Engineering Letters, 14:1, EL_14_1_21 (Advance online publication: 12 February 2007)

A Page-Classification Approach to Web Usage Semantic Analysis

Jean-Pierre Norguet[1], Benjamin Tshibasu-Kabeya[2], Gianluca Bontempi[2], Esteban Zimányi[1]

[1]Department of Computer & Network Engineering
Université Libre de Bruxelles, CP 165/15
50 Avenue F.D. Roosevelt, 1050 Brussels, Belgium
e-mail: {jnorguet,ezimanyi}@ulb.ac.be

[2]Machine Learning Group, Département d'Informatique
Université Libre de Bruxelles, CP 212
Boulevard du Triomphe, 1050 Brussels, Belgium
e-mail: {btshibas,gbonte}@ulb.ac.be

Abstract

With the emergence of the World Wide Web, analyzing and improving Web communication has become essential to adapt the Web content to the visitors' expectations. Web communication analysis is traditionally performed by Web analytics software, which produce long lists of page-based audience metrics. These results suffer from page synonymy, page polysemy, page temporality, and page volatility. In addition, the metrics contain little semantics and are too detailed to be exploited by organization managers and chief editors, who need summarized and conceptual information to take high-level decisions. To obtain such metrics, we propose to classify the Web site pages into categories representing the Web site topics and to aggregate the page hits accordingly. In this paper, we show how to compute and visualize these metrics using OLAP tools. To solve the page-temporality issue, we propose to classify the versions of the pages using support vector machines. To validate our approach, we perform experiments on real data with SQL Server OLAP Analysis Service, the R statistical tool, and our prototype WASA-PC. Finally, we compare our results against directory-based metrics and concept-based metrics.

Keywords: World Wide Web, Web usage mining, data mining, machine learning, page classification.

1 Motivations and Related Work

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 [18]. 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 [9], 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 [21]. 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 [18]. A potential contribution to Web analytics was attempted with reverse clustering analysis [15], 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 is proposed in the IUNIS algorithm of the Information Scent model [3]. This algorithm 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 [4], no Web usage mining research project has proposed a satisfying solution to provide site-wide summarized and conceptual audience metrics.

In previous papers, we have shown how to obtain concept-based metrics by analyzing the text content output by Web servers. In this approach, we successively mine the output Web pages [11], count the term weights in the output and online Web pages [13], and aggregate the term-based metrics with respect to an ontology representing the Web site knowledge domain [12]. The resulting metrics give an indication of the visitors' consultation, presence, and interest into the Web site topics and subtopics. These metrics prove extremely summarized and conceptual but depend on the ontology coverage of the Web site domain knowledge. As ontology completion is a difficult operation to automate [6], Web sites for which a satisfying ontology cannot be produced manually or semi-manually need an alternate approach. In this paper, we propose to enrich the ontology with Web site pages instead of Web site terms.

The paper is structured as follows. In Section 2, we introduce the idea of classifying the Web site pages into a taxonomy representing the Web site topics. Then, we explain how to aggregate the page hits along the taxonomy in order to obtain category-based audience metrics. Then, we formalize these metrics and we explain how to compute them using OLAP tools. In Section 3, we describe our experiments on real data with SQL Server Analysis Service, the R statistical tool, and our prototype WASA-PC. We show some examples of queries and visualizations, and we validate the results against concept-based and directory-based metrics. Finally in Section 4, we discuss the limitations implied by page granularity and the possible improvements.

2 Category-Based Audience Metrics

For a given Web site to analyze, we choose a taxonomy or ontology that models the Web site knowledge domain. The taxonomy entries should represent the hierarchy of the Web site topics. For each topic in the taxonomy, we classify the Web site pages that fit into the corresponding category (Figure 1). As in most taxonomies, the terms are hierarchically linked by a relationship of type *part of*, *is a kind of*, or *is a* [19], the audience of the lower topics contributes to the communication of the upper topics. Therefore, and as the number of page hits can be retrieved from the Web site logs [12], category-based hits can be obtained by hierarchical aggregation of the page hits from the leaves up to the taxonomy root.

Category-based hits can be formalized as follows. For a mining period between days d_1 and d_2 and a given category C_i in the taxonomy, the number of hits for the C_i category is given by the following recursive expression, where C_j are the subcategories of C_i and p_{ij} are the pages classified into C_i:

$$\begin{aligned} \text{Hits}(C_i, d_1, d_2) \quad := \quad & \sum_{C_j} \text{Hits}(C_j, d_1, d_2) \\ & + \sum_{d=d_1}^{d_2} \sum_{p_{ij}} \text{Hits}(p_{ij}, d). \quad (1) \end{aligned}$$

Practically, hierarchical aggregation of the page-based metrics into category-based metrics can be computed and visualized using OLAP tools. The computation of

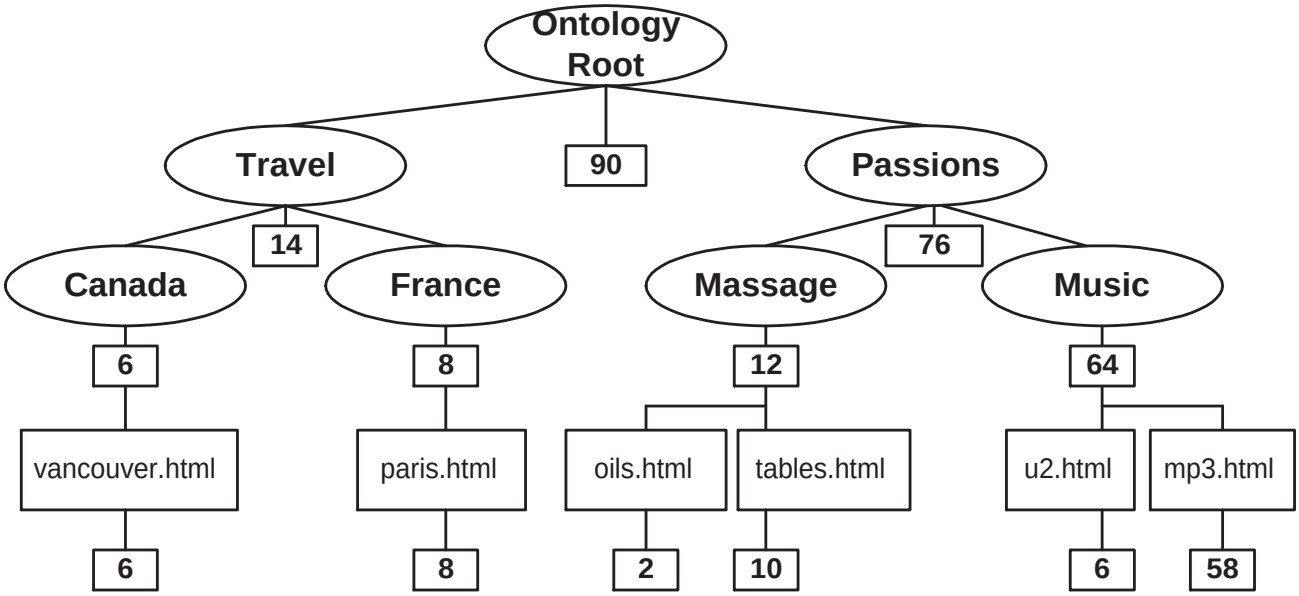

Figure 1: Classified Web pages in categories and page hits aggregation.

Equation 1 with OLAP tools requires a multidimensional model with two dimensions: Time and Taxonomy (Figure 2). The taxonomy dimension should be designed as a *parent-child dimension* to support taxonomies with any number of levels in each branch [8]. The time dimension, hereby schematized, can be designed from an aggregation of days, weeks, months, years, etc. [12]. The cube fact table must contain the daily page hits, which can be computed from the Web logs. The measure to define in the cube is the number of hits. After the cube has been introduced and processed in the OLAP tool, category-based hits can be extracted and visualized with any OLAP client, like Microsoft Excel PivotChart (see Section 3).

To take the page temporality into account, we use a *content journal* to keep track of the page content evolution [12]. Practically, a content journal records the history of the Web site pages, including the online periods and the publishing URIs. The analyzer can therefore retrieve from the content journal the content of any Web page sent to the client, based on the request datetime and URI. If the analysis period is long, classifying the content journal pages can be overwhelming. In this case, an automatic classifier can be used. Automatic classifiers require a training phase on an annotated document set. An example of document set can be the latest snapshot of the Web site pages. As the content of Web sites usually expands rather than contracts, this snapshot should ensure a good coverage of the knowledge domain. This should improve the classification of the content journal pages.

3 Experimentation

To test our approach, we developed a prototype called WASA (Figure 3). WASA stands for Web Audience Semantic Analysis and PC for Page Classification. The prototype implements content journaling, Web log parsing, daily page hits counting, and persistence storage into a temporary MySQL database. WASA is written in the Java language and is composed of more than 12,000 lines of code. Page classification is carried out by using a support vector machine algorithm [20] implementated by the R statistical tool [7]. The classification results provided by R are integrated into WASA-PC, which prepares the fact table and transfers it into SQL Server for OLAP analysis. In SQL Server OLAP Analysis Service, we introduce the OLAP cube of the multidimensional model described in Section 2. After cube processing, the metrics are aggregated and can be queried from Microsoft Excel to produce the various visualizations.

In our case study, we considered a Web site with three main topics: {computer science, travel, leisure}. The Web site contains about 300 pages and receives about 100 page requests a day. The Web site content is very rich, with an average of 1,000 words per page. The analysis period is one year, from January to De-

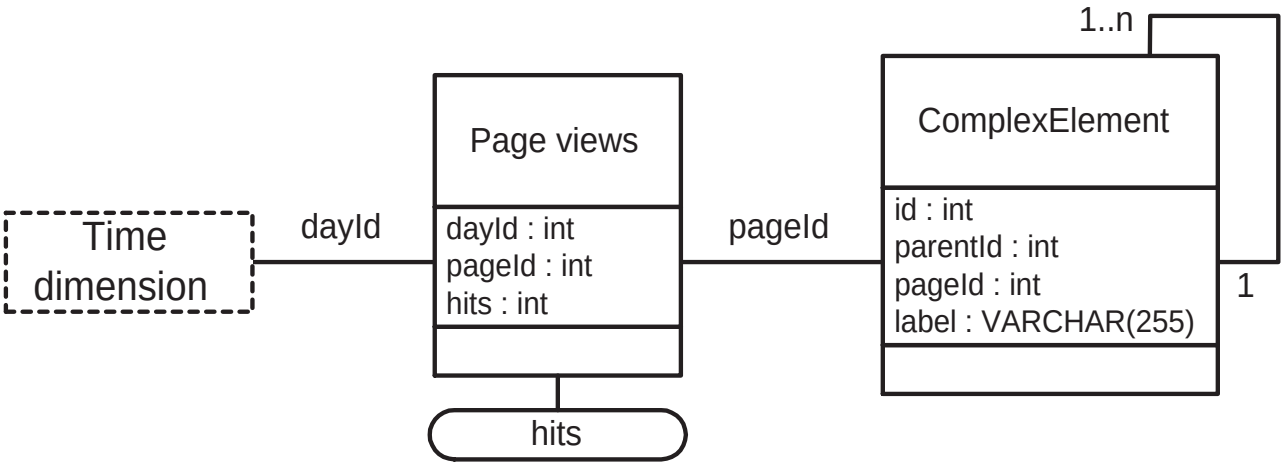

Figure 2: Multidimensional model for category-based hits.

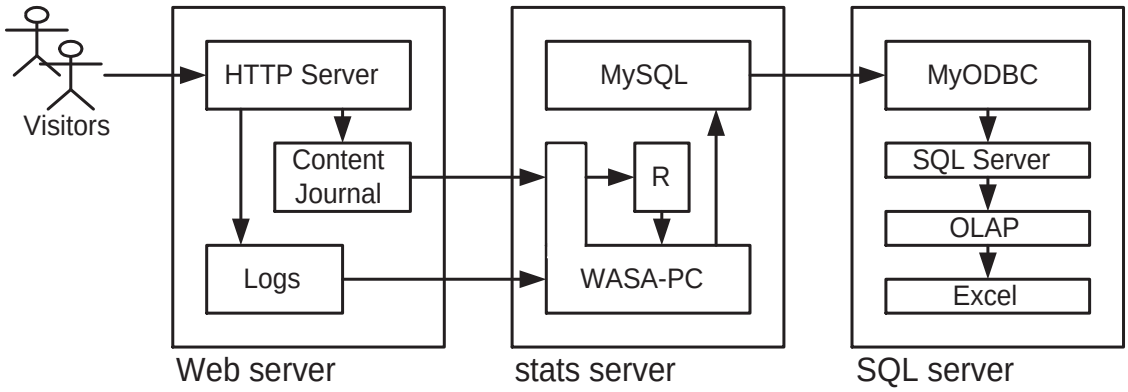

Figure 3: Experimental configuration.

cember. The content journal over the analyzed year contains 2,700 files. These files have been submitted to unformatting, tokenization, stopword-removal, stemming, and term-weighting operations [1], in order to feed the classification phase properly. To train the classifier, a 283-page snapshot of the latest content has been extracted from the content journal and each page has been manually classified into one of the following four categories: {computer science, travel, leisure, others}. The 13-fold cross validation shows an accuracy of 63%. The confusion matrix returned by R is given in Table 1. After the learning phase, the classifier has categorized the rest of the 2,700 content journal pages.

To validate our approach against existing software, we compared our results against Google Analytics, a popular Web analytics tool. Although WASA-PC and Google Analytics results are very different, there is a particular case of Web site where the Google Analytics results are comparable to the WASA-PC results. Indeed, if the Web site directories match the taxonomy topics, the hits by directories obtained by Google Analytics should be comparable to the hits by category obtained by WASA-PC.

To further verify the results, we also ran the test case with WASA-SA, which produces concept-based metrics from term hierarchies [13]. For the purpose of the test case, a custom taxonomy of 1,150 terms has been built manually. The taxonomy contains the main topics and subtopics of the Web site, including the three topics {computer science, travel, leisure}. Given the term-level precision of the results produced by WASA-SA, these will be considered as reference for the following comparisons.

Predictions	Real categories			
	Others	Computer science	Travel	Leisure
Others	21	0	2	0
Computer science	0	30	0	0
Travel	2	0	60	0
Leisure	8	6	11	143

Table 1: Confusion matrix for the cross validation of the 283-page training set.

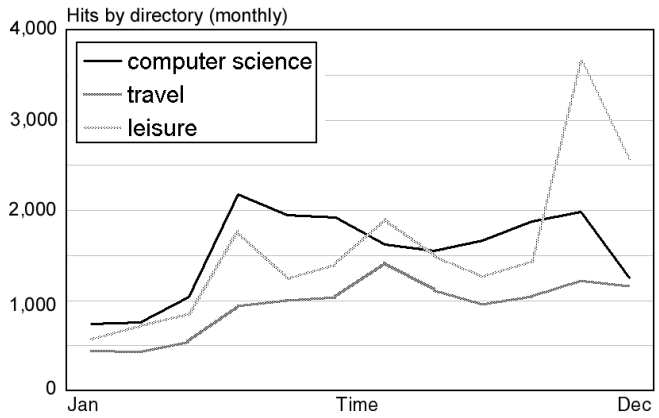

Figure 4: Directory-based hits. Results obtained with Google Analytics.

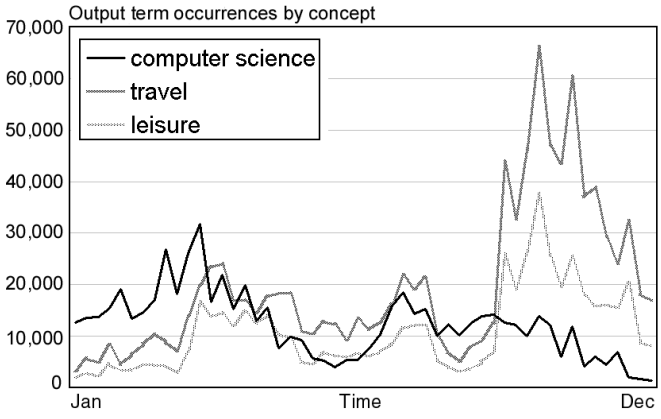

Figure 5: Concept-based hits. Results obtained with WASA-SA.

To compare these three result sets, we produced a directory-based graph with Google Analytics (Figure 4), a concept-based graph with WASA-SA (Figure 5), and a category-based graph with WASA-PC (Figure 6). The audience metrics are represented for the main three topics {computer science, travel, leisure}. The y-axis values are not shown as we are only interested in the relative values between the curves. The experimental conditions of the result sets are identical.

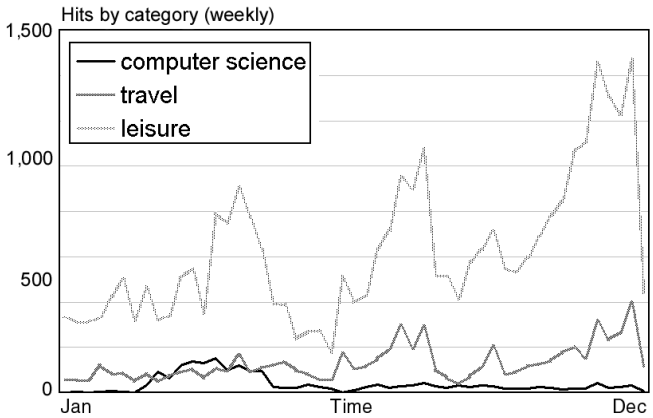

Figure 6: Category-based hits. Results obtained with WASA-PC.

By looking at the three graphs, we can see common peaks by the months of April, July, and November. According to the Web logs, these peaks can be explained by external events. The April peak is due to the referral link from a computer science online magazine. The July peak is due to the holiday consultation of leisure-travel reports. The November peak is due to the referral link from a music search engine. This similarity between the graphs, confirmed by the Web logs, shows the global validity of the three approaches.

During the first trimester, the computer science curve in Google Analytics shows lower values than in the WASA-SA graph. According to the Web logs, the difference is due to the first-trimester success of several computer science pages located outside the computer science directory. In this case, the directory granularity does not handle synonymous pages properly. As the structure of Web sites is generally rigid, page syn-

onymy is a major issue in directory-based metrics. In contrast, concept-based and category-based metrics are independent from the Web site structure and do not suffer from page synonymy.

During the November peak, the travel curve in Figure 5 is significantly higher than in Figure 4. This success of the travel concept can be explained by the number of world regions cited in the music pages. This difference between the graphs shows the limitation implied by page polysemy and the superiority of term granularity with respect to this aspect. In the case of category-based metrics, page polysemy can be solved by multiple classification [14].

Finally, in Figure 6, we can see that the three curves are similar in shape as in the other graphs, but are placed at different levels. In particular, the computer science and travel curves are placed lower, while the leisure curve is placed higher. By looking back at the confusion matrix in Table 1, we can see that several computer-science and travel pages have been classified in the leisure category. These misclassifications decrease the computer science and travel hits while increasing the leisure hits. The level differences in the curves are as important as the misclassified pages have received many hits. The classification of popular pages is therefore critical for the quality of the results.

4 Conclusions and Future Work

In this paper we presented our solution to answer the need for summarized and conceptual audience metrics in Web analytics. We first described the classification of pages into ontology entries representing the Web site topics. We have seen that hierarchical aggregation of the page hits along the ontology taxonomical structure can provide topic-based audience metrics. Then, we have shown how to compute the hierarchical aggregation using OLAP tools. For this computation, we have presented a multidimensional model that includes the time dimension as well as a parent-child taxonomy dimension that is able to represent page classifications with any number of levels in each branch. The result is a number of audience metrics for each of the Web site topics represented in the taxonomy. To validate our approach, we have run experiments on real data with SQL Server OLAP Analysis Service, the R statistical tool, and our prototype WASA-PC. Finally, we have compared our category-based metrics with the directory-based metrics obtained with Google Analytics, and the concept-based metrics obtained with WASA-SA.

Given their precision, the concept-based metrics obtained with WASA-SA have been taken as reference. These metrics require an ontology that contains all the significant terms of the Web site; such an ontology is in most cases not available. In these cases, directory-based and category-based metrics are possible alternatives. To be valid, directory-based metrics require the Web site structure to precisely match the Web site topics. In addition, they suffer from page synonymy and page polysemy, and do not solve the page-temporality and page-volatility issues. Category-based metrics do not suffer from these limitations and do not depend on ontology availability, but are very sensitive to misclassifications. However, as the training set is representative of the Web site knowledge domain and is extremely similar to the test set, the misclassification ratio can be limited significantly by adapting the classification process.

To improve the classification process, our future work will consider decision-tree classifiers [10] and ontology-based reasoning [5]. The combination of these techniques should be better adapted to the particularities of the training and test sets than support vector machines. Also, we anticipate that category-based metrics would encounter several limitations before their wide adoption. These limitations include page polysemy, term polysemy, training set availability, and data volume in high-traffic dynamic Web sites. In our future work, we will therefore consider as respective insights: multiple classification [14], word sense disambiguation [17], classifier optimization using external training sets [2], and statistical inference from page samples [16].

Besides these limitations, category-based metrics — like the other topic-based metrics — prove extremely summarized and conceptual. As a consequence, category-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. Category-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 the editors' writing tasks. As decisions at higher levels in the organization are more effective, category-based metrics should significantly contribute to extending Web analytics results.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *VLDB J.*, 7(3):163–178, 1998.

[3] E. H. Chi, P. Pirolli, K. Chen, and J. E. Pitkow. Using information scent to model user information needs and actions and the web. In *Proc. of the SIGCHI on Human Factors in Computing Systems*, pages 490–497, 2001.

[4] F. M. Facca and P. L. Lanzi. Mining interesting knowledge from weblogs: a survey. *Data Knowl. Eng.*, 53(3):225–241, 2005.

[5] H. Johan, D. Perrotta, R. Steinberger, and A. Varfis. Document classification and visualisation to support the investigation of suspected fraud. In *Proc. of the 4th European Conf. on Principles and Practice of Knowledge Discovery in Databases, PKDD*, 2000.

[6] A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

[7] J. Maindonald and J. Braun. *Data Analysis and Graphics Using R*. Cambridge University Press, Cambridge, 2003. ISBN 0-521-81336-0.

[8] E. Malinowski and E. Zimányi. OLAP hierarchies: A conceptual perspective. In *Proc. of the 16^{th} Int. Conf. on Advanced Information Systems Engineering, CAiSE'04*, LNCS 3084, pages 477–491. Springer-Verlag, 2004.

[9] J.G. March, H.A. Simon, and H.S. Guetzkow. *Organizations*. Cambridge Mass. Blackwell, 2nd edition, 1983.

[10] T. M. Mitchell. *Machine Learning*. McGraw-Hill Higher Education, 1997.

[11] J. P. Norguet and E. Zimányi. Topic-based audience metrics for internet marketing by combining ontologies and output page mining. In *Proc. of the Int. Conf. on Intelligent Agents, Web Technology and Internet Commerce, IAWTIC*. IEEE Computer Society, 2005.

[12] J. P. Norguet, E. Zimányi, and R. Steinberger. Improving web sites with web usage mining, web content mining, and semantic analysis. In *Proc. of the 32nd Int. Conf. on Current Trends in Theory and Practice of Computer Science, SOFSEM*. Springer-Verlag, 2006.

[13] J. P. Norguet, E. Zimányi, and R. Steinberger. Semantic analysis of web site audience. In *Proc. of the 21th ACM Symposium on Applied Computing, SAC*. Assoc. for Computing Machinery, 2006.

[14] A. M. Ráez, L. A. Ureña López, and R. Steinberger. Adaptive selection of base classifiers in one-against-all learning for large multi-labeled collections. In *Proc. of the 4th Int. Conf. on Advances in Natural Language Processing, EsTAL*, pages 1–12, 2004.

[15] S. A. Ríos, J. D. Velásquez, E. S. Vera, H. Yasuda, and T. Aoki. Using SOFM to improve web site text content. In *Proc. of the 1st Int. Conf. on Advances in Natural Computation, ICNC, Part II*, pages 622–626, 2005.

[16] V.K. Rohatgi. *An Introduction to Probability Theory and Mathematical Statistics*. John Wiley & Sons, 1976.

[17] M. Sanderson. Word sense disambiguation and information retrieval. In *Proc. of the 17th Int. Conf. on R&D in IR, SIGIR*, pages 142–150, 1994.

[18] J. Srivastava, R. Cooley, M. Deshpande, and T. Pang-Ning. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2), 2000.

[19] G. Stumme and A. Maedche. FCA-MERGE: Bottom-up merging of ontologies. In *Proc. of the 17th Int. Joint Conf. on Artificial Intelligence, IJCAI*, pages 225–234, 2001.

[20] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, 1995.

[21] U. Wahli, J.P. Norguet, J. Andersen, N. Hargrove, and M. Meser. *Websphere Version 5 Application Development Handbook*. IBM Press, 2003.