doi:10.1093/comjnl/bxs023

Capsule Reviews

FAIROUZ KAMAREDDINE

The Capsule Reviews are intended to provide a short succinct review of each paper in the issue in order to bring it to a wider readership. The Capsule Reviews were compiled by Fairouz Kamareddine. Professor Kamareddine is an Associate Editor of *The Computer Journal* and is based in the Department of Mathematical and Computer Sciences at Heriot-Watt University, Edinburgh, UK.

SocLaKE: Social Latent Knowledge Explorator. GRZEGORZ KUKLA, PRZEMYSAW KAZIENKO, PIOTR BRODKA AND TOMASZ FILIPOWSKI

The authors argue that the capability to quickly find the correct answer for a specific question is crucial to every organization but that most of the information needed to answer queries remain in the workers' minds as latent knowledge. Hence, the authors develop the so-called Social Latent Knowledge Explorator (SocLaKE), which uses the organization's social networks (SNs) and the information about people's areas of expertise to create the list of recommended people who might know the answer for a given question or may know another person who knows the solution. After an introduction to the problem description and the related work, the general concept of (SocLaKE) is given along with the query propagation model and the recommendations in query propagation. The recommendation is defined as a finite subset of options available to the particular user. A set of coefficients is needed to calculate the probability of finding the right answer to the question stated in a particular community (SN) and supported by SocLaKE via recommendations. The authors give a list of coefficients that need to be estimated and how this can be done. A set of recommendation strategies has been examined and a number of simulation experiments are set up whose results are given for small and large networks.

An Improved Contextual Advertising Matching Approach based on Wikipedia Knowledge. ZONGDA WU1, GUANDONG XU, YANCHUN ZHANG, PETER DOLOG AND CHENGLANG LU

Generally, Web advertising consists of textual ads that are short text messages usually marked as sponsored links. Textual Web advertising is either sponsored search (which selects ads based on users' search queries) or contextual advertising (which selects ads based on the contextual closeness embedded in pages). Contextual advertising supports various types of Web sites and plays an essential role in the market value of the web. The most prevalent payment strategy for textual ads is Pay-Per-Click (PPC), where the advertisers pay a certain amount to the publisher and the ad platform for each user's click on the ads. To maximize the PPC payment strategy for a page, it is preferable to select textual ads that are relevant to the context of the page rather than to simply placing generic ads. However, it is difficult to accurately judge the relevance of the ad to a generic page due to the semantic analysis capability of textual documents. Wikipedia-based matching was proposed to enhance the semantic representation of text documents. Since Wikipedia-based matching too chooses a set of appropriate articles from Wikipedia to construct the intermediate semantic reference model for ads/pages, it is difficult to match all the articles for an ad/page. To deal with the trade-off between effectiveness and efficiency, the authors propose a new Wikipedia-based matching approach, called Selective Wikipedia Matching, and study three selective matching strategies, aimed at matching the most relevant reference Wikipedia articles to an ad/page rather than all the articles.

Univariate Decision Tree Induction using Maximum Margin Classification. OLCAY YILDIZ

Decision trees are well-known machine learning algorithms. The nodes of decision trees contain the attributes to test and the leaves show the decisions made. The type of the decision tree is based on the type of the split. This paper concentrates on univariate decision trees where the split is based on one attribute. In particular, this paper considers continuous univariate margin decision trees where there are two children to each internal node and where, for each continuous attribute, the best split is found using convex optimization. A novel decision tree classifier is given, which finds the best split for each attribute at each decision node using convex optimization. First, the univariate margin tree (UMT) is introduced along with the proposed algorithm for finding the best split at each decision node of the UMT. The performance of the proposed UMT algorithm is compared with C4.5 and LDT in terms of the generalization error, accuracy, tree size and model convexity. A total of 47 data sets are used in the experimental tests, 36 of which come from

the UCI repository and the remaining 11 are bioinformatics cancer data sets.

Exploiting Wikipedia Knowledge for Conceptual Hierarchical Clustering of Documents. GERASIMOS SPANAKIS, GEORGIOS SIOLAS AND ANDREAS STAFYLOPATIS

Traditional clustering algorithms are usually based on the bag-of-words (BOW) approach for representing documents and this faces a number of disadvantages. This paper attempts to enhance performance in a clustering task by enriching and compressing document representation. An efficient document clustering technique is introduced using knowledge extracted from Wikipedia to create the so-called document concepts to replace a BOW. A conceptual hierarchical clustering (CHC) technique is implemented to produce conceptual clustering informative of the content of the documents. After a review of the related work, the document representation model using Wikipedia is presented. The concept extraction method from Wikipedia, the concept disambiguation process and the representation of the document are introduced. The document representation is via a vector representation where each component corresponds to the importance of each concept in the document. Thereafter, the CHC method is introduced and provides a cluster description based on the Wikipedia concepts extracted from the corpus examined. The effectiveness of the method is compared with two standard document clustering techniques (HAC and k-NN). Different experiments are carried out and conclusions made re the runtime of the algorithm its complexity.

Spatial Sampling for Image Segmentation. MARIANO RIVERA, OSCAR DALMAU, WASHINGTON MIO AND ALONSO RAMIREZ-MANZANARES

Image segmentation consists in partitioning an image into regions with similar characteristics. It is however task- and user-dependent. In this paper, the authors present a novel framework for probabilistic image segmentation based on the maximum likelihood estimator where the likelihood in lack of multiple observations is improved. After an introduction to the basic notations used, the likelihood based on spatial samples is introduced and it is shown that if the image to segment is composed of an assemble of relative large regions, then spatial samples (neighborhood of pixels) can alleviate the lack of multiple observations for each pixel. Since an accurate segmentation depends on selecting a pixel neighborhood such that its majority belongs to the correct class, the neighborhood selection is addressed in detail. Then, the algorithm for estimating likelihoods based on multiple spatial samples is presented and its versatility is demonstrated through a number of applications and experiments. In particular, four applications are considered: interactive image segmentation that integrates color and texture clues, stereo-disparity estimation, denoising and multi-tensor field restoration.

Multi-objective Evolutionary Optimization of Training and Topology of Recurrent Neural Networks for Time-Series Prediction. HIDEKI KATAGIRI, ICHIRO NISHIZAKI, TOMOHIRO HAYASHIDA AND TAKANORI KADOMA

The paper starts from the observations that, when using neural networks (NNs), (1) it is necessary to select the appropriate network model and its topology, and that this requires a high computational cost and a hard decision process since a network with a low number of computing units may not correctly learn, whereas a large number of computing units may produce over-learning and (2) a small-size network may be more desirable for easing model understanding and hardware implementation. A number of approaches have been proposed in the literature for training and optimizing the size of NNs. This paper focuses on recurrent NNs (RNNs) and attempts to overcome the drawbacks of existing methods. The authors propose a more efficient evolutionary multi-objective optimization method for training RNNs through past data and for optimizing their topology. In particular, the authors discuss the drawbacks of an algorithm proposed by Delgado *et al*. for optimizing through the evolutionary process a particular kind of RNN, and propose a new algorithm for optimizing a more general type of RNN model. In the proposed algorithm, the authors focus on the intensive exploration of the solution area with small training errors, instead of the global exploration of the Pareto frontier as in the earlier algorithm. This is done through (a) a local optimal solutions preservation strategy, (b) an elite preservation strategy and (c) a self-adaptive mutation probability setting. To verify the effectiveness of the proposed model, its performance is compared with that of the model by Delgado *et al*. through numerical experiments using nine time-series prediction benchmark instances.

Discovering Community of Lingual Practice for Matching Multi-lingual Tags from Folksonomies. JASON J. JUNG

Folksonomy (or social tagging) consists of a set of tags used in tagging activities to manage users' resources in online social networks and to implement collective intelligence (CI). Co-occurrence patterns between entities (e.g. users, tags and resources) in a folksonomy system can be used to discover and exploit meaningful patterns that are implicitly hidden. However, the tags may be in different languages, which makes it difficult for the CI-based system to find useful patterns. This paper addresses this issue by analyzing a multi-lingual folksonomy that is written in several different languages to investigate what kind of CI can be established among users who speak different languages and among multi-lingual users. First, the author discusses how to formalize multi-lingual folksonomy generated by multiple users and how to discover meaningful co-occurrence patterns between multi-lingual tags in the folksonomy. Then, the author explains how to build a community of practice to organize a group of people who

share common interests and skills. Thereafter, the author moves on to demonstrate a tag-based IR system with multi-lingual folksonomies that realizes lingual gaps between tags in the folksonomy, and reduces the lingual gaps by using a relevant lingual practice. This proposed scheme is then evaluated by collecting a large amount of tags and resources from two information sources (social bookmarking and photo sharing). Multi-lingual tag matching is evaluated by choosing 10 major languages to build a simplified folksonomy and to test the proposed scheme with this folksonomy. Multi-lingual resource retrieval is evaluated by inviting 18 students (9 Korean domestic and 9 international students) in two foreign literature departments (French and German) to measure their satisfaction on the proposed tag-based IR. Finally, the author discusses how to identify tag semantics and how to match tags, how to search by tag matching, and how to support community-based social collaboration, and ends with concluding remarks that also touch on the limitations of the proposed study.

Neural Network-Based Approach for Predicting Trust Values Based on Non-uniform Input in Mobile Applications. MUHAMMAD RAZA, FAROOKH KHADEER HUSSAIN AND OMAR KHADEER HUSSAIN

The authors focus on the need to have appropriate techniques for trust management that capture the dynamic factors and variability in the trust values since these can be used to make an informed decision in a business interaction. In particular, the authors focus on trust modeling, which is one of the key strategies in trust management. Since trust modeling is classified by 'trust determination' and 'trust prediction', the authors give an overview of the existing work on trust values prediction, its confusion with trust determination as well as the problems to be addressed in the paper. Since trust prediction deals with uncertainty and existing approaches for trust predication cannot handle non-uniform inputs, the authors state the need for a mechanism/method to: (1) predict trust for any given input data series (uniform/non-uniform, stationary/non-stationary, seasonal, noisy, etc.); (2) to predict the future trust values accurately and (3) to capture the different trends or variations present in the input series and utilize them to model the outputs. This paper proposes a neural network-based approach for predicting trust values for any given entity and taking into account the dynamic nature of trust. This is done through a multi-layer feed forward network (MLFFN) for trust prediction based on uniform and non-uniform input series. For the approach to work at optimal capacity, it is fine-tuned and adapted to give the desired results. Different types of input data are considered to determine the point at which the proposed MLFFN approach for trust value prediction works best with the maximum accuracy. Four different data sets have been collected and used for the experiments. Each data set represents a different data series pattern: non-uniform stationary data series, non-uniform seasonal series, non-uniform trend series and noisy data series. Experiments, simulation results and analysis are discussed in detail.