Special issue on Advances in Web Intelligence

Stefan Rüger[a,**], Vijay Raghavan[b,*], Irwin King[c,*], Jimmy Xiangji Huang[d,*]

[a]Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom
[b]The Center for Advanced Computer Studies, University of Louisiana at Lafayette, P.O. Box 44330, Lafayette, LA 70504-4330, USA
[c]Department of Computer Science & Engineering, 908 Ho Sing Hang Engineering Building, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, SAR
[d]School of Information Technology, York University, Toronto, Ontario, Canada M3J 1P3

Abstract

We summarize the scientific papers of the special issue "Advances in Web Intelligence,"which will appear in the Neurocomputing journal, Volume 76, Issue 1 (2012)- published by Elsevier. These papers are substantially extended from original contributions to the Web Intelligence 2010 conference held in Toronto, Canada, in September 2010.

Keywords: special issue, web intelligence

1. Introduction

Web intelligence is commonly seen as a combination of applied artificial intelligence and information technology in the context of the web with a view to study and characterize emerging — or design new — products and services of the internet. The Web Intelligence conference, held every year between 2001 and 2010 with the exception of 2002, has recognized these new directions and provides a scientific forum for researchers and practitioners to further topics such as web intelligence foundations; world wide wisdom web (W4); web information retrieval and filtering; semantics and ontology engineering; web mining and farming; social networks and ubiquitous intelligence; knowledge grids and grid intelligence; web agents; web services; intelligent human-web interaction; web support systems; intelligent e-technology; and other related areas.

Web Intelligence 2010 received 313 submissions, of which 51 regular papers were accepted. The papers in this special issue reflect the trends at Web Intelligence 2010 and focus on particularly strong contributions to the conference, of which we invited 16 to the special issue. These authors since then expanded their Web Intelligence 2010 contribution significantly for the benefit of the Neurocomputing readership guided by an independent, critical peer review process that ultimately accepted 9 submissions. The areas to which the papers of this special issue contribute can broadly be characterized into content analysis (Section 2), social media and network analysis (Section 3) and machine learning for web intelligence (Section 4).

2. Content analysis

The paper "Multimodal Representation, Indexing, Automated Annotation and Retrieval of Image Collections via Non-negative Matrix Factorization" by Caicedo et al. proposes a novel method of analyzing and generating multimodal image representations that integrate both visual features in images and their associated text information such as descriptions, comments, user ratings and tags. Experiments using Corel and Flickr data sets have demonstrated the advantage of non-negative matrix factorization with asymmetric multimodal representation over other approaches such as direct matching and singular value decomposition.

Krestel and Fankhauser's paper "Personalized Topic-Based Tag Recommendation" proposes an approach for personalized tag recommendation that combines tags derived via the probabilistic model of a web resource with those obtained from the user. The paper investigates simple language models as well as Latent Dirichlet Allocation as alternatives for modeling the resource content. Experiments on a real world dataset show that personalization improves tag recommendation, and the proposed approach significantly outperforms state-of-the-art approaches, such as FolkRank.

The next paper "On Optimization of Expertise Matching with Various Constraints" by Tang et al. studies mechanisms of assigning experts to a set of items such as submitted papers to a conference or to-be-reviewed products under constraints, for example, that the overall workload of individual experts is balanced and that each item has a certain number of reviews by senior and less senior experts. The paper formulates the expertise matching problem in a general constraint-based optimization framework that links the problem to a convex cost flow, which promises an optimal solution under various constraints. Tang et al. also propose an online matching algorithm that incorporates immediate user feedback. Experimental results validate the effectiveness of the proposed approach in the cases of reviewer to conference paper assignment and teacher to course assignment.

[*]Guest editor
[**]Managing guest editor

Reprinted from Neurocomputing, Vol 76, Issue 1 (2012), 48–49, Copyright (2012), with permission from Elsevier

3. Social media and network analysis

The paper "Characteristics of Information Diffusion in Blogs, in Relation to Information Source Type" by Kazama et al. introduces information diffusion properties to analyze the dynamics of blogs based on constructed subgraphs for information recommendation and ranking. The work focuses on three types of basic structures: information scattering, information gathering, and information transmission structures. With these information diffusion properties, the work is able to represent various social media characteristics and provide priority to different types of information sources.

"A Framework For Joint Community Detection Across Multiple Related Networks" by Comar et al. utilizes non-negative matrix factorization that combines information from multiple networks in order to identify communities and learn the correspondences among these networks simultaneously. The method has shown good performance over other approaches such as normalized cut and matrix factorization with experiments done on both synthetic as well as real-world wikipedia and digg data sets.

Largillier and Peyronnet demonstrate in their paper "Webspam Demotion: Low Complexity Node Aggregation Methods" a mechanism to lower the ranking of webspam, which are undesirable web pages that were created with the sole purpose of influencing link-based ranking algorithms to promote a particular target page. Webspam techniques evolve all the time, but almost inevitably they create a specific linking architecture around the target page to increase its rank. Largillier and Peyronnet study the effects of node aggregation of the well-known PageRank algorithm in presence of webspam. Their lightweight node aggregation methods aim to construct clusters of nodes that can be considered as a sole node in the PageRank computation. Experimental results show the promise of the presented webspam demotion approach.

4. Machine learning for web intelligence

Yan et al. propose in their paper "Semi-Supervised Dimensionality Reduction for Analyzing High-Dimensional Data with Constraints" a novel technique to address the problems of inefficient learning and costly computation in coping with high-dimensional data. The approach, termed Dual Subspace Projections, embeds high-dimensional data in an optimal low-dimensional space, which is learned with a few user-supplied constraints and the structure of input data. The method overcomes the model overfitting problem by simultaneously preserving both the structure of original high-dimensional data and user-specified constraints. Experiments on real datasets from multiple domains demonstrate that significant improvement in learning accuracy can be achieved via their dimensionality reduction technique, even with only a few user-supplied constraints.

Ramirez et al.'s paper "Topic Model Validation" considers the problem of performing external validation of the semantic coherence of topic models. Ramirez et al. generalize the Fowlkes-Mallows index, a clustering validation metric, for the case of overlapping partitions and multi-labeled collections rendering it suitable for assessing topic modeling algorithms. They also propose probabilistic metrics inspired by the concepts of recall and precision and show how these can be applied to validate and compare other soft and overlapping clustering algorithms.

In their paper "Modeling and Predicting the Popularity of Online Contents with Cox Proportional Hazard Regression Model", Lee et al. propose a framework, which can be used for modeling and predicting the popularity of discussion forum threads based on initial observations of how the thread evolves and the number of comments. The underlying approach is rooted in survival analysis, which models the survival time until an event of a failure or death. Lee et al. model the lifetime of discussion threads and the number of comments that the contents receives, with a set of explanatory and externally observable factors, using the Cox proportional hazard regression model, which divides the distribution function of the popularity metric into two components: one which is explained by a set of observable factors, and another, a baseline survival distribution function, which integrates all the factors not taken into account. The methodology is validated with two datasets that were crawled from two different discussion fora.

Acknowledgements

The guest editors wish to thank the Editor-in-Chief Tom Heskes for giving us the opportunity to present the Advances in Web Intelligence; the reviewers, who have helped to select the very best papers from Web Intelligence 2010 and gave valuable feedback on the extended work in the papers of this special issue within a very short time-frame; the journal staff for a professional production; and, not least, the authors, who carry this field with much enthusiasm and scientific rigor.

References

Caicedo, J. C., BenAbdallah, J., Gonzalez, F. A., Nasraoui, O., 2012. Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. Neurocomputing 76 (1), 50–60.

Comar, P. M., Tan, P.-N., Jain, A. K., 2012. A framework for joint community detection across multiple related networks. Neurocomputing 76 (1), 93–104.

Kazama, K., Imada, M., Kashiwagi, K., 2012. Characteristics of information diffusion in blogs, in relation to information source type. Neurocomputing 76 (1), 84–92.

Krestel, R., Fankhauser, P., 2012. Personalized topic-based tag recommendation. Neurocomputing 76 (1), 61–70.

Largillier, T., Peyronnet, S., 2012. Webspam demotion: Low complexity node aggregation methods. Neurocomputing 76 (1), 105–113.

Lee, J. G., Moon, S., Salamatian, K., 2012. Modeling and predicting the popularity of online contents with cox proportional hazard regression model. Neurocomputing 76 (1), 134–145.

Ramirez, E. H., Brena, R., Magatti, D., Stella, F., 2012. Topic model validation. Neurocomputing 76 (1), 125–133.

Tang, W., Tang, J., Lei, T., Tan, C., Gao, B., Li, T., 2012. On optimization of expertise matching with various constraints. Neurocomputing 76 (1), 71–83.

Yan, S., Bouaziz, S., Lee, D., Barlow, J., 2012. Semi-supervised dimensionality reduction for analyzing high-dimensional data with constraints. Neurocomputing 76 (1), 114–124.