

Improving Web Sites with Web Usage Mining, Web Content Mining, and Semantic Analysis

Jean-Pierre Norguet¹, Esteban Zimányi¹, and Ralf Steinberger²

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

² European Commission – Joint Research Centre
Via E. Fermi 1, T.P. 267
21020 Ispra (VA), Italy
email: Ralf.Steinberger@jrc.it

Abstract. With the emergence of the World Wide Web, Web sites have become a key communication channel for organizations. In this context, analyzing and improving Web communication is essential to better satisfy the objectives of the target audience. Web communication analysis is traditionally performed by Web analytics software, which produce long lists of audience metrics. These 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 decisions. Our solution to obtain such conceptual metrics is to analyze the content of the Web pages output by the Web server. In this paper, we first present a list of methods that we conceived to mine the output Web pages. Then, we explain how term weights in these pages can be used as audience metrics, and how they can be aggregated using OLAP tools to obtain concept-based metrics. Finally, we present the concept-based metrics that we obtained with our prototype WASA and SQL Server OLAP tools.

1 Introduction

The ease and speed with which information exchange and business transactions can be carried out over the Web has been a key driving force in the rapid growth of the Web and electronic commerce. In this context, improving Web communication is essential to better satisfy the objectives of both the Web site and its target audience, and Web usage mining [17], a relatively new research area, has gained more attention. The strategic goals of Web usage mining are prediction of the user's behaviour within the site, comparison between expected and actual Web site usage, and adjustment of the Web site with respect to the interests of its users. Web analytics [19] is the part of Web usage mining that has the most emerged in the corporate world. Web analytics focuses on improving Web communication by mining and analyzing Web usage data to discover interesting metrics and usage patterns.

From the huge amount of usage data collected by Web servers, Web analytics software produce many detailed reports. The usefulness of these reports depends on the report viewers in the organization. While Web designers are interested in detailed reports, organization managers are only interested in summary dashboards that show the number of visitors and a list of the most viewed pages, and the Web site chief editor needs concept-based results to redefine the publishing rules. In addition, the temporal evolution of the Web site content and the volatility of the scripted pages are not considered by Web analytics software.

To solve these issues, our approach aims at analyzing the content of the Web pages output by the Web server in order to obtain concept-based metrics. In Section 2, we present a list of methods that we conceived to mine the output pages, whatever the Web site technologies. In Section 3, we describe the content processing that we apply to the pages. From the term occurrences in the pages, we define the term-based consultation and we discuss some results obtained with our prototype WASA. Then, we group the terms into meaningful concepts using the concept hierarchies of ontologies. By the means of hierarchical aggregation, we define a set of concept-based metrics and we compute them with OLAP tools. In Section 4, we present and discuss the results obtained with our prototype WASA and SQL Server OLAP. In Section 5, we describe the results exploitation process. In Section 6, we expose the limitations of the metrics and our future work. Finally, in Section 7 we discuss how our approach compares with related work and we conclude in Section 8.

2 Output Page Mining

The first step in our approach is to mine the Web pages that are output by the Web server. We have conceived a number of methods, each of them being located at some point in the Web environment.

- In the Web server, log files can be coupled to a content journal that stores the evolution of the Web site content.
- In the Web server, a plugin can store the pages after they have been sent to the browser.
- On an Ethernet wire, a network monitor can capture the TCP/IP packets and reassemble the Web pages.
- On the client machine, an embedded program can run inside the page and send the content to a mining server.

This makes a number of mining methods that can be used alone or in combination. Each method has its advantages. Log file parsing combined with content journaling is easy to setup, runs in batch, and offers good performance. Dynamic Web sites require the use of a Web server plugin, a network monitor, or a client-side miner. Web server plugins are usually installed in secure Web sites, and network monitors elsewhere because of the lower risk. For the pages composed on the client-side, like XML/XSL pages, a client-side miner is required. As far as we can see, this set of methods can mine the pages output from any Web site.

3 Concept-Based Audience Metrics

Once the output pages have been mined, they can be processed in order to extract meaningful content. This processing is well-known in information retrieval [2]. Content processing includes unformatting, tokenization, stopword removal, stemming, and term selection. From P_d the set of Web pages mined during day d , content processing ultimately produces a list of stemmed terms s_i that appear with a frequency which we call $Consultation(s_i, d)$. The consultation represents the number of times the term s_i has been displayed on visitors' screens during day d . To neutralize the fluctuation of the metrics along time, the consultation can be divided by the total number of pages views. In this sense, term-based consultation is similar to the term frequency in the vector model [16].

To experiment this notion of consultation, we developed a prototype called WASA.³ We ran WASA for the academic year 2003-2004 on our department's Web site cs.ulb.ac.be, which contains about 2,000 Web pages and receives an average of 100 page requests a day. The result is a list of 30,000 terms and their daily consultation. Term-based consultation is very promising but the list of terms is too long and suffers from polysemy and synonymy. These observations call for the grouping of terms into meaningful concepts.

The main difficulty in grouping the terms is to define groups that match semantic fields for the human mind. Such groups can be found in *ontologies* [7]. If we define an ontology $\Omega := (S, r_0, R, \sigma)$ where S is a set of terms, r_0 is a partial order relation on S , R is a set of relation names, and $\sigma \rightarrow \mathcal{P}(S \times S)$ is a function, then meaningful term groups can be found in the hierarchy obtained by restricting the ontology (S, r_0, R, σ) to (S, r_0) [20]. For each term s_i in S , we define the associated concept C_i as the aggregation of the term s_i and its subterms s'_j in the hierarchy: $C_i := \{s_i\} \cup \{\dots, s'_j, \dots\}$. The consultation of a concept is the sum of the consultation of the term and of the consultation of the subterms:

$$Consultation(C_i, d) := Consultation(s_i, d) + \sum_{s'_j} Consultation(s'_j, d) \quad (1)$$

If a term is a leaf in the hierarchy, it has no subterms and therefore $C_i = \{s_i\}$. In this case, $Consultation(C_i, d) = Consultation(s_i, d)$. As the term consultation is known, the consultation of the concepts can be recursively aggregated from the leaf terms up to the root. Similarly, we define the presence of a concept by adding the frequency of the terms and of the subterms in the online Web pages during day d :

$$Presence(C_i, d) := Presence(s_i, d) + \sum_{s'_j} Presence(s'_j, d), \quad (2)$$

with $Presence(s_i, d) = \int_d Presence(s_i, t) dt$. The interest into a concept is defined as the division of the two:

$$Interest(C_i, d) := \frac{Consultation(C_i, d)}{Presence(C_i, d)} \quad (3)$$

³ WASA stands for Web Audience Semantic Analysis

Recursive aggregation of the term-based metrics into concept-based metrics can be computed by OLAP tools. Our multidimensional model (OLAP cube) is represented in Figure 1. The notation used in the figure was introduced in [9]. In our cube, we define two dimensions: Time and Ontology. The time dimension has two important levels: Week and Day. Metrics by week neutralize the weekly patterns, which contain insignificant information. More levels can be added depending on the needs (year, months, quarters, ...). The ontology dimension is modeled as a *parent-child dimension* to support ontologies with any number of levels. Other dimensions could be added like physical geography, site geography, Web geography, pages, users, internal referrers, external referrers, or other variations of the time dimension. The cube fact table contains daily term consultation and presence, which are provided by our prototype WASA. The cube measures are consultation, presence, and interest, where the interest measure is a calculated member defined as the division of the first two measures.

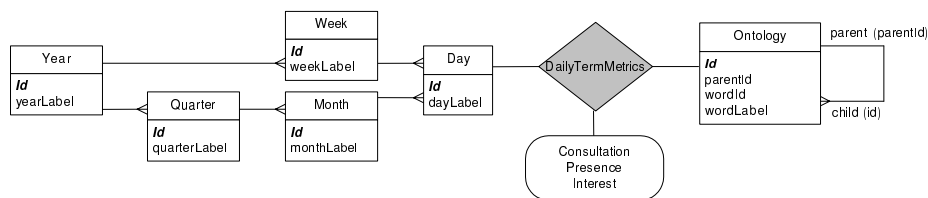


Fig. 1. OLAP cube with two dimensions: Time and Ontology.

4 Experimentation

To test our approach, we introduced our cube into SQL Server, along with the audience data computed by our prototype WASA for our department's Web site and the ACM classification. After cube processing, queries can be formulated on any combination of dimensions and measures. For example, if we display the ontology dimension vertically and the metrics horizontally, we can expand the concepts to see detailed results of the subconcepts (Figure 2). The cube can be queried and browsed with the SQL Server built-in module, from a Microsoft Excel PivotTable, or from any OLAP client like Mondrian/JPivot. With Microsoft Excel, we can produce a variety of charts to visualize cube-queried results. For example, we produced a multi-line chart where each curve represents the visitors' consultation of the top ACM concepts (Figure 3). This chart is easy to relate to the problem domain. For example, Computing Methodologies, Software, and Information Systems rank in the top, as many students follow these courses. Also, a peak of interest in Theory of Computation can be observed at the beginning of the academic year, when the 1st-year students starts following the corresponding course in the computers room. Finally, the average consultation falls down during the various periods of examination: August-September,

Level 02		Data		
- (Level 03)		Consultation	Presence	Interest
Computer Applications		2708.459184	583792.6282	46.39419981
Computer Systems Organization		12817.90958	2931785.46	43.72048964
Computing Methodologies		22518.35586	10057223.39	22.39023136
Computing Milieux		11996.16535	3672462.928	32.66517751
Data		6579.186259	1133266.697	58.05505689
General Literature		336.0899529	70145.62587	47.91317331
Hardware		5241.491734	1446859.252	36.23169539
Information Systems		22756.03891	5449485.45	41.75814235
	Database Management	5800.808073	950737.2194	61.01376863
	Information Interfaces and Presentation	5743.398621	2150755.4	26.70410416
	Information Storage and Retrieval	2400.738092	1287511.649	18.64632521
	Information Systems Applications	1700.907825	350728.1293	48.49647584
	Models and Principles	38401.88853	10189217.85	37.68875011
Information Systems Total *		7684.084683	1502851.178	51.13004385
Mathematics of Computing		18059.03443	3586113.109	50.64066641
Software		10783.28339	1674443.972	64.39918904
Theory of Computation		137125.9489	36827962.09	37.23419412
Grand Total *				

Fig. 2. Browsing the ACM classification and associated metrics in SQL Server.

January-February, and May-June. We can also produce a bar chart representing

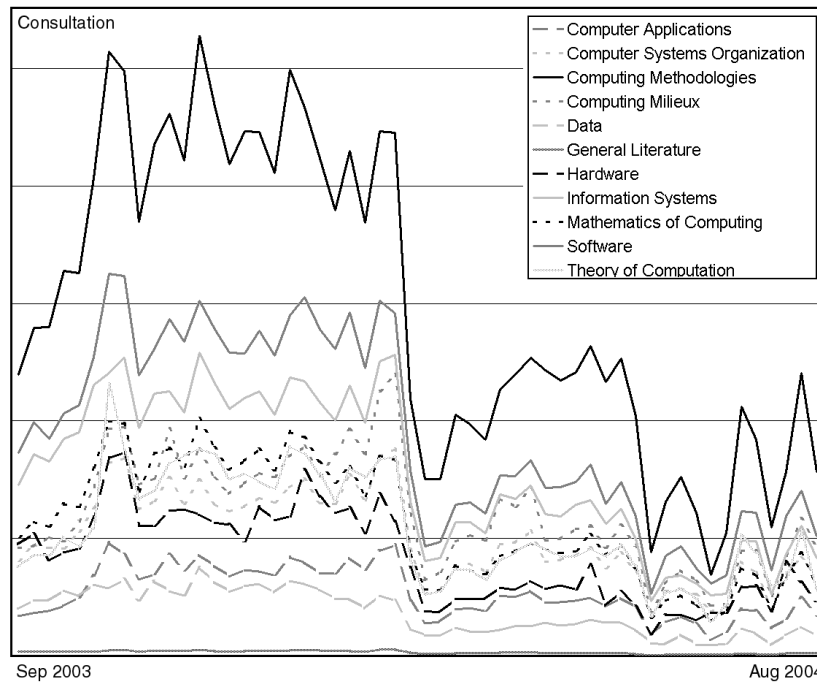


Fig. 3. Consultation of the ACM classification top concepts on the cs.ulb.ac.be Web site during the 2003-2004 academic year.

the various metrics for each of the top concepts. This kind of chart allows to compare the metrics of the various concepts, as well as the different metrics together. For example, we produced a chart for our department's Web site and the

ACM classification (Figure 4). The top 3 consulted concepts are: (1) Information Systems, (2) Computing Methodologies, and (3) Software. However, these concepts are major topics in the Web site, which is confirmed by high presence values. Therefore, high consultation values are not representative of the visitors' interest, what is indicated by low interest values. The top 3 concepts of interest are: (1) Theory of Computation, (2) Data, and (3) Mathematics of Computing. We can see that the ranking of the concepts can dramatically change according to the considered metrics, and that these should be interpreted carefully.

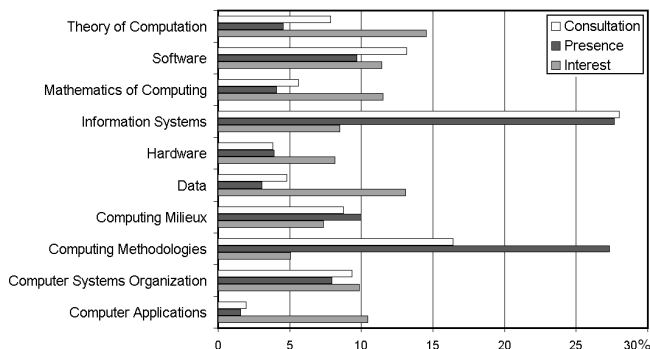


Fig. 4. Consultation, presence and interest metrics for the top concepts in the ACM classification.

To test the influence of the ontology on concept-based metrics, we ran WASA on our (computer science) department's Web site with two ontologies: Eurovoc, the European Commission's thesaurus, and the ACM classification. Eurovoc knowledge domain is extremely generic, from sociology to science, while the ACM classification knowledge domain is focused on computer science. With more than 5 times less terms than Eurovoc, the ACM classification covers much better the Web site knowledge domain. This coverage can be quantified by the percentage r_{Ω} of the ontology terms that appear in the output pages:

$$r_{\Omega} := \frac{\text{card}(S \cap P_{max})}{\text{card}(S)} \quad (4)$$

where P_{max} is the set of distinct terms in the output pages mined during the maximal period of time. For our department's Web site, the ACM classification coverage is 16% while the Eurovoc coverage is only 0.75%. This indicates how the meaning of the results improves with the ontology coverage of the Web site knowledge domain. A similar problem with Eurovoc has been observed in [18].

5 Exploitation

As concept-based metrics are extremely intuitive, they can be exploited at the highest levels of the organization, in order to take more effective decisions [10].

As concept-based metrics target different roles than classical Web analytics software, the exploitation process must be re-organized. With concept-based metrics, the chief editor and the sub-editors define the concepts relevant to the Web site knowledge domain. The tool administrator encodes these concepts into WASA, which generates the concept-based metrics reports. These reports are distributed to the organization manager and to the chief editor. With concept-based metrics, the organization manager is provided with an intuitive view of what messages are delivered through the Web site. He can then redefine the organization strategy according to the visitors' interests, adapt the other communication channels, and eventually request the chief editor to better adapt the Web site communication to the organization objectives. The chief editor on his part can redefine the publishing orders, dispatch the reports to the sub-editors, and redefine the writing tasks (Figure 5).

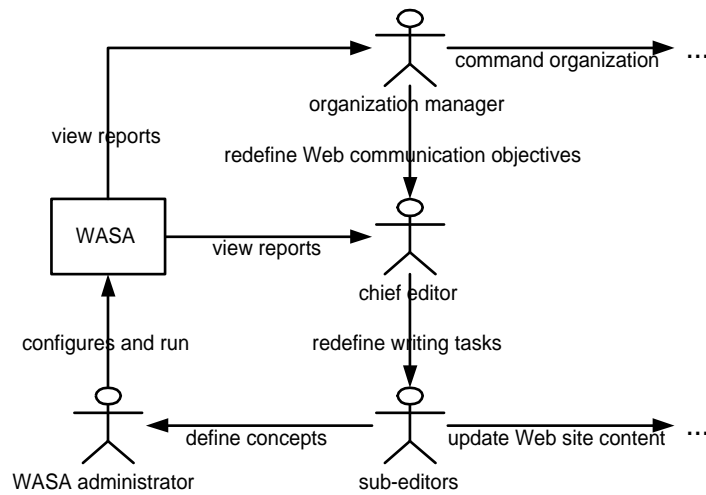


Fig. 5. Concept-based metrics exploitation life cycle.

6 Future Work

Our future work will aim to study the benefits of improving ontology coverage. First, we will evaluate a manual approach. The researchers in our department will enrich the ACM classification with terms of the department's Web site. Each researcher will browse the Web pages under his/her responsibility and select the most relevant terms of his domain knowledge. At the end, the chief editor will validate the enrichments. This method will ensure an optimal improvement of the ontology coverage, the effect of which will be evaluated by running WASA on

the enriched ontology. Furthermore, this manual enrichment will be compared against automatic and semi-automatic techniques.

The results obtained by our approach will be validated against WebTrends in a particular case of Web site where the results obtained by WebTrends should be comparable to those obtained by WASA. Indeed, if the Web site directories match the ontology concepts, the hits by directories obtained by WebTrends should be comparable to the interest by concept obtained by WASA.

Although the complexity of our algorithms are linear, we will test the scalability of our prototype WASA on our university's Web site,⁴ which contains a very high number of pages (about 50,000) and receives a very high number of page requests (about 200,000 a day).

Finally, variations of the metrics inspired from the vector model [16], as well as evaluators for ontology coverage of Web site knowledge domain [8], should be experimented.

7 Related Work

In the recent years, Web analytics software have shown little evolution. The most interesting feature introduced is page grouping with respect to a concept. For example, in the most popular Web analytics tool WebTrends, the pages can be grouped into *content groups*, which can be defined either by enumeration or regular expressions over the URI [21]. In the subsequent content-groups report, WebTrends shows the score of each content group, computed by aggregating the hits of the composing pages. The report is more intuitive than the page-views report, but the quality of the results depends on the page-grouping operation, which is not assisted by the software. Also, the temporal evolution of the pages remains ignored. Finally, content groups are groups of entire pages, with no finer-grained data units. Another attempt in the corporate world to consider more semantics has been to map back-end product data to an id parameter in URLs, like in IBM Tivoli Web Site Analyzer [12]. However this solution is limited to e-commerce Web sites and remains site specific.

In the research world, the closest approach to ours is reverse clustering analysis [15], an algorithm based on self-organizing feature maps. This algorithm integrates Web Usage Mining and Web Content Mining by integrating Web logs and Web site text. The result of this integration is a list of pages representing the most popular Web pages in the site. The pages are prioritized with regard to a score called "square vicinity". Although the results help to improve the content of a Web site, the approach suffers from a list of drawbacks. First, the text content which is part of the analysis process does not appear in the results, which are consequently prived from the corresponding intuitivity. Second, although the page list is limited to a fraction of the Web site, it remains proportional to the site size and can therefore lack summarization. Third, the technique handles static Web sites only, which excludes the many dynamic Web sites from being

⁴ <http://www.ulb.ac.be/>

analyzed. Finally, the experimental performances and the algorithm complexity do not guarantee the scalability of the approach.

The Information Scent model aims to infer an intuitive representation of the user need from the user actions [5]. In particular, the IUNIS algorithm inputs a sequence of visited pages and outputs a weighted vector of terms describing the visitor's interest. These vectors are quite intuitive, but can be very vague without context. In addition, the analysis is more user-centric than site-centric. Finally, the scalability of the algorithm is not proven; the cited paper presents results for single visits over a few pages, and does not discuss performance, which makes it unclear how the algorithm can handle 50,000 visits over 50,000 pages.

Most of the other Web usage mining research efforts have focused on other research paths, like usage pattern analysis [4], personalization [11], system improvement [1], site structure modification [13], marketing business intelligence [3], and usage characterization [14]. In these research paths, Web analytics concerns have been mostly left aside.

Finally, many other research projects are somehow related to Web usage mining, as unveiled in a recent survey [6]. To the best of our knowledge, our approach is the first to analyze the content of output Web pages to provide site-wide concept-based metrics in order to represent the user needs of any Web site, whatever the Web server technologies.

8 Conclusion

In this paper we have presented our solution to answer the need for summarized and conceptual audience metrics in Web analytics. We first described the various techniques that we conceived to mine the Web pages output by a Web server, showing a set of combinable options that should be applicable for any Web server. Then, we defined three term-based metrics: consultation, presence, and interest. We have seen that these metrics are much more interesting if the terms are grouped into meaningful concepts. Our first experiments on automated term-grouping algorithms showing disappointing results, we reuse the term groups that are naturally present in the concept hierarchies of ontologies. OLAP tools can be used to aggregate the term-based metrics into concept-based metrics. The OLAP cube can be queried from any OLAP-enabled visualization interface. According to our first experiments with the WASA prototype and SQL Server, concept-based metrics prove intuitive enough to support the decision-making process of Web site editors and organization managers. The condition to the wide adoption of concept-based metrics in Web analytics software is the generalization of custom ontologies for Web sites. The availability of large generic ontologies and the development of ontology enrichment techniques and tools, as well as the growing interest into the Semantic Web, should fill the gap with the continuous and growing development of suitable ontologies.

References

1. C. C. Aggarwal and P. S. Yu. On disk caching of web objects in proxy servers. In *Proc. of the 6th Int. Conf. on Information and Knowledge Management, CIKM'97*, pages 238–245, 1997.
2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.
3. A. G. Büchner and M. D. Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. *SIGMOD Record*, 27(4):54–61, 1998.
4. M.-S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. *IEEE Trans. Knowl. Data Eng.*, 8(6):866–883, 1996.
5. 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.
6. F. M. Facca and P. L. Lanzi. Mining interesting knowledge from weblogs: a survey. *Data Knowl. Eng.*, 53(3):225–241, 2005.
7. D. Fensel. *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, 2000.
8. A. Lozano-Tello and A. Gómez-Pérez. Ontometric: A method to choose the appropriate ontology. *J. Database Manag.*, 15(2):1–18, 2004.
9. 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.
10. J.G. March, H.A. Simon, and H.S. Guetzkow. *Organizations*. Cambridge Mass. Blackwell, second edition, 1983.
11. B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on Web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.
12. M. Moeller, C. Cicaterri, A. Presser, and M. Wang. *Measuring e-business Web Usage, Performance, and Availability*. IBM Press, 2003.
13. M. Perkowitz and O. Etzioni. Towards adaptive web sites: Conceptual framework and case study. *Artif. Intell.*, 118(1-2):245–275, 2000.
14. P. Pirolli and J. E. Pitkow. Distributions of surfers' paths through the world wide web: Empirical characterizations. *World Wide Web*, 2(1-2):29–45, 1999.
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 First Int. Conf. on Advances in Natural Computation, ICNC 2005, Part II*, pages 622–626, 2005.
16. G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
17. J. Srivastava, R. Cooley, M. Deshpande, and T. Pang-Ning. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD*, 1(2), 2000.
18. R. Steinberger, B. Pouliquen, and C. Ignat. Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. In *Proc. of the 4th Slovenian Language Technology Conf., Information Society 2004*, 2004.
19. J. Sterne. *Web Metrics: Proven Methods for Measuring Web Site Success*. John Wiley & Sons, 2002.
20. G. Stumme and A. Maedche. Fca-merge: Bottom-up merging of ontologies. In *Proc. of the 17th Int. Joint Conf. on Artificial Intelligence, IJCAI 2001*, pages 225–234, 2001.
21. U. Wahli, J.P. Norguet, J. Andersen, N. Hargrove, and M. Meser. *Websphere Version 5 Application Development Handbook*. IBM Press, 2003.