Mining web data: Techniques for understanding the user behavior in the Web.

By Juan D. Velásquez
PhD University of Tokyo
Assistant Professor
Department of Industrial Engineering
University of Chile
jvelasqu@dii.uchile.cl
http://wi.dii.uchile.cl/

A correct web site structure and content should help the users to find what they are looking for. However, sometimes the web site structure is complex, hiding the information and causing a "lost in hyperspace" feeling to the user. On the other hand, when the web site contains simple context, like free text only, it cannot be attractive for the user.

How can we prepare the correct web site structure and content in the right moment for the right user? The answer is not simple and, for the moment, there are only approximations to a possible final solution.

It seems to be that the key is in the understanding of the user behavior in a web site and using this knowledge, construct systems for personalizing the site, i.e., to adapt its structure and content for a particular user.

Web Mining techniques have contributed to the analysis of data originated in the Web, also called web data. By applying these techniques, significant patterns about the user behavior and his/her preferences can be discovered and used to personalize the web site.

In this tutorial, we will review the main web mining techniques used in the extraction of knowledge about the user behavior in the web, with emphasis on using hybrid and computational intelligence techniques for web mining. Some real-world applications will be presented.

Acknowledgement

This work has been funded partially by the Millenium Scientific Nucleus on Complexes Engineering Systems.

Content

1. Motivation
2. Web data
    a. Web operation
    b. Web site page content.
    c. Web site hyperlinks structure.

    d. Web site logs.
    e. Data preprocessing and cleaning.
3. Web Mining
    a. Web content mining.
    b. Web structure mining.
    c. Web usage mining.
4. Applications
    a. Web personalization
    b. Understanding the user behavior.
    c. Improving the web site structure and content.
    d. Intelligent web site.
5. Summary

References

[Abraham03] A. Abraham and V. Ramos. Web usage mining using artificial ant colony clustering and genetic programming. In *Procs. Int. Conf. CEC03 on Evolutionary Computation*, pages 1384–1391. IEEE Press, 2003.

[Amitay00] E. Amitay and C. Paris. Automatically summarizing web sites: Is there any wayaround it? In *Procs. of the 9th Int. Conf. on Information and Knowledge Management*, pages 173–179, McLean, Virginia, USA, 2000.

[Apte04] C. Apte, F. Damerau, and S. Weiss. Automatedlearning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3):233–251, 2004.

[Asirvatham05] A. P. Asirvatham. Web page categorization based on document structure, 2003. http://www.iiit.net/students/stnd_pdfs/kranthi.pdf Last accessed 9/13/2005.

[Ball67] G.B. Ball and D.J. Hall. A clustering technique for summarizing multivariate data. *Behavioral Science*, 12:153–155, 1967.

[Berendt01] B. Berendt and M. Spiliopoulou, Analysis of navigation behavior in web sites integrating multiple information, The VLDB Journal, 9, 56-75, 2001.

[Bestavros95] A. Bestavros. Using speculation to reduce server loadand service time on the www. In *In Proc. 4th Int. Conf. ACM International Conference onInformation and Knowledge Management*, pages 782–786, Baltimore, Maryland, USA, 1995.

[Borges99] J. Borges and M. Levene. Data mining of user navigation patterns. In *of the Web Usage Analysis and User Proling Workshop*, pages 31–36, San Diego, USA, 1999.

[Boving04] K. B. Bøving and J. Simonsen. Http log analysis as an approach to studying the use of web-based information systems. *Scandinavian Journal of Information Systems*, 16:145–174, 2004.

[Brusilovsky96] P. Brusilovsky, Methods and Techniques of Adaptive Hypermedia, User Modeling and User-Adapted Interaction, 6(2-3), 87-129, 1996.

[Buyukkokten01] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: text summarization for web browsing on handheld devices. In *Procs. Int. of the 10th Int Conf in World Wide Web*, pages 652–662, Hong Kong, 2001.

[Chakrabarti98a] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Resource compilation by analyzing hyperlink structure and associated text. In Procs. Int.Conf. World-Wide Web conference, pages 65–74, Brisbane, Australia, 1998.

[Chakrabarti98b] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In Procs. Int.Conf. of the ACM SIGMOD, pages 307–318, Seatle, WA, USA, 1998.

[Chakrabarti99]S. Chakrabarti, B.E. Dom, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the web's link structure. IEEE Computer, 32(8), August 1999.

[Chan00] P.K. Chan. Constructing web user profiles: A non-invasive learning approach. In *WEBKDD '99: Revised Papers from the International Workshop on Web Usage Analysis and User Profiling*, pages 39–55, London, UK, 2000. Springer-Verlag.

[Chen98] M.S. Chen, J.S. Park, and P.S. Yu. Efficient data miningfor path traversal patterns. *IEEE Trans. on Knowledge and Data Engineering*, 10(2):209–221, 1998.

[Coenen00] F. Coenen, G. Swinnen, K. Vanhoof and G. Wets, A framework for self adaptive websites: tactical versus strategic changes, Procs. in 4th PAKDD Pacific-Asia Conference, April, 1-6, 2000.

[Cooley99] R. Cooley, B. Mobasher and J. Srivastava, Data preparation for mining world wide web browsing patterns, Journal of Knowlegde and Information Systems, 1, 5-32, 1999.

[Dellmann04] F. Dellmann, H. Wulff, and S. Schmitz. Statistical analysis of web log files of a german automobile producer. Technical report, Fachoch schule M*u*ster, University of Applied Sciences, February 2004.

[Eirinaki03] M. Eirinaki and M. Vazirgannis. Web mining for web personalization. *ACM Transactions on Internet Technology*, 3(1):1–27, 2003.

[Feldman95] R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In *Procs Int. Conf. First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pages 112–117, Montreal, Canada, 1995.

[Fisher87] D.H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.

[Fisher96] D.H. Fisher. Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Inteligence Research*, pages 147–179, 1996.

[Flake02] G. Flake, S. Lawrence, C. Giles, and F Coetzee. Self-organization of the web and identification of communities. Computer, 35(3):66–71, 2002.

[Ford56] L.R. Ford, D.R Fulkerson, and R.L. Rivest. Maximal flow through a network.Canadian Journal of Mathematics, 8:399–404, 1956.

[Hahn00] U. Hahn and I. Mani. The challenges of automatic summarization. *IEEE Computer*, 33(11):29–36, 2000.

[Hay04] B. Hay, G. Wets, and K. Vanhoof. Mining navigation patterns using a sequence alignment method. *Knowledge and Information Systems*, 6(2):150–163, 2004.

[Hinneburg03] A. Hinneburg and D.A. Keim, Tutorial: Advances in clustering and applications, ICDM Int. Conf.", 2003.

[Honkela97] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Websom - self-organizing maps of document collections. In *In Proc. of Workshop on Self-Organizing Maps WSOM'97*, pages 310–315, 1997.

[Jang97] J. Jang, C. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing A Computational Approach to Learning and Machine Intelligence*. Prentice Hall, 1997.

[Joshi00] A. Joshi and R. Krishnapuram, On Mining Web Access Logs, Proc. of the 2000 ACM SIGMOD, 63-69, 2000.

[Joachims97] T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In Procs. of the Fifteenth Int. Conf. Joint Conf. on Artificial Intelligence, pages 770–775, 1997.

[Karypis99] G. Karypis, E.-H. E.H. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32(8):68–75, 1999.

[Kilfoil03] M. Kilfoil et al., Toward an Adaptive Web: The State of the Art and Science, Procs. Annual Conference on Communication Networks & Services Research, 119-130, 2003.

[Kleinberg99] J. M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, Journal of ACM, 46(5), pages 604-632, 1999.

[Kumar02] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web and social networks. Computer, 35(11):32–36, 2002.

[Koutri04] M. Koutri, N. Avouris, and S. Daskalaki. *Adaptable and Adaptive Hypermedia Systems*, chapter A survey on web usage mining techniques for web-based adaptive hypermedia systems. Idea Publishing Inc., Hershey, 2004.

[Kwon03] O. Kwon and J. Lee. Text categorization based on k-nearest neighbor approach for web site classification. Information Processing & Management, 39(1):25–44, 2003.

[Lewis92] D. D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In In Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval, pages 37--50, Kobenhavn, DK, 1992.

[Lu03] Z. Lu, Y.Y. Yao and N. Zhong, Web Intelligence, Springer-Verlag, Berlin,2003.

[Mani99] I. Mani and M.T. Maybury. *Advances in automatic text summarization*. MIT Press, Cambridge, Mass., 1999.

[Mika04] P. Mika. Social networks and the semantic web. In WI-2004 In Procs. Int. Conf. of the IEEE/WIC/ACM on Web Intelligence, pages 285–291. IEEE Press, 2004.

[McCallum98] A. McCallum and K. Nigam. A comparision of event models for naive bayes text classification. In *In AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[Mobasher00] B Mobasher, R. Cooley and J. Srivastava, Automatic personalization based on Web usage mining, Communications of the ACM, 43(8), 142-151, 2000.

[Mobasher01] B. Mobasher, B. Berendt and M. Spiliopoulou, Tutorial: KDD for Personalization, KDD Conference, 2001.

[Mortazavi01] B. Mortazavi-Asl. Discovering and mining user web-page traversal patterns. Master's thesis, Computing Science, Simon Fraser Univ., Canada, 2001.

[Nielsen99] J. Nielsen, User interface directions for the web, Communications of ACM, 42(1), 65-72, 1999.

[Ngu97] D.S.W. Ngu and X. Wu. Sitehelper: A localized agent that helps incremental exploration of the world wide web. *Computer Networks and ISDN Systems: The International Journal of Computer and Telecommunications Networking*, 29(8):1249–1255, 1997.

[Page98] L. Page, S. Bring, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Tech. rep., Computer Systems Laboratory, Stanford University, Stanford, CA, USA, 1998.

[Perkowitz98] M. Perkowitz and O. Etzioni, Adaptive Web Sites: Automatically Synthesizing Web Pages, In Procs. of the 15th National Conference on Artificial Intelligence, 727-732, 1998.

[Pierrakos03] D. Pierrakos, G. Paliouras, C. Papatheodorou and C. Spyropoulos, Web usage mining as a tool for personalization: A survey, User Modeling and User Adapted Interaction, 13(4): 311-372, 2003.

[Quinlan93] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[Runkler03] T. A. Runkler and J. Bezdek, Web Mining with Relational Clustering, Int. Journal of Approximate Reasoning, 32(2-3), 217-236, 2003.

[Salton75] G. Salton, A. Wong and C. S. Yang, A vector space model for automatic indexing, Communications of the ACM 18(11):613--620, 1975.

[Sebastiani99] F. Sebastiani. A tutorial on automated text categorisation. In Procs. of ASAI-99, 1st Argentinean Symposium on ArtificialIntelligence, pages 7–35, Buenos Aires, Argentina, 1999.

[Schwarzkopf01] E. Schwarzkopf. An adaptive web site for the um 2001 conference. In *Procs. UM2001 Workshop on User Modeling, Machine Learning and Information Retrieval*, pages 77–86, November 2001.

[Schutze95] H. Schutze, D. A. Hull, and J. O. Pedersen. A comparison of classifiers and document representations for the routing problem. In Proceedings of SIGIR-95,18th ACM International Conference on Research and Development in Information Retrieval, pages 229–237, Seattle, US, 1995.

[Spiliopoulou99] M. Spiliopoulou, L.C. Faulstich, and K. Winkler. A data miner analyzing the navigationalbehaviour of web users. In *Procs. of workshop on Machine Learning in User Modeling of the ACAI'99*, pages 588–589, 1999.

[Staab05] S. Staab, P. Domingos, P. Mika, J. Golbeck, L. Ding, T. Finin, A. Joshi, A. Nowak, and R.R. Vallacher. Social networks applied. IEEE Intelligent Systems, 20(1):80–93, 2005.

[Strehl00] A. Strehl and J. Ghosh. Value-based customer grouping from large retail datasets. In *In Procs. of SPIE Conf. on Data Mining and Knowledge Discovery*, pages 33–42, Orlando, Florida, USA, 2000.

[Srivastava00] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[Theodoridis99] S. Theodoridis and K. Koutroumba, Pattern Recognition, Academic Press, 1999.

[Velasquez03] J. D. Velásquez, H. Yasuda, T. Aoki, R. Weber and E. Vera, Using self organizing feature maps to acquire knowledge about visitor behavior in a web site", Lecture Notes in Artificial Intelligence, 2773(1): 951-958, 2003.

[Velasquez04a] J. D. Velásquez, H. Yasuda, T. Aoki and R. Weber, A new similarity measure to understand visitor behavior in a web site, IEICE Transactions on Information and Systems, Special Issues on Information Processing Technology for web utilization, E87-D(2): 389-396,2004.

[Velasquez04b] J. D. Velásquez, P.A. Estévez,H. Yasuda, T. Aoki and E. Vera, Intelligent Web Site: Understanding the Visitor Behavior, Lecture Notes in Computer Science, 3213(1): 140-147, 2004.

[Velasquez05a] J. D. Velásquez, R. Weber, H. Yasuda and T. Aoki, Acquisition and maintenance of knowledge for web site online navigation suggestions, IEICE Transactions on Information and Systems, E88-D(5):993–1003, May 2005.

[Velasquez05b] J. D. Velásquez, S. Ríos, A. Bassi, H. Yasuda and T. Aoki, Towards the identification of keywords in the web site text content: A methodological approach, International Journal of Web Information Systems", 1(1):11-15,2005.

[Willet88] P. Willet. Recent trends in hierarchical document clustering: a clisitcal review. *Information Processing and Management*, 24:577–597, 1988.

[Wong01] C. Wong, S. Shiu, and S. Pal. Mining fuzzy association rules for web access case adaptation. In *In Workshop on Soft Computing in Case-Based Reasoning Research and Development, Fourth Int. Conf. on Case-Based Reasoning (ICCBR 01)*, 2001.

[Xiao01] J. Xiao, Y. Zhang, X. Jia, and T. Li. Measuring similarity of interests for clustering web-users. In *ADC '01: Proceedings of the 12th Australasian conference on Database technologies*, pages 107–114, Washington, DC, USA, 2001.

[Yuan04] F. Yuan, H.Wu, and G. Yu. Web users classification using fuzzy neural network. *Lecture Notes in Computer Science*, 3213(1):1030–1036, 2004.