UNIVERSITÄT LEIPZIG

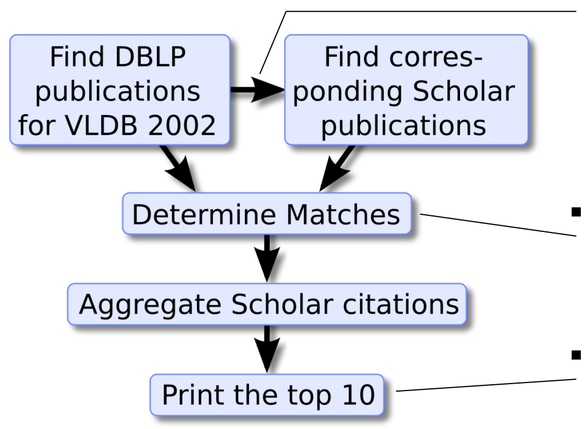
WETSUIT: An Efficient Mashup Tool for Searching and Fusing Web Entities

Stefan Endrullis, Andreas Thor, Erhard Rahm
Database Group, University of Leipzig
http://dbs.uni-leipzig.de

Motivation

What are the top cited papers of the VLDB 2002?

- Tasks involved:

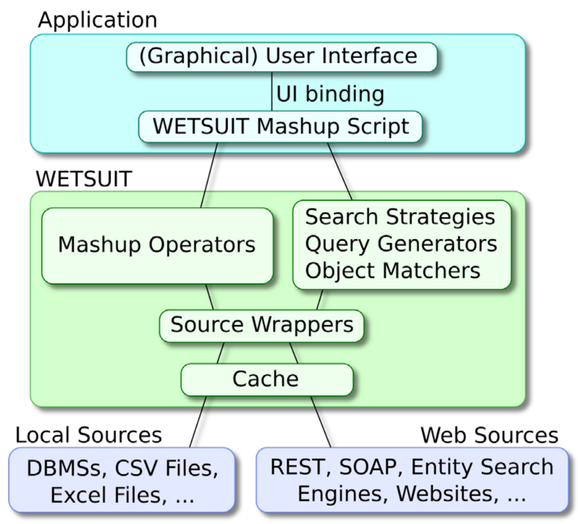


- Requires efficient entity search:
  e.g. start with one venue query, refine with a keyword query for each DBLP publication
- Requires entity resolution:
  e.g. compare pub. titles and authors
- Present intermediate results to reduce waiting time for user

WETSUIT

What is WETSUIT?

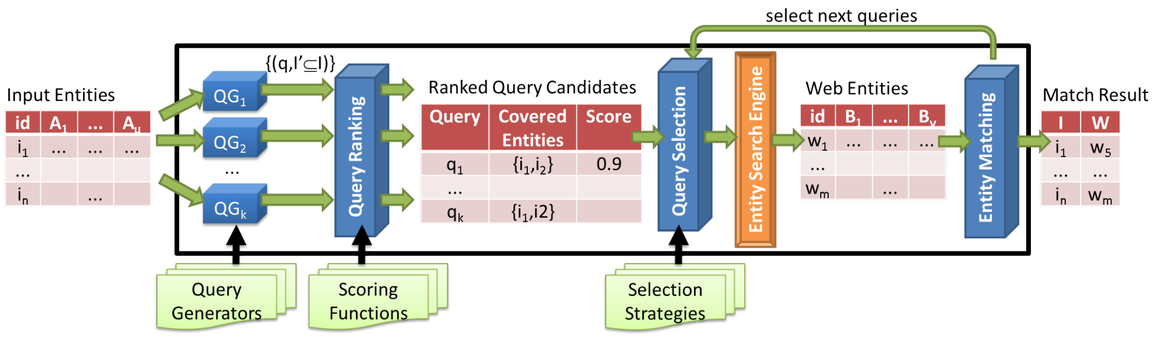
- WETSUIT: Web EnTity Search and fUsIon Tool
- Efficient and powerful mashup framework



System architecture

Features

- Efficient execution of mashup workflows
  - Efficient and effective entity retrieval from entity search engines using **advanced search strategies** which exploit multiple query generators
  → may increase recall while reducing the number of search queries



  - **Pipeline parallelism**: all workflow operators work in parallel as long as they have input entities to process
  - **Data parallelism**: each operator may process multiple entities simultaneously

Interaction of operators in mashup workflows

- **Entity resolution** to deal with dirty data
- **Fast presentation of first results**
  - By streaming partial results immediately to the next operators
  → **Result refinement in background**
  - Classical blocking operators (e.g. filterTop, groupBy, aggregate) stream preliminary results, i.e., they may update or revoke them later
- Mashups definition language **embedded in Scala**
  - Concise and well-readable syntax
  - **High extensibility** due to Scala and WETSUIT's object oriented design
  - Simple integration of existing Java/Scala libraries in WETSUIT mashups (e.g. for entity extraction from HTML documents) due to Java Bytecode compatibility
  - Well supported by development tools (e.g. Eclipse, Intellij IDEA, maven, ant)
- **Automatic GUI generation** based on workflow definition
- **Open Source** (GNU Affero General Public License (AGPL) Version 3)

Mashup Operators

- Set-based; support streaming of partial and preliminary results

Selected mashup operators of WETSUIT

Operator	Definition
inputOne (*label*) inputMany (*label*)	asks the user for one/many input value(s); *label* denotes the caption of the GUI component
selectOne (*label*) selectMany (*label*)	displays intermediate results and asks the user to select one/many of them
outputOne (*label*) outputMany (*label*)	presents results
map (*f*)	maps each input entity to a new entity using the mapping function *f*
flatMap (*f*)	maps each input entity to a set of new entities using the mapping function *f*
filter (*cond*)	filters the input entities based on the filter condition *cond*
groupBy (*groupAttr*)	groups the input entities by *groupAttr*; groupBy has to be followed by filterTop or aggregateValue
filterTop (*n*) by (*orderAttr*)	for each group/whole input set: filters the top *n* elements on *orderAttr*
aggregateValue (*value*) via (*agg*, *[rev]*)	for each group/whole input set: aggregates *value* using aggregation function *agg* and its reverse operation *rev* (*rev* is needed if input entities can be revoked)
findAt (*ese*)	searches for the input entities at entity search engine *ese* using an implicitly defined search strategy
matchWith set_2 using (*matcher*)	matches the input entities against entities of set_2 by using the match strategy *matcher*
union, intersect, diff	classical set operators
join set_2 on (*theta*)	θ-join of two input sets using the binary function *theta* as join condition
buffer (*minDelay*, *maxDelay*)	buffers new entities and entity revokes for at least *minDelay* and at most *maxDelay* milliseconds

Mashup Example: Online Citation Service

Screenshot

Journal/conference: VLDB 2002 [Search]

Select a journal/conference

FULL TITLE	TITLE	TYPE	YEAR
VLDB 2002, Proceedings of 28th International Conference on Very Large Data I	VLDB	Conference	2002
Efficiency and Effectiveness of XML Tools and Techniques and Data Integratior	EEXTT	Conference	2003

DBLP pubs with citations

CITATIONS	TITLE	AUTHORS	YEAR
966	Approximate Frequency Counts over Data Stre	Gurmeet Singh Manku, Rajeev Motwa	2002
949	Streaming Queries over Streaming Data.	Sirish Chandrasekaran, Michael J. Fr	2002
921	COMA - A System for Flexible Combination of S	Hong Hai Do, Erhard Rahm	2002
853	Monitoring Streams - A New Class of Data Mai	Donald Carney, Ugur Çetintemel, Mit	2002
694	XMark: A Benchmark for XML Data Managemei	Albrecht Schmidt, Florian Waas, Mart	2002
641	Hippocratic Databases.	Rakesh Agrawal, Jerry Kiernan, Ram	2002

Workflow definition

```
inputOne ("Journal/conference: ", "VLDB 2002")
// search for corresponding DBLP journals or conferences
flatMap (name => Dblp.Jonf.where("full_title like ?", "%"+name+"%"))
selectOne ("Select one journal/conference")
flatMap (_.publications)
// search for the selected DBLP publications at Google Scholar
findAt (Scholar)   // returns a set of correspondences (DBLP pub., GS pub., similarity)
// for each Scholar publication find the best matching DBLP publication
groupBy (_.range) filterTop 1 by (_.sim)
// summarize the Scholar citation counts for each DBLP publication
groupBy (_.domain) aggregateValue (_.range.citations) via (_+_, _-_)
outputMany ("DBLP pubs with citations")
```

Related Work

- S. Endrullis, A. Thor, and E. Rahm. Entity Search Strategies for Mashup Applications. In ICDE, 2012.
- A. Thor and E. Rahm. CloudFuice: A Flexible Cloud-Based Data Integration System. In ICWE, 2011.
- S. Endrullis, A. Thor, and E. Rahm. Evaluation of Query Generators for Entity Search Engines. In Int. Workshop on USETIM, 2009.

Printed by Universitätsrechenzentrum Leipzig