Document clustering as a record linkage problem

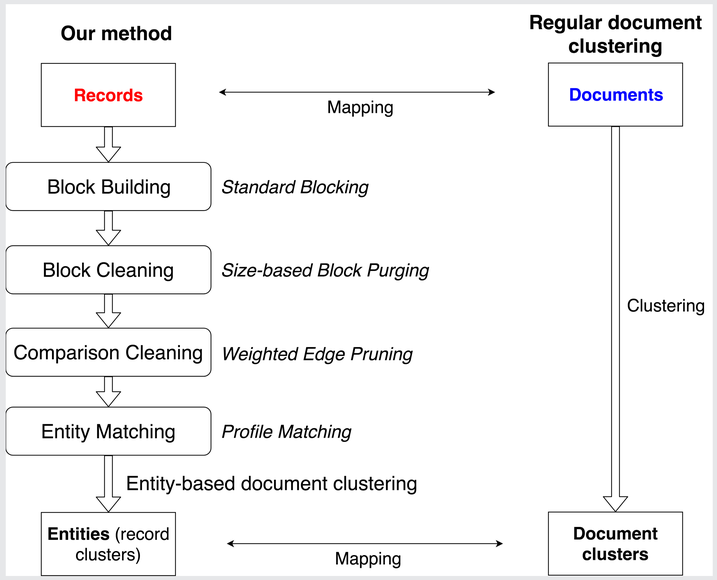
Nikiforos Pittaras, George Giannakopoulos, Leonidas Tsekouras
Institute of Informatics and Telecommunications, N.C.S.R. "Demokritos", Greece

Iraklis Varlamis
Department of Informatics and Telematics, Harokopio University of Athens, Greece
varlamis@hua.gr

Contributions

- Formulation of document clustering as a record linkage problem
- Introduction of large-scale record comparison approaches to document clustering
- Better scalability using Record Linkage blocking techniques
- Evaluation on diverse entity and text-based representations, multiple datasets and metrics

Process Flow



Proposed method

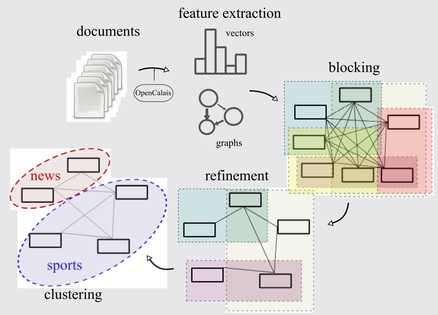
- JedAI Record Linkage toolkit, a modular entity matching software solution.
- Blocking and comparison pruning in the similarity graph.
- Multiple features, both vector-based (Bag of Words, TF-IDF, Bag of N-Grams) and graph-based (word/character N-Gram Graphs) text representation approaches.
- Named-entity, text-based and mixed document representations.

Information and Representation

- Vector Space Model (VSM)
  - Bag Of Words / n-grams, $n \in \{1, 3\}$
  - Frequency histogram / TF-IDF weighting
  - Cosine similarity
- Graph Based Model
  - Word / character n-gram graphs ($n = 3$)
  - Normalized Value Similarity (NVS)
- Mixed Graph Model
  - Word n-gram graphs ($n = 1$) for entities, top TF-IDF words
  - Normalized Value Similarity (NVS)

Example

- A JedAI record linkage document clustering workflow example.



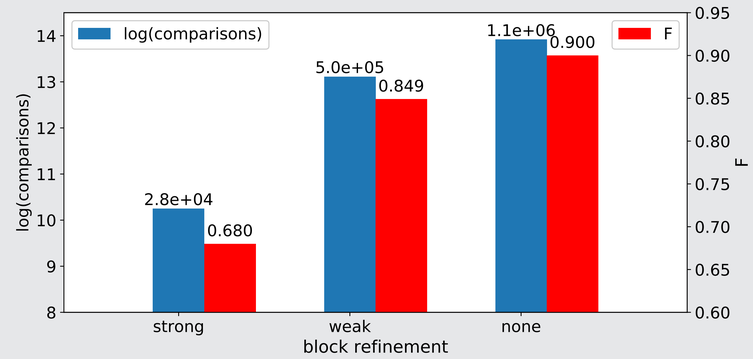
D1: **Trump** says ready to meet **Iran**'s **Rouhani**
D2: **Graham** ready to meet with **Trump** to start an initiative for health care.
D3: **Dimon** says **Trump**'s tax cut and deregulation have 'accelerated growth'.
D4: **Trump** lawyer **Jamie Dimon**: Deregulation is not a crime, it will lead to growth.
D5: **Graham** says health care initiative with **Dimon** and **Trump** may start.*
*All names and statements are fictional.

Experimental evaluation

- Datasets
  - 20Newsgroups (7.5K English texts, 20 topics)
  - Multiling'15 (15 event classes, 400 EN, ES, FR texts)
- Entity Extraction: OpenCalais API
- Clustering Evaluation Measures: Micro F-measure, Clustering Precision (CPr), Penalized Clustering Precision ($PCPr$)

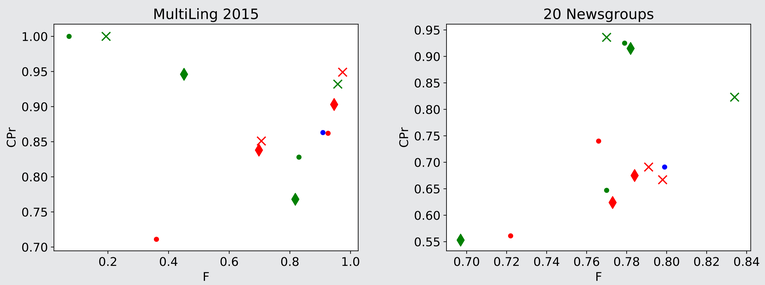
Results

- Time complexity



- Clustering performance

Symbol (representation)	Color (Information used)
◇: BoW	Red: entities only
×: TF-IDF	Green: text only
●: N-gram graphs	Blue: entities + text



(a) MultiLing 2015 (b) 20 Newsgroups

Findings

- **Blocking offers significant speedup** (1 to 2 orders of magnitude fewer comparisons) with 5.6% to 24.4% performance drop.
- **TF-IDF unigrams** achieve the best scores with both entities (Multiling) and text (20Newsgroups), in terms of F-score
- **TF-IDF trigrams** and **N-Gram graph word trigrams** achieve the best scores with text, in terms of CPr.