The TIJAH XML Information Retrieval System*

Henk Ernst Blok[1], Vojkan Mihajlović[1], Georgina Ramírez[2],
Thijs Westerveld[2], Djoerd Hiemstra[1], Arjen P. de Vries[2]

[1] University of Twente
PO Box 217, 7500 AE
Enschede, The Netherlands

[2] Centrum voor Wiskunde en Informatica (CWI)
P.O. Box 94079, 1090 GB
Amsterdam, The Netherlands

[1] {h.e.blok, v.mihajlovic, d.hiemstra}@utwente.nl
[2] {georgina.ramirez, thijs.westerveld, arjen.de.vries}@cwi.nl

Categories and Subject Descriptors:
H.3.3:[**Information search and retrieval**]: Retrieval models; H.2.3:[**Languages**]: Query languages; H.2.4:[**Systems**]: Multimedia databases, Query processing, Textual databases

General Terms: Experimentation, Verification

Keywords: Information retrieval, databases, structured documents, XML, region algebra

Not many XML information retrieval (IR) systems exist that allow easy addition of and switching between different IR models. Especially for the scientific environment where building a system takes a lot of time and keeps researchers away from the real work, i.e., investigating what is the most effective IR model, a platform that would provide this functionality would be ideal. For this reason we developed such an XML IR system. It is centered around a logical algebra, named score region algebra (SRA), that enables transparent specification of IR models for XML databases (see [1] for more details).

The transparency is achieved by a possibility to instantiate various retrieval models, using abstract score functions within algebra operators, while logical query plan and operator definitions remain unchanged. Our algebra operators model three important aspects of XML IR: element relevance score computation, element score propagation, and element score combination. To implement a new IR model, one only needs to provide definitions for these abstract function classes.

To illustrate the usefulness of our algebra our demo system supports several, well known IR scoring models (e.g., Language Models, Okapi, and tf.idf), combined with different score propagation and combination functions. The user can select which model to use at run time.

Following good practice in database systems design, our prototype system has a typical three-layered architecture. (1) The *conceptual layer* takes a NEXI [3] query expression as input, e.g.,

```
//article[about(.,java)]
  //sec[about(.,implementing threads)]
```

This is passed through a filter to standardize/sanitize (e.g., stemming and stopping) it, and produces an SRA expression. (2) The *logical layer* takes the SRA expression, e.g.,

```
R0 := SELECT_NODE("Root");
R1 := SELECT_NODE("article");
R2 := R1 CONTAINED_BY R0;
R3 := SELECT_TERM("java");
R4 := R2 P_CONTAINING_T R3;
R5 := R4 SCALE 1.000000;
R6 := SELECT_NODE("sec");
R7 := R6 P_CONTAINED_BY R5;
R8 := SELECT_TERM("implementing");
R9 := R7 P_CONTAINING_T R8;
R10 := R9 SCALE 1.000000;
R11 := SELECT_TERM("threads");
R12 := R7 P_CONTAINING_T R11;
R13 := R12 SCALE 1.000000;
R14 := R10 P_AND R13;
```

Here each `Rxx` is a set of scored regions, i.e., XML elements/terms with their score. Note that behavior of the `SELECT_TERM` and `P_CONTAINING_T` operators in many cases will not be strict selection and strict containment and depends on the used IR model.

Next the SRA expression is translated into a physical expression, taking into account the user's choice of which IR model to use. Due to the modular setup, changing retrieval models is straightforward: just plug in a different translation module. Building a new translation module – to capture yet another IR model to experiment with – takes less than half an hour. (3) The *physical layer* executes the query. It is based on a low-level physical DB engine, i.e., MonetDB [2]. At this level the score region algebra operators are implemented as a set of procedures in Monet Interpreter Language (MIL).

For more info on the TIJAH system and the CIRQUID project please visit `http://www.cs.utwente.nl/~cirquid/`.

*This research was performed as part of the Complex Information Retrieval Queries in a Database (CIRQUID) project, funded by the Netherlands Organization of Scientific Research (NWO) under grand no. 612.061.210.

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

1. REFERENCES

[1] V. Mihajlović, H. E. Blok, D. Hiemstra, and P. M. G. Apers. Score Region Algebra: Building a Transparent XML-IR Database. In *ACM Conference on Information and Knowledge Management (CIKM '05)*, Oct. 2005.

[2] MonetDB. http://monetdb.cwi.nl/.

[3] A. Trotman and B. Sigurbjörnson. Narrowed Extended XPath I (NEXI). In *International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX '04)*, volume 3493 of *Lecture Notes in Computer Science*, Jan. 2005.