



Guest Editorial: Text and Web Mining

AH-HWEE TAN

Nanyang Technological University, Blk N4, 2A-13 Nanyang Avenue, Singapore 639798

asahtan@ntu.edu.sg

PHILIP S. YU

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

psyu@us.ibm.com

Text mining and web mining are two interrelated fields that have received a lot of attention in recent years. Text mining [1, 2] is concerned with the analysis of very large document collections and the extraction of hidden knowledge from text-based data. Web mining [3] refers to the analysis and mining of all web-related data, including web content, hyperlink structure, and web access statistics. Among the three aspects of web mining, text mining is most closely related to web content mining. However, whereas text mining deals with text documents in general, such as emails, letters, reports, and articles, that exist in both intranet and internet environment, web content mining is primarily concerned with the materials on the web only. Dealing with free-form unstructured and semi-structured text, text mining can be envisaged as an immediate extension of data mining or *knowledge discovery from databases* [2]. Web content mining, on the other hand, covers a wider scope of dealing with rich multimedia contents, including text, image, audio, and video, intermixed with HTML formatting tags and hyper-links. Nevertheless, as text constitutes a large portion of web content, text mining is still recognized by many as a key enabling technology for web resource management and mining.

Text and web mining are both technically interesting and commercially relevant. A good number of companies including high-tech start-ups and established players, such as Verity, Autonomy, Megaputers, Microsoft, and IBM, have released a range of text and web mining related products and services since a few years back. The various functions supported include search and retrieval, document navigation/exploration, text analysis,

and knowledge management. We have also witnessed a convergence of interests from many established academic fields, including statistics, pattern recognition, machine learning, database, data mining, natural language processing, and computational linguistic into text and web mining. To name a few, some of the well-known efforts in the research communities include World Wide Knowledge Base (Web->KB) [5] and WebWatcher [6] by the Text Learning Group at Carnegie Mellon University, Natural Language Learning [7] research by the Machine Learning Group at the University of Texas at Austin, and WebBase [8], that has culminated the PageRank Algorithm [9] as used in *Google*, at Stanford University.

This special issue contains eight technical papers selected by a panel of over thirty international experts through a rigorous peer review process. The articles include three openly solicited papers as well as expanded versions of five papers presented at the International Workshop on Text and Web Mining, held in conjunction with the Sixth Pacific Rim International Conference on Artificial Intelligence (PRICAI'2000) in Melbourne Convention Centre, Australia on 28 August 2000. The papers collected here focus primarily on text and web content mining, covering such topics as document retrieval, text/web categorization, tagging, schema extraction, clustering, and information discovery.

The first two articles focus on the problem of document retrieval. *Genetic Mining of HTML Structures in Web-Document Retrieval* by Kim and Zhan exploits the inherent HTML structure in web document to facilitate document retrieval. Tan et al., on the other hand,

present a novel approach to text retrieval from document images based on word shape analysis. We include two articles on document clustering that coincidentally are both based on a popular class of unsupervised neural networks known as Self-Organizing Map (SOM) [10]. Rauber and Merkl describe a SOM-based digital library and discuss the representational issues of topics and genres. Lee and Yang present a framework for performing multilingual text mining based on Self-Organizing Maps. On the supervised learning aspect, we have a paper by He, Tan, and Tan that reports benchmark comparisons of three state-of-the-machine learning methods for Chinese document categorization.

While the first five articles focus on techniques for organizing textual information, the remaining three papers investigate the problem of extracting information and knowledge from documents. The contributed paper of Velardi and Missikoff describes various text mining techniques to automatically enrich a domain ontology. Carchiolo, Longheu, and Malgeri propose a method for extracting logical schema from the Web. The last paper of the issue describes an application of text mining to medical decision support. By formulating rule mining as a categorization problem, Loh, Oliveira, and Gameiro present a method for constructing automatic decision systems from patient records.

We hope this issue provides a snapshot of some of the latest advancement in the fields of text and web mining. More collections of text and web mining related articles can be found in the form of workshop proceedings of several major conferences, such as the KDD'2000 Workshop on Text Mining [11] and the PRICAI'2000 Workshop on Text and Web Mining [12]. We believe text and web mining will continue to gain importance in the years to come. Several promising research directions include content personalization, multilingual content mining, and web resource management/mining. We also expect to see more domain-specific applications of text and web mining technologies to building personalized/customized vertical portals and competitive intelligence systems. We hope you enjoy reading the special issue and look forward to more exciting development in the fields of text and web mining.

Acknowledgments

We would like to thank the PRICAI'2000 Text and Web Mining workshop program committee members and many other anonymous referees who have spent precious time in reading numerous manuscripts and

providing many invaluable comments to the authors. We also thank all authors who have contributed to the Text and Web Mining workshop and this journal issues for their contributions.

References

1. R. Feldman and I. Dagan, "KDT—knowledge discovery in texts," in *Proceedings of the First Annual Conference on Knowledge Discovery and Data Mining (KDD)*, 1995.
2. M.A. Hearst, "Untangling text data mining," in *Proceedings of ACL'99*, 1999, pp. 20–26.
3. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in *Advances in Knowledge Discovery and Data Mining*, MIT Press: Cambridge, MA, 1996, pp. 1–38.
4. R. Kosala and H. Blockeel, "Web mining research: A survey," *SIGKDD Explorations*, vol. 2, no. 1, pp. 1–15, 2000.
5. M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery, "Learning to extract symbolic knowledge from the world wide web," in *Proceedings, AAAI-98*, 1998.
6. T. Joachims, D. Freitag, and T. Mitchell, "WebWatcher: A tour guide for the world wide web," in *Proceedings, IJCAI'97*, 1997.
7. R.J. Mooney and C. Cardie, *Symbolic Machine Learning for Natural Language Processing*, A tutorial at ACL, 1999.
8. J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke, "Web-Base: A repository of web pages," in *Proc. of 9th Int. World-Wide Web Conf.*, 2000, pp. 277–293.
9. S. Brin and L. Page, "The Anatomy of a large-scale hypertextual web search engine," in *7th International World Wide Web Conference*, 1998.
10. T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, V. Paatero, and A. Saarela, "Self organization of a massive document collection," *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, vol. 11, no. 3, pp. 574–585, 2000.
11. A.-H. Tan and P.S. Yu (eds.), *Proceedings, PRICAI'2000 Workshop on Text and Web Mining*, Melbourne, 2000.
12. M. Grobelnik, D. Mladenic, and N. Milic-Frayling (eds.), *Proceedings, SIGKDD'2000 Workshop on Text Mining*, Boston, 2000.



Ah-Hwee Tan is an Associate Professor in the School of Computer Engineering at Nanyang Technological University. He was a Research Manager and Senior Member of Research Staff at the Kent Ridge Digital Labs, Laboratories for Information Technology, and Institute for Infocomm Research, where he led R&D projects in knowledge discovery, document analysis, and information mining.

He received his Ph.D. in Cognitive and Neural Systems from Boston University in 1994. Prior to that, he obtained his Bachelor of Science (First Class Honors) (1989) and Master of Science (1991) in Computer and Information Science from the National University of Singapore. He is an editorial board member of *Applied Intelligence* and a member of Singapore Computer Society, ACM, and ACM SIGKDD.



Philip S. Yu received the B.S. Degree in E.E. from National Taiwan University, the M.S. and Ph.D. degrees in E.E. from Stanford

University, and the M.B.A. degree from New York University. He is with the IBM Thomas J. Watson Research Center and currently manager of the Software Tools and Techniques group. His research interests include data mining, Internet applications and technologies, database systems, and multimedia systems. Dr. Yu has published more than 290 papers in refereed journals and conferences. He holds or has applied for 234 US patents.

Dr. Yu is a Fellow of the ACM and a Fellow of the IEEE. He is the Editor-in-Chief of *IEEE Transactions on Knowledge and Data Engineering*. He is also an associate editor of *ACM Transactions on the Internet Technology* and that of *Knowledge and Information Systems*. He is a member of the IEEE Data Engineering steering committee and is also on the steering committee of IEEE Conference on Data Mining. In addition to serving as program committee member on various conferences, he was the program co-chair of the 11th Intl. Conference on Data Engineering and the general chair of the 14th Intl. Conference on Data Engineering. He has received several IBM and external honors including Best Paper Award, 2 IBM Outstanding Innovation Awards, an Outstanding Technical Achievement Award, 2 Research Division Awards and the 65th plateau of Invention Achievement Awards. Dr. Yu is an IBM Master Inventor.