



Web Information Extraction

Short Introduction to the Proseminar

<http://education.dbai.tuwien.ac.at/wie/WS05/>

Wolfgang Gatterbauer
Vienna, October 7, 2005

Web Information Extraction

Proseminar Wissenschaftliches Arbeiten, 181.081, WS 2005/06



Webpage: <http://education.dbai.tuwien.ac.at/wie/WS05/>

Place and time:

- Every Friday, 10h s.t. – 12h in SE 184/2
- October 7, 2005 – December 16, 2005 (11 weeks)

Supervisors:

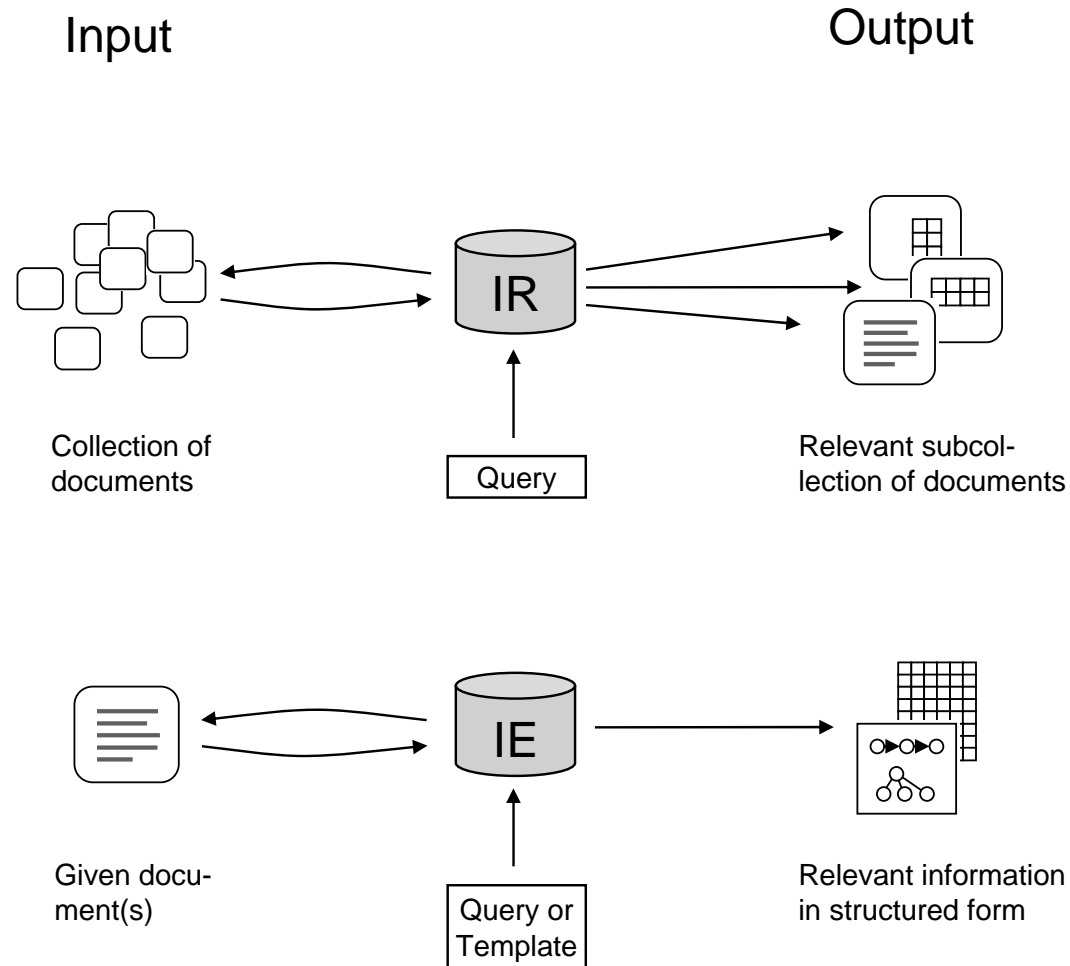
- Prof. Georg Gottlob
- Dr. Marcus Herzog

Responsible and contact person:

- Wolfgang Gatterbauer
- [gatter @ dbai.tuwien.ac.at](mailto:gatter@dbai.tuwien.ac.at)

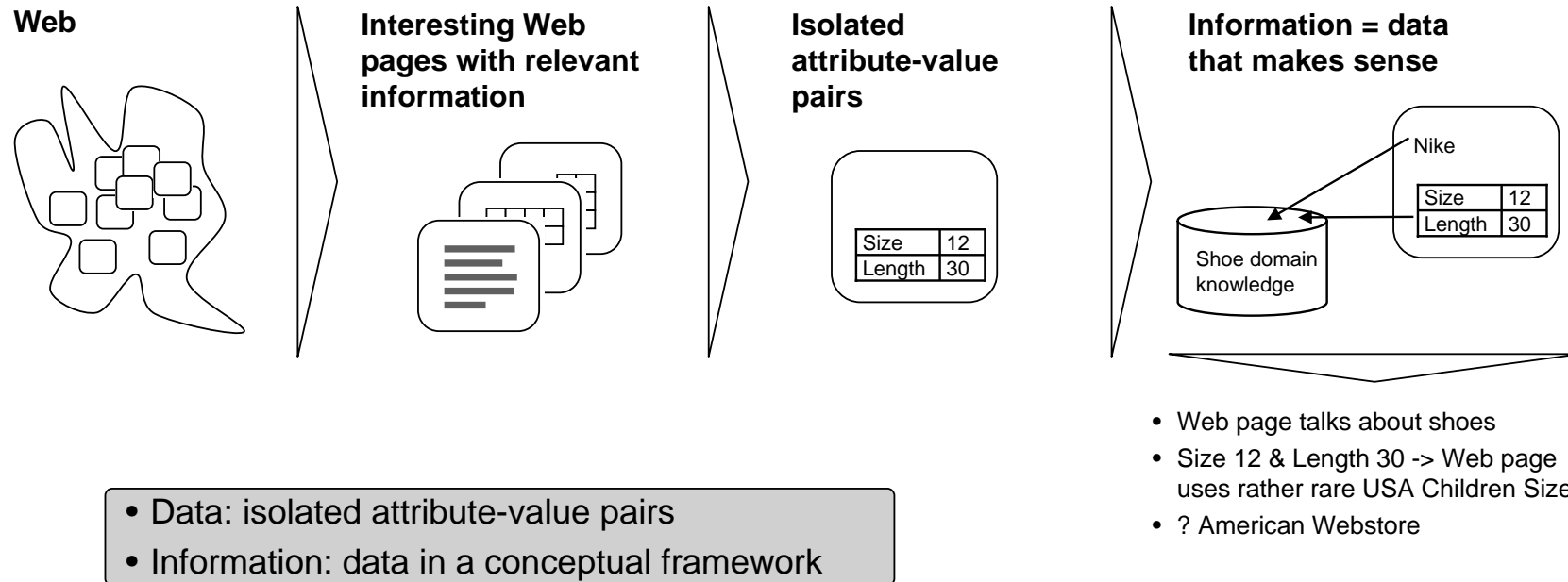
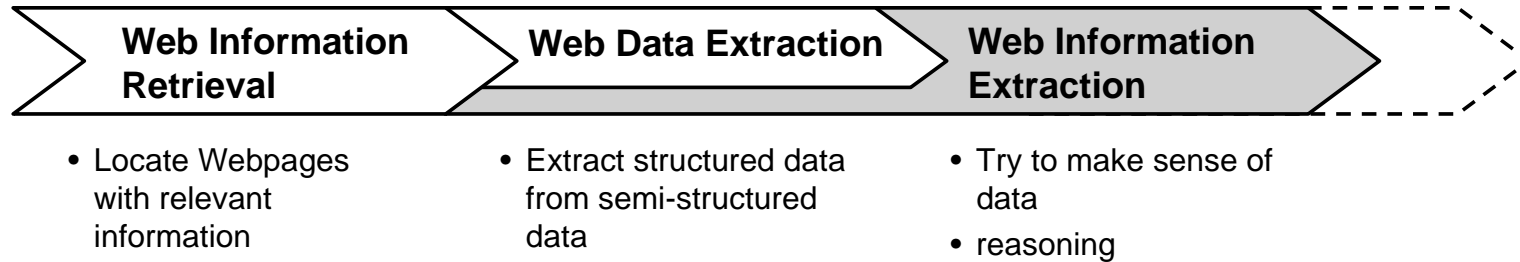
Whereas IR retrieves documents out of a collection of documents, IE extracts relevant information from such documents

INFORMATION RETRIEVAL VS. INFORMATION EXTRACTION



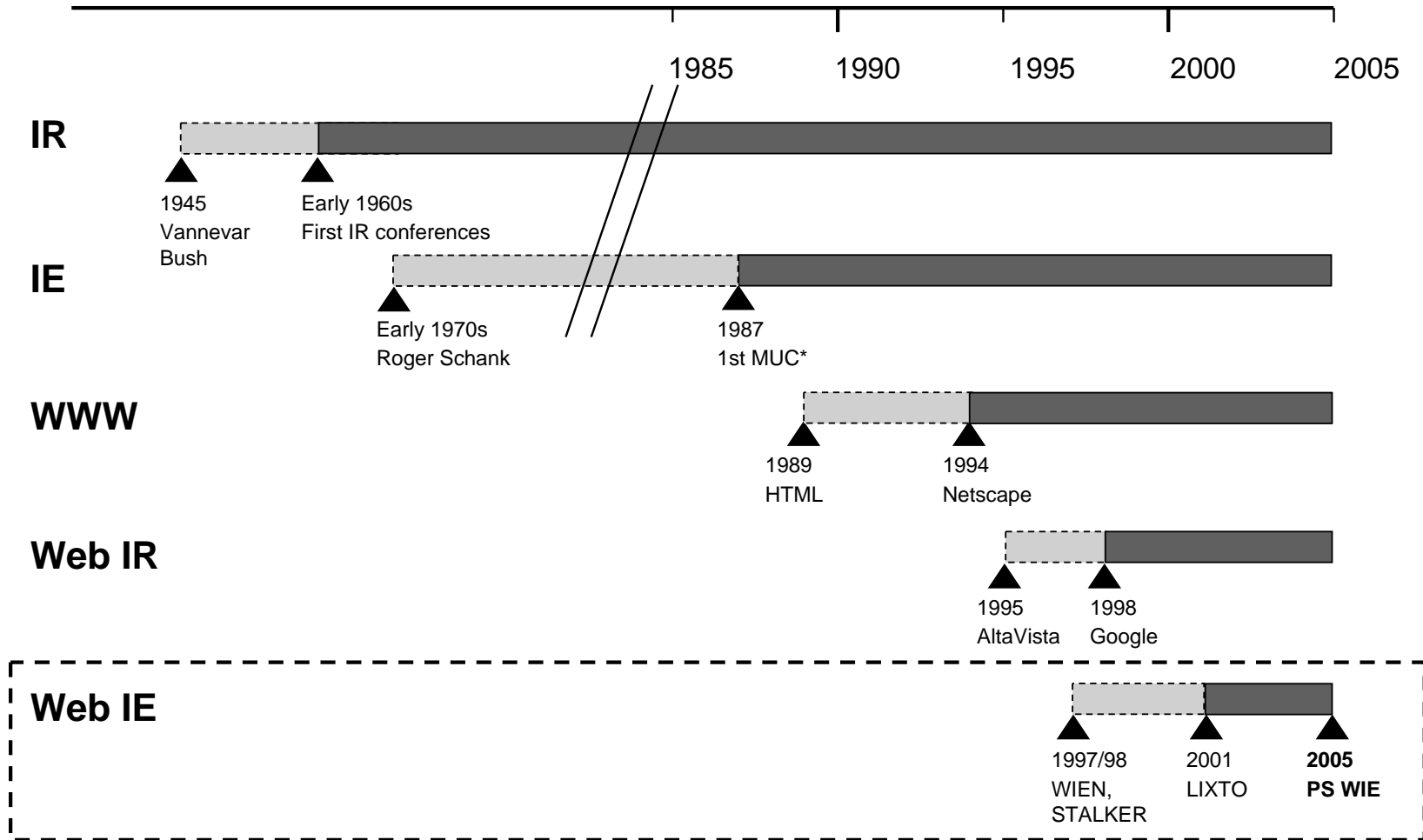
In the context of the Web, IR and IE can be seen as consecutive processes

WEB INFORMATION EXTRACTION – A PERSONAL VIEW



This Proseminar is going to focus on Web Information Extraction (WIE), a rather young research area

WEB INFORMATION EXTRACTION – TIMELINE



* MUC (Message Understanding Conferences): “Analyzing free text, identifying events of a specified type, and filling a data base template with information about each such events.”

Source: <http://education.dbai.tuwien.ac.at/wie/WS05/>

RELATED LITERATURE

- [Bau01]: R. Baumgartner, S. Flesca, and G. Gottlob. Visual web information extraction with Lixto. In Proc. of VLDB, 2001.
- [Bus45]: Vannevar Bush. As we may think. The Atlantic Monthly, July 1945.
- [Eik99]: Line Eikvil, Information Extraction from World Wide Web - A Survey, Rapport Nr. 945, July, 1999. ISBN 82-539-0429-0
- [GrS96]: R. Grishman, B. Sundheim. Message Understanding Conference 6: A Brief History. COLING-96, Copenhagen, Denmark, Europe. 1996.
- [KWD97]: N. Kushmerick, D.S. Weld, R. Doorenbos. Wrapper Induction for Information Extraction. IJCAI-97, Nagoya, 1997.
- [MBK00]: C.T. Meadow, B.R. Boyce, D.H. Kraft. Text Information Retrieval Systems. Second edition, Academic Press, 2000.
- [MMK99]: I. Musela, S. Minton, C.A. Knoblock. Hierarchical wrapper induction for semistructured information sources. Agents-99, Seattle, WA, USA, 1999.