# W3QS: A Query System for the World-Wide Web

**David Konopnicki** **Oded Shmueli**

Computer Science Department, Technion, Haifa, Israel
*http://www.cs.technion.ac.il/~konop*

***Proceeding of 21th VLDB Conference, 1995***

1996.2.5.

## Content

- **Introduction**
- **Query Language**
- **System Architecture**
- **Conclusion**

# Introduction

- **World-Wide Web**
  - it can be viewed as a gigantic database (mostly **read-only**)
  - the browsers enable roaming through sites of interests, however, they are **not** query processors
  - the current status is analogous to that of a huge **file system**, or a **document retrieval system**, with many indexes but without a convenient facility for **querying** this information



# Introduction

- **Difficulties in Searching**
  - there is no reliable road map for the WWW
    - **constantly growing size**
  - it is difficult to analyze obtained information
    - **heterogeneous data format**
  - it is difficult to search for information related to the organization of the hypertext
    - **uncertain semantic of hyperlink**

## Introduction

- **Critique of Search Engines**
  - ☐ replication of information
  - ☐ most appropriate for text data
  - ☒ they do**not** capture the hypertext structure of data
  - ☐ indexes become repidly obsolete
- **Critique of Navigating Tools**
  - ☒ they do**not** provide information retrieval facility
  - ☐ the graph representation can be useful for only a very small portions of the whole WWW



## Introduction

- **Main Goal of this work**
  - ☐ design and construct a high-level query language for **locating**, **filtering** and **presenting** WWW-held information
    - **specify the syntax and semantics of a high-level SQL-like query language**
    - **provide a view maintenance facility at a much higher level than robot maintained indexes**
    - **provide advanced display facilities for gathered information**

## Query Language

- **Content Queries**
  - ☐files that have strict inner structure
    - **BibTex files, UNIX environment files...**
  - ☒semi-structured files that contain formatting code
    - **Latex files, HTML files...**
  - ☐raw files
    - **pure text files, image and sound files...**
- **SQLCOND**
  - ☐state conditions about contents or join conditions
    - **use the file name, and the *file* UNIX utility**



## Query Language

- **Query Example 1.**

  SELECT **cp n3/* result**; FROM (n1, l1), l2, n3;
  WHERE **SQLCOND**(n1.**format** = HTML) AND
  (l1.**REL** = "example") AND (n3.**name** = "*.gif");
- **Structure-specifying Queries**
  - ☒a pattern definition sub-language is used
    - **graph pattern, pattern[-specifying] graph**
  - ☐finding sets of WWW nodes that satisfy a pattern
  - ☒form completion

# Query Language

- **Query Example 2.**

```
SELECT cp n2/* result;
FROM n1, l1, n2;
WHERE n1 in ImportantIndex.url;
        FILL n1.form AS IN ImportantIndex.fil
            WITH keyword="A. Einstein";
        SQLCOND (n2.format = Latex) AND
                (n2.author = "A. Einstein");
```

## Query Language

- **Query Example 3.**

```
SELECT CONTINUOUSLY SQLPRINT n2.url;
FROM n1, l1, n2;
WHERE n1 in ...;
        FILL n1.form AS IN...;
        RUN learnformat IF n1.form
              UNKNOWN IN ImportantIndex.fil;
        SQLCOND (n2.format = Latex)....;
EVALUATED EVERY week;
```



## Query Language

- **Query Example 4.**

```
SELECT cp art/* result;
FROM ind, l1, chap, l2, ref, l3, art;
WHERE SQLCOND (ind.url =
          "http://cs.technion.ac.il/BookIndex.html")
              AND (chap.url = /.Chapter-1.html/)
               AND (l2.HREF = /.#$l3.NAME/);
USING BFS;
```

# System Architecture

- **Main Modules**
  - ☐query processor
  - ☐remote search program
  - ☒libraries
    - **algorithm library, condition library, format library**
- **RSP Construction Kit Functions**
  - ☐structure-specifying graph functions
  - ☐access functions of WWW
  - ☒extension to inter-RSP communication

# Conclusion

- **Current Status**
  - complete the design of query language
  - construct a prototype system **W3QS**
- **Future Works**
  - utilize robot-constructed indexes in query processing optimization
  - include a view maintenance facility in **W3QS**
  - provide advanced display facilities for extracted information

# Conclusion

- **Contributions**
  - form completion
    - **robot programming, machine learning**
  - deal with changing information
    - **database view, information filtering**
  - flexible query definition
    - **invoke external UNIX programs in condition clause**
    - **allow unbounded-length paths in query definition**