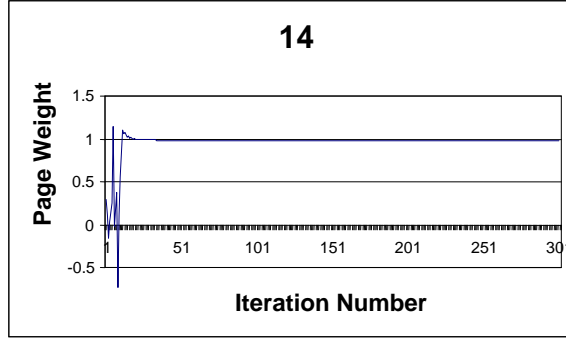
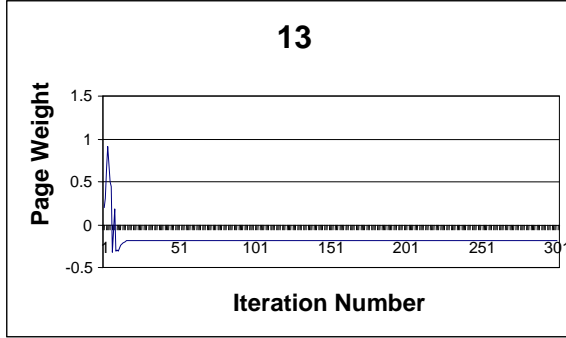
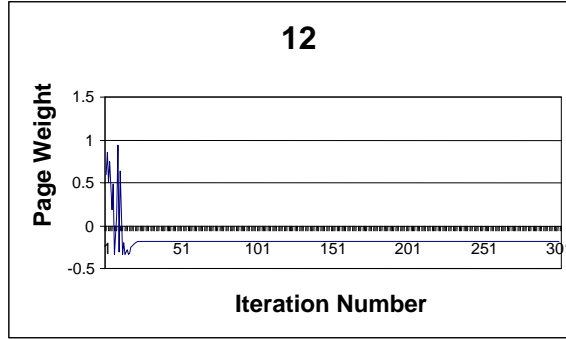
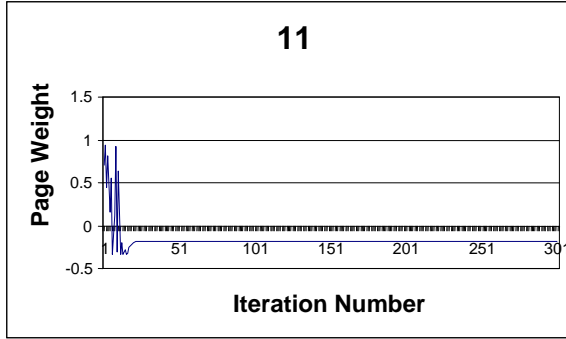
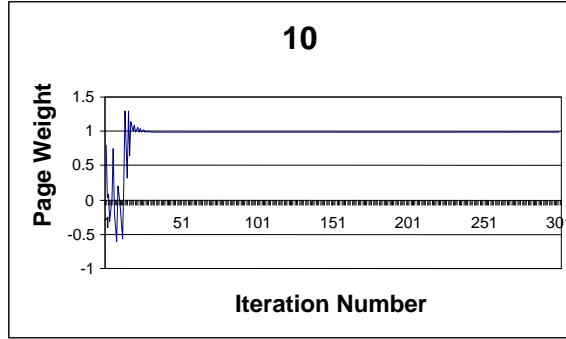
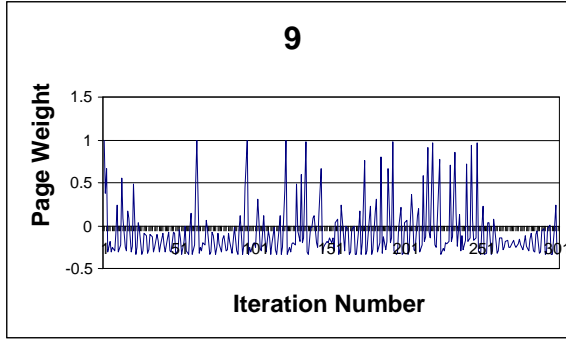
Thanks are due to Dr. Satish R. Shetye, Director, National Institute of Oceanography, Dona Paula, Goa – 403 004 for permitting this work and Dr. V.S.N. Murthy, Scientist-In-Charge, National Institute of Oceanography, Regional Centre, Visakhapatnam, India for the encouragement. (*NIO Contribution No: 4395*)

References:

1. Amanda Spink, Term Relevance feedback and mediated database searching: implications for information retrieval practice and system design, *Information Process Management*, Vol, 31:2, pp 161-171, 1995.
2. Xiaofeng He, Hongyuan Zha, Chris H.Q Ding, Horst D Simon, Web document clustering using hyperlink structures; *Computational Statistics & Data Analysis* Vol 41, pp 19-45, (2002)
3. Agosti. M, Crestani. F and Grandenigo. G, Towards data modeling in information Retrieval, *Journal of information science*, 25 (6): 307-319, 1989.
4. Lifantsev Maxim; A System for Collaborative web resource categorization and Ranking (Available at <http://www.ecsl.cs.sunysb.edu/~maxim/Papers/SCWRCR/>)
5. Schettini Raimondo, Carla Brambilla, Claudio Cusano: Content-Based Classification of Digital Photos. *Multiple Classifier Systems 2002*: 272-281
6. Donald H. Kraft, T. Lee: Stopping rules and their effect on expected search length. *Information Process. Management*. 15(1): 47-58 (1979)
7. Lakshminarayana S; Numbering of Class of Trees and Search Algorithms for World Wide Web; Ph.D thesis, Andhra University, 2003, pp 1-85.
8. Kleinberg. J.M Authoritative sources in a hyperlinked environment, *Journal of the ACM*, 46(5-7) pp 604-632, 1999.
9. Chakrabarti. S., Byron Dom, Prabhakaran Raghavan, Sridhar Rajagopalan, David Gibson, Jon Klienberg,. Automatic Resource Compilation by analyzing hyperlink structure and associated text. In *Proceedings of the seventh international World Wide Web conference*, Brisbane, Austria. (published in *Computer Networks and ISDN Systems*, Vol 30 (1-7), Article FP7), 1998.
10. Borodin Allan, Gareth O. Roberts Rosenthal J.S. and Tsaparas P: Finding authorities and hubs from link structures on the World Wide Web: In the *Proceedings of 10th International World Wide Web conference*, May 1-5, 2001, Hongkong (www10.org.hk)
11. Ross, N. & Wolfram, D. End User Searching on the Internet: An Analysis of Term Pair Topics Submitted to the Excite Search Engine. *Journal of the American Society for Information Science*, 51(10), 2000, pp 949-958.
12. George Meghabghab, Discovering authorities and hubs in different topological web graph structures, *Information Processing and Management*, Vol 38 (2002), 111-140.
13. Lakshminarayana. S, Viswas Chavan & Arvind Ghosh, World wide Web – A graphical approach, *Proceedings of National Conference of IEEE (Internet for India)*, Hyderabad, India, pp 21-23, 1997.
14. Karen Sparck Jones, Information retrieval and artificial intelligence, *Artificial Intelligence*, Vol 114 (1999) 257-281.

15. <http://buffy.eecs.berkeley.edu/IRO/Summary/02abstracts/nikravesh.1.html>
16. Liem Tran and Lucien Duckstein; Comparison of fuzzy numbers using fuzzy distance measure, *Fuzzy sets and Systems*, Vol 130(3) 16 September 2002, Pages 331-341
17. Paul R. Cohen and Edward A. Feigenbaum. 1982. *The Handbook of Artificial Intelligence*, Volume III. Los Altos, CA.: William Kaufmann. pp 1-74.
18. Brin, S. & Page, L. (1998). The anatomy of a large scale hyper-textual web search engine. In *Proceedings of the seventh World Wide Web conference*, Brisbane, Australia.
19. Frei. H.P. and Stieger. D. The use of semantic links in hypertext information retrieval, *Information Processing and Management*, Vol 31 (1995), 1-13.
20. Lakshminarayana. S, Dynamic ranking with $n + 1$ dimensional vector space models: An alternative search mechanism for World Wide Web, *Journal of the American Society for Information Science and Technology* Vol 54(3), 2003, pp: 274-275.
21. Kodla and Dianne; *ACM Transactions on information systems*, Vol 16, No 4, October 1998, pp 322-346
22. Kwok. K.L; Application of Neural Networks to Information Retrieval. In *Proceedings of the International Joint Conference on Neural Networks*, Vol 2 623-626, Washington, USA, January 1990, IEEE.
23. McBrayn. O.A., (1994) GENVL & WWW: tools for taming the Web. In *proceedings of the first international World Wide Web conference*.
24. Gupta S.C. and Kapoor V.K., *Fundamentals of Mathematical Statistics*, (1983) Sultan Chand & Sons, New Delhi.
25. Agosti. M., Colotti. R and Gradenico. G, Issues of data modeling in information retrieval. *Electronic publishing* Vol 4(4), 219-237 (December 1991).
26. Croft. W.B and Thompson. R.H. I³R: A new approach to the design of document retrieval system, *Journal of the American Society for Information Science*, 38(6):389-404, 1987.
27. Weaver. M.T., France. R.K., Chen. Q.F. and Fox. E.A., Using a frame based language for information retrieval, *International Journal of intelligence systems* 4(3): 223-258, 1989.
28. Risvik Knut Magne and Rolf Michelsen, *Search Engines and Web dynamics; Computer Networks*, Vol 39, pp 289-302, 2002.





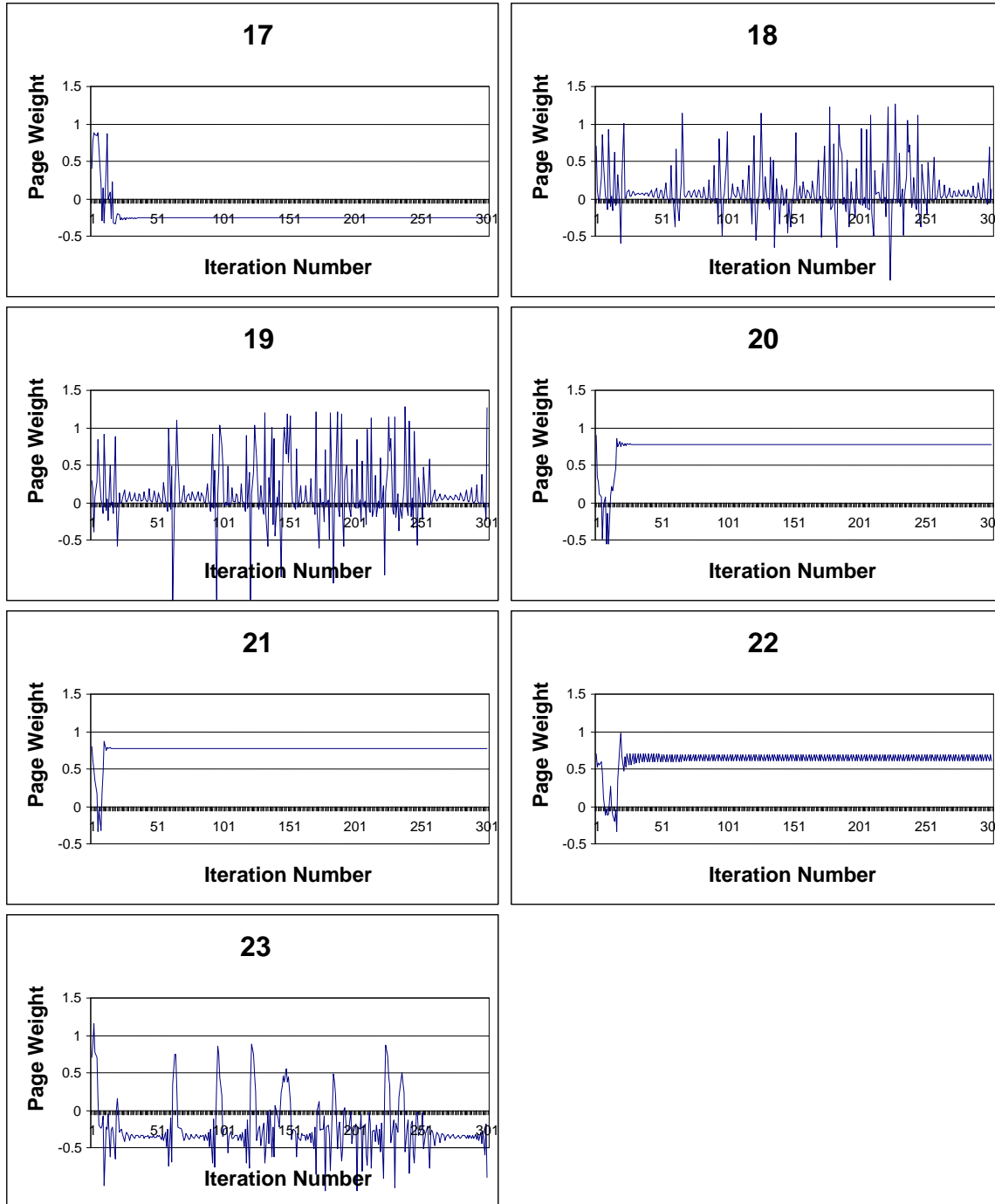


Fig:1 Page Weights of 23 pages for 300 iterations