International Journal of Foundations of Computer Science

Editorial Notes on the Special Issue on "Mining the Web"

Charles X. Ling
Department of Computer Science, The University of Western Ontario
London, Ontario N6A 5B7, Canada
ling@csd.uwo.ca

Nick Cercone
Department of Computer Science, University of Waterloo
Waterloo, Ontario N2L 3G1, Canada
ncercone@uwaterloo.ca

Received (received date)
Revised (revised date)
Communicated by Editor's name

1. Summary

Data mining is often defined as the discovery of high-level knowledge from massive datasets. The World Wide Web (WWW) is massive comprising billions of web pages, gigabytes of user logs, and millions of daily data transitions and email exchanges. Therefore, data mining applied to the WWW, or Web Mining, is an important and fascinating area for research and applications. The purpose of this Special Issue is to advance our knowledge and techniques of data mining applied to the web.

Knowledge discovery is more than simple data collection and statistical analyses. Data mining emphasizes the process of learning, with machine learning algorithms for performance improvement. Thus, deploying a crawler to obtain all web pages containing the phrase "call for papers" is not considered as "Mining the Web" by our definition, while an intelligent crawler that learns to become more expert in finding web pages containing "call for papers" (see the paper by Lorincz et. al.) satisfies our definition. The papers published in this special issue all have explicit components of learning and self-improvement.

From the 12 submitted papers, 5 were selected to be included in this special issue, after careful reviews were performed and revisions were made. We hope that this Special Issue advances research and knowledge on issues concerned with "Mining the Web".

The first paper, by Lorincz et. al., reports on aspects about intelligent crawlers (or focussed crawlers) for specific topics and special interests. Searching databases collected from this kind of crawlers will, presumably, return more relevant information much faster than traditional search engines such as Google and Altavista. However, how does an intelligent crawler find all the relevant web pages without crawling through and downloading all (or most) web pages on the Internet, as does Google? The central idea of the paper is to **learn** to recognize if the crawler is moving closer to target pages or not, and if it is not, to eliminate crawling over and downloading those pages.

Their system first uses previously classified web documents to train classifiers (several Support Vector Machines) to assign scores to web documents for their interestingness, or how close they are to the target pages. The system also learns weights for combining several SVMs using reinforcement learning during crawling. Thus, the system becomes better, by reinforcement, upon downloading pages that are actually promising because reinforcement improves crawling decision making.

Lorincz et. al.'s experiments demonstrate that the hit ratio for the adaptive crawlers is much higher than that of a focussed crawler without on-line learning (adaptation).

Estivill-Castro and Yang's work reports novel clustering algorithms applied to the visit paths of web visitors. The goal is to cluster web visitors into similar clusters (such as ones with similar demographic features) so web companies can understand their visitors better. Almost all previous clustering algorithms are quadratic in terms of n, the size of the training examples. Their algorithm, belonging to the family of partition clustering algorithms, finds the partition that optimizes a homogeneity measure defined in terms of some distance measure in $O(nlogn)$ time. They then discuss two versions of the algorithm, the crisp and the non-crisp ones, and prove that both versions converge.

Experiments show that the new algorithms are several orders of magnitude faster than the previous matrix-based clustering algorithm on public web logs. Simulations on artificial datasets show that the new algorithms reconstruct the clusters more accurately than do previous methods.

It should be noted that their clustering algorithms are not restricted to web paths; they can be applied to other feature-vectors that may arise in different applications.

The paper by Gao et. al. presents a web page recommendation system that learns from previous user's visit sequences, and predicts the consequent pages that a current user is likely to visit. Previous work has been done for similar problems, but their work contains several features. First, their learning algorithm is a supervised classification algorithm which aims at achieving a high predictive accuracy. Second, the system is real-time and dynamic. The current user clicks and feedbacks are also used to train the recommendation engine. Third, with careful use of the k-NN algorithm, updating training examples is effortless, yet the testing time is still within control. Overall, the system produces a good predictive accuracy and has a relatively fast response time.

As the authors indicated, if the system can distinguish the "viewed page" (that users actually want to read) and the "transitional pages" (through which users browse to obtain the viewed pages), and only recommend the viewed pages, the value of the recommendation will be further improved.

Ng and Ho's paper concerns direct marketing on the web: posting the advertisements to customers who are actually interested in them. The authors designed an Intelligent Advertisement Agent (IAA) that consists of four components. First, deriving and describing advertisement categories. Second, discovering user's interests (profile) when they browse web pages. This is done with data mining applied to the user web logs. A variation of the association learning algorithm is designed to discover the access patterns of individual users that would indicate their interests. The third component compares the advertisement categories with the discovered users' profile. Both are represented by Gaussian curves so the comparison can be done with ease. The fourth component matches advertisements to users.

Evaluations on the IAA system is carried out with human users and with some web companies. The overall predictive accuracy is about 70%; that is, about 70% of tested users are interested in the most favourable advertisement category that the system discovered.

The paper by Zhong is about Web Intelligence (WI), which integrates AI techniques and the web. Ontology, or concept hierarchy, is an important knowledge representation format in AI. It facilitates reasoning and inference, sharing of information, and semantic understanding. Therefore, ontology is useful for web knowledge representation and manipulation.

The paper describes an iterative process of automatic construction of ontologies. The major steps include text classification (supervised learning from partially labelled examples), concept relationship analysis, generation of ontology (using a variant of Hopfield networks), and refinement of ontology.

We trust and hope that the readers of this special issue will benefit from the research described herein. If merely to stimulate interest and debate, then we will have satisfied one major goal of this special collection. If readers carry out further research based on ideas that germinate from reading this collection, then we are truly blessed.