International Journal of Distributed Sensor Networks, 5: 89, 2009
Copyright © Taylor & Francis Group, LLC
ISSN: 1550-1329 print / 1550-1477 online
DOI: 10.1080/15501320802575062

A Markov Prediction Model Based on Page Hierarchical Clustering

YAO YAO[1,2], LEI SHI[1,2], and ZHANHONG WANG[3]

[1]Henan Provincial Key Lab on Information Network, Zhengzhou, China
[2]School of Information Engineering, Zhengzhou University, Zhengzhou, China
[3]Department of Computer Science of Xinyang Normal University, Xinyang, China

The Markov prediction model is the basis of Web prefetching and personalized recommendation. It can be used to extract connotative Web link hierarchy. The visualized site structure can not only help users understand the relationships between the pages they have visited, but also suggest where they can go next. But the existence of a large amount of Web objects results in data redundancy and model hugeness. Therefore, how to mine and improve the link structure of a website has become a chief problem and it has positive meanings for prefetching.

This paper presents an improved method that simplifies the topology structure of a website and extracted the conceptual link hierarchy which can make the organization clearly and legibly. First, the Markov Tree is constructed for the reason that a more capable mechanism for representing past activity in a form usable for prediction is a Markov Tree. In this case the Markov chain model can be defined as a three-tuple (A, S, P), where A is the collection of operation, S is the state space consisting of all the states in a link structure, and P is the one-step transition probability matrix. The transition probability matrix is calculated based on the Markov tree. Second, an algorithm is given to extract the hierarchical tree from the above matrix. The website link hierarchy (WLH) is obtained accordingly. A WLH only contains a trunk link which is a hyperlink from a page on a higher conceptual level to a page on its adjacent lower conceptual level. With the levels increment, there must be more and more pages in each level. It may blur the structure of the website. In order to tackle the problem, a clustering algorithm is proposed to cluster conceptually-related pages on same levels based on their in-link and out-link similarities, which are measured by the concept of weighted Euclidean distance. After the pages in WLH have been clustered, WLC can be constructed. Finally, the simplified model will be used for Web page prediction. Three parameters, i.e. precision, recall, and PRS have been employed to measure the performance in the experiments. Experiments based on two real Web log data demonstrate the efficiency of the proposed method, which can not only have good overall performance and clustering effect but also keep the relative higher prediction accuracy and recall.

Keywords Markov prediction model; Website link hierarchy structure (WLH); Website conceptual link hierarchy (WLC); Link similarity; clustering

Address correspondence to Lei Shi, Zhengzhou University, School of Information Engineering, No. 100 of Science Road, Zhengzhou, Henan, 450001, China. E-mail: shilei@zzu.edu.cn

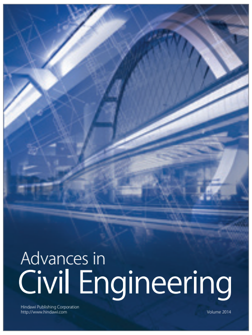
Journal of Engineering

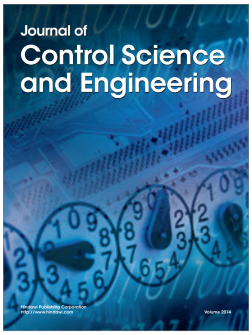
The Scientific World Journal

International Journal of Rotating Machinery

Journal of Sensors

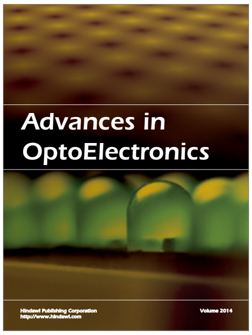
International Journal of Distributed Sensor Networks


Advances in Civil Engineering


Journal of Control Science and Engineering

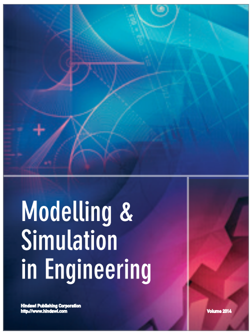
Journal of Robotics

Journal of Electrical and Computer Engineering

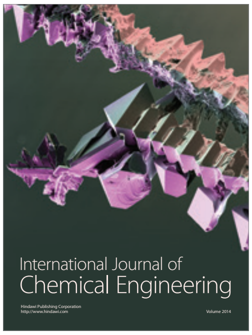

Advances in OptoElectronics

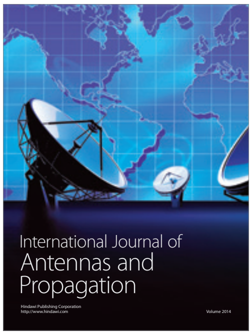
VLSI Design

International Journal of Navigation and Observation

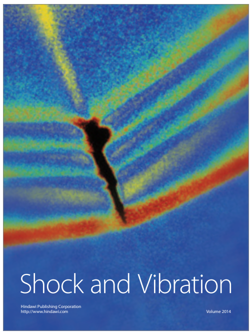

Modelling & Simulation in Engineering

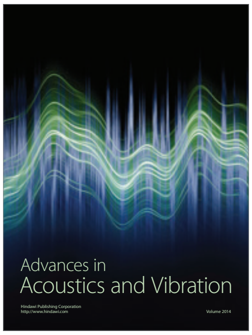
International Journal of Aerospace Engineering


International Journal of Chemical Engineering


International Journal of Antennas and Propagation

Active and Passive Electronic Components


Shock and Vibration


Advances in Acoustics and Vibration