Université Libre de Bruxelles
Faculté des Sciences Appliquées
Département Informatique et Réseaux

Semantic Analysis in Web Usage Mining

Jean-Pierre Norguet

Promoteur: Pr. Esteban Zimányi

Thèse présentée en vue de l'obtention du grade académique de
Docteur en Sciences Appliquées
Année académique 2005-2006

Abstract

With the emergence of the Internet and of the World Wide Web, the Web site has become a key communication channel in organizations. To satisfy the objectives of the Web site and of its target audience, adapting the Web site content to the users' expectations has become a major concern. In this context, Web usage mining, a relatively new research area, and Web analytics, a part of Web usage mining that has most emerged in the corporate world, offer many Web communication analysis techniques. These techniques include prediction of the user's behaviour within the site, comparison between expected and actual Web site usage, adjustment of the Web site with respect to the users' interests, and mining and analyzing Web usage data to discover interesting metrics and usage patterns. However, Web usage mining and Web analytics suffer from significant drawbacks when it comes to support the decision-making process at the higher levels in the organization.

Indeed, according to organizations theory, the higher levels in the organizations need summarized and conceptual information to take fast, high-level, and effective decisions. For Web sites, these levels include the organization managers and the Web site chief editors. At these levels, the results produced by Web analytics tools are mostly useless. Indeed, most of these results target Web designers and Web developers. Summary reports like the number of visitors and the number of page views can be of some interest to the organization manager but these results are poor. Finally, page-group and directory hits give the Web site chief editor conceptual results, but these are limited by several problems like page synonymy (several pages contain the same topic), page polysemy (a page contains several topics), page temporality, and page volatility.

Web usage mining research projects on their part have mostly left aside Web analytics and its limitations and have focused on other research paths. Examples of these paths are usage pattern analysis, personalization, system improvement, site structure modification, marketing business intelligence, and usage characterization. A potential contribution to Web analytics can be found in research about reverse clustering analysis, a technique based on self-organizing feature maps. This technique integrates Web usage mining and Web content mining in order to rank the Web site pages according to an original popularity score. However, the algorithm is not scalable and does not answer the page-polysemy, page-synonymy, page-temporality, and page-volatility problems. As a consequence, these approaches fail at delivering summarized and conceptual results.

An interesting attempt to obtain such results has been the Information Scent algorithm, which produces a list of term vectors representing the visitors' needs. These vectors provide a semantic representation of the visitors' needs and can be easily interpreted. Unfortunately, the results suffer from term polysemy and term synonymy, are visit-centric rather than site-centric, and are not scalable to produce. Finally, according to a recent survey, no Web usage mining research project has proposed a satisfying solution to provide site-wide summarized and conceptual audience metrics.

In this dissertation, we present our solution to answer the need for summarized and conceptual audience metrics in Web analytics. We first described several methods for mining the Web pages output by Web servers. These methods include content journaling, script parsing, server monitoring,

network monitoring, and client-side mining. These techniques can be used alone or in combination to mine the Web pages output by any Web site. Then, the occurrences of taxonomy terms in these pages can be aggregated to provide concept-based audience metrics. To evaluate the results, we implement a prototype and run a number of test cases with real Web sites.

According to the first experiments with our prototype and SQL Server OLAP Analysis Service, concept-based metrics prove extremely summarized and much more intuitive than page-based metrics. As a consequence, concept-based metrics can be exploited at higher levels in the organization. For example, organization managers can redefine the organization strategy according to the visitors' interests. Concept-based metrics also give an intuitive view of the messages delivered through the Web site and allow to adapt the Web site communication to the organization objectives. The Web site chief editor on his part can interpret the metrics to redefine the publishing orders and redefine the sub-editors' writing tasks. As decisions at higher levels in the organization should be more effective, concept-based metrics should significantly contribute to Web usage mining and Web analytics.

Acknowledgements

The writing of this dissertation has been directed by professor Esteban Zimányi. His support, his comments, and his sense of balance made possible the realization of this work. I am also grateful for the patience he has deployed to educate me, not only to scientific research but also to criticism, and to another perception of reality. I also thank the members of my committee for their attention and comments. The committee was composed of Pr. Hugues Bersini as president, Pr. Roel Wuyts as secretary, Pr. Esteban Zimányi as dissertation director, Pr. Gianluca Bontempi as president of the accompaniment committee, and Pr. Marie-Christine Rousset from University of Grenoble as external expert. In particular, I want to acknowledge the brillance and mastering of Pr. Bersini in his role of committee president, the outstanding quality of the technical input provided by Pr. Bontempi in the pre-defense review process, the precision and quality of modification requests suggested by Pr. Wuyts, and the elogious committee report written by Pr. Rousset about my dissertation. I am very grateful to all of you.

During the writing of my dissertation, I have received valuable help and support from my colleagues. Jean-Michel Dricot has provided me with outstanding support and empathy in the hardest moments, as well as with his valuable knowledge in the scientific world. The natural of his scientific cast of mind, his generosity with his co-workers, and his strong intuition with respect to others' feelings make Jean-Michel Dricot an incredibly-valuable colleague and a precious element among the University members. Sabri Skhiri dit Gabouje has been a very close co-worker; we shared day-to-day difficulties and we provided each other mutual support. Olivier Samyn and Louis Jacomet helped me jumpstarting the work and contributed to the friendly atmosphere necessary to feel comfortable during the first workdays. Elzbieta Malinowski and Mohammed Minout have shared the psychological difficulties of finishing a PhD, and provided me with their technical help on data mining. Natascha Vanderheyden has been perfect at her secretary job, with a dedication that honours her. Marie-Ange Remiche, Joël Cannau, Pierre Stadnik, Johnny Tsheke Shele, Inès Gam, Stéphane Faulkner, and Stéphane Dehousse, have also been valuable co-workers, sharing support, thoughts, and comments. My casual coworkers Ueli Wahli, Jonas Andersen, Markus Meser, Nicole Hargrove, Benjamin Tshibasu-Kabeya, Bruno Pouliquen, and Ralf Steinberger have helped me in achieving great things during the writing of this dissertation. Gérard Materna contributed a valuable piece of work to the output page mining prototype mod_trace_output. Philippe Vincke and Yves De Smet helped with the page sampling issues. Marie-Paule Lefranc's post-doctoral position offer in her Montpellier immunogenetics lab has been a key driving inspiration and a strong motivation for achieving this dissertation. Meeting so many valuable persons during my doctoral period has been very inspiring.

I also have been continuously supported by my family and friends. My motivation to achieve this work would not have been so strong without the unanimous encouragements of Valérie Leclercq, Sylvie Suzor, Linda Tempels, Sophie Bourlon, Valérie Procès, Claire Garnier, Franck Seynave, Pascal Dormal, Jean-Claude Idée, Pascal Masset, Patrick Defauw, Claudia Ucros, Donatienne Croonenberghs, Henri Verbert, Nathalie Ardizzone, Magali Auquier, Jean-Michel Reniers, Brigitte Reniers, Jonathan Shin, Marie Navez, Alpesh Patel, Davy Paindaveine, Frank Vandenberghen, Frédéric Bourgeois, Jean-Louis Willems, Michel De Hennin, Serge Vidal, Carole Graf, Thierry Joly, Paolo Torsella, Anne Huybrechts, Michel Hardy, and Florence Bovesse. My roommates Elodie Manera, Silvia Cervantes, Raphaël Micheli, Stale Engen, Victor Fernandez Soriano, Sandrine Segalini, Verena Grossgasteiger, Cristina Valentini, and Teresa Calleja contributed on their part to the quietness and delightness of the atmosphere at home.

To all, thank you for your help in the achievement of this Ph.D. dissertation!

Contents

List of Figures