UDC 004.032.026

P. Tymoshchuk
L’viv Polytechnic National University,
CAD Department

INFORMATION RETRIEVAL FROM DATA SETS VIA ANALOGUE NEURAL CIRCUIT OF MAXIMAL VALUE IDENTIFICATION FROM SIGNAL SET

Using the analogue neural circuit of maximal value signals from signal set identification is proposed for information retrieval in data sets. The circuit is fast, it has simple structure and can be implemented in a modern hardware. A resolution of the circuit is theoretically infinite and it is not dependent on a value of its parameter. An average time necessary for trajectory convergence of the circuit state variable to a steady state is not dependent on a dimension of input data. The results of numerical experiments obtained on the base of the data set provided by PageRank algorithm are presented. These results give witness of the circuit using for information search in data sets.

Key words: information retrieval, analogue neural circuit, hardware, computational complexity, resolution, state variable trajectory.

Для інформаційного пошуку у наборах даних запропоновано використання аналогової нейронної схеми максимального значення сигналів з набору сигналів ідентифікації. Схема є доволі швидкою, має просту структуру і її можна застосовувати у сучасному технічному забезпеченні. Розширення схеми є теоретично нескінченним і не залежить від значення її параметрів. У середньому час для траєкторії зближення змінної стану схеми до стаціонарного стану не залежить від величини введених даних. Наведено результати численних експериментів, які отримали на основі набору даних, наданих алгоритмом PageRank. Ці результати свідчать про використання схеми для інформаційного пошуку у наборах даних.

Ключові слова: інформаційний пошук, аналогова нейронна схема, технічне забезпечення, складність обчислення, розширення, траєкторія стану змінних величин.

1. Introduction

It is known that the techniques for information retrieval from data sets play a very important role as the size of the world-wide web exceeded 800 million pages in 1999 to 11.5 billion in 2005, and possibly more than 30 billion nowadays [1], [2]. A most promising work in utilizing the link structure of the web for improving the quality of search results may be PageRank, an iterative algorithm that determines the importance of a web page based on the significance of its parent pages [2], [3]. This led to many impressive works in the past decade, such as analyzing the efficiency [4], doing computational experiments [5], [6], improving the efficiency and effectiveness [7], [8] and further analysis on social networks [9], [10]. It was found out that a main bottleneck to large scale network search engine is not calculating the weighting coefficients but the quick sorting of those coefficients. This problem is the “Top-K” problem with L = 1 list of numbers, which is defined as follows: given a list of real numbers, find the top K scoring ones.

Many attempts on solving the “Top-K” problem efficient has been done in the past, e.g. [11] - [16]. Particularly, the algorithm Quicksort which has O(N log N) computational complexity on average is used for fast sorting of search results [17]. However, the growth of the size of internet has been far more rapid than that of computing speed in the past several decades and will remain in the future. Requirements to the time of obtaining the search results are also increasing. Therefore if the selection process has to be operated in real time or the number of inputs is large, parallel algorithms and hardware implementation of sorters are desirable for “Top-K” problem solving.

The problem of fast sorting can be effectively solved by using K-winners-take-all (KWTA) neural circuits [18] - [20] which find largest/smallest among set of unknown signals. Such circuits have essentially parallel structure. As known, KWTA circuit which is generalization of winners-take-all network, selects k largest among N input data, where 1 ≤ K < N [21]. KWTA has been shown to be a computationally powerful operation compared with standard neural network models of threshold logic gates [22], and has been widely used in various applications, such as decoding [23], feature extraction [24], signal processing [25] [26], etc.

In this paper, an analogue KWTA neural circuit which has a single state variable is used for obtaining solution of “Top-K” problem for information retrieval in data sets [27]. Experimental results are presented which indicates the potential efficiency of using the KWTA circuit for information retrieval.

2. An information search in data sets

As it has been pointed out, there are basically two main parts in internet information retrieval, one is calculating the weight of all the pages or data and the other is finding out the most ”wanted” K results with higher weighting coefficients and show them in a very short time. Let us use the PageRank algorithm for obtaining solutions of first part of the problem. We assume that after each crawl of the web, the ranking vector is computed only once. The values of its elements can then be used to change the ranking of search results. This means that PageRank algorithm does not need to be run most of the time when there happens to be a searching request. Note that many impressive works have been done to accelerate the calculation of PageRank [28].

The PageRank algorithm results can be sorted using, for instance, algorithm Quicksort which has O(N log N) computational complexity on average. Note that in internet searching, instead of sorting all of the pages, usually only the ten or twenty most ”interested” or ”related” pages with higher weights need to be figured out as soon as possible and shown to the costumers. Such the problem can be formulated as the problem of selecting K most important among N results, where $1 \le K < N$ [20], [29]. Let us assume that the weight vector of N pages obtained by PageRank algorithm is an input vector of KWTA neural circuit. Then one can find K the most important pages on the base of output vector of such the circuit. In order to solve such problem we use an analogue neural circuit which finds largest among signal set described by the following state equation:

$$\dot{x} = -\alpha \begin{cases} x, & \text{if} \quad E(x) > 0; \\ 0, & \text{if} \quad E(x) = 0; \\ x - A, & \text{if} \quad E(x) < 0, \end{cases} \tag{1}$$

where $x \in \Re$ is a state variable (scalar dynamical shift of inputs), $0 \le x_0 \le A$ is an initial condition,

$$E(x) = K - \sum_{k=1}^{N} S_k(x) \tag{2}$$

is a residual function,

$$S_k(x) = \begin{cases} 1, & \text{if} \quad r_{n_k} - x > 0; \\ 0, & \text{if} \quad r_{n_k} - x \le 0 \end{cases} \tag{3}$$

Is a step function, α is a constant parameter (decaying coefficient), which is used to control the convergence speed of state variable trajectory to KWTA state, $r = \left(r_{n_1}, r_{n_2}, ..., r_{n_N}\right)^T$ is a weight vector, obtained by PageRank algorithm, as it is supposed, with different (unequal) values of elements, ordered in a descending order of magnitude satisfying the following inequalities:

$$\infty > r_{n_1} > r_{n_2} > \cdots > r_{n_N} > 0, \tag{4}$$

$r_{n_k} = a_{n_k} - a_{\min}$, $n_1, n_2, ..., n_N$ are numbers of the first, second and so on up to N-th largest weight, $b_{n_k} = r_{n_k} - x$ is an output data of KWTA circuit, $k = 1,2,..., N$, $1 < N < \infty$, $1 \le K < N$, $A = a_{\max} - a_{\min}$, $a_{\min}$ and $a_{\max}$ are minimal and maximal possible values of inputs correspondingly [27]. Note that the circuit described by equation (1) has computational complexity O(N).

On a base of the state equation (1), corresponding analogue structural-functional neural circuit can be designed. Such the circuit is characterized computational complexity O(N) and it has simple architecture. The circuit can be implemented in a modern hardware using such traditional components as analogue summers, switchers, integrator and sources of constant voltage or current. Computational complexity of the circuit is less than that of other analogs. The circuit resolution ability if theoretically infinite and does not depend on its parameter value, i. e. if input data are distinct, the circuit can always identify them. Since the present circuit can operate correctly with any finite initial conditions, it requires neither a periodical resetting for repetitive processing of input sets, nor corresponding analogue supervisory circuit, nor spending additional processing time. This simplifies the hardware and decreases the convergence time to the KWTA operation [27].

3. Computer simulation results

In order to demonstrate a quality of functioning the KWTA neural circuit described by the state equation (1) for solving a problem of information retrieval in a global network, let us consider concrete example with results of numerical experiments, obtained for real data set. For this purpose, we design corresponding program in the codes of high-performance language of technical computing Matlab. To run such programs we use a 1.81 GHz desktop PC.

Example. Let us assume that after fulfilling search in a global network there was computed a page weight vector and its element values can be used for ranking of the search results. We suppose that it is necessary to find K=10 pages with largest values of weights among 500 pages. Let us use as an input vector of the model (1) a weight vector r_{n_k}, k=1,2,3,…,500 with elements uniformly distributed on interval [0,15000] obtained by using PageRank with the model parameters $\alpha = 1000$, $x_0 = 0$. We employ the variable order Adams-Bashforth-Moulton solver of non-stiff differential equations ODE113 with relative and absolute error tolerances equal to 10^{-5}. The simulation results presented in Fig. 1 show that steady state of output signals $b_{n_k} = r_{n_k} - x$, k=1,2,3,…,10 which indicate on pages with largest PageRank weights is achieved after the convergence time less than 0.005 s. For comparison, the time of such problem solving by using one of the most fast and simple analogue KWTA network described in [29] is larger than 0.1 s. Thus it is observable that the time needed on searching problem solving on the base of continuous-time model of analogue KWTA neural circuit (1) is by one order less than that of comparable analog from [29].

Note that in contrast to competitive analog presented in [29], either the average time or the average number of iterations needed for the state variable trajectories to converge to steady states do not depend on the problem size N that is an important advantage. It means that the KWTA neural circuit can perform equally fast as for small-scale as for large-scale dataset information retrieval.

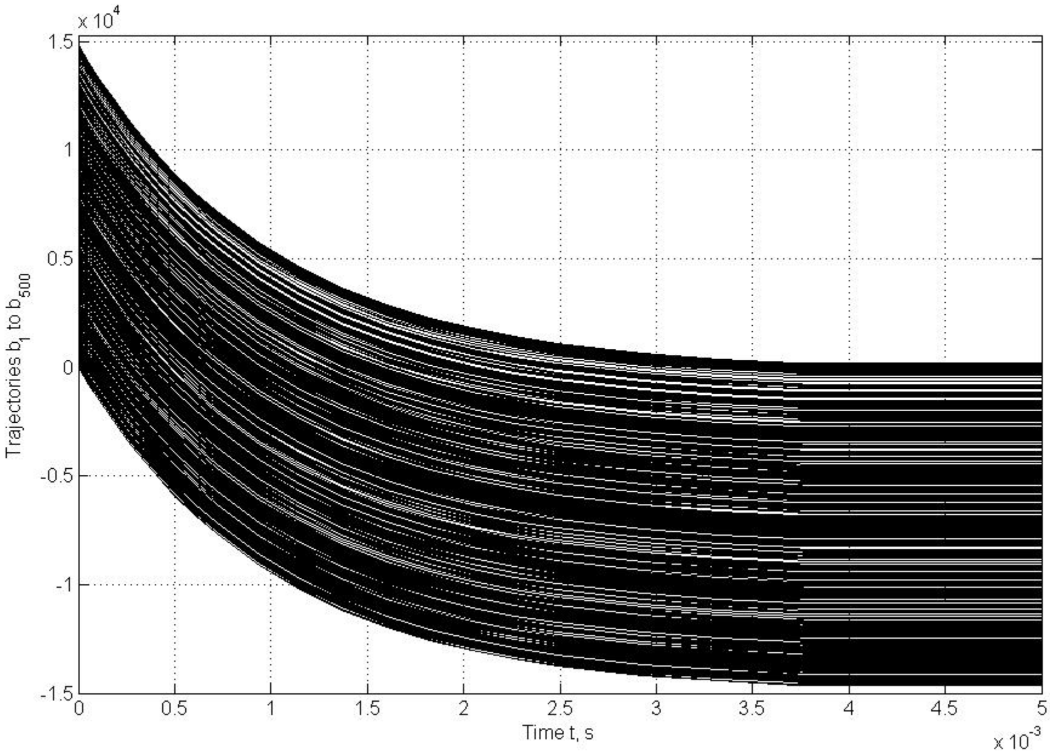

Рис. 1. Trajectories of output signals $b_{n_k} = r_{n_k} - x$ *, k=1,2,...,500*
of the neural circuit described by the state equation (1)

Only simulation results of information retrieval by software are presented in the example above. As known, such simulation take longer time than searching algorithms implemented in corresponding hardware. Therefore it can be predicted that with analogue hardware implementation, the KWTA neural circuit model (1) would perform an information retrieval process faster.

4. Conclusions

In the paper analogue neural circuit described by equation (1) is proposed to use for "Top-K" problem solving for information retrieval in data sets. Experimental results based on real world data set obtained after running the PageRank algorithm are presented. The proved superior performance of the KWTA circuit is demonstrated by the simulation results. An important characteristics of the model shown by experiment is that the state variable trajectories of the KWTA circuit converge equally fast facing a small-scale and large-scale problems. This indicates on a perspective of using an analogue neural circuit of finding largest among signal set for problem solving of information retrieval in data sets.

1. S. Lawrence and C. L. Giles, "Accessibility of information on the web," Nature, vol. 400, pp. 107–109, 1999. 2. S. Brin, and L. Page, "The anatomy of a large-scale hypertextual web search engine," in Proceedings of 7th International World Wide Web Conference, 1998. 3. L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the web," Technical report, Stanford University, 1998. 4. T. H. Haveliwala, "Efficient computation of PageRank," Stanford Univ. Technical Report, 1999. 5. T. H. Haveliwala and S. Kamvar, "The second eigenvalue of the google matrix," Stanford Univ. Technical Report, 2003. 6. A. Arasu, J. Novak, J. Tomlin, and J. Tomlin, "PageRank computation and the structure of the web: experiments and algorithms," in Proceedings of 11th International World Wide Web Conference, pp. 107–117, 2002. 7. G. M. Del Corso, A. Gull, and F. Romani, "Fast PageRank computation via a sparse linear system," Internet Mathematics, vol. 2, no. 3, pp. 251–273, 2005. 8. S. Kamvar, T. Haveliwala, C. Manning, and G. Golub, "Extrapolation methods for accelerating PageRank computations," in Proceedings of 12th International World Wide Web Conference 2003. 9. A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in Proceedings of 7th ACM SIGCOMM Conference on Internet Measurement, pp. 29–42, 2007. 10. J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in Proceedings of 15th International Conference on Knowledge Discovery and Data Mining, 2009. 11. G. N. Frederickson and D. B. Johnson, "Generalized selection and ranking," in Proc. of the 12th STOC, pp. 420–428, 1980. 12. M. Kendall and J. D. Gibbons, "Rank correlation methods," Edward Arnold, London, 1990. 13. R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top-k lists," SIAM J. Discrete Math, vol. 17, pp. 134–160, 2003. 14. N. Mamoulis, M. Yiu, K. Cheng, and D. W. Cheung, "Efficient top-k aggregation of ranked inputs," ACM Transactions on Database Systems, vol. 32, no. 3, article 19, 2007. 15. P. Hall and M. G. Schinek, "Inference for the top-k rank list problem," in Proceedings in Computational Statistics, pp. 433 –444, 2008. 16. K. Henderson and T. Eliassi-Rad, "Solving the top-k problem with fixed-memory heuristic search," Technical report, Lawrence Livermore National Laboratory, 2009. 17. Z. Guo and J. Wang, "Information retrieval from large data sets via multiple-winners-take-all", in : Proc. ISCAS, 2011, pp. 2669–2672. 18. T. M. Kwon and M. Zervakls, "KWTA networks and their applications," Multidimensional Syst. Signal Process., vol. 6, no. 4, pp. 333–346, 1995. 19. J. Wang, "Analysis and design of an analog sorting network," IEEE Trans. Neural Netw., vol. 6, no. 4, pp. 962–971, Jul. 1995. 20. Wang, J.: Analysis and design of a k-winners-take-all model with a single state variable and the Heaviside step activation function. IEEE Trans. on Neural Networks 9, 1496-1506 (2010). 21. E. Majani, R. Erlanson and Y. Abu-Mostafa, "On the k-winners-take-all network," Advances in Neural Information Processing Systems, vol. 1, pp. 634–642, 1989. 22. W. Maass, "On the computational power of winner-take-all," Neural Comput., vol. 12, pp. 2519–2535, 2000. 23. R. Erlanson and Y. Abu-Mostafa, "Analog neural networks as decoders," Advances in Neural Information Processing Systems, vol. 1, pp. 585– 588, 1991. 24. A. Yuille and D. Geiger, "Winner-take-all networks," The Handbook of Brain Theory and Neural Networks (2nd ed.), MIT Press Cambridge, MA, pp. 1228–1231, 2003. 25. A. Fish, D. Akselrod, and O. Yadid-Pecht, "High precision image centroid computation via an adaptive k-winner-take-all circuit in conjunction with a dynamic element matching algorithm for star tracking applications," Analog Integrated Circuits and Signal Processing, vol. 39, pp. 251–266, 2004. 26. A. K. J. Hertz, and R. G. Palmer, "Introduction to the Theory of Neural Computation," Redwood City, CA: Addison-Wesley, 1991. 27. Тимощук П.В. Модель аналогової нейронної схеми ідентифікації найбільших сигналів // Комп'ютерні системи та мережі. – 2012. – № 745. – С. 180–185. (Вісн. Нац. ун-ту "Львівська політехніка"). 28. The Google Search Engine: Commercial search engine founded by the originators of PageRank. Located at http://www.google.com/. 29. Z. Guo and J. Wang, "Information retrieval from large data sets via multiple-winners-take-all", in: Proc. ISCAS, 2011, pp. 2669–2672.