Xcerpt and visXcerpt: Integrating Web Querying

Sacha Berger François Bry Tim Furche
University of Munich, Institute for Informatics, http://www.ifi.lmu.de/

Xcerpt [2] and visXcerpt [1], cf. http://xcerpt.org/, are Web query languages related to each other in an unusual way: Xcerpt is a *textual* query language, visXcerpt is a *visual* query language obtained by rendering Xcerpt query programs. Furthermore, Xcerpt and visXcerpt, short (vis)Xcerpt, have been conceived for querying both standard Web data such as XML and HTML and Semantic Web data such as RDF and Topic Maps.

This paper describes a demonstration focusing on three aspects of (vis)Xcerpt. First its core features, especially the pattern-oriented queries and answer-constructors, its rules or views, and its specific language constructs for incomplete specifications. Incomplete specifications are essential for retrieving semi-structured data. Second, the integrated querying of standard Web and Semantic Web data to ease the accessing of the two kinds of data in a same query program. Third, the complementary and integrated nature of the two languages.

Setting of the Demonstration. In the demonstration, prototypes of both, the textual query language Xcerpt and its visual rendering visXcerpt are demonstrated in parallel on the same examples. Both prototypes rely on the same run time system for evaluating queries, but differ in rendering: visXcerpt provides a two-dimensional *rendering* of textual Xcerpt programs implemented using mostly HTML and CSS. Additionally, the visual prototype provides an interactive environment for editing visXcerpt queries, as well as for data, query, and answer browsing.

Excerpts from DBLP[1], and from a computer science taxonomy form the base for the scenario considered in the demonstration. DBLP is a collection of bibliographic entries for articles, books, etc. in the field of Computer Science. DBLP data are representatives for standard Web data using a mixture of rather regular XML content combined with free form, HTML-like information. A small Computer Science taxonomy has been built for the purpose of this demonstration. Very much in the spirit of SKOS [3], this is a lightweight ontology based on RDF and RDFS. Combining such an ontology as metadata with the XML data of DBLP is a foundation for applications such as community based classification and analysis of bibliographic information using interrelations between researchers and research fields. Realizing such applications is eased by using the integrated Web and semantic Web query language (vis)Xcerpt that also allows reasoning using rules.

Technical Content of the Demonstration. The use of query and construction patterns in (vis)Xcerpt is presented, both for binding variables in query terms and for reassembling the variables in so-called construct terms. The variable binding paradigm is that of Datalog, i.e. the programmer specifies patterns (or terms) including variables. Special interactive behavior of variables in visXcerpt highlights the relation between variables in query and construct terms. Arguably, pattern based querying and constructing together with the variable binding paradigm make complex queries easier to specify and read. This is demonstrated by online query authoring and refactoring.

To cope with the semistructured nature of Web data, (vis)Xcerpt query patterns use a notion of incomplete term specifications with optional or unordered content specification. This feature distinguishes (vis)Xcerpt from query languages like Datalog and query interfaces like "Query By Example" [4]. Simple, yet powerful textual and visual constructs of incompleteness are presented in the demonstration.

An important characteristic of (vis)Xcerpt is its rule-based nature: (vis)Xcerpt provides rules very similar to SQL views. Arguably, rules or views are convenient for a logical structuring of complex queries. Thus, in specifying a complex query, it might ease the programming and improve the program readability to specify (abstract) rules as intermediate steps—very much like procedures in conventional programming. Another aspect of rules is the ability, to solve simple reasoning tasks. Both aspects of rules are needed for the demonstration scenario.

Referential transparency and answer closedness are essential properties of Xcerpt and visXcerpt, surfacing in various parts of the demonstration. They are two precisely defined traits of the rather vague notion of "declarativity". Referential transparency means that within a definition scope, all occurrences of an expression have the same value, i.e., denote the same data. Answer-closedness means that replacing a sub-query in a compound query by a possible single answer always yields a syntactically valid query. Referentially transparent and answer-closed programs are easy to understand (and therefore easy to develop and to maintain), as the unavoidable shift in syntax from the data sought for to the query specifying this data is minimized.

A novelty of the visual language visXcerpt is how it has been derived from the textual language: as a rendering without changing the language constructs and the runtime system for query evaluation. This rendering is mainly achieved via CSS styling of the constructs of the textual language Xcerpt. The authors believe that this approach to twin textual and visual languages is promising, as it makes those languages easy to learn—and easy to develop. The first advantages is highlighted in the demonstration by presenting both languages side-by-side.

[1] http://www.informatik.uni-trier.de/~ley/db/

This research has been funded by the European Commission and by the Swiss Federal Office for Education and Science within the 6th Framework Programme project REWERSE number 506779 (cf. http://www.rewerse.net/).

References.

[1] S. Berger, F. Bry, S. Schaffert, and C. Wieser. Xcerpt and visXcerpt: From Pattern-Based to Visual Querying of XML and Semistructured Data. In *29th Intl. Conf. on Very Large Data Bases*, 2003.

[2] S. Schaffert and F. Bry. Querying the Web Reconsidered: A Practical Introduction to Xcerpt. In *Extreme Markup Languages*, 2004.

[3] W3C. *Simple Knowledge Organisation System (SKOS)*, 2004.

[4] Moshé M. Zloof. Query-by-Example: A Data Base Language. *IBM Systems Journal*, 16(4):324–343, 1977.