Online market intelligence (OMI), and especially competitive intelligence for product pricing, is a very important application area for Web data extraction. However, OMI presents nontrivial challenges to data extraction technology. Sophisticated and highly parameterized navigation and extraction tasks are required. Data cleaning ‘on the fly’ is necessary in order to identify identical products from different suppliers. It must be possible to smoothly define data flow scenarios that merge and filter streams of extracted data from several Web sites and to store the resulting data in a data warehouse, where the data are subjected to market intelligence analytics. Finally, the system must be highly scalable in order to allow extraction and processing of massive amounts of data in a short period of time.

Online market intelligence. According to Wikipedia, Market Intelligence (MI) is “the information relevant to a company’s markets, gathered and analyzed specifically for the purpose of accurate and confident decision-making in determining market opportunity, market penetration strategy, and market development metrics.” Market intelligence comprises competitive intelligence as a special case. Online market intelligence (OMI) covers all aspects of MI that are related to online information sources and, predominantly, to the Web. Given that most information on pricing, product availability, store locations, and so on, is available on the Web, OMI is currently becoming the most important form of market intelligence. Currently, almost every large retail company has OMI needs for marketing and pricing. Most, but not all applications of OMI are in the area of competitive intelligence. Here are five examples of typical OMI needs. The first three are in the area of competitive intelligence, while the last two are not.

- An electronics retailer would like to get a comprehensive overview of the marked in the form of a dashboard displaying daily information on price developments including shipping costs, pricing trends, and product mix changes by segment, product, geographical region, or competitor.
- A supermarket chain wishes to be informed continually about their competitors’ product prices. Moreover, they want to be immediately informed in case a competing supermarket chain issues a special offer or promotion. They need to react very quickly to price changes or new special discounts in order to maintain their competitive position. They also want to be informed as soon as new products appear on the market.
- An online travel agency offering a best price guarantee needs to know at which prices the packages they offer are sold over the Web by competing travel agencies. Moreover, they want to be informed about the average market price of each travel product they feature.
- A road construction corporation would like to be informed about new public tenders in a dozen of different countries or states.
- Once a week, the house price index (reflecting the average prices for various property types) of a country is updated on the website of the country’s national statistics agency. Variations of this house price index immediately trigger changes in the value of shares of various industries related to the real estate market. A hedge fund wants to anticipate the house price variations by one day. To achieve this, the hedge fund needs to obtain online market intelligence by aggregating data such as asking prices, number of listings, and average time on market, from the web pages of real-estate agencies.

All these applications have in common that they require massive amounts of Web data from a relatively limited number of sources and that they are rather time critical. Moreover, in several cases aggregate data rather than raw data are needed.

OMI and Web data extraction. It is quite obvious that the main approach for solving OMI problems such as those mentioned consists in the use of Web data extraction technology. Moreover, given that the relevant sources are usually known, and
that the OMI applications require a complete coverage of the sources and a very high level of data precision, manual or semi-automatic wrapper generators for the relevant sources are preferable to generic Web harvesting tools. The latter may be used as additional tools, for example, in order to discover new sources.

Standard data extraction technology alone does not suffice for addressing advanced OMI needs such as:

(i) The possibility of defining sophisticated and highly parameterized navigation and extraction tasks;
(ii) the possibility to perform data cleaning tasks, e.g. to identify identical products from different suppliers;
(iii) features for smoothly defining data flow scenarios that merge and filter streams of extracted data stemming from several Web sites and store the resulting data into a data warehouse;
(iv) means for carrying out market intelligence analytics and for presenting the results in form of appropriate dashboards to the users;
(v) features enabling the highly scalable extraction and further processing of massive amounts of data in a short time.

Benedikt and Gottlob are, moreover, interested in data cleaning issues. Dr. Dan Olteanu is currently investigating aspects of probabilistic databases that are also closely related to data extraction and data cleaning.

Co-operation with industry. The group would welcome co-operation with industrial partners from the UK to carry out interesting projects on Web data extraction, online market intelligence, and data cleaning. Projects could range from consultancies and applicability studies to the implementation of prototypes or software pilots in order to evaluate the technological benefits of these technologies in a particular domain.

Software tools used. As a software solution, we currently use the Lixto data extraction tools whose initial version was developed at TU Vienna by Prof. Gottlob and his students before Gottlob moved to Oxford University. These tools have been further developed and are now commercialised by the Austrian company Lixto, a spin-off of TU Vienna. This advanced web data extraction technology, has been designed to access, augment and deliver content and structured data from web applications that utilise client-side processing techniques such as JavaScript, AJAX and dynamic HTML.

This allows for fast reactions to changes in the online channels. On top of the data store, enterprise-class reporting infrastructure can be used to provide all necessary reports and analytics.

Scalability. Internet Cloud Computing is used to improve the scalability and processing power needed for massive data extraction tasks.

Contact: Prof. Georg Gottlob
email: georg.gottlob@comlab.ox.ac.uk