Web Science – **Investigating the Future** of Information and Communication

Social Computing for Libraries: Data De-Duplication Through the Crowd

Claudiu S. Firan, Mihai Georgescu, **Wolfgang Nejdl**

Bielefeld Conference 2012

The FreeSearch System

- Search engine for digital libraries
- Simple to use interface
- Intuitive functionalities
- Easily scalable

- Now with focus on
  Duplicate detection and duplicate merging
  using collaborative intelligence

http://freesearch.isearch-it-solutions.net

The FreeSearch System

Data Sources

- DBLP (2M documents)
  - Good quality metadata
  - No abstracts, no fulltext
  - No duplicates
- TIBKat (2M documents)
  - Several languages
  - Some duplicates
- CiteSeer (1.2M documents)
  - Citation information
  - Some duplicates
- BibSonomy (0.5M documents)
  - User generated metadata
  - Many duplicates

Active Users

- About 200 unique users Mo-Fr
- 11% from Germany

- Improvement through regular usage analysis

- Provide regular users with the possibility to
  - Clearn up their own publications
  - Adapt results to their needs

Search history, Favorites

Free Search
Toward Computer Science Ideas

information retrieval

Query Syntax
Favorites
Search History
Options & About

The original query is translated and searched as:

- (information retrieval) OR (Information-retrieval) OR (regăsirea de informații)

Automatic query translation

For more specific results, try to search:

- information as In or Publisher
- retrieval as In or Publisher

Field suggestions

Faceted search

- Type
- Language
  - English (6358)
- Year
- By
  - Jian-Yun Nie (44)
  - C. J. van Rijsbergen (42)
  - Fabio Crestani (42)
  - Norbert Fuhr (40)
  - Gerard Salton (39)
  - [more]
- In
- Publisher
- Source

- Topic
  - International Workshop Biomedical Information Retrieval
  - Medical
  - Music Information Retrieval
  - Query
  - Mobile
  - Applications
  - Content
  - Data Fusion
  - Machine Learning

Topical clustering

6,352 (6,358 including duplicates) results for *information retrieval* sorted by most recent

Question Answering and Multi-search Engine ... rmation Retrieval.

By: Fernando S. Peregrino, David Tomás, Fer...
In: CICLing (2), 2012

[More] [Full Text] [Bibtex] [Google] [+Favorites] [Similar]

Intuitive fields: "by", "in"

Ensemble Approach for Cross Language Information Retrieval.

Authors: Dinesh Mavaluru, R. Shriram, W. Aisha Banu
Book: CICLing (2) Pg. 274-285 [contents]
Year: 2012
Language: English
Type: inproceedings (conf)
Source: DBLP

[Less] [Full Text] [Bibtex] [Google] [+Favorites] [Similar]

Social bookmarking

iTrust: Trustworthy Information Publication, Search and Retrieval.

By: P. Michael Melliar-Smith, Louise E. Moser, Isai Michel Lombera, Yung-Ting Chuang
In: ICDCN, 2012

[More] [Full Text] [Bibtex] [Google] [+Favorites] [Similar]

A static technique for fault localization using character n-gram based information retrieval model.

By: Sangeeta Lal, Ashish Sureka

Faceted Search

- Users can drill down on document type, language, year, persons, venue, publisher, general tags and data source

Topical Clustering

- Displays topics generated from the result documents
- Instant topic generation, similar to facets

Intuitive Fields: “by”, “in”

- Author, editor, contributor → by
- Book, conference, venue, series, journal, year → in
- “in 2012” → “year:2012”; “before 2011” → “year<2011”

Field Suggestions

- Suggest specific fields to search in
- “bielefeld” → “in:bielefeld”

Search History & Favorites

- All past queries appear in the search history
- Auto-completion of past queries (in addition to author names)

Social Bookmarking

- Sharing on over 330 services using AddThis
- Integration with the BibSonomy literature sharing system

Automatic Query Translation

- User query translated to find international documents in different languages of interest
- Using Microsoft Translator API

Spelling Correction

- Did you mean … ?

BibTex Export

- For every publication
- For all publications in Favorites

Data De-Duplication based on algorithms and humans

Duplicated results are grouped together, only one is shown, and the number of versions.

For each result a query for similar documents can be made. The results will be split into similar and duplicates.

Users have the opportunity to correct the duplicates detected directly via the interface

In order to improve the performance of duplicate detection crowd sourcing is used via Amazon Mechanical Turk.

Similar Publications vs. Additional Versions: User Feedback

Correction/ Validation of detected duplicates

Detected Duplicate

Possibility to mark similar publications as duplicates

Similar publication

Why Finding Entities in Wikipedia is Difficult, Sometimes

By: Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, Ralf Krestel, Wolfgang Nejdl
In: Information Retrieval, 2010

[More] [Bibtex] [Google] [+Favorites] [Similar]

Additional versions:

[This is NOT a duplicate] [YES, this is a duplicate] [Show Diff]
Why finding entities in Wikipedia is difficult, sometimes.

Authors: Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, Ralf Krestel, Wolfgang Nejdl
Journal: Inf. Retr. Vol. 13 No. 5 Pg. 534-567 [Contents]
Year: 2010
Language: English
Type: journal (article)
Source: DBLP

[Less] [Full Text] [Bibtex] [Google] [+Favorites] [Similar]

Similar publications:

[Mark as a duplicate]
A Model for Ranking Entities and Its Application to Wikipedia.

Authors: Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, Ralf Krestel, Wolfgang Nejdl
Book: LA-WEB Pg. 29-38 [Contents]
Year: 2008
Language: English
Type: conference (inproceedings)
Source: DBLP

[Less] [Full Text] [Bibtex] [Google] [+Favorites] [Similar]

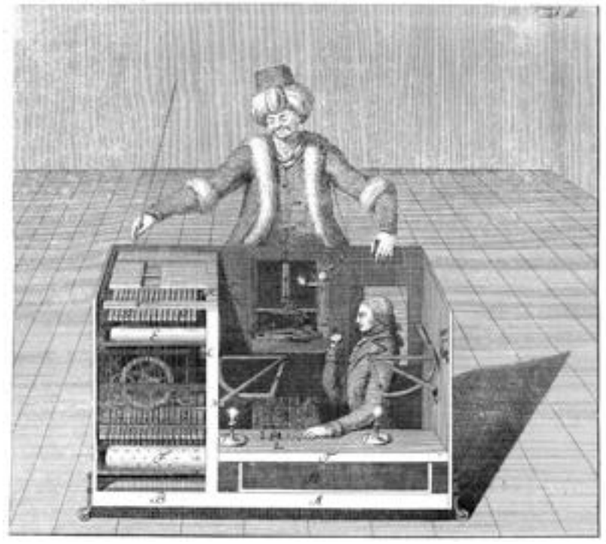
Duplicate Scorer Algorithm

- Combines multiple text similarity features
- Levenshtein distance; Jaccard similarity; field-sensitive matching

- Combining user feedback and algorithm results
  1. Current user feedback
  2. Majority feedback
  3. Similarity score

Social Computing using amazon mechanical turk™ Artificial Artificial Intelligence

- Crowdsourcing Internet Marketplace
- Enables the co-ordination of human intelligence to perform tasks that computers are still unable to do.
- Mechanical Turk Requesters and Workers
- Micro-jobs called "HITs" (human intelligence tasks)
- Microtasks posted on MTurk pay about $2,000 per day in total



De-Duplication using Amazon Mechanical Turk

The documents that are around the threshold, for which the assignment as duplicates is not certain will be sent to AMT, so that humans can decide.

A HIT is composed of 5 pairs of publications, and the users on AMT have to classify them as duplicates or not. There has to be an agreement between 3 of the users on a pair.

Using the crowd for duplicates detection

The output from crowd sourcing via AMT is used directly, and to improve our automatic duplicates detection algorithm

Find the parameter choice that gives the biggest overlap between the output of the algorithm and the human feedback.

The performance of our scorer improves with each resolved HIT

Data Merging

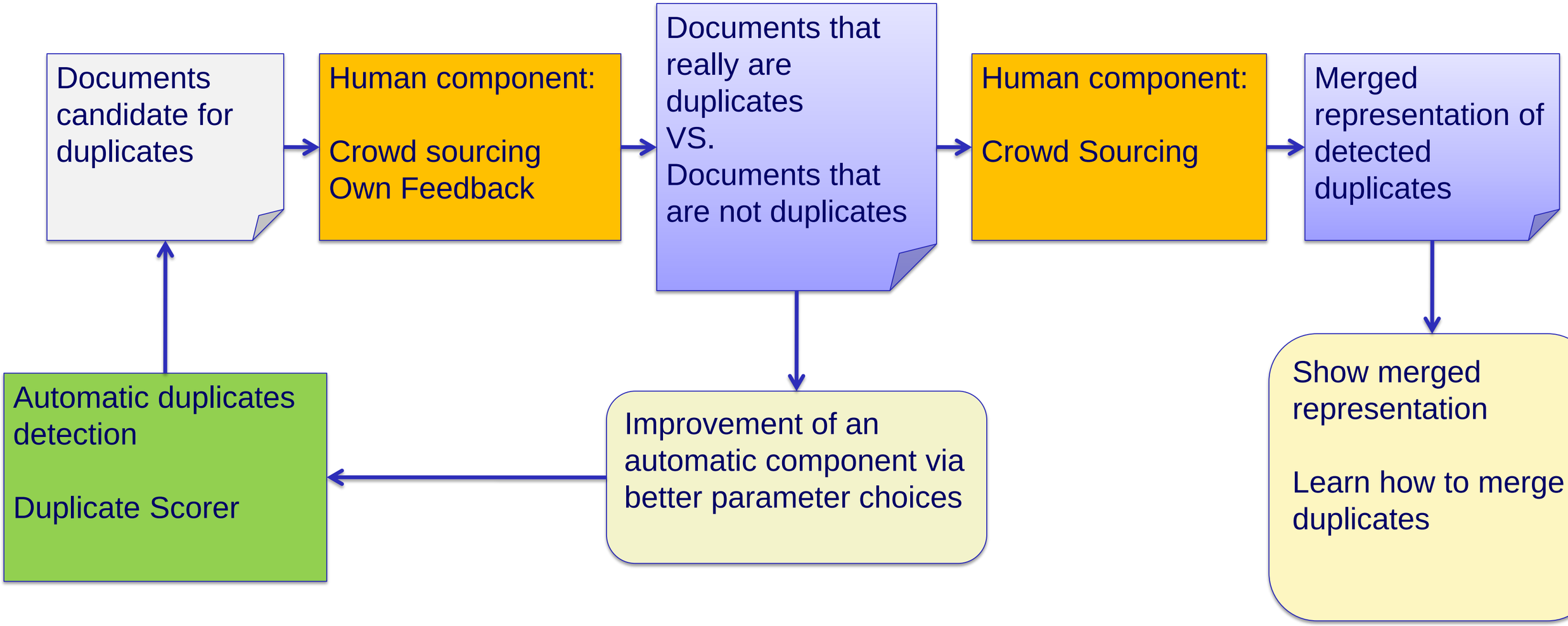
We also use the crowd to create a merged representation of two publications detected to be duplicates.

Find which fields are more relevant, have most differences, can be easily merged.

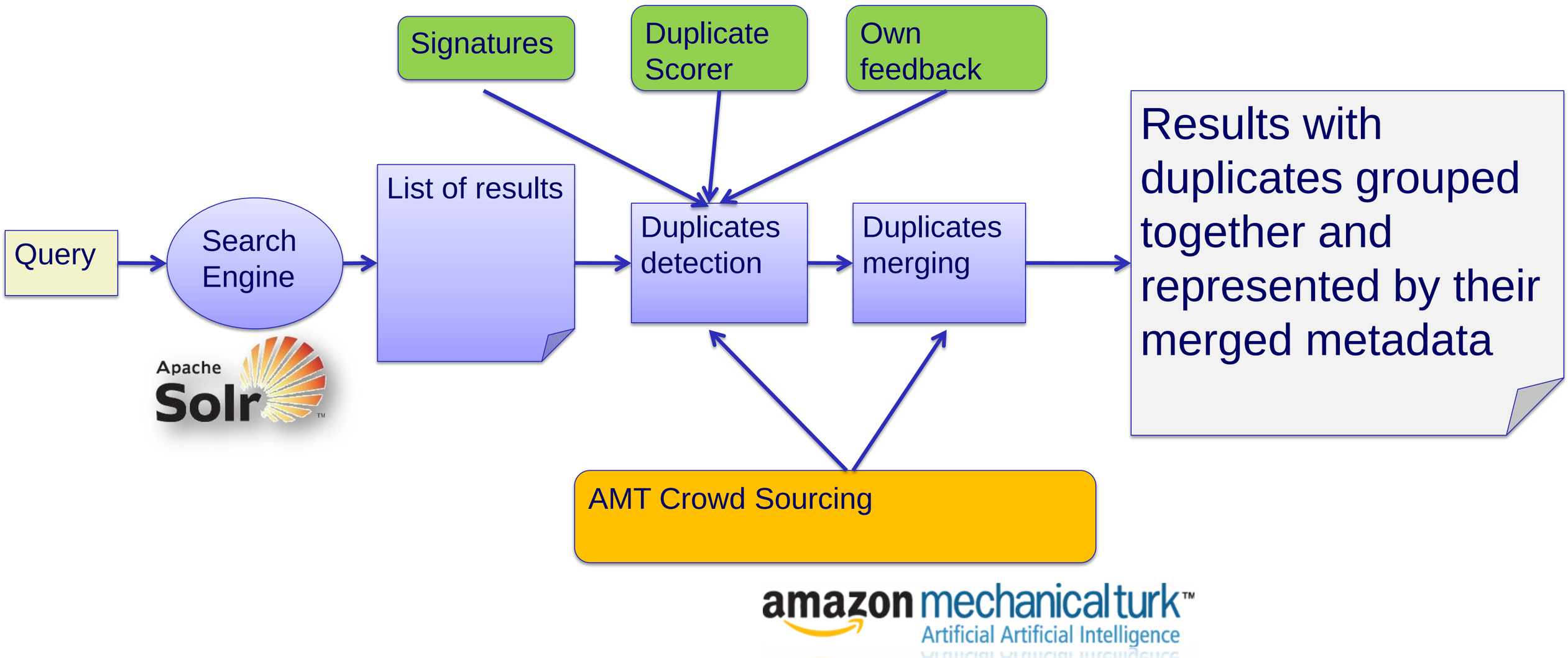
AMT users are presented with 2 pairs of publications that are supposed to be duplicates, and they have to do the merging.

Developing an algorithm that can learn from the human input to do the merging automatically is an obvious next step.

Crowd Sourcing: Human in the Loop



Duplicate Detection Workflow



http://freesearch.isearch-it-solutions.net