Information Retrieval Techniques for Pattern Matching: Managing and Searching Textual and XML Information in 21st Century Applications

Riccardo Martoglia

Book published by Lambert Academic Publishing

Abstract.

Information is the main value of Information Society. The recent developments in computing power and telecommunications, along with the constant drop of Internet access costs and data management and storing, created the right conditions for the global diffusion of the Web and, more generally, of new research tools able to analyze information and their contents. Depending on the particular application scenario and on the type of information that has to be managed and searched, different techniques need to be devised. In this book, the author deals with the two most common types of information: plain text, discussed in the first part, and semi-structured data, in particular XML documents, deeply discussed the second part. The detailed analysis of approximate matching, duplicate document detection, exact, approximate and semantic query answering, multi-version document management and personalized access techniques offered in this book will guide Information Technology professionals and users in effectively and efficiently managing information and knowledge, thus answering the increasingly complex Information needs of most 21st century applications.

Contents

Bibliography

[1] Secure Hash Standard. Technical Report FIPS PUB 180-1, U.S. Department of Commerce/-National Institute of Standards and Technology, 1995.

[2] R. Agrawal, C. Faloutsos, and A. N. Swami. Efficient Similarity Search In Sequence Databases. In *Proc. of 4th International Conference on Foundations of Data Organization and Algorithms (FODO 1993)*, 1993.

[3] R. Agrawal, K.-I. Lin, H. S. Sawhney, and K. Shim. Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. In *Proc. of 21th International Conference on Very Large Data Bases (VLDB 1995)*, 1995.

[4] E. Amitay, R. Nelken, W. Niblack, R. Sivan, and A. Soffer. Multi-resolution disambiguation of term occurrences. In *Proc. of the 12th Conference on Information and Knowledge Management (CIKM 2003)*, 2003.

[5] J. Artiles, A. Penas, and F. Verdejo. Word Sense Disambiguation based on term to term similarity in a context space. In *Proc. of Senseval-3*, 2004.

[6] F. Baader, I. Horrocks, and U. Sattler. Description logics for the semantic web. *Künstliche Intelligenz*, 16(4), 2002.

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

[8] R. A. Baeza-Yates and G. H. Gonnet. A Fast Algorithm on Average for All-Against-All Sequence Matching. In *Proc. of the International Workshop and Symposium on String Processing and Information Retrieval (SPIRE 1999)*, 1999.

[9] R. A. Baeza-Yates and G. Navarro. A Faster Algorithm for Approximate String Matching. In *Combinatorial Pattern Matching, 7th Annual Symposium*, 1996.

[10] T. Baldwin and H. Tanaka. The Effects of Word Order and Segmentation on Translation Retrieval Performance. In *Proc. of the 18th International Conference on Computational Linguistics (COLING 2000)*, 2000.

[11] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proc. of 18th IJCAI Conference*, 2003.

[12] R. Baumgartner, S. Flesca, and G. Gottlob. Visual Web information extraction with Lixto. In *Proc. of the Twenty-seventh Int. Conference on Very Large Data Bases*, 2001.

[13] B. Becker, S. Gschwind, T. Ohler, B. Seeger, and P. Widmayer. An asymptotically optimal multiversion b-tree. *VLDB J.*, 5(4), 1996.

[14] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5), 2001.

[15] Philip A. Bernstein and Erhard Rahm. On Matching Schemas Automatically. Technical Report MSR-TR-2001-17, Microsoft Research (MSR), 2001.

[16] S. Boag, D. Chamberlin, M. F. Fernández, D. Florescu, J. Robie, and J. Siméon. XQuery 1.0: An XML Query Language. W3C Working Draft, 2003.

[17] P. Bouquet, L. Serafini, and S. Zanobini. Semantic coordination: a new approach and an application. In *Proc. of the 2nd ISWC Conference*, 2003.

[18] L. Bowker and M. Barlow. Bilingual concordancers and translation memories: A comparative evaluation. In *Proc. of the 2nd International Workshop on Language Resources for Translation Work, Research and Training (LR4Trans-II 2004)*, 2004.

[19] D. Braga and A. Campi. A Graphical Environment to Query XML Data with XQuery. In *Proc. of the 4th Intl. Conference on Web Information Systems Engineering*, 2003.

[20] M. M. Breunig, H. P. Kriegel, P. Kröger, and J. Sander. Data Bubbles: Quality Preserving Performance Boosting for Hierarchical Clustering. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 30(2), 2001.

[21] D. Bricklin. Copy Protection Robs the Future. http://www.bricklin.com/robfuture.htm.

[22] Sergev Brin, James Davis, and Hector Garcia-Molina. Copy Detection Mechanisms for Digital Documents. In *Proc. of the 1995 ACM SIGMOD International Conference on Management of Data (SIGMOD 1995)*, 1995.

[23] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the Web. *Computer Networks and ISDN Systems*, 29(8-13), 1997.

[24] R.D. Brown. Example-Based Machine Translation in the Pangloss Systems. In *Proc. of 16th International Conference on Computational Linguistics*, 1996.

[25] N. Bruno, N. Koudas, and D. Srivastava. Holistic twig joins: optimal XML pattern matching. In *Proceedings of the 2002 International Conference on Management of Data (SIGMOD 2002)*, 2002.

[26] A. Budanitsky and G. Hirst. Semantic distance in wordnet: an experimental, application-oriented evaluation of five measures. In *Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, 2001.

[27] D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Y. Vardi. View-Based Query Processing for Regular Path Queries with Inverse. In *Proc. of the Nineteenth ACMSIGMOD-SIGACT-SIGART (PODS-00)*, 2000.

[28] D. Castelli and P. Pagano. A System for Building Expandable Digital Libraries. In *Proc. of the Third ACM/IEEE-CS Joint Conference on Digital Libraries*, 2003.

[29] K.-P. Chan and A. W.-C. Fu. Efficient time series matching by wavelets. In *Proc. of the 15th International Conference on Data Engineering (ICDE 1999)*, 1999.

[30] T. Chen, J. Lu, and T. Wang Ling. On boosting holism in xml twig pattern matching using structural indexing techniques. In *Proc. of the ACM SIGMOD*, 2005.

[31] S. Chien, V. J. Tsotras, and C. Zaniolo. Efficient schemes for managing multiversionxml documents. *VLDB J.*, 11(4), 2002.

[32] S.-Y. Chien, Z. Vagena, D. Zhang, V. Tsotras, and C. Zaniolo. Efficient structural joins on indexed XML documents. In *Proceedings of 28th International Conference on Very Large Data Bases (VLDB 2002)*, 2002.

[33] A. Chowdhury, O. Frieder, and D. Grossman. Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems*, 20(2), 2002.

[34] E. Chvez and G. Navarro. A metric index for approximate string matching. In *Proc. of the 5th Latin American Symposium on Theoretical Informatics*, 2002.

[35] P. Ciaccia and M. Patella. Searching in Metric Spaces with User-defined and Approximate Distances. *Trans. on Database Systems (TODS)*, 4(27), 2002.

[36] P. Ciaccia, M. Patella, and P. Zezula. M-Tree: An efficient access method for similarity search in metric spaces. In *Proc. of 23rd International Conference on Very Large Data Bases (VLDB)*, 1997.

[37] P. Ciaccia and W. Penzo. Relevance ranking tuning for similarity queries on xml data. In *Proc. of the VLDB EEXTT Workshop*, 2002.

[38] R. Cilibrasi and P.M.B. Vitanyi. Automatic meaning discovery using Google. Technical report, University of Amsterdam, 2004.

[39] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In *Proc. of the 13th World Wide Web Conference (WWW 2004)*, 2004.

[40] A. Cobbs. Fast Approximate Matching Using Suffix Trees. In *Proc. of the 6th International Symposium on Combinatorial Pattern Matching (CPM 1995)*, 1995.

[41] P. Collins and P. Cunningham. Adaptation Guided Retrieval in EBMT: A Case-Based Approach to Machine Translation. In *Proc. of the 3rd European Workshop on Advances in Case-Based Reasoning, (EWCBR 1996)*, 1996.

[42] L. Cranias, H. Papageorgiou, and S. Piperidis. A Matching Technique In Example-Based Machine Translation. In *Proc. of the 15th International Conference on Computational Linguistics (COLING 1994)*, 1994.

[43] V. Crescenzi, G. Mecca, and P. Merialdo. RoadRunner: automatic data extraction from data-intensive web sites. In *Proc. of the 2002 ACM SIGMOD Int. Conference on Management of Data (SIGMOD 2002)*, 2002.

[44] F. Currim, S. Currim, C. Dyreson, and R. T. Snodgrass. A Tale of Two Schemas: Creating a Temporal Schema from a Snapshot Schema with τXSchema. In *Proc. of EDBT*, Heraklion, Greece, 2004.

[45] C. De Castro, F. Grandi, and M. R. Scalas. Semantic interoperability of multitemporal relational databases. In *Proc. of ER*, 1993.

[46] Atril Deja Vu - Translation Memory and Productivity System. Home page http://www.atril.com.

[47] P.F. Dietz. Maintaining Order in a Linked List. In *Proceedings of 14th Anual ACM Symposium on Theory of Computing (STOC 1982)*, 1982.

[48] H. Do, S. Melnik, and E. Rahm. Comparison of schema matching evaluations. In *Proc. of the 2nd Int. Workshop on Web Databases*, 2002.

[49] H. Do and E. Rahm. COMA – A system for flexible combination of schema matching approaches. In *Proc. of the 28th Conference on Very Large Data Bases (VLDB 2002)*, 2002.

[50] A. Doan, P. Domingos, and A. Y. Halevy. Reconciling schemas of disparate data sources: a machine-learning approach. *SIGMOD Record*, 30(2), 2001.

[51] B. Dorr, P. Jordan, and J. Benoit. A survey of current research in machine translation. *Advances in Computers*, 49, 1999.

[52] The "Semantic web techniques for the management of digital identity and the access to norms" PRIN Project Home Page. http://www.cirsfid.unibo.it/eGov03.

[53] Marc Ehrig and Alexander Maedche. Ontology-focused crawling of web documents. In *Proc. of the ACM SAC*, 2003.

[54] R. T. Snodgrass et al. *The TSQL2 Temporal Query Language*. Kluwer Academic Publishing, New York, 1995.

[55] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast Subsequence Matching in Time-Series Databases. In *Proc. of the 1994 ACM SIGMOD International Conference on Management of Data (ICMD 1994)*, 1994.

[56] D. Gao and R. T. Snodgrass. Temporal slicing in the evaluation of xml queries. In *Proc. of VLDB*, Berlin, Germany, 2003.

[57] R. Giegerich, F. Hischke, S. Kurtz, and Enno Ohlebusch. A General Technique to Improve Filter Algorithms for Approximate String Matching. In *Proc. of the 4th South American Workshop on String Processing*, 1997.

[58] F. Grandi and F. Mandreoli. Temporal modelling and management of normative documents in xml format. *Data Knowl. Eng.*, 54(3), 2005.

[59] F. Grandi, F. Mandreoli, P. Tiberio, and M. Bergonzini. A temporal data model and management system for normative texts in xml format. In *Proc. of the 15th ACM Intl' Workshop on Web Information and Data Management (WIDM)*, New Orleans, LA, November 2003.

[60] F. Grandi, F. Mandreoli, P. Tiberio, and M. Bergonzini. A temporal data model and system architecture for the management of normative texts. In *Proc. of the 11th Natlional Conference on Advanced Database Systems (SEBD)*, Cetraro, Italy, June 2003.

[61] L. Gravano, P.G. Ipeirotis, H.V. Jagadish, N. Koudas, S. Muthukrishnan, and D. Srivastava. Approximate String Joins in a Database (Almost) for Free. In *Proc. of 27th International Conference on Very Large DataBases (VLDB 2001)*, 2001.

[62] T. Grust. Accelerating XPath location steps. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD 2002)*.

[63] T. Grust, M. Van Keulen, and J. Teubner. Staircase Join: Teach a Relational DBMS to Watch its (Axis) Steps. In *Proceedings of 29th International Conference on Very Large Data Bases (VLDB 2003)*, 2003.

[64] T. Grust, M. Van Keulen, and J. Teubner. Accelerating XPath evaluation in any RDBMS. *ACM Transactions on Database Systems*, 29(1), 2004.

[65] S. Guha, H. V. Jagadish, N. Koudas, D. Srivastava, and T. Yu. Approximate XML joins. In *Proc. of ACM SIGMOD*, 2002.

[66] N. Heintze. Scalable Document Fingerprinting. *Second Usenix Workshop on Electronic Commerce*, 1996.

[67] I. Horrocks and P. F. Patel-Schneider. Reducing owl entailment to description logic satisfiability. In *Proc. of ISWC 2003*, 2003.

[68] N. Ide and J. Veronis. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24(1), 1998.

[69] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall Inc., 1988.

[70] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3), 1999.

[71] C. S. Jensen, C. E. Dyreson, and (Eds.) et al. The Consensus Glossary of Temporal Database Concepts - February 1998 Version. In O. Etzion, S. Jajodia, and S. Sripada, editors, *Temporal Databases — Research and Practice*. Springer-Verlag, 1998. LNCS No. 1399.

[72] H. Jiang, W. Wang, H. Lu, and J. Xu Yu. Holistic Twig Joins on Indexed XML Documents. In *Proceedings of 29th International Conference on Very Large Data Bases (VLDB 2003)*, 2003.

[73] T. Kahveci and A. K. Singh. Variable Length Queries for Time Series Data. In *Proc. of the 17th International Conference on Data Engineering (ICDE 2001)*, 2001.

[74] L.G. Khachiyan. A Polynomial Algorithm in Linear Programming. *Doklady Akademii Nauk SSSR*, 244, 1979.

[75] Seung-Kyum Kim and Sharma Chakravarthy. Modeling time: Adequacy of three distinct time concepts for temporal data. In *Proc. of 12th International Conference on the Entity-Relationship Approach (ER'93)*, Arlington, TX, December 1993. LNCS No. 823.

[76] C. Koch, S. Scherzinger, N. Schweikardt, and B. Stegmaier. FluXQuery: An Optimizing XQuery Processor for Streaming XML Data. In *Proc. of 30th International Confl on Very Large Data Bases (VLDB 2004), August 31 – September 3, 2004, Toronto, Canada*, 2004.

[77] R. Kraft, F. Maghoul, and C. C. Chang. Y!Q: Contextual Search at the Point of Inspiration. In *Proc. of CIKM 2005*, Bremen, Germany, 2005.

[78] S. H. Kwok. Watermark-based copyright protection system security. *Communications of the ACM*, 46(10), 2003.

[79] O. Lassila and R. Swick. Resource Description Framework (RDF) model and syntax specification. W3C Working Draft WD-rdf-syntax-19981008, 1998.

[80] S. Lawrence, K. Bollacher, and C. Lee Giles. Indexing and Retrieval of Scientific Literature. In *Proc.s of 8th International Conference on Information and Knowledge Management (CIKM 1999)*, 1999.

[81] C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An electronic lexical database*. MIT Press, 1998.

[82] C. Leacock, M. Chodorow, and G. A. Miller. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1), 1998.

[83] J. Litman. Digital Copyright and the Progress of Science. *ACM SIGIR Forum*, 36(2), 2002.

[84] J. Madhavan, P. A. Bernstein, A. Doan, and A. Y. Halevy. Corpus-based schema matching. In *Proc. of 21st International Conference on Data Engineering (ICDE 2005)*, 2005.

[85] J. Madhavan, P. A. Bernstein, and E. Rahm. Generic Schema Matching with Cupid. In *Proc. of the 27th Conference on Very Large Data Bases (VLDB 2001)*, 2001.

[86] F. Mandreoli, R. Martoglia, and E. Ronchetti. Versatile structural disambiguation for semantic-aware applications. In *Proc. of the 14th International Conference on Information Knowledge and Management (CIKM 2005)*, 2005.

[87] F. Mandreoli, R. Martoglia, and E. Ronchetti. Strider: a versatile system for structural disambiguation. In *Proc. of the 10th International Conference on Extending Database Technology (EDBT 2006)*, 2006.

[88] F. Mandreoli, R. Martoglia, and E. Ronchetti. Supporting temporal slicing in xml databases. In *Proc. of the 10th International Conference on Extending Database Technology (EDBT 2006)*, 2006.

[89] F. Mandreoli, R. Martoglia, E. Ronchetti, P. Tiberio, F. Grandi, and M. R. Scalas. Personalized access to multi-version norm texts in an egovernment scenario. In *Proc. of the International Conference on E-Government (DEXA EGOV 2005)*, 2005.

[90] F. Mandreoli, R. Martoglia, and P. Tiberio. A Syntactic Approach for Searching Similarities within Sentences. In *Proc. of the 11th ACM Conference of Information and Knowledge Management (CIKM 2002)*, 2002.

[91] F. Mandreoli, R. Martoglia, and P. Tiberio. Searching Similar (Sub)sentences for Example Based Machine Translation. In *Proc. of the 10th Convegno su Sistemi Evoluti per Basi di Dati (SEBD 2002)*, 2002.

[92] F. Mandreoli, R. Martoglia, and P. Tiberio. Exploiting multi-lingual text potentialities in EBMT systems. In *Proc. of the 13th IEEE International Workshop on Research Issues in Data Engineering: Multi Lingual Information Management (RIDE-MLIM 2003)*, 2003.

[93] F. Mandreoli, R. Martoglia, and P. Tiberio. A Document Comparison Scheme for Secure Duplicate Detection. *International Journal of Digital Libraries*, 4(3), 2004.

[94] F. Mandreoli, R. Martoglia, and P. Tiberio. Approximate Query Answering for a Heterogeneous XML Document Base. In *Proc. of the 5th International Conference on Web Information Systems Engineering (WISE 2004)*, 2004.

[95] J. M. Martinez. MPEG-7 standard overview, ISO/IEC JTC1/SC29/WG11 N6828. http://www.chiariglione.org/mpeg/standards/mpeg-7.

[96] O. Mason. QTag, a probabilistic parts-of-speech tagger. http://web.bham.ac.uk/O.Mason/software/tagger/.

[97] I. Dan Melamed. Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. In *Proc. of the Third Workshop on Very Large Corpora (WVLC3)*, 1995.

[98] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity Flooding: A Versatile Graph Matching Algorithm and ist Application to Schema Matching. In *Proc. of the 18th International Conference on Data Engineering (ICDE 2002)*, 2002.

[99] A. O. Mendelzon, F. Rizzolo, and A. A. Vaisman. Indexing temporal xml documents. In *Proc. of VLDB*, 2004.

[100] G.A. Miller. WordNet: A Lexical Database for English. In *CACM 38*, 1995.

[101] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five Papers on WordNet. Technical report, Princeton University's Cognitive Science Laboratory, 1993.

[102] M. Nagao. *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle*. Nato Publications, 1984.

[103] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1), 2001.

[104] G. Navarro and R. A. Baeza-Yatesa. Very Fast and Simple Approximate String Matching. *Information Processing Letters*, 72, 1999.

[105] G. Navarro and R. A. Baeza-Yatesa. New and faster filters for multiple approximate string matching. *Random Structures and Algorithms*, 20(1), 2002.

[106] Y. Papakonstantinou, A. Gupta, and L. M. Haas. Capabilities-Based Query Rewriting in Mediator Systems. *Distributed and Parallel Databases*, 6(1), 1998.

[107] Y. Papakonstantinou and V. Vassalos. Query rewriting for semistructured data. In *Proc. of the ACM SIGMOD*, 1999.

[108] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, 2002.

[109] ParaConc - Multilingual Concordancer. http://www.athel.com.

[110] T. Bach Pedersen, C. S. Jensen, and C. E. Dyreson. Extending practical pre-aggregation in on-line analytical processing. In *Proc. of VLDB*, 1999.

[111] M. Peim, E. Franconi, N. W. Paton, and C. A. Goble. Query Processing with Description Logic Ontologies Over Object-Wrapped Databases. In *Proc. of the 14th International Conference on Scientific and Statistical Database Management*, 2002.

[112] Owl plugin for protégé. http://protege.stanford.edu/plugins/owl/, 2004.

[113] P. Resnik. Disambiguating Noun Groupings with Respect to WordNet Senses. In *Proc. of the Third Workshop on Very Large Corpora*, 1995.

[114] C. Rick. A New Flexible Algorithm for the Longest Common Subsequence Problem. Technical report, University of Bonn, Computer Science Department IV, 1994.

[115] N. Rishe, J. Yuan, R. Athauda, S. Chen, X. Lu, X. Ma, A. Vaschillo, A. Shaposhnikov, and D. Vasilevsky. Semantic Access: Semantic Interface for Querying Databases. In *Proc. of the 26th Conference on Very Large Data Bases (VLDB 2000)*, 2000.

[116] S. Rodotà. Introduction to the "one world, one privacy" session. In *Proc. of 23rd Data Protection Commissioners Conference, Paris, France.*

[117] S. Sassatelli. Approssimazione semantica per routing di interrogazioni in un PDMS. Master thesis, Università degli studi di Modena e Reggio Emilia, 2004/2005.

[118] S. Sato and M. Nagao. Toward Memory-based Translation. In *Proc. of the 13th International Conference on Computational linguistics (COLING 1990)*, 1990.

[119] T. Schlieder and Felix Naumann. Approximate tree embedding for querying XML data. In *Proc. of ACM SIGIR Workshop On XML and Information Retrieval*, 2000.

[120] X. Shen, B. Tan, and C. X. Zhai. Implicit user modeling for personalized search. In *Proc. of CIKM 2005*, Bremen, Germany, 2005.

[121] N. Shivakumar and H. Garcia-Molina. SCAM: A Copy Detection Mechanism for Digital Documents. In *Proc. of the 2nd International Conference on Theory and Practice of Digital Libraries*, 1995.

[122] N. Shivakumar and H. Garcia-Molina. Building a scalable and accurate copy detection mechanism. In *Proc. of the 1st ACM International Conference on Digital Libraries (ICDL 1996)*, 1996.

[123] Web services activity. W3C Consortium, http://www.w3.org/2000/xp/Group/, 2004.

[124] H. Somers. Review Article: Example-based Machine Translation. *Machine Translation*, 14(2), 1999.

[125] E. Sumita and H. Iida. Experiments and Prospects of Example-based Machine Translation. In *Proc. of the 29th Annual Meeting of the Association for Computational Linguistics (ACL 1991)*, 1991.

[126] E. Sutinen and J. Tarhio. On Using q-gram Locations in Approximate String Matching. In *Proc. of 3rd Annual European Symposium*, 1995.

[127] E. Sutinen and J. Tarhio. Filtration with q-samples in Approximate String Matching. In *Proc. of the 7th annual Symposium on Combinatorial Pattern Matching*, 1996.

[128] A. Tagarelli and S. Greco. Clustering Transactional XML Data with Semantically-Enriched Content and Structural Features. In *Proc. of the 5th International Conference on Web Information Systems Engineering (WISE 2004)*, 2004.

[129] I. Tatarinov and A. Halevy. Efficient Query Reformulation in Peer Data Management Systems. In *Proc. of ACM SIGMOD*, 2004.

[130] A. Theobald and Gerhard Weikum. The index-based XXL search engine for querying XML data with relevance ranking. *Lecture Notes in Computer Science*, 2287, 2002.

[131] M. Theobald, R. Schenkel, and G. Weikum. Exploiting Structure, Annotation, and Ontological Knowledge for Automatic Classification of XML Data. In *Proc. of the WebDB Workshop*, 2003.

[132] H. S. Thompson, D. Beech, M. Maloney, and N. Mendelsohn. XMLSchema. W3C Recommendation, 2001.

[133] Trados Team Edition - Translation Memory Technologies. Home page http://www.trados.com.

[134] E. Ukkonen. Approximate String Matching with q-grams and Maximal Matches. *Theoretical Computer Science*, 92(1), 1992.

[135] J. S. Vitter. An Efficient Algorithm for Sequential Random Sampling. *ACM Transactions on Mathematical Software*, 13(1), 1987.

[136] W. Y. Arms. *Digital Libraries*. The MIT Press, 2000.

[137] P. Zezula, G. Amato, F. Debole, and F. Rabitti. Tree Signatures for XML Querying and Navigation. In *Proceedings of the XML Database Symposium (XSym 2003)*, 2003.

[138] P. Zezula, F. Mandreoli, and R. Martoglia. Tree Signatures and Unordered XML Pattern Matching. In *Proceedings of 30th International Conference on Current Trends in Theory and Practice of Informatics (SOFSEM 2004)*, 2004.

[139] C. Zhang, J. F. Naughton, D. J. DeWitt, Q. Luo, and G. M. Lohman. On supporting containment queries in relational database management systems. In *Proc. of ACM SIGMOD*, 2001.

[140] D. Zhang, V. J. Tsotras, and B. Seeger. Efficient temporal join processing using indices. In *Proc. of ICDE*, 2002.

[141] K. Zhang, R. Statman, and D. Shasha. On the editing distance between unordered labeled trees. *Information Processing Letters*, 42(3), 1992.

[142] J. Zhou and J. Sander. Data bubbles for non-vector data: Speeding-up hierarchical clustering in arbitrary metric spaces. In *Proc. of 29th International Conference on Very Large Data Bases (VLDB)*, 2003.