Information Retrieval from Large Data Sets via Multiple-winners-take-all

Zhishan Guo
Department of Mechanical and Automation Engineering
The Chinese Univ. of Hong Kong, Shatin, N.T., Hong Kong
Email: zsguo@mae.cuhk.edu.hk

Jun Wang
Department of Mechanical and Automation Engineering
The Chinese Univ. of Hong Kong, Shatin, N.T., Hong Kong
Email: jwang@mae.cuhk.edu.hk

***Abstract*— Recently, a continuous-time k-winners-take-all (kWTA) network with a single state variable and a hard-limiting activation function and its discrete-time counterpart were developed. These kWTA networks have proven properties of finite-time global convergence and simple architectures. In this paper, the kWTA networks are applied for information retrieval, such as web search. The weights or scores of pages in two real-world data sets are calculated with the PageRank algorithm, based on which experimental results of kWTA networks are provided. The results show that the kWTA networks converge faster as the size of the problem grows, which renders them as a promising approach to large-scale data set information retrieval problems.**

I. Introduction

The techniques for information retrieval from large data sets play a very important role as the size of the world-wide web exceeded 800 million pages in 1999 [1] to 11.5 billion in 2005 [2], and possibly more than 30 billion nowadays. 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 importance of its parent pages [3] [4]. This led to many impressive works in the past decade, such as analyzing the efficiency [6], doing computational experiments [7] [8], improving the efficiency and effectiveness [9] [10] and further analysis on social networks [11] [12]. It has been pointed out that because of the large eigengap of the modified adjacency matrix, the values of the PageRank eigenvector are fast to approximate, which indicates that only a few iterations are needed [7]. As a result, 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. [13] [14] [15] [16] [17] [18].

As a generalization of the winner-take-all (WTA) operation, the k-winners-take-all (kWTA) is to select the k largest inputs out of n inputs $(1 \leq k < n)$ [19]. kWTA has been shown to be a computationally powerful operation compared with standard neural network models of threshold logic gates [20], and has been widely used in various applications, such as decoding [21], feature extraction [22], signal processing [23] [24], etc.

When the number of inputs is large or the selection process has to be operated in real time, parallel algorithms and hardware implementation are desirable. Many kWTA networks have been proposed during the past two decades; e.g. [25][26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39]. In particular, a kWTA model with Heaviside step activation function was developed [39], where its global stability and finite-time convergence are proven with derived upper and lower bounds. Its architectural complexity is almost (if not yet) the simplest due to its single state variable and single design parameter.

In this paper, the single-state kWTA network is adopted to accelerate the response speed of large-scale network search engines. Experimental results on two data sets will be shown, which indicates the potential efficiency by involving the kWTA network into large-scale network search engines.

The remainder of the paper is organized as follows. The related preliminaries and model descriptions are presented in Section II. Experimental results are presented in Section III. Finally, Section IV concludes the paper.

II. Preliminaries

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 find out the most "wanted" k results with higher weighting coefficients and show them in a very short time. In this work, the traditional well known PageRank algorithm is used for the first part and then the proposed kWTA network is implemented to solve the second part.

We assume that after each crawl of the web, the ranking vector is computed only once, and the values can then be used to influence the ranking of search results [5], which indicate that PageRank algorithm do not need to be run most of the time when there happens to be a searching request. Moreover, many impressive works has been done to accelerate the calculation of PageRank. But for the second part, it is never easy to find out the 10 or 20 pages with biggest weights among millions or even billions of pages in a very short time. The growth of the size of internet has been far more rapidly than that of

The work described in the paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Grants CUHK417209E and CUHK417608E.

978-1-4244-9474-3/11/$26.00 ©2011 IEEE

computing speed in the past several decades and will remain in the future. The proposed kWTA network [39] has a simplest architecture which makes it easy to implement on hardware. Digital circuits is also designable based on its discrete-time counterpart [36]. It has been pointed out that as the problem size n increases, either the average convergence time of the state variable or the average number of iterations needed for the of state variable to converge decreases [37]. As a result, for large-scale dataset information retrieval, kWTA network would perform better than current methods.

1) The PageRank Algorithm: The essential idea of the PageRank algorithm is if page j has a link to page l, then the author of j is implicitly conferring some importance to page l [4]. PageRank of j represents its importance and thus k confers a importance of $1/o_j$ to l, where o_j is the outdegree of page j. This simple idea leads to the following fixpoint computation that yields the rank vector r over all of the pages on the web. Let n be the number of pages and assign all pages with the initial rank value $1/n$. Let $\mathcal{S}_l$ represent the set of pages pointing to l. In each iteration i, the ranks r_i are propagate as follows:

$$\forall l, r_{i+1}(l) = \frac{1-p}{n} + \sum_{j \in \mathcal{S}_l} p \cdot r_i(j)/o_j \tag{1}$$

where parameter $p < 1$ is usually included as a dump coefficient is to make the Markov chain regular.

In PageRank algorithm, for the pages that has no links towards any other pages, we assume them linking to all other ones. The iterations continue until Rank stabilizes to within some threshold, and the final r is the PageRank vector over the web. This r then plays the role of the input vector in kWTA network, and thus it is easy to find out the most wanted k pages from the output.

2) The kWTA Network: Generally speaking, the kWTA operation can be defined by the following binary function

$$x_i = f(u_i) = \begin{cases} 1, & \text{if } u_i \in \{k \text{ largest elements of } u\}, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $u = (u_1, u_2, \ldots, u_n)^T$ is the input vector and $x = (x_1, x_2, \ldots, x_n)^T$ is the output vector.

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, which turns out to be a k-Winners-take-all problem.

A kWTA network with a single state variable was developed [39] with the following equations:

- State equation

$$\epsilon \frac{dy}{dt} = \sum_{i=1}^{n} x_i - k, \tag{3}$$

- Output equation

$$x_i = g(r_i - y), \quad i = 1, \cdots, n, \tag{4}$$

where $y \in \Re$ is the state variable, r_i are the weights from PageRank algorithm, x is the output vector of kWTA network with its elements being 0 or 1, and $g(\cdot)$ is the Heaviside step activation function defined as:

$$g(z_i) = \begin{cases} 0, & z_i \leq 0, \\ 1, & z_i > 0. \end{cases} \tag{5}$$

A discrete-time counterpart of the kWTA model is as follows:

- State equation

$$y(t+1) = y(t) + \beta \left(\sum_{i=1}^{n} x_i(t) - k \right), \tag{6}$$

- Output equation

$$x(t) = g(r - ey(t)), \tag{7}$$

where $\beta > 0$ is the step size, e is the unit vector, and integer t represents the step number.

To guarantee the globally convergent, β should also be no bigger than the minimum difference of inputs $\min_{1 \leq i,j \leq n} (r_i - r_j)$. For web searching, usually the minimum resolution of r_i is zero. It has also been pointed out that for such case, a gradual value reduction of β will always work well out any persistent oscillation and reach convergence eventually.

The global stability and finite-time convergence of the kWTA networks (3)(4)(6)(7) are proved with derived upper and lower bounds in [39].

III. Experimental results

In order to show the performance of the kWTA network in solving network searching problems, experimental results on several real world data sets are stated in this section.

The first data set is a tiny but very typical one with only 7 pages and 17 links between them as Figure 1 shows. The PageRank weight of each page and link is also provided after running the algorithm with $p = 1$, which indicates no dump coefficient is added in this toy example.

The transient behaviors of $r_i - y$ of the discrete-time kWTA network in (6) and (7) is depicted in Fig. 2 with $k = 3$ and $\beta = 10^{-3}$. The output vector $x = g(r - ey) = [1, 1, 0, 0, 1, 0, 0]^T$ indicates that Page $1, 2$, and 5 are the ones with higher PageRank weight, which is obviously correct for this example in Figure 1.

The second example is a heterogeneous film-director-actor-writer network, which is crawled from Wikipedia under the category of English-language films in [12], where there are $34,279$ pages in total with $142,426$ relationships between the heterogeneous nodes. Figure 3 shows part of the square adjacency matrix of example 2, where a dot on the ith column and the jth row represents that there is a directed link pointed to the jth page from the ith one. The visible part in the following figure is a rectangle one because there is no input links for some pages, and as a result elements in part of this adjacency matrix is 0.

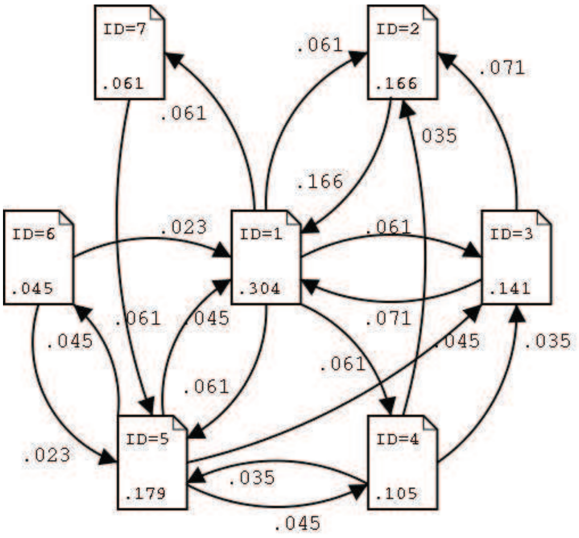

Fig. 1. Global view of the links and pages of data set 1 (adapted from PageRank - Wikipedia, http://en.wikipedia.org/wiki/File:Linkstruct3.svg).

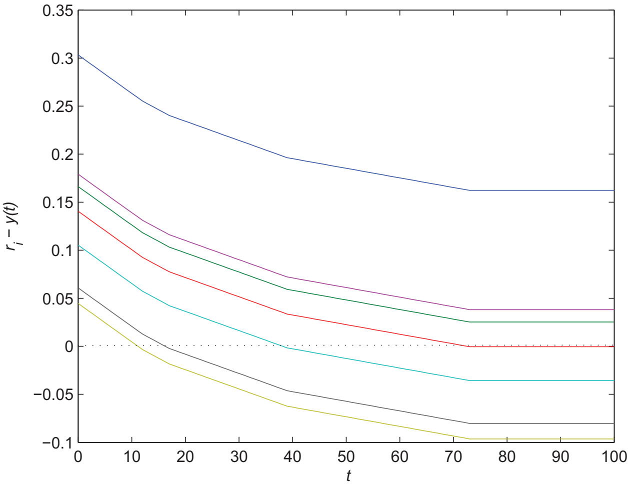

Fig. 2. Transient behaviors of $r - ey(t)$ in the discrete-time kWTA network for example 1, where $k = 3$, $\beta = 10^{-3}$, and $y(1) = 0$.

PageRank algorithm has been run under the parameter of $p = 0.85$, and Fig. 4 shows the transient behaviors of $r_i - y$ of the discrete-time kWTA network in (6) and (7) for example 2. As there are too many unimportant pages of which stable values of the transient curves are below 0, only 100 randomly selected ones are drawn in blue lines in Fig. 4. It is obvious that only the 10 red lines converges to a positive value, which indicates that the kWTA network has automatically "choose" 10 pages. The answer to this query $[3111, 3869, 4058, 4621, 6938, 8974, 10341, 11502, 13320, 15326]^T$ can be easily achieved from the sparse representation of the output vector $x = g(r - ey(t))$, where 10 of the elements are nonzero, and its correctness has been verified manually.

Although only software simulation has been done in this work, which take longer simulation time than current network

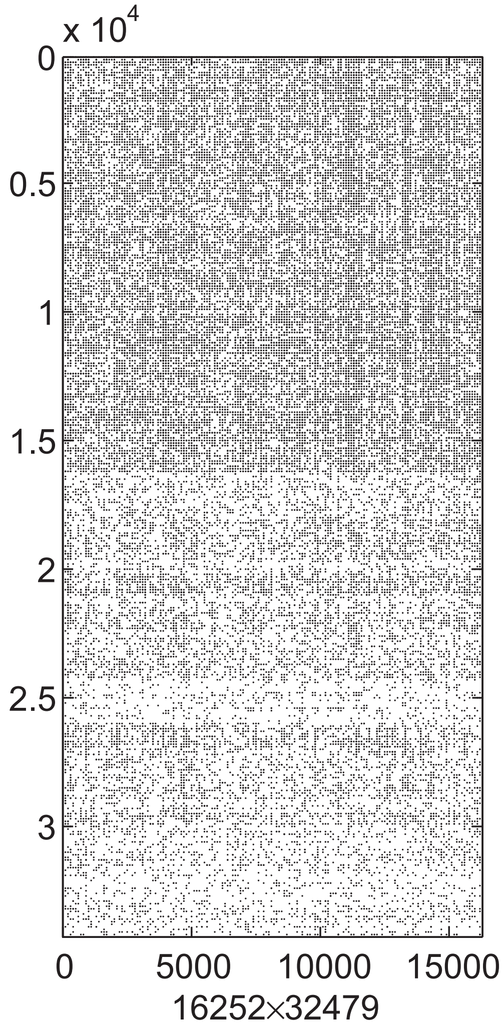

Fig. 3. Global view of the adjacency matrix for pages in example 2.

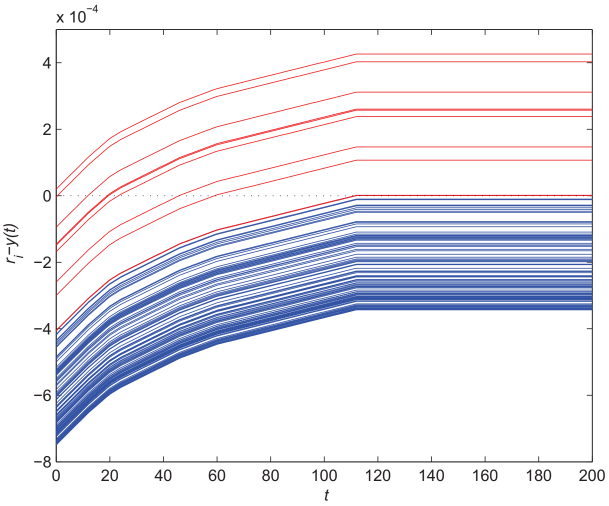

Fig. 4. Transient behaviors of $r - ey(t)$ in the discrete-time kWTA network for example 2, where $k = 10$, $\beta = 10^{-6}$, and $y(1) = 10^{-3}$.

searching algorithms, it can be predicted that with hardware implementation, kWTA network would perform much better when facing larger-scale data set than current algorithms.

IV. Conclusions

In this paper, the kWTA networks are applied for large-scale data set information retrieval. Experimental results based on two real world data sets are shown after running the PageRank algorithm. The proved superior performance of the kWTA network is also demonstrated by the simulation results. A most important characteristic of the model shown by experiment to the second example is that kWTA network converge fast enough when facing a large-scale problem, which indicates them as promising keys to web site information retrieval problems.

References

[1] S. Lawrence and C. L. Giles, “Accessibility of information on the web,” *Nature*, vol. 400, pp. 107-109, 1999.

[2] A. Gulli and A. Signorini, “The indexable web is more than 11.5 billion pages,” *Proceedings of 14th International Conference on World Wide Web*, Special interest tracks and posters, 2005.

[3] 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.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: bringing order to the web,” Technical report, Stanford University, 1998.

[5] The Google Search Engine: Commercial search engine founded by the originators of PageRank. Located at http://www.google.com/.

[6] T. H. Haveliwala, “Efficient computation of PageRank,” Stanford Univ. Technical Report, 1999.

[7] T. H. Haveliwala and S. Kamvar, “The second eigenvalue of the google matrix,” Stanford Univ. Technical Report, 2003.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] G. N. Frederickson and D. B. Johnson, “Generalized selection and ranking,” in *Proc. of the 12th STOC*, pp. 420-428, 1980.

[14] M. Kendall and J. D. Gibbons, “Rank correlation methods,” *Edward Arnold*, London, 1990.

[15] R. Fagin, R. Kumar, and D. Sivakumar, “Comparing top-k lists,” *SIAM J. Discrete Math*, vol. 17, pp. 134-160, 2003.

[16] 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.

[17] P. Hall and M. G. Schinek, “Inference for the top-k rank list problem,” in *Proceedings in Computational Statistics*, pp. 433-444, 2008.

[18] K. Henderson and T. Eliassi-Rad, “Solving the top-k problem with fixed-memory heuristic search,” Technical report, Lawrence Livermore National Laboratory, 2009.

[19] 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.

[20] W. Maass, “On the computational power of winner-take-all,” *Neural Comput.*, vol. 12, pp. 2519-2535, 2000.

[21] R. Erlanson and Y. Abu-Mostafa, “Analog neural networks as decoders,” *Advances in Neural Information Processing Systems*, vol. 1, pp. 585-588, 1991.

[22] 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.

[23] 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.

[24] A. K. J. Hertz, and R. G. Palmer, “Introduction to the Theory of Neural Computation,” Redwood City, CA: Addison-Wesley, 1991.

[25] W. Wolfe, D. Mathis, C. Anderson, J. Rothman, M. Gottler, G. Brady, R. Walker, G. Duane, and G. Algghband, “K-winner networks,” *IEEE Trans. Neural Networks*, vol. 2, pp. 310-315, 1991.

[26] J. Wang, “Analogue winner-take-all neural networks for determining maximum and minimum signals,” *Int. J. Electron.*, vol. 77, no. 3, pp. 355-367, 1994.

[27] K. Urahama, and T. Nagao, “K-winners-take-all circuit with o(n) complexity,” *IEEE Transactions on Neural Networks*, vol. 6, no. 3, pp. 776-778, 1995.

[28] J. Yen, J. Guo, and H. Chen, “A new k-winners-take-all neural network and its array architecture,” *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 901-912, 1998.

[29] B. Sekerkiran and U. Cilingiroglu, “A CMOS k-winners-take-all circuit with O(N) complexity,” *IEEE Trans. Circuits and Systems*, vol. 46, pp. 1-5, 1999.

[30] B. A. Calvert and C. Marinov, “Another k-winners-take-all analog neural network. IEEE Trans. Neural Netw., 11, 829-838 (2000)

[31] Marinov, C. and Calvert, B.: “Performance analysis for a k-winners-take-all analog neural network: basic theory,” *IEEE Transactions on Neural Networks*, vol. 14, no. 4, pp. 766-780, 2003.

[32] C. A. Marinov and J. J. Hopfield, “Stable computational dynamics for a class of circuits with O(N) interconnections capable of KWTA and rank extractions,” *IEEE Trans. Circuits Syst. I*, vol. 52, pp. 949-959, 2005.

[33] S. Liu and J. Wang, “A simplified dual neural network for quadratic programming with its KWTA application,” *IEEE Trans. Neural Netw.*, vol. 17, no. 6, pp. 1500-1510, 2006.

[34] Q. Liu and J. Wang, “Two k-winners-take-all networks with discontinuous activation functions,” *Neural Networks*, vol. 21, no. 2-3, pp. 406-413, 2008.

[35] X. Hu and J. Wang, “An improved dual neural network for solving a class of quadratic programming problems and its k-winners-take-all application,” *IEEE Trans. Neural Networks*, vol. 19, pp. 2022-2031, 2008.

[36] Q. Liu, J. Cao, and J. Liang, “A discrete-time recurrent neural network with one neuron for k-winners-take-all operation,” in: *Proc. of 6th Intl. Sym. on Neural Networks*, Springer LNCS, vol. 5551, pp. 272-278, 2009.

[37] J. Wang, and Z. Guo, “Parametric Sensitivity and Scalability of k-Winners-take-all Networks,” in *Proc. of 7th Intl. Sym. on Neural Networks*, Springer LNCS, vol. 6063, pp. 77-85, 2010.

[38] Q. Liu, C. Dang, J. Cao, “A novel recurrent neural network with one neuron and finite-time convergence for k-winners-take-all operation,” *IEEE Trans on Neural Networks*, vol. 21, pp. 1140-1148, 2010.

[39] J. Wang, “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*, vol. 21, pp. 1496-1506, 2010.