THE 18TH EUROPEAN CONFERENCE ON MACHINE LEARNING
AND
THE 11TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE
OF KNOWLEDGE DISCOVERY IN DATABASES

PROCEEDINGS OF THE GRAPH LABELLING WORKSHOP AND WEB SPAM CHALLENGE

GRAPHLAB'07

September 17, 2007

Warsaw, Poland

Editors:
Carlos Castillo
Yahoo! Research, Spain
Brian D. Davison
Lehigh University, USA
Ludovic Denoyer and Patrick Gallinari
LIP6 - University Pierre et Marie Curie, France

Preface

Topics

The workshop focus is on the **Graph labelling problem**. The goal of the graph labelling task is to automatically label the nodes of a graph (with or without content information on the nodes of the graph). The generic task has a lot of applications in many different domains: Web spam detection, Social networks,.....

The scope of this workshop is the development of new models for graph labelling and all the applications where the data can be represented as a graph :

- Generic graph labelling models
- Tree annotation models
- Models for large graph
- Web spam detection
- XML annotation
- Wiki, Blog, Web retrieval
- Web classification and clustering
- Social networks

The workshop particularly focuses on a key application which is **Web Spam detection** where the goal is to label the nodes (the Web pages or Web hosts) of a graph as spam or not spam. This workshop presents the results obtained by different research teams on the second phase of the PASCAL WebSpam Challenge[1].

The workshop has been opened to any submission concerning theoretical models or large size applications of graph labelling with a particular focus on internet graphs.

Graph Labelling

Many domains and applications are concerned with complex data composed of elementary components linked according to some structural or logical organization. These data are often described as a graph where nodes and links hold information. In Computer Vision for example, a picture can be described by a graph which corresponds to the organization of the different regions of the pictures – each node of the graph corresponding to a region. In the text domain, the diffusion of new data formats like XML and HTML has considerably changed the domains of Information Retrieval. On the Web, documents are organized according to a graph structure where nodes correspond to the Web page and edges correspond to the hyper links between the pages. Moreover, Web pages are also structured documents containing both a content information and a logical information encoded by the HTML tags, and can be viewed as labelled trees. Other application domains concerned with graph data include image processing, multimedia (video), natural language processing, social networks, biology, etc. Handling structured

[1] http://webspam.lip6.fr

data has become a main challenge for these domains and different communities have been developing for some years their own methods for dealing with structured data. The ML community should be a major actor in this area. Graph labelling which consists in labelling all the vertices of a graph from a partial labelling of the graph vertices has been identified as a generic ML problem with many fields of application.

Web Spam detection

Web spam detection is becoming a major target application for web search providers. The Web contains numerous profit-seeking ventures that are attracted by the prospect of reaching millions of users at a very low cost. There is an economic incentive for manipulating search engine's listings by creating pages that score high independently of their real merit. In practice such manipulation is widespread, and in many cases, successful.

Traditional IR methods assumed a controlled collection in which the authors of the documents being indexed and retrieved had no knowledge of the IR system and no intention of manipulating its behaviour. On Web-IR, these assumptions are no longer valid, specially when searching at global scale.

Almost every IR algorithm is prone to manipulation in its pure form. A ranking based purely on the vector space model, for instance, can be easily manipulated by inserting many keywords in the document; a ranking based purely on counting citations can be manipulated by creating many meaningless pages pointing to a target page, and so on.

Of course, ideally the search engine administrators want to stay ahead of the spammers in terms of ranking algorithms and detection methods. Fortunately, from the point of view of the search engine, the goal is just to alter the economic balance for the would-be spammer, not necessarily detecting 100% of the Web spam. If the search engine can maintain the costs for the spammers consistently above their expected gain from manipulating the ranking, it can really keep Web spam low.

The adversarial Information Retrieval on the Web (AIRWeb) series of workshops was started in 2005 by the academic community. Many existing heuristics for detection are often specific to a specific type of spam and can not be used if a new Web spam technique appears. We need to propose new models able to learn to detect any type of Web Spam and that can be adapted quickly to new unknown spam techniques. **Machine learning methods are the key to achieve this goal.**

Carlos Castillo
Brian D. Davison
Ludovic Denoyer
Patrick Gallinari

Workshop Organization

Workshop Chairs

Carlos Castillo (Yahoo! Research)
Brian D. Davison (Lehigh University)
Ludovic Denoyer (LIP6 - University Pierre et Marie Curie)
Patrick Gallinari (LIP6 - University Pierre et Marie Curie)

ECML/PKDD Workshop Chair

Marzena Kryszkiewicz (Warsaw University of Technology)

Workshop Program Committee

Kumar Chellapilla
Brian D. Davison
Ludovic Denoyer
Dennis Fetterly
Patrick Gallinari
Remi Gilleron
Marco Gori
Mark Herbster
Massimiliano Pontil
Juho Rousu
John Shawe Taylor
Alessandro Sperduti
Tanguy Urvoy

Table of Contents

Semi-Supervised Learning: A Comparative Study for Web Spam and Telephone User Churn

András A. Benczúr, Károly Csalogány, László Lukács, and Dávid Siklósi

Informatics Laboratory
Computer and Automation Research Institute
Hungarian Academy of Sciences
11 Lagymanyosi u, H-1111 Budapest
and
Eötvös University, Budapest
{benczur, cskaresz, lacko, sdavid}@ilab.sztaki.hu
http://datamining.sztaki.hu/

Abstract. We compare a wide range of semi-supervised learning techniques both for Web spam filtering and for telephone user churn classification. Semi-supervised learning has the assumption that the label of a node in a graph is similar to those of its neighbors. In this paper we measure this phenomenon both for Web spam and telco churn. We conclude that spam is often linked to spam while honest pages are linked to honest ones; similarly churn occurs in bursts in groups of a social network.

Full article in PDF

A Fast Method to Predict the Labeling of a Tree

Sergio Rojas Galeano and Mark Herbster

[1] Department of Computer Science
University College London
Gower Street, London WC1E 6BT, UK
{M.Herbster
[2] S.Rojas}@cs.ucl.ac.uk

Abstract. Given an n vertex weighted tree with (structural) diameter $S_{\mathbf{G}}$ and a set of ℓ vertices we give a method to compute the corresponding $\ell \times \ell$ Gram matrix of the pseudoinverse of the graph Laplacian in $O(n + \ell^2 S_{\mathbf{G}})$ time. We discuss the application of this method to predicting the labeling of a graph. Preliminary experimental results on a digit classification task are given.

Full article in PDF

A Semi-Supervised Approach for Web Spam Detection using Combinatorial Feature-Fusion

Ye Tian, Gary M. Weiss, and Qiang Ma

Department of Computer and Information Science
Fordham University
441 East Fordham Road
Bronx, NY 10458
{tian,gweiss,ma}@cis.fordham.edu

Abstract. This paper describes a machine learning approach for detecting web spam. Each example in this classification task corresponds to 100 web pages from a host and the task is to predict whether this collection of pages represents spam or not. This task is part of the 2007 ECML/PKDD Graph Labeling Workshop's Web Spam Challenge (track 2). Our approach begins by adding several human-engineered features constructed from the raw data. We then construct a rough classifier and use semi-supervised learning to classify the unlabelled examples provided to us. We then construct additional link-based features and incorporate them into the training process. We also employ a combinatorial feature-fusion method for "compressing" the enormous number of word-based features that are available, so that conventional machine learning algorithms can be used. Our results demonstrate the effectiveness of semi-supervised learning and the combinatorial feature-fusion method.

Full article in PDF

Web Spam Challenge 2007 Track II - Secure Computing Corporation Research

Yuchun Tang, Yuanchen He, Sven Krasser, and Paul Judge

Secure Computing Corporation
4800 North Point Parkway, Suite 300
Alpharetta, GA 30022, USA http://www.trustedsource.org

Abstract. To discriminate spam Web hosts/pages from normal ones, text-based and link-based data are provided for Web Spam Challenge Track II. Given a small part of labeled nodes (about 10%) in a Web linkage graph, the challenge is to predict other nodes' class to be spam or normal. We extract features from link-based data, and then combine them with text-based features. After feature scaling, Support Vector Machines (SVM) and Random Forests (RF) are modeled in the extremely high dimensional space with about 5 million features. Stratified 3-fold cross validation for SVM and out-of-bag estimation for RF are used to tune the modeling parameters and estimate the generalization capability. On the small corpus for Web host classification, the best F-Measure value is 75.46% and the best AUC value is 95.11%. On the large corpus for Web page classification, the best F-Measure value is 90.20% and the best AUC value is 98.92%.

Full article in PDF

Semi-supervised classification with hyperlinks

Jacob Abernethy[1] and Olivier Chapelle[2]

[1] Department of Computer Science, UC Berkeley
[2] Yahoo! Research

Full article in PDF

Webspam detection via Semi-Supervised Graph Partitioning

Chris Biemann and Hans Friedrich Witschel

{biem|witschel}@informatik.uni-leipzig.de

Full article in PDF

SpamChallenge 2007 - Track II: France Telecom RD Submissions

Pascal Filoche, Tanguy Urvoy, and Marc Boullé

France Telecom RD, Lannion, France

Full article in PDF

Author Index