

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 brilliance 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 Tshibas-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 Bournon, 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 Re-

niers, Jonathan Shin, Marie Navez, Alpesh Patel, Davy Paindaveine, Frank Vandenberghe, 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

1	Introduction	1
1.1	Motivations	1
1.2	Scope and Goals	8
1.3	Web Usage Mining in the World Wide Web Evolution	9
1.3.1	Web Usage Mining Genesis in the Early Web	9
1.3.2	The Document Web	10
1.3.3	The Application Web	10
1.3.4	The Service Web	11
1.3.5	The Semantic Web	11
1.4	Related Research Domains	12
1.4.1	Web Mining	12
1.4.2	Web Usage Mining and Web Analytics	13
1.4.3	Web Content Mining	14
1.4.4	Web Structure Mining	14
1.4.5	Web Site Management	15
1.4.6	Web Application Development	15
1.4.7	Information Retrieval	16
1.4.8	Business Intelligence	16
1.5	Main Contributions of the Dissertation	16
1.6	Contents	18
2	Preliminaries	21
2.1	Web Communication Analysis Model	21
2.2	Web Data Flow	23
2.3	Transaction Metadata	24
2.4	Log Formats	25
2.5	Web Environment Data Sources	26
2.5.1	Client Level Collection	27
2.5.2	Network Monitoring	27
2.5.3	Server Monitoring	28
2.5.4	Server Level Collection	29
2.5.5	Proxycache Level Collection	29
2.6	Issues	29
2.6.1	Internal Hits	29
2.6.2	Data Volume	30
2.6.3	IP Multiplexing	30

2.6.4	HTTP Parameters	31
2.6.5	Web Services	31
2.6.6	Multimedia and Embedded Content	31
2.6.7	Privacy Issues	31
2.7	Summary	32
3	State of the Art	35
3.1	Web Usage Mining Overview	35
3.2	Web Analytics Software	38
3.2.1	Page-Based Metrics	38
3.2.2	Page-Group Metrics	39
3.2.3	Results Exploitation in Organizations	40
3.2.4	Data Analysis Process	43
3.3	Summary	43
4	Output Page Mining	45
4.1	Web Environment	45
4.2	Web Server Logging and Content journaling	46
4.3	Server Monitoring	49
4.4	Network Monitoring	50
4.5	Client-Side Mining	50
4.6	Summary	51
5	Lexical Analysis	53
5.1	Content Processing	53
5.1.1	Unformatting	54
5.1.2	Terms Tokenization	54
5.1.3	Stopword Removal	55
5.1.4	Stemming	56
5.1.5	Term Selection	56
5.2	Term-Based Metrics	56
5.2.1	Formalization	57
5.2.2	Term-Based Metrics	58
5.2.3	Computation	59
5.2.4	Evaluation	60
5.3	Vector Model Refinements	60
5.4	Page Visibility	61
5.5	Data Volume	61
5.5.1	Page Sampling	62
5.6	Summary	63
6	Semantic Analysis	65
6.1	Ontologies	65
6.2	Concept-Based Metrics	66
6.3	Ontology Sources	68
6.3.1	Thesauri	68
6.3.2	Databases	69

6.3.3	UML Models	70
6.3.4	Public Ontologies	70
6.4	Hierarchical Ontologies Durability	73
6.5	Representation Languages	74
6.5.1	XML	75
6.5.2	RDF	75
6.5.3	RDFS	77
6.5.4	OWL	80
6.6	Summary	81
7	Hierarchical Aggregation with OLAP	83
7.1	Business Intelligence and OLAP	83
7.2	The Multidimensional Model	84
7.2.1	Ontology Dimension	84
7.2.2	Time Dimension	85
7.2.3	Querying the Cube	86
7.3	Visualization	86
7.4	Weekly Cycles	88
7.5	More Dimensions	88
7.6	Summary	90
8	Implementation	91
8.1	Architecture	91
8.1.1	Daily Batch	93
8.1.2	Content journaling (WASA-CJ)	93
8.1.3	Consultation Analysis (WASA-CA)	96
8.1.4	Dynamic Traces (WASA-DC)	97
8.1.5	Lexical Analysis (WASA-LA)	97
8.2	Technologies	98
8.2.1	Programming Language	98
8.2.2	Web Interface	99
8.2.3	Database Connectivity	100
8.2.4	File Transfer	100
8.2.5	Stemming	101
8.2.6	Ontology Support	102
8.2.7	Log Line Parsing	104
8.3	WASA Framework	105
8.3.1	Persistence	105
8.3.2	Connection Management	106
8.3.3	Error Handling	107
8.3.4	Application Configuration	108
8.4	Summary	112
9	Experimental Results	113
9.1	Experimental Environment	113
9.2	Visualization	114
9.3	Stemming and Multiwords	115

9.3.1	Parameter Sets	116
9.3.2	Results and Discussions	117
9.3.3	Performance	118
9.4	Ontology Coverage	119
9.5	Scalability	119
9.6	Validation	121
9.7	Exploitation	122
9.8	Summary	124
10	Conclusions and Future Directions	125
10.1	Ontology Coverage	125
10.2	Weighted Taxonomy from Structured Document Set	129
10.3	General-Scope Public Corpora	130
10.4	Page Classification	131
10.4.1	Early Experiments	135
10.4.2	Observations	136
10.4.3	Conclusions	138
10.5	Electric Similarity Measure	138
10.6	Client-Side Behaviour	139
10.7	Multilingualism	140
10.8	Web Application Debugging	143
10.9	Summary	146
10.10	Conclusion	146
A	Operations in SQL Server OLAP	149
A.1	Importing the data into SQL Server	149
A.2	Defining the Cube	149
A.3	Cube Visualization in Microsoft Excel	151
B	Collaborative Ontology Enrichment Instructions	155
C	Stopwords	159
C.1	English Stopwords	159
C.2	French Stopwords	160

List of Figures

1.1	Organization diagram.	2
1.2	ULB Web site search page.	3
1.3	Search results for "ULB".	3
1.4	Search results for "RESULB".	4
1.5	A page about data mining.	4
1.6	Another page about data mining.	5
1.7	Different concepts on the ULB's homepage.	6
1.8	The ULB's homepage one month later.	7
1.9	Evolution of the World Wide Web.	9
1.10	Web mining domains.	13
2.1	The Shannon-Weaver general model.	21
2.2	Shannon-Weaver developed model.	22
2.3	Push diffusion.	23
2.4	Pull diffusion.	23
2.5	HTTP transaction.	24
2.6	URL/URI.	25
2.7	Web usage data gathering locations.	26
2.8	Proxycache.	30
3.1	Page views.	39
3.2	Content groups.	40
3.3	Content groups report.	40
3.4	Web analytics process lifecycle.	41
3.5	WebTrends report-viewing interface.	42
3.6	Report distribution means.	42
3.7	Data analysis process.	43
4.1	Data mining points in Web environment.	46
4.2	Content journal class diagram.	46
4.3	File trace model.	47
4.4	File trace state diagram.	47
4.5	Client Level Collection.	50
5.1	Content processing phases.	53
5.2	Eurovoc multiwords indexing.	55
5.3	Relations between corpora.	58

5.4	Static presence curve.	60
6.1	A simple ontology about food.	66
6.2	Food ontology concept hierarchy.	66
6.3	Example of audience metrics tree for terms of the food ontology (hits over presence).	68
6.4	Example of audience metrics for concepts of the food ontology (hits over presence, over interest, over accrued interest).	68
6.5	Biochamps.be ERD (UML format).	69
6.6	IMDB Movie table.	70
6.7	UML class diagram matching the food ontology.	71
6.8	Biochamps.be UML class diagram.	71
6.9	Knowledge model class-to-object representation.	71
6.10	UML model to ontology format transformation.	72
6.11	Hierarchical structure schema (RDFS).	78
6.12	Ontology Viewing in KAON.	79
7.1	Business Intelligence process.	84
7.2	Time and Ontology Dimensions.	84
7.3	Snow-flake dimension for the food ontology.	85
7.4	Parent-child dimension for the food ontology.	85
7.5	Time dimensions example.	86
7.6	Time dimension tables for days and weeks in SQL Server.	86
7.7	Expanding a concept in Microsoft Excel PivotTable to see detailed audience results.	87
7.8	Sample visualization of audience results over a time period.	87
7.9	Visualization of daily audience results (nervous variations).	88
7.10	Temporal curve set with integration of weekly cycles.	89
8.1	WASA system architecture.	92
8.2	Daily batch sequence of operations (overview).	94
8.3	WASA-CJ sequence of operations.	95
8.4	WASA-CA sequence of operations.	96
8.5	WASA-LA sequence of operations.	98
8.6	FTP, SFTP, and HTTP in the layered OSI model.	101
8.7	Stemming in Lucene.	101
8.8	Lucene stemmers.	102
8.9	Analysis ontology objects.	103
8.10	Food ontology element (OWL).	103
8.11	Importing partial food ontology from Jena statement list.	104
8.12	Access log line parsing.	105
8.13	WASA persistence framework.	106
8.14	Classical persistent OO software architecture.	106
8.15	Thread-based connection management system.	107
8.16	WASA exceptions hierarchy.	108
8.17	Application initialization in daily batch.	111
9.1	Experimental environment architecture.	114
9.2	Evolution of visitors' interest into the ACM classification top concepts on the cs.ulb.ac.be Web site during the 2003-2004 academic year.	115

9.3	Audience metrics for the top concepts in the ACM classification.	116
9.4	Term separation in ontology.	117
9.5	Term separation and stemming in ontology.	117
9.6	Stemming in ontology.	117
9.7	Parameter set summary chart.	119
9.8	Audience metrics for the enriched ACM classification top concepts.	120
9.9	Directory hits for wasa.ulb.ac.be during year 2003. Results obtained with Google Analytics.	122
9.10	Concept consultation of wasa.ulb.ac.be during year 2003. Results obtained with WASA.	122
9.11	Concept-based metrics exploitation life cycle.	123
10.1	Partial food ontology.	126
10.2	Food-selling Web site: vegetable list.	126
10.3	Food-selling Web site: fruits list.	127
10.4	Multidimensional model for weighted concept hierarchies (time dimension schematized).	130
10.5	Reduced multidimensional model for weighted concept hierarchies (time dimension schematized).	131
10.6	Classified page views hierarchically aggregated.	133
10.7	Adapted multidimensional model (the time dimension is schematized).	133
10.8	Reduced multidimensional model (the time dimension is schematized).	134
10.9	Multidimensional model for category-based hits with a content journal.	134
10.10	Experimental configuration.	135
10.11	Directory-based hits. Results obtained with Google Analytics.	136
10.12	Concept-based hits. Results obtained with WASA-SA.	137
10.13	Category-based hits. Results obtained with WASA-PC.	137
10.14	Initial electric circuit.	139
10.15	Improved electric circuit.	140
10.16	Multilingual text extraction algorithm.	142
10.17	Multilingual ontology (first levels of the food ontology; in English and in French).	142
10.18	Multilingual ontology integration.	142
10.19	Overview of the Rational Unified Process (excerpt from [56]).	144
10.20	Top URIs.	145
10.21	Top content groups.	145
A.1	Cube Editor in SQL Server.	150
A.2	Cube results.	151
A.3	Empty PivotChart.	153
A.4	Pie PivotChart.	153
A.5	Bar PivotChart.	153
A.6	Line PivotChart.	153
B.1	Web Browser and file explorer in collaboration.	157
B.2	Enriched ACM classification compared in UltraEdit text editor.	158