INFORMATION PAGE OF DOCTOR THESIS

Subject: **Mining Sequence Data For Webpage Access Prediction**

Code: **9.48.01.04**

Ph.D. Candidate: **Nguyen Thon Da**

Supervisor 1: **Tan Hanh, Ph.D**

Supervisor 2: **Nguyen Hoang Duy, Ph.D**

Ph.D. Academic institute: **Vietnam Posts and Telecommunications Institute of Technology**.

THESIS CONTRIBUTIONS

- **The first contribution:** Propose a solution in order to design and build sequence databases for Webpage access prediction. Four datasets collecting from real-time Weblog files consist of periwinklelecottages.com, palmviewsanibel.com, devqa.robotec.co.il, inees.org. The problem desired to address is that how to create a sequence database from a collection of Weblog files. The main idea of the above approach is described as follows: In the Weblog dataset, find an array containing different user IPs and an array containing different visited links. With regard to every different User IP, there is a group of distinct visited links in time order. These groups are sequences in the desired sequence database. Furthermore, by analyzing the characteristics of Weblog data such as access time, access links, the thesis introduces how to transform Weblog data into sequence databases by means of a parallel algorithm and a non-parallel algorithm.

- **The 2nd contribution:** Propose a solution in order to reduce the prediction time for Webpage access prediction. Five sequence databases are used to perform. They include two sequence databases created from Weblog datasets (collected palmviewsanibel.com and inees.org) and three sequence databases collected from click-stream datasets like KOSARAK, FIFA and MSNBC. The problem desired to address is that how to predict a page next to a given sequence *S* in a given sequence database with a good time performance. To deal with this issue, the thesis proposes five major steps: (i) Input a sequence database *SDB* and a sequence *S*; (ii) Discard sequences in SDB that do not contain elements of the sequence S. With regard to sequences in SDB that contain elements of the sequence S, remove sequences in SDB that only contains elements of the sequence S at the

last position of them. This solution will reduce the size of the original sequence database. Rely on this solution, the prediction time of new sequence database (reduced-size database) is faster than that of original sequence database (non-reduced-size database). With respect to datasets collect from Weblog files, experimental result on palmviewsanibel.com dataset shows that the prediction time of the proposed model is up to 2.7 times faster than that of typical model without accuracy. Similarly, on inees.org dataset shows that the prediction time of the proposed model is nearly 2 times faster than that of typical model without accuracy. With regard to click-stream datasets, experimental result on the FIFA dataset shows that the prediction time of the proposed model is up to 3 times faster than that of typical model without accuracy. In a similar way, on the KOSARAK dataset (up to 30 times), on the MSNBC dataset (up to 103 times). Therefore, making the prediction on click-stream datasets is much more effective than that on datasets collected from Weblog files.

- **The 3rd contribution:** Propose a solution in order to increase the accuracy for Webpage access prediction. The thesis uses three sequence databases to perform this solution. They are sequence databases collected from click-stream datasets: KOSARAK, FIFA and MSNBC. Based on PageRank's properties and the CPT, the problem desired to address is that how to find a page next to a given sequence S in a given sequence database with a good solution in terms of accuracy. The thesis proposes 5 core steps to solve this problem include: (i) Input a sequence database, (ii) Convert links into nodes for a graph database, (iii) Calculate PageRank for each node, (iv) Calculate average of PageRank for each sequence, (v) Get rid of sequences that have low average of PageRank provided that the accuracy of reduced-size sequence database is still higher than that of original sequence database (has not reduced yet). Experimental results indicates that the proposed solution offers a generally higher accuracy than the common approach from the literature while also being more consistent across the various datasets. In particularly, on the MSNBC dataset, when reducing the size of the original sequence database (removed sequences that have the low average of PageRank) up to 50%, the accuracy increased up to 25%; on the FIFA dataset, when reducing the size of the original sequence database up to 15%, the accuracy increased up to 0.013%; and on the KOSARAK dataset, when reducing the size of the original sequence database up to 30%, the accuracy increased up to 0.027%.

- **The 4th contribution:** Propose a model that combines between increasing the accuracy and reducing the prediction time. The largest sequence database (KOSARAK) is used for performing this solution. By means of using cross check algorithm K-Folder-Validation (with K = 10), the KOSARAK dataset was separated into 10 random parts. Every part includes 90% data used for training and 10% remain data used for testing (predicting). Experimental results proved that when reducing the size of the original sequence database up to 34% (using the 3rd contribution), the average accuracy of proposed solution still remain better than that of common approach. Next, using 66% the size of original sequence database (removed bad and redundant data by PageRank algorithm) to predict by means of the solution introduced in the 2nd contribution, the thesis indicates that the average accuracy increased up to 0.0621% and the average prediction time of proposed model is 80 times faster than that of common model.

APPLICATION AND FUTURE WORK

- Discover more deeply about sequence prediction to develop novel algorithms aiming to solve better issues related to Web page access prediction.
- Besides, significant challenges of Big Data include capturing data, data storage, data analysis, search, sharing, transfer, visualization and so on. Therefore, Big Data often includes data with sizes that exceed the capacity of common software. Because of this, sequence prediction on Big Data is still an open issue and give us many big problems to be solved. In the future, the improvement of the thesis is that how to address effectively sequence prediction issue on Big Data in terms of time execution and accuracy.

Supervisor 1 **Candidate Ph.D**

Tan Hanh Ph.D **Nguyen Thon Da**

Supervisor 2

Pham Hoang Duy Ph.D