

Category-Based Audience Metrics for Web Site Content Improvement using Ontologies and Page Classification

Jean-Pierre Norguet¹, Benjamin Tshibas-Kabeya²,
Gianluca Bontempi², Esteban Zimányi¹

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 automatic classifiers.

1 Motivations and Related Work

With the emergence of the Internet, Web sites have become key communication channels in organizations. To satisfy the objectives of the Web site, 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 [13]. However, Web usage mining and

Web analytics suffer from significant drawbacks when it comes to support the decision-making process at higher levels in the organization.

Indeed, according to organizations theory [6], 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 [15]. 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 [13]. A potential contribution to Web analytics was attempted with reverse clustering analysis [10], 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 [2]. 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 [3], no Web usage mining research project has proposed a satisfying solution to provide site-wide summarized and conceptual audience metrics.

In this paper, we propose to automatically classify the Web site pages into ontology categories. 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. Finally, we formalize these metrics and we explain how to compute them using OLAP tools.

2 Category-Based Audience Metrics

Given a 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* [14], the audience of the lower topics contributes to the communication of the upper topics. As the number of page hits can be retrieved from the Web site logs [8], category-based hits can be obtained by hierarchical aggregation of the page hits from the leaves up to the taxonomy root.

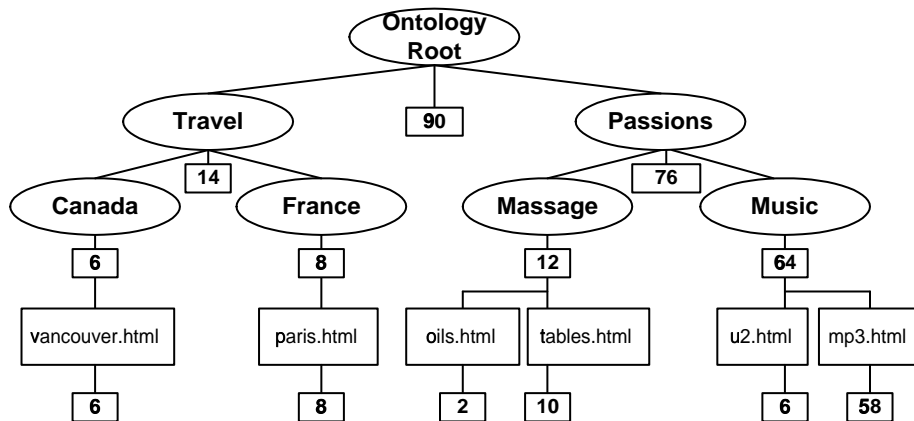


Fig. 1. Classified Web pages in categories and page hits aggregation.

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 :

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

Practically, hierarchical aggregation of the page-based metrics into category-based metrics can be computed and visualized using OLAP tools. The computation of 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 [5]. The time dimension, hereby schematized, can be designed from an aggregation of days, weeks, months, years, etc. [8]. 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.

To take the page temporality into account, we use a *content journal* to keep track of the page content evolution [8]. Practically, a content journal records the history of the Web site pages, including the online periods and the publishing

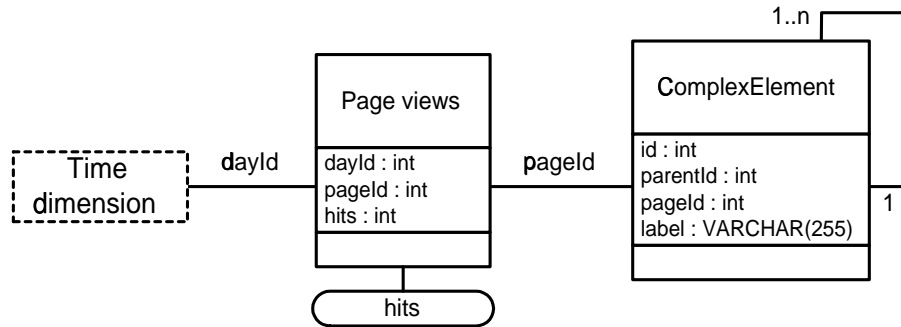


Fig. 2. Multidimensional model for category-based hits.

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 Conclusions and Future Work

As category-based metrics are summarized and conceptual, they 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.

The main limitation of category-based metrics is their dependency on proper page classification. To improve the classification process, our future work will consider other classification techniques like ontology-based reasoning [4] and decision-tree classifiers [7]. 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 [9], word sense disambiguation [12], classifier optimization using external training sets [1], and statistical inference from page samples [11].

References

1. 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.
2. 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.
3. F. M. Facca and P. L. Lanzi. Mining interesting knowledge from weblogs: a survey. *Data Knowl. Eng.*, 53(3):225–241, 2005.
4. 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.
5. E. Malinowski and E. Zimányi. OLAP hierarchies: A conceptual perspective. In *Proc. of the 16th Int. Conf. on Advanced Information Systems Engineering, CAiSE'04*, LNCS 3084, pages 477–491. Springer-Verlag, 2004.
6. J.G. March, H.A. Simon, and H.S. Guetzkow. *Organizations*. Cambridge Mass. Blackwell, 2nd edition, 1983.
7. T. M. Mitchell. *Machine Learning*. McGraw-Hill Higher Education, 1997.
8. 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.
9. 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.
10. 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.
11. V.K. Rohatgi. *An Introduction to Probability Theory and Mathematical Statistics*. John Wiley & Sons, 1976.
12. 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.
13. 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.
14. 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.
15. U. Wahli, J.P. Norguet, J. Andersen, N. Hargrove, and M. Meser. *Websphere Version 5 Application Development Handbook*. IBM Press, 2003.