

# Semantic Analysis of Web Site Audience

Jean-Pierre Norguet\*, Esteban Zimányi  
Université Libre de Bruxelles, CP165/15  
Lab. of Computer and Network Engineering  
Av. F. D. Roosevelt 50  
1050 Brussels, Belgium  
{jnorguet,ezimanyi}@ulb.ac.be

Ralf Steinberger  
European Commission – Joint Research Centre  
Via E. Fermi 1, T.P. 267  
21020 Ispra (VA), Italy  
<http://www.jrc.it/langtech>  
Ralf.Steinberger@jrc.it

## 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 mine the content of the Web pages output by the Web server. For a given taxonomy covering the Web site knowledge domain, we compute the term weights in the output pages and we aggregate them using OLAP tools, in order to obtain concept-based metrics representing the audience of the Web site topics. To demonstrate how our approach solves the cited problems, we actually compute concept-based metrics with SQL Server OLAP Analysis Service and our prototype WASA for a number of case studies. Finally, we validate our results against a popular Web analytics tool.

## Keywords

World Wide Web, Web analytics, Semantic Web, Web usage mining, Data Mining

## 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

\*Jean-Pierre Norguet's work has been funded by a FNRS Research Fellow Grant.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC'06 April 23-27, 2006, Dijon, France

Copyright 2006 ACM 1-59593-108-2/06/0004 ...\$5.00.

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 [11]. 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 [7], 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 management and the Web site chief editor. At these levels, the results produced by Web analytics tools are mostly useless. Indeed, most reports 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 hits give the Web site chief editor conceptual results, but these are limited by several problems like page synonymy (several pages contain the same concept), page polysemy (a page contains several concepts), page temporality, and page volatility. These limitations therefore make Web analytics tools mostly useless to this problem domain.

Web usage mining research projects have mostly left Web analytics aside and have focused on fertile research paths like usage pattern analysis, personalization, system improvement, site structure modification, marketing business intelligence, and usage characterization [11]. A potential contribution to the problem domain 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 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. An interesting attempt to solve these problems is proposed in the IUNIS algorithm of the Information Scent model [2]. This algorithm produces a list of term vectors representing the users' needs, which can be easily interpreted. On the other hand, the results are visit-centric rather than site-centric, suffer from term polysemy and term synonymy, and the algorithm scalability is unclear. 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.

To answer the need of such metrics, our approach aims at analyzing the Web content output by Web servers. Indeed, so far, little or no interest has been shown in the content of the output pages. This disinterest is explained by the lack of techniques to mine the output Web pages and by the high number of pages to analyze afterwards [11]. In Section 2, we present the methods that we conceived to mine the output pages: content journaling, script parsing, server monitoring, network monitoring, and client-side mining. These methods should allow to mine the output pages of any Web site. In Section 3, we explain how term weights in these pages can be aggregated with respect to a taxonomy representing the Web site domain knowledge domain in order to obtain audience metrics representing the consultation, presence, and visitors' interest into the Web site topics. In Section 4, we present and discuss the results obtained with SQL Server OLAP Analysis Service and our prototype WASA on several case studies. In particular, we compare different metrics, we show some interesting visualizations, we study the effect of the taxonomy knowledge domain, and we validate our approach against Urchin, a popular Web analytics tool. Finally in Section 5, we describe the results exploitation process, we expose the limitations of the approach, and we present some insights of solutions for future work.

## 2. OUTPUT PAGE COLLECTION

In our approach, we start by collecting the Web pages output from the Web server. This operation differs from the classical Web usage data collection [3, 11], as actual page content is collected in addition to the HTTP transaction metadata. Output page collection also differs from Web site content mining [1, 11], as the pages collected are those output by the Web server instead of the online pages stored on the Web server file system. To collect the output pages, we therefore conceived a number of methods which can be used alone or in combination:

**Web logs and content journaling** is the easiest method to setup and requires the least amount of storage and computation. A content journal stores the content of the online pages and their temporal evolution during the mining period. Coupled to the Web server logs, the content journal allows to retrieve the content of any output page. The drawback of this method is it supports static pages only.

**Server monitoring** is an efficient method for collecting dynamic pages. A server monitor runs inside the Web server and stores the output pages after they have been sent to the client. The main drawback of server monitoring is the instability introduced in the Web server.

**Network monitoring** is independent from the Web server. A network monitor runs in network-promiscuity mode on the same Ethernet subnet as the Web server. It captures and reassembles the TCP/IP packets exchanged with the Web clients. Network monitoring is CPU-intensive because all transferred files must be reassembled before they can be mined or ignored. Also, encrypted communications cannot be monitored.

**Client-side mining** is performed by a page-embedded program. When the page is displayed in the visitor's browser, the program sends the page content to a mining server. This method can mine most kinds of Web

pages including those composed on the client side like XML/XSL pages. However, client-side mining suffers from its visibility and from its vulnerability [8].

All these methods used alone or in combination should provide the necessary means for collecting the output pages of any Web site. This solves the page-temporality and page-volatility problems.

## 3. CONCEPT-BASED AUDIENCE METRICS

For the given Web site to analyze, we choose a taxonomy that models the Web site knowledge domain. The top terms in the taxonomy should represent the Web site main topics. Thus, for each taxonomy term, the term weight [1] in the output pages gives an indication of the term consultation by the visitors during the mining period. If the Web site is mostly static, the term weight in the online pages gives an indication of the term presence on the site. Term consultation and term presence are two interesting metrics but suffer from polysemy and synonymy problems.

These problems can be overcome by aggregating the term metrics along the taxonomy. Indeed in most taxonomies, the terms are hierarchically linked by a relationship of type *is a* or *part of* [14]. In these taxonomies, the audience of the subterms contributes to the communication of the concepts denoted by the superterms. For example, the consultation of the "fruit" concept would include the occurrences of "strawberry" and "apple". Therefore, the audience metrics aggregation from the leaves up to the taxonomy root gives an indication of the audience obtained by the Web site concepts. Furthermore, the consultation-to-presence ratio gives an indication of the visitors' *interest* into the concepts. If the top terms in the taxonomy represent the Web site main concepts, the corresponding consultation, presence, and interest metrics can be used as conceptual audience measures.

These metrics can be formalized as follows. For a mining period between days  $d_1$  and  $d_2$  and a given concept  $C_i$  defined as the union of the term  $s_i$  and of its subterms in the taxonomy, the consultation and presence metrics can be formalized as follows.

$$\text{Consultation}(C_i, d_1, d_2) = \sum_{s_j \in C_i} \sum_{d=d_1}^{d_2} w_j(d) \quad (1)$$

$$\text{Presence}(C_i, d_1, d_2) = \sum_{s_j \in C_i} \int_{d_1}^{d_2} w'_j(t) dt \quad (2)$$

where  $w_j(d)$  is the term weight of term  $s_j$  in the output pages mined during day  $d$  and  $w'_j(t)$  is the term weight of term  $s_i$  in the online pages at time  $t$ . If between  $d_1$  and  $d_2$  these pages have been online during a time  $\Delta t_k$ , the integral in Equation 2 is equal to  $\sum_{p_k} w'_{jk} \Delta t_k$ , where  $w'_{jk}$  is the weight of term  $s_j$  in page  $p_k$ . This expression can be computed easily.

Practically, hierarchical aggregation of the term-based metrics into concept-based metrics can be easily computed and visualized using OLAP tools. The computation of Equations 1 and 2 with OLAP tools requires a multidimensional model with two dimensions: Time and Taxonomy [9]. The taxonomy dimension should be designed as a *parent-child dimension* to support taxonomies with any number of levels in each branch [6]. The cube fact table must contain the

daily term metrics, which can be computed by content processing and term counting in the output and online pages. The measures to define in the cube are consultation, presence, and interest. The interest measure can be defined as a calculated member dividing the first two measures. After the cube has been introduced and processed in the OLAP tool, the concept-based audience metrics can be visualized with any OLAP client like Microsoft Excel PivotChart.

## 4. EXPERIMENTATION

To test our approach, we developed a prototype called WASA (Figure 1). WASA stands for Web Audience Semantic Analysis. The prototype implements output page mining from Web logs and content journaling (see Section 2) and analyzes the output and online pages to produce the daily consultation and presence metrics for each term of a given taxonomy. The prototype is written in the Java language and is composed of 10,000 lines of code. The metrics are stored in a MySQL database and transferred into SQL Server for OLAP analysis. In SQL Server OLAP Analysis Service, we introduce the OLAP cube representing the multidimensional model described in Section 3. After cube processing, the metrics are aggregated and can be queried from Microsoft Excel to produce the various visualizations.

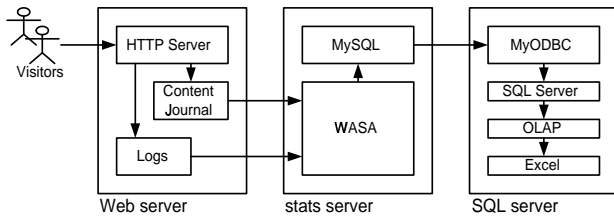


Figure 1: Experimental configuration.

### 4.1 Visualization

In our first case study, we analyzed <http://cs.ulb.ac.be>, our computer science laboratory’s Web site, which contains about 2,000 Web pages and receives an average of 100 page requests a day. The taxonomy was extracted from the ACM classification, which contains 1230 hierarchically-linked terms. The mining period is the academic year 2003-2004.

We first produced a multi-line chart where each curve represents the visitors’ consultation of the top ACM concepts (Figure 2). Computing Methodologies, Software, and Information Systems rank in the top, which is not surprising as these domains are the subject of major lectures. Also, a peak of interest in Theory of Computation can be observed at the beginning of the academic year, when the first-year students start following the corresponding lessons in the computers room. Finally, the average consultation falls down during the academic holiday periods: January-February and July-August. As we can see, this kind of chart can be intuitively related to the problem domain.

To compare the various metrics, we also produced a bar chart representing the metrics for each of the top concepts (Figure 3). The top concepts are Information Systems, Computing Methodologies, and Software. However, these concepts are very present in the Web site, which is confirmed

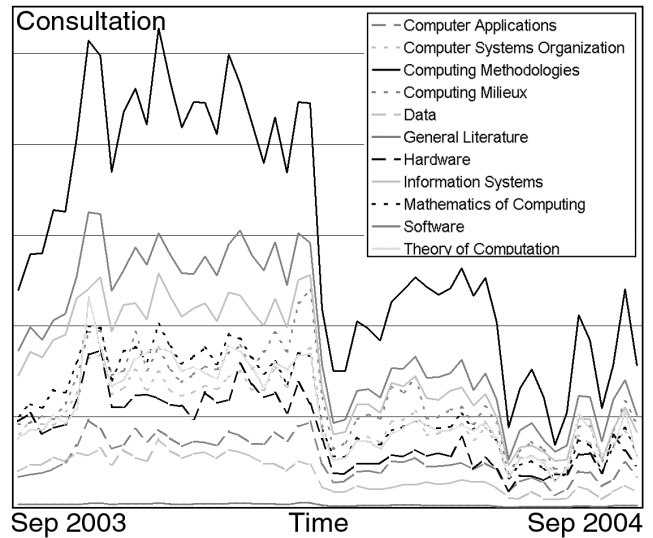


Figure 2: Consultation of the ACM classification top concepts on the [cs.ulb.ac.be](http://cs.ulb.ac.be) Web site during the 2003-2004 academic year.

by high presence values. Therefore, high consultation values are not representative of the visitors’ interest, for which low interest values can be observed. The interesting concepts are Theory of Computation, Data, and Mathematics of Computing. By comparing the consultation and interest in this example, we can see that the considered metrics can dramatically change the ranking of the concepts and should be interpreted carefully.

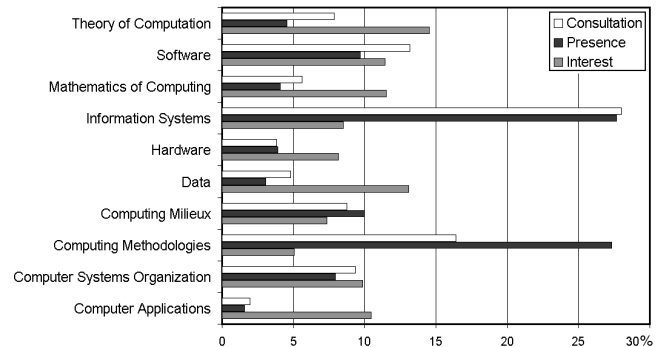


Figure 3: Audience metrics for the ACM classification top concepts.

### 4.2 Taxonomy Coverage

To test the influence of the taxonomy on concept-based metrics, we made the same experiments with Eurovoc, the European Commission’s thesaurus [12]. Eurovoc contains a taxonomy of 6650 terms, and its domain knowledge include all the European Commission’s fields of interest. These include a broad range of domains, from sociology to science, while the ACM classification knowledge domain is focused on computer science. Although Eurovoc contains about five times more terms than the ACM classification, it offers a poor coverage of the computer science domain. Therefore,

the results obtained with Eurovoc are difficult to relate to the Web site knowledge domain. This kind of problem is typical of very conceptual taxonomies like Eurovoc [12]. This shows how the choice of the taxonomy is important for the results interpretation.

As a natural continuation of the Eurovoc experiment, we studied the benefits of improving taxonomy coverage with respect to the Web site knowledge domain. To evaluate what results can be obtained with an optimal taxonomy enrichment, our department's staff enriched the ACM classification with terms of the Web site. This manual method ensures an optimal improvement of the taxonomy coverage. If we define the taxonomy coverage as the number of taxonomy terms that appear in the output Web pages, our enrichment operations have increased the coverage from 70 to 90, that is an increase of about 30%.

The effect of this enrichment has been evaluated by running WASA with the enriched taxonomy on our department's Web site. With regard to the enriched taxonomy, the top consulted concepts are Software, Computing Methodologies, and Information Systems, while the interesting concepts are Mathematics of Computing, Computing Methodologies, and Software (Figure 4). By comparing these results with those obtained with the raw ACM classification (Figure 3), we can see the importance of the taxonomy knowledge domain in the interpretation of the results.

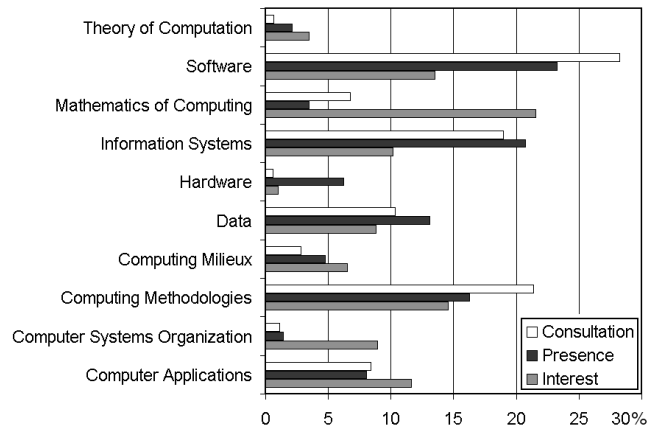


Figure 4: Audience metrics for the enriched ACM classification top concepts.

### 4.3 Validation

To validate our approach against existing software, we compared our results against Urchin, a popular Web analytics tool. Although WASA and Urchin results are very different, there is a particular case of Web site where the Urchin results are comparable to those obtained by WASA. Indeed, if the Web site directories match the taxonomy concepts, the hits by directories obtained by Urchin should be comparable to the interest by concept obtained by WASA.

To verify this, we ran the tests on <http://wasa.ulb.ac.be>, a personal Web site where the directories have been structured with respect to the Web site concepts. For the purpose of the case study, the Web site author has manually built a custom taxonomy containing the main concepts and sub-concepts. The Web site main concepts include computer science, travel, and passions. The passions concept is sub-

divided into music, chess, cinema, and well-being. The taxonomy contains about 1150 terms in total. The Web site contains about 200 pages and receives about 100 page requests a day. The mining period is 2003.

To compare the results, we produced a directory-based graph with Urchin (Figure 5) and a concept-based graph with WASA (Figure 6) representing the audience metrics for the main three concepts: computer science, travel, and passions. By looking at the two graphs, we can see common peaks by the months of March and November. The March peak is due to the referral link from a computer science online magazine, while the November peak is due to the referral link from a music search engine.

The concept graph in the first trimester of the year shows a predominance of the computer science concept, which cannot be seen for the computer science directory. According to the Web logs, this predominance is due to the success of various computer science pages located outside the computer science directory and linked by computer science sites like <http://www.linux.org>. The dispersion of the pages within the site is rigid because the referral links pointing to these belong to external sites and are not under direct control. This difference between the two graphs shows the limitations implied by page synonymy and by directory structure rigidity. In contrast, concept-based metrics do not suffer from these limitations and are therefore superior with respect to those aspects.

Another difference can be observed during the November peak, where the travel concept outperforms the passions concepts, while the passions directory clearly outperforms the travel directory. The 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 the concept granularity.

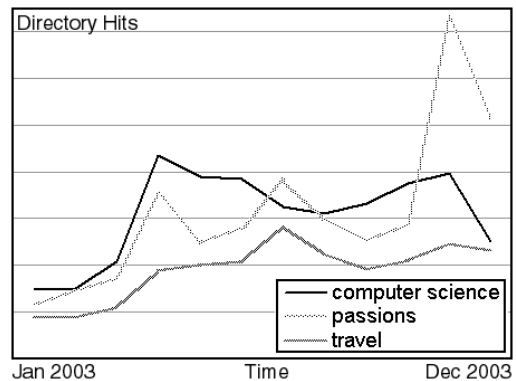
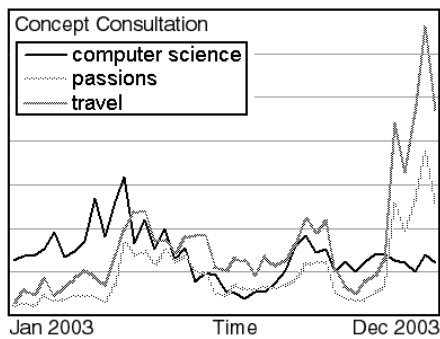


Figure 5: Directory hits for wasa.ulb.ac.be during year 2003. Results obtained with Urchin.

## 5. CONCLUSION 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 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



**Figure 6: Concept consultation of wasa.ulb.ac.be during year 2003. Results obtained with WASA.**

output by any Web site.

Then, we have seen that aggregating the occurrences of taxonomy terms in these pages can provide audience metrics for the Web site concepts. According to the first experiments on real data 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 analytics.

Experiments on real Web sites with several taxonomies like Eurovoc and the ACM classification have shown the importance of the considered metric (consultation, presence, interest) and of the taxonomy coverage of the Web site knowledge domain. Also, comparing our prototype results with a popular Web analytics tool validates our approach while demonstrating the superiority of concept-based metrics over directory-based and page-based metrics. Indeed, these metrics suffer from directory structure rigidity, page synonymy, and page polysemy. This calls for the adoption of concept-based metrics in Web analytics tools.

A limitation to the wide adoption of concept-based metrics is the lack of custom taxonomies for Web sites. To overcome this limitation, we will explore automatic and semi-automatic taxonomy enrichment techniques [5]. In our future work, we will also apply further text analysis techniques to the Web site pages. These techniques will include geo-coding, clustering, date recognition, and organization/person name identification [13]. The overall analysis will provide a multi-facetted vector representation which we will integrate in our multidimensional model. We will also add other dimensions like Web topology and Web site structure. We will evaluate the influence of these additional dimensions by running similar experiments as in this paper. Finally, variations of the metrics inspired from the vector model [1] as well as evaluators for taxonomy coverage [4] should be experimented.

## 6. REFERENCES

- [1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.
- [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] A. Lozano-Tello and A. Gómez-Pérez. Ontometric: A method to choose the appropriate ontology. *J. Database Manag.*, 15(2):1–18, 2004.
- [5] A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.
- [6] E. Malinowski and E. Zimányi. OLAP hierarchies: A conceptual perspective. In *Proc. of the 16<sup>th</sup> Int. Conf. on Advanced Information Systems Engineering, CAiSE'04*, LNCS 3084, pages 477–491. Springer-Verlag, 2004.
- [7] J.G. March, H.A. Simon, and H.S. Guetzkow. *Organizations*. Cambridge Mass. Blackwell, second edition, 1983.
- [8] 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*, 2005.
- [9] 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*, 2006.
- [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 First Int. Conf. on Advances in Natural Computation, ICNC 2005, Part II*, pages 622–626, 2005.
- [11] 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.
- [12] 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. B of the 7th Int. Multiconference on Language Technologies, IS 2004*, 2004.
- [13] R. Steinberger, B. Pouliquen, and C. Ignat. Navigating multilingual news collection using automatically extracted information. In *Proc. of the 27th Int. Conf. on Information Technology Interfaces, ITI 2005*, 2005.
- [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.