Proseminar: XML zur Informationsintegration

Prof. Dr. Katharina Morik
Lehrstuhl Informatik VIII

June 30, 2008

Abstract

Die Informatik durchdringt inzwischen fast alle Bereiche des Arbeits- und Alltagslebens. Eine Flut von Daten wird an den verschiedensten Stellen gesammelt, die Anzahl der Seiten des World Wide Webs wurde 2005 auf 11,5 Milliarden geschätzt [5] und darüber hinaus werden unzählige Dokumente in Firmen und Privathaushalten genutzt. Dem allgemeinen Vorhandensein von Daten und Dokumenten steht eine eingeschränkte Verfügbarkeit gegenüber: nur jeweils für eine konkrete Anwendung entwickelte spezielle Programme machen bestimmte Daten für einzelne Anwendungen verfügbar. Durch Metadaten (Schemata) versucht man, eine Verallgemeinerung zu erreichen. XML ist ein Formalismus für Metadaten zu unstrukturierten Daten, ähnlich wie Relationenschemata Metadaten für Datenbanken sind. Was aber, wenn jeder ein anderes Schema verwendet?

Das Problem der Informationsintegration wird seit mehr als 10 Jahren in der Datenbankliteratur ausführlich behandelt [12, 11]. Dabei werden die einzelnen Datenbanken als Sichten auf eine intendierte globale Datenbank betrachtet. Auf der Basis logischer Regeln wird eine globale Anfrage durch Mediatoren in solche an die einzelnen Sichten überführt. Die Beschränkung auf eine einheitliche logische Modellierung aller beteiligter Datenbanken wurde durch den Abgleich von unterschiedlichen Modellen (Schema Matching) abgeschwächt [9]. Die Einbeziehung anderer Modelle, wie etwa in XML ausgedrückt und für Dokumente verwendet, erweiterte das Schema Matching (z.B. [8]). Die Ausnutzung von XML zur Anfrage an eine Dokumentensammlung (Information Retrieval) verdeutlicht die Analogie [4]. Ansätze zum Lernen unterstützen das Schema Matching [3]. Dabei geht es letztlich um die Ähnlichkeit von Bezeichnern [1]. Ob sich zwei Namen auf das selbe Objekt beziehen, kann auch aus weiteren Informationsquellen gelernt werden [7]. Das WWW kann für die Informationsintegration genutzt werden [2].

1 Vorgehen

In dem Proseminar soll zunächst der Begriff der Informationsintegration, wie er für relationale Datenbanken eingeführt wurde, geklärt werden. Dann kann

die Einbeziehung von anderen Datensammlungen, deren Schemata in XML ausgedrückt sind, in Analogie gesetzt werden. Dazu sehen wir uns XML auch praktisch an [10, 6]. Anhand verschiedener XML-Schemata für Referate versuchen wir ein einfaches Schema-Matching. Dies schafft dann das Verständnis für lernende Ansätze.

Die Studierenden lernen ein wichtiges Gebiet der Informatik kennen. In Referaten üben sie das Verstehen eines Fachtextes und seine Präsentation. In praktischen Übungen lernen sie XML kennen, so dass das Verständnis der Forschungstexte durch eigene praktische Erfahrung erleichtert wird.

References

[1] Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48, New York, NY, USA, 2003. ACM.

[2] Kevin Chen-Chuan Chang, Bin He Zhang, and Zhen. Mining semantics for large scale integration on the web: evidences, insights, and challenges. *SIGKDD Explorations Newsletter*, 6(2):67–76, 2004.

[3] William W. Cohen and Jacob Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480, New York, NY, USA, 2002. ACM.

[4] Norbert Fuhr. Xml information retrieval and information extraction. In F. Franke, G. Nakhaeizadeh, and I. Renz, editors, *Text Mining. Theoretical Aspects and Applications*. Physica, Heidelberg, 2003.

[5] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903, New York, NY, USA, 2005. ACM.

[6] Evan Lenz. *XSLT 1.0 kurz und gut.* O'Reilly, 2006.

[7] Martin Michalowski, Snehal Thakkar, and Craig A. Knoblock. Automatically utilizing secondary sources to align information across sources. *AI Mag*, 26(12):33–44, 2005.

[8] Rachel A. Pottinger and Philip A. Bernstein. Merging models based on given correspondences. In *vldb'2003: Proceedings of the 29th international conference on Very large data bases*, pages 862–873. VLDB Endowment, 2003.

[9] Erhard Rahm and Philip A. Bernstein. A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal*, 10(4):334–350, 2001.

[10] Simon St. Laurent and Michael Fitzgerald. *XML kurz und gut*. O'Reilly, 6 edition, 2006.

[11] Jeffrey D. Ullman. Information Integration Using Logical Views. In *Proceedings of the 6th International Conference on Database Theory (ICDT)*, pages 19–40, London, UK, 1997. Springer-Verlag.

[12] Gio Wiederhold. Mediators in the Architecture of Future Information Systems. *IEEE Computer*, 25(3):38–49, 1992.