Adaptive Page Ranking with Neural Networks

Franco Scarselli
University of Siena
Siena, Italy
franco@ing.unisi.it

Sweah Liang Yong
University of Wollongong
Wollongong, Australia
sly56@uow.edu.au

Markus Hagenbuchner
University of Wollongong
Wollongong, Australia
markus@uow.edu.au

Ah Chung Tsoi
Australian Research Council
Canberra, Australia
ahchung.tsoi@arc.gov.au

ABSTRACT

Recent developments in the area of neural networks provided new models which are capable of processing general types of graph structures. Neural networks are well-known for their generalization capabilities. This paper explores the idea of applying a novel neural network model to a web graph to compute an adaptive ranking of pages. Some early experimental results indicate that the new neural network models generalize exceptionally well when trained on a relatively small number of pages.

Categories and Subject Descriptors: E.1 Data Structures: Graphs and networks. H.3.3.b Information Search and Retrieval: Information Filtering. I.2.6 Learning: Connectionism and Neural Nets.

General Terms: Algorithm, Experimentation

Keywords: Adaptive Page Rank, Neural Networks, Graph Processing

1. INTRODUCTION

The hyperlink information of the world wide web is often exploited for ranking purposes. Ranking algorithms are used by search engines to sort URLs according to their relevance to user queries. The most common approach is PageRank [1]. It considers a link from page *p* to page *q* as an endorsement of the quality of page *q*. Formally the PageRank (PR_n) of a page n is defined as:

$$PR_n = d \sum_{u \in pa[n]} \frac{PR_u}{h_u} + (1-d)\,, \tag{1}$$

where $pa[n]$ is the set of parents of page n, h_u is the outdegree of page u, and $d \in [0,1]$ is a damping factor [1].

This approach is global in the sense that all pages in the web are treated equally. However, from a user's point of view, the ranking may not be optimal. For example, when one is searching for "Amazon", one might expect to see the results of "Amazon" the online bookstore, "Amazon" the river in South America or the name of a local theme park. A user may be more interested in pages on the local theme park than the online bookstore, or geography, however, pages on the local theme park may be ranked much lower than those concerning the online bookstore, or geography using Equation 1.

Much work was performed on how to modify the PageRank method [1] to reflect the interests of the user. This is sometimes referred to as "adaptive ranking" of web pages. Existing methods may be classified as: **(A)** Bias PageRank with a constraint vector (e.g. [5]), **(B)** Use web logs or past users behaviors (e.g. [2]), **(C)** Cluster users into different demographic group with precomputed biased PageRank (e.g. [3]).

Little work has been done on employing neural network models for adaptive page ranks. This is mainly due to a lack of established neural network models capable of graph processing.

Recent developments saw the introduction of a new class of neural network models, called *Graph Neural Networks* (GNN), which is capable of processing general types of graphs [4]. GNNs can be used for classification of web pages. In fact, the web can be represented as a graph where nodes embody pages and arcs denote hyperlinks. Nodes may have numeric labels which contain an encoding of the page content. For example, we used as a node label the classification of a page using the Bayes classifier. The function implemented by a GNN is learned from a set of positive and negative examples.

In this paper, we present a specialized version of GNN which is particularly suited for web page ranking problems. A GNN attaches to each page n a vector $\boldsymbol{x}_n \in \mathbb{R}^s$, called the *state*, which collects a representation of the page. The state $\boldsymbol{x}_n$ is defined as the solution of a system of equations:

$$\boldsymbol{x}_n = \sum_{u \in \boldsymbol{pa}[n]} \boldsymbol{A}_{n,u}\boldsymbol{x}_u + \boldsymbol{b}_n \tag{2}$$

where $\boldsymbol{pa}[n]$ is the set of parents of n. For each page n, the rank $r_n \in \mathbb{R}$ is defined using the solution $\boldsymbol{x}_n$ of Equation 2:

$$r_n = c_n^T \boldsymbol{x}_n \tag{3}$$

The superscript T denotes the transpose of a vector. Here, the vectors $c_n \in \mathbb{R}^s$, $\boldsymbol{b}_n \in \mathbb{R}^s$ and the matrix $\boldsymbol{A}_{n,u} \in \mathbb{R}^{s \times s}$ are parameters defined by the outputs of three respective multilayered feedforward neural networks. Those networks compute the parameters using the labels, i.e. the page content. For example, $\boldsymbol{b}_n = \rho(\boldsymbol{l}_n)$ where ρ is the function implemented by a multilayered feedforward neural network and $\boldsymbol{l}_n$ the label of page n.

Note that Equations (2) and (3) implicitly assume that the rank of page n depends on its content and on the pages that have a hyperlink pointing to n. Such an assumption is similar to the one adopted in Google's PageRank [1], with the difference that PageRank considers only the web connectivity and not their page contents. The rank defined by Equations (2) and (3) can be considered as a parameterized version of PageRank. The parameters may be obtained by the minimization of an error function which represents a set of desired constraints. The constraints may be in the form of $r_n = t_n$ (rank of page n should be t_n, where t_n is a given value) or in the form $r_n \geq r_u$ (page n should have a higher rank than page u) and should be produced by an interface that captures the desires of a search engine administrator.

2. EXPERIMENT RESULTS

We evaluated the GNN model on the WT10G dataset distributed by CSIRO, Australia. This dataset is a snapshot of a portion of

Copyright is held by the author/owner.
WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

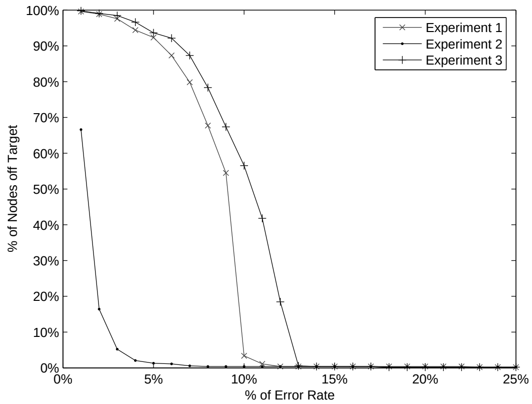

Figure 1: Results of GNN trained on second set of experiments

the World Wide Web which pays special attention to the connectivity between the web pages. There are 1,692,096 pages in this dataset. In the experiments, we select a subgraph G of WT10G as the training set. This training set [1] consists of both positive and negative examples; positive examples are a small number of pages which have been manually labelled as belonging to certain topic of interest to the user, e.g., "Sports", "Surgery"; while the negative examples are randomly selected from other topics which are unrelated to the topic of interest. The test dataset in all experiments conducted consists of the entire WT10G dataset. In all the experiments GNN is trained with 5 hidden neurons, 1 neuron for the state and 1 output neuron.

2.1 Hard Target Experiments

In the first set of experiments, we use the PageRank [1] to generate the targets. The aim here is to double the rank of pages associated with "Sports" on the entire WT10G dataset, while leaving the other pages unrelated to "Sports" close to their PageRanks. In this case, we chose 20 pages on the topic "Sports", assigned their ranks as $2 \times PR$, where PR is the PageRank as computed using Equation (1). We chose an additional 3980 pages randomly from pages which are unrelated to "Sports" to make up the training dataset. It is observed from Table 1 that the GNN generalizes very well from the small training dataset to the entire WT10G dataset.

	Experiment 1	Experiment 2	Experiment 3
on target	99.30%	99.58%	99.94%

Table 1: A node is considered on target with $\pm 5\%$ error from the target value. All 3 experiments are conducted on the same topic "Sports", but with different selection of pages in the training dataset, while keeping other training parameters constant.

The second set of experiments were conducted with the aim of imposing the constraints that pages related to either "Sports" or "Surgery" must double their respective PageRank values, while pages related to both topics "Sports" and "Surgery" as well as all other pages unrelated to either of these topics have their rank values the same as those provided by Equation (1).

The results are shown in Figure 1. Experiments 1 and 3 appear to have larger errors than Experiment 2. However, results of Experiment 1 and 3 need to be interpreted cautiously. A deeper analysis has shown that the pages with the largest errors are those having the smallest ranks. In fact, those pages have a smaller effect on the error function and are not important becuase they usually are not returned on user queries.

2.2 Soft Target Experiments

In this experiment, our aim is to increase the rank of all pages in WT10G dataset related to the topic "Sports" relative to those pages related to "Surgery". In other words, we do not specify quantitatively by how much pages on "Sports" must be relative to those on the topic "Surgery". The results are shown in Figure 2.

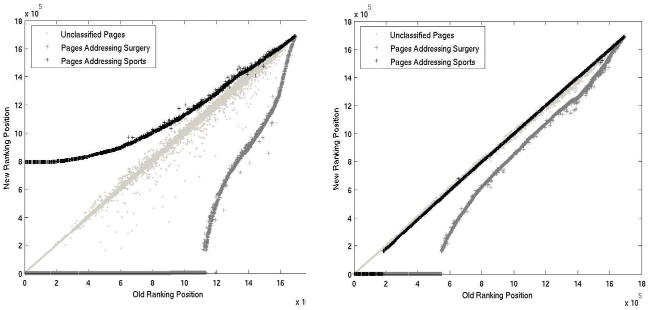

Figure 2: The result of original and current position of page using soft target constraints.

These experiments show that GNN can be used to bias PageRank in favor of the topic of interest. A sportsman who search for "pain killer" would obtain the results of medication for sports rather then medication for surgery.

It is observed from the right hand graph of Figure 2 that GNN may produce alternative solutions. Instead of raising the rank of pages related to "Sports" to higher than the PageRank as in Equation (1), it forces the rank of those related to "Surgery" lower than "Sports" while maintaining the rank of pages related to "Sports" close to their PageRank values.

3. CONCLUSIONS

It has been demonstrated through some simple experiments that the GNN is capable of generalizing from a relatively small number of training examples to the entire WT10G dataset. These preliminary results are very encouraging and motivate us for a more thorough investigation into its practical applications to the World Wide Web. A particular focus of future research is on overcoming issues with the computational complexity of the proposed approach such that an application to the World Wide Web becomes possible.

4. ACKNOWLEDGMENTS

The second and third author wish to acknowledge financial support from the Australian Research Council in the form of a Discovery Project grant.

5. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[2] L. Fitzpatrick and M. Dent. Automatic feedback using past queries: Social searching? In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 306–313. ACM Press, 1997.

[3] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the eleventh international conference on World Wide Web*, pages 517–526. ACM Press, 2002.

[4] F. Scarselli, A. C. Tsoi, M. Gori, and M. Hagenbuchner. A new neural network model for graph processing. Tech. rep., DII 1/05, University of Siena, 2005 Aug. 2004.

[5] A. C. Tsoi, G. Morini, F. Scarselli, M. Hagenbuchner, and M. Maggini. Adaptive ranking of web pages. In *Proceedings of the twelfth international conference on World Wide Web*, pages 356–365. ACM Press, 2003.

[1] The experiments shown use just 4000 pages as a training set.