Data Min Knowl Disc (2012) 24:473–477
DOI 10.1007/s10618-012-0257-y

Guest editorial: special issue on a decade of mining the Web

Myra Spiliopoulou · Bamshad Mobasher · Olfa Nasraoui · Osmar Zaiane

Received: 25 January 2012 / Accepted: 2 February 2012 / Published online: 3 March 2012

Abstract As editors of the Special Issue on a Decade of Mining the Web, we provide a brief overview of how Web mining evolved from the first Web mining workshop (WEBKDD'99) till today. We then introduce the papers of the special issue. Each of them is in a domain of Web mining research; it contains a survey of the past and a vision for the future.

Keywords Web mining · Clickstream mining · Social Web mining · Semantic Web mining · Web mining and privacy · Web mining and recommenders

1 Mining the Web in the nineties and in the new millennium

When the first workshop on mining the Web, WEBKDD'99 (Masand and Spiliopoulou 2000), took place at the KDD'99, the International Conference on Knowledge Discovery and Data Mining, not only the challenges and the opportunities of the Web were

M. Spiliopoulou (✉)
Otto-von-Guericke University Magdeburg, Magdeburg, Germany
e-mail: myra@iti.cs.uni-magdeburg.de

B. Mobasher
DePaul University, Chicago, IL, USA
e-mail: mobasher@cs.depaul.edu

O. Nasraoui
University of Louisville, Louisville, KY, USA
e-mail: olfa.nasraoui@louisville.edu

O. Zaiane
University of Alberta, Edmonton, AB, Canada
e-mail: zaiane@ualberta.ca

very different from nowadays. Even the terminology was different: the term "Social Web" did not exist, the term "Semantic Web" was emerging, a "weblog" was not a blog but the log of a web server. A major aim of Web mining was to understand what users want and to help them perform simple tasks inside a web site. Mining goals included finding navigation patterns and preventing disorientation (Baumgarten et al. 1999; Fu et al. 1999; Spiliopoulou et al. 1999), predicting a user's next page request and calibrating server load (Lan et al. 1999), promoting purchases as the users move through an e-shop (Chan 1999; Getoor and Sahami 1999; Lee et al. 1999). Preparing web server log data for mining and discovering patterns from them was a complex process that required new solutions for cleaning and modeling the data and for learning from them (Cooley et al. 1999). Scholars soon realized that the Web and the analysis of Web activity also poses threats to personal privacy, and proposed the first solutions (Broder 1999).

The infancy of the Web did not last long: core technologies matured soon, e-business models emerged and were put to the test of time. By the time the WEBKDD'05 workshop (Nasraoui et al. 2006) took place, recommendation engines had turned to a flagship application of web mining, and lead to solutions for core problems like mining with incomplete data (Grcar et al. 2005; Schickel-Zuber and Faltings 2005) and personalization to the user demands (DeLong et al. 2005; Kim and Chan 2005), but also to emerging problems, like protecting recommendation engines from malicious influence (Mobasher et al. 2005). The role of semantics in Web mining was already gaining momentum (Berendt 2005; Grcar et al. 2005), and so did the awareness that models learned in the Web must be adapted to change. Web log server data were now perceived as a *stream*: the term "clickstream" emerged and gradually replaced the "web (server) log". As the need for adaptive algorithms became apparent, solutions were proposed for the adaptation of user profiles (Suryavanshi et al. 2005) and for detecting and adapting to change in general (Aggarwal and Yu 2005; Lu et al. 2005).

During the first decade of the new millennium, people's perception of the Web changed: the Web is not only a network where users can acquire information, purchase goods and obtain services; it is also a social environment where users can interact with others, and where each one can actively upload content and share information. The Social Web brought forward new challenges, many of which required theoretical underpinnings from other disciplines—social science (which deliver theories) and physics (which deliver models for the simulation of dynamic environments). In 2007, WEBKDD took place jointly with the first workshop on social network analysis SNAKDD (Zhang et al. 2007), allowing participants that studied the Web from different perspectives to share experiences and methods.

When the 10th WEBKDD workshop (Nasraoui et al. 2008) closed the series in 2008, the core Web mining technologies had reached maturity. Web mining embodied the topics of knowledge discovery for recommendation engines, model adaptation for user profiling, understanding communities and monitoring their evolution, modeling and interpreting user search. Privacy and protection from malicious activity were still mission-critical issues but must be dealt with in additional contexts (as e.g. in Castillo et al. 2008). As scholars, practitioners and users agree that the Web changed our ways of acquiring information, sharing experience, making friends and helping other people, the topics of Web mining became too disperse for a single forum. At the beginning of

the second decade of the new millennium, there were already established conferences on recommendation engines, on social network analysis, and on the Semantic Web.

2 The special issue on a decade of mining the Web

We have designed this special issue as a wrap-up of the past and an agenda for the ongoing decade. We solicited contributions that should combine an overview of past achievements and a vision of the future. The emphasis was less on a bright new solution to one of the many open challenges of Web mining, and more on a wider look at a broad topic. Each of the papers in the next pages addresses a Web mining topic under this light.

Two aspects of mining the Social Web are investigated in Tsytsarau and Palpanas (2012), Papadopoulos et al. (2012): active participation and community formation.

Web users are active participants of the Web world: they upload content and link it with tags and opinions. The survey of Tsytsarau and Palpanas (2012) is on the analysis of opinionated data. Such data items are subjective—they exhibit positive or negative polarity, and often express the writer's sentiment. The survey covers methods and applications of opinion mining, solutions for the extraction of opinions from text and for polarity assessment. A remarkable aspect of the survey is the discussion on the evolution of research on sentiment analysis and opinion mining over the years. Furthermore, the address discuss two topics of particular interest in commerce applications—spam and contradictory opinions on the same subject (or product).

Web users join social platforms and establish friendship relations to other users. Papadopoulos et al. (2012) survey community detection in the Web. They discuss definitions of the term "community", since different definitions require different community discovery algorithms. They elaborate on methods for detecting and monitoring communities as the network evolves. They focus particularly on the issue of algorithm performance, but they also elaborate on different applications of community detection, e.g. user profiling, event detection and tag disambiguation.

The analysis of social networks may deliver insights towards the attitude of people towards an product, a piece of news or any other kind of resource. Hence, modern recommendation engines take into account which items are similar to each other in terms of content or with respect to people's opinions, which users have similar preferences, deliver similar opinions or are friends of each other. Obviously, different models of user or item similarity lead to different kinds of recommendations. Hence, some recommenders attach an explanation to each suggestion they make. Papadimitriou et al. (2012) survey recommendation engines on the kind of information they consider when making a recommendation.

E-business companies exploit the advances of Web mining to identify influential users in social networks, to monitor opinions on products, to learn and adapt user profiles, to formulate personalized, context-aware recommendations. However, understanding the customer and establishing a profitable, long-term relationship to customers goes beyond the aforementioned tasks. Customer Relationship Management is discussed by Alex Tuzhilin (2012) in the invited contribution on CRM and Web mining. His overview contributes to understanding what CRM is and what its

objectives and demands are; this is a prerequisite for exploiting the potential of Web mining methods.

How to model a customer, or an arbitrary Web user? Intuitively, a user is an individual with rich semantics—has own properties and is linked to other users. Rettinger et al. (2012) study learning on (Web) entities that have rich semantics. They survey different approaches to knowledge representation and methods for relational learning. They point out that the semantic representation of individuals and their relationships in the Semantic Web require dedicated and elaborate learning approaches, and discuss first machine learning algorithms for the Semantic Web.

What are people doing in the Web? Next to being active members of social platforms, people use the Web to satisfy different kinds of information: facts (e.g. radius of earth), reviews on and comparisons of products, breaking news about ongoing events etc. Agosti et al. (2012) survey research advances on information acquisition and on learning from the interaction of a user with a search engine. They discuss query log mining, recommendation of query terms, mono-lingual and cross-lingual query expansion in two contexts—Web search engines and search engines of conventional digital libraries. They elaborate on methods for analyzing query logs, and stress the importance of performing such an analysis in a repeatable, verifiable way.

All applications that involve analysis of user activity in the Web deliver some insights on what people think, like, want and are. This gives raise to privacy considerations and concerns. The invited contribution on privacy in the Web mining context by Bettina Berendt (2012) stresses that privacy preservation is much more than hiding information. She discusses different definitions of privacy, distinguishes among privacy-as-hiding, privacy-as-control and privacy-as-practice, and places these distinct aspects into the context of knowledge discovery from the Web.

Acknowledgments We thank all authors who submitted to this special issue. We are indebted to the reviewers for their thorough work and for many constructive comments to the authors. Cordial thanks go also to the journal editor-in-chief Dr. Geoff Webb for all his support and advice. Particular thanks go to Gayathri Balasubramanian from the Springer editorial team for her tireless support and troubleshooting throughout the submission and reviewing rounds of the special issue, and to Melissa Fearon for the efficient coordination of the publication process.

References

Aggarwal C, Yu P (2005) On clustering techniques for change diagnosis in data streams. In: WEBKDD'05 postworkshop volume (see Nasraoui et al. 2006), pp 139–157

Agosti M, Crivellari F, Di Nunzio GM (2012) Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. Data Min Knowl Discov. doi:10.1007/s10618-011-0228-8

Baumgarten M, Büchner AG, Anand, SS, Mulvenna MD, Hughes JG (1999) Navigation pattern discovery from internet data. In: WEBKDD'99 postworkshop volume (see Masand and Spiliopoulou 2000), pp 70–87

Berendt B (2005) Using and learning semantics in frequent subgraph mining. In: WEBKDD'05 postworkshop volume (see Nasraoui et al. 2006), pp 18–38

Berendt B (2012) More than modeling and hiding: towards a comprehensive view of Web mining and privacy. Data Min Knowl Discov. doi:10.1007/s10618-012-0254-1

Broder A (1999) Data mining, the Internet, and privacy. In: WEBKDD'99 postworkshop volume (see Masand and Spiliopoulou 2000), pp 51–69

Castillo C, Corsi C, Donato D, Ferragina P, Gionis A (2008) Query-log mining for detecting polysemy and spam. In: WEBKDD'08 workshop volume in ACM digital library (see Nasraoui et al. 2008), pp 29–42
Chan PK (1999) Constructing web user profiles: a non-invasive learning approach. In: WEBKDD'99 post-workshop volume (see Masand and Spiliopoulou 2000), pp 34–50
Cooley R, Tan PN, Srivastava J (1999) Discovery of interesting usage patterns from web data. In: WEB-KDD'99 postworkshop volume (see Masand and Spiliopoulou 2000), pp 160–179
DeLong C, Desikan P, Srivastava J (2005) USER: user-sensitive expert recommendations for knowledge-dense environments. In: WEBKDD'05 postworkshop volume (see Nasraoui et al. 2006), pp 77–95
Fu Y, Sandhu K, Shih MY (1999) A generalization-based approach to clustering of web usage sessions. In: WEBKDD'99 postworkshop volume (see Masand and Spiliopoulou 2000), pp 16–33
Getoor L, Sahami M (1999) Using probabilistic relational models for collaborative filtering. In: WEBKDD'99 online notes volume
Grcar M, Mladenic D, Fortuna B, Grobelnik M (2005) Data sparsity issues in the collaborative filtering framework. In: WEBKDD'05 postworkshop volume (see Nasraoui et al. 2006), pp 58–76
Kim HR, Chan PK (2005) Personalized search results with user interest hierarchies learnt from bookmarks. In: WEBKDD'05 postworkshop volume (see Nasraoui et al. 2006), pp 158–176
Lan B, Bressan S, Ooi BC (1999) Making web servers pushier. In: WEBKDD'99 postworkshop volume (see Masand and Spiliopoulou 2000), pp 108–122
Lee J, Podlaseck M, Schonberg E, Hoch R, Gomory S (1999) Analysis and visualization of metrics for online merchandizing. In: WEBKDD'99 postworkshop volume (see Masand and Spiliopoulou 2000), pp 123–138
Lu L, Dunham M, Meng Y (2005) Mining significant usage patterns from clickstream data. In: WEBKDD'05 postworkshop volume (see Nasraoui et al. 2006), pp 1–17
Masand B, Spiliopoulou M (eds) (2000) Advances in web usage mining and user profiling: proceedings of the WEBKDD'99 workshop. LNAI, vol 1836. Springer, Berlin
Mobasher B, Burke R, Williams C, Bhaumik R (2005) Analysis and detection of segment-focused attacks against collaborative recommendation. In: WEBKDD'05 postworkshop volume (see Nasraoui et al. 2006), pp 96–118
Nasraoui O, Zaïane O, Spiliopoulou M, Mobasher B, Masand B, Yu PS (eds) (2006) Advances in web mining and web usage analysis—7th international workshop on knowledge discovery on the web, WebKDD 2005, revised papers. LNAI, vol 4198. Springer, Berlin
Nasraoui O, Spiliopoulou M, Zaïane O, Srivastava J, Mobasher B (eds) (2008) 10th international workshop on knowledge discovery on the web, WEBKDD'08: 10 years of knowledge discovery on the web. ACM, Las Vegas. In conjunction with the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2008)
Papadimitriou A, Symeonidis P, Manolopoulos Y (2012) A generalized taxonomy of explanations styles for traditional and social recommender systems. Data Min Knowl Discov. doi:10.1007/s10618-011-0215-0
Papadopoulos S, Kompatsiaris Y, Vakali A, Spyridonos P (2012) Community detection in social media: Performance and application considerations. Data Min Knowl Discov. doi:10.1007/s10618-011-224-z
Rettinger A, Lösch U, Tresp V, d'Amato C, Fanizzi N (2012) Mining the Semantic Web. Statistical learning for next generation knowledge bases. Data Min Knowl Discov. doi:10.1007/s10618-012-0253-2
Schickel-Zuber V, Faltings B (2005) Overcoming incomplete user models in recommendation systems via an ontology. In: WEBKDD'05 postworkshop volume (see Nasraoui et al. 2006), pp 39–57
Spiliopoulou M, Pohle C, Faulstich LC (1999) Improving the effectiveness of a web site with web usage mining. In: WEBKDD'99 postworkshop volume (see Masand and Spiliopoulou 2000), pp 139–159
Suryavanshi BS, Shiri N, Mudur SP (2005) Adaptive web usage profiling. In: WEBKDD'05 postworkshop volume (see Nasraoui et al. 2006), pp 119–138
Tsytsarau M, Palpanas T (2012) Survey on mining subjective data on the web. Data Min Knowl Discov. doi:10.1007/s10618-011-0238-z
Tuzhilin A (2012) Customer relationship management and Web mining: the next frontier. Data Min Knowl Discov. doi:10.1007/s10618-012-0256-z
Zhang H, Spiliopoulou M, Mobasher B, Giles CL, McCallum A, Nasraoui O, Srivastava J, Yen J (eds) (2007) Advances in web mining and web usage analysis—9th international workshop on knowledge discovery on the web, WebKDD 2007, and 1st international workshop on social network analysis, SNA-KDD 2007, Aug 2007, San Jose, USA. Revised papers. LNAI, vol 5439. Springer, Berlin