Linked Movie DataBase

Oktie Hassanzadeh
Mariano Consens
University of Toronto

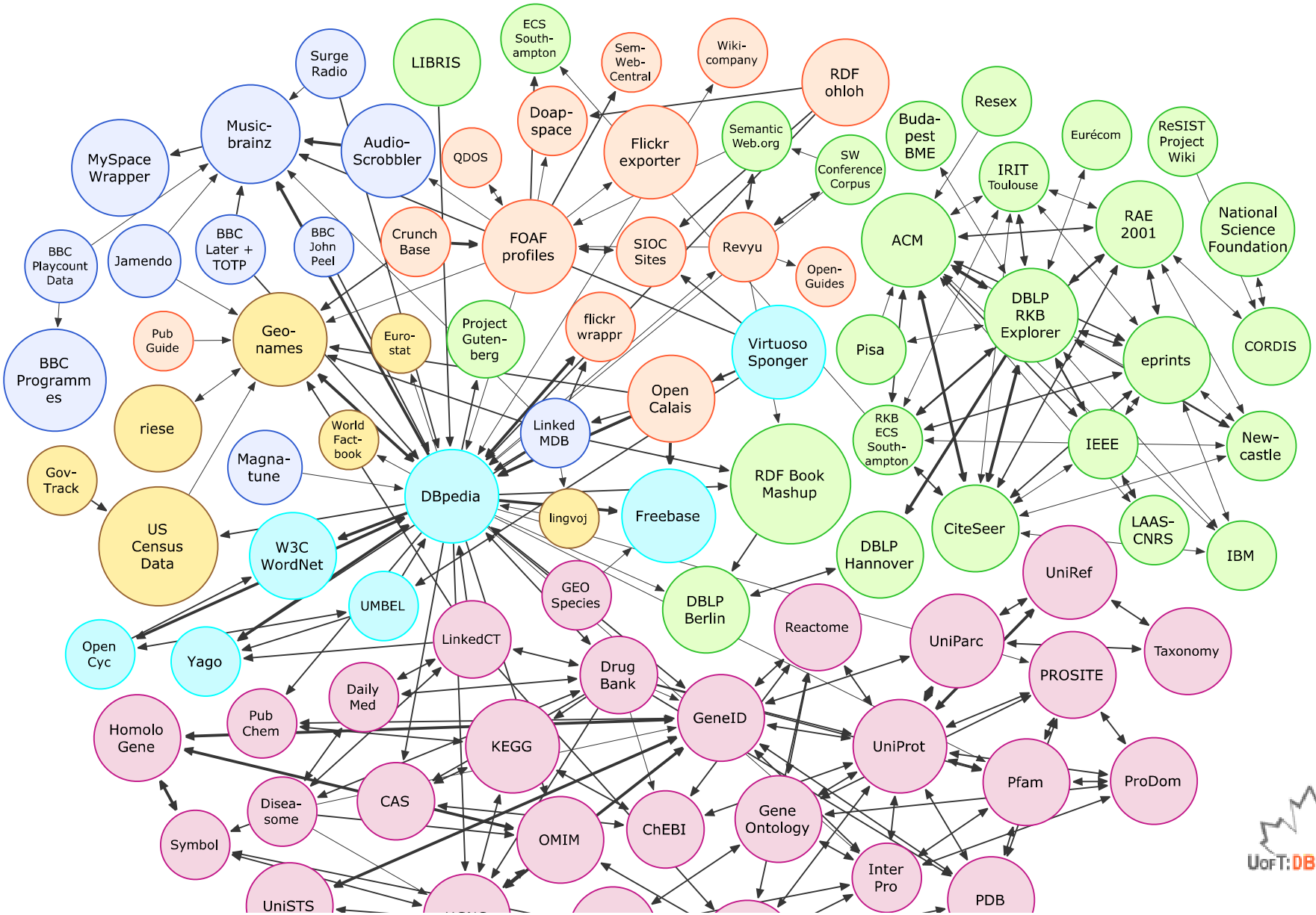
April 20th, 2009
Madrid, Spain

Presentation at the Linked Data On the Web (LDOW) 2009 Workshop

LinkedMDB

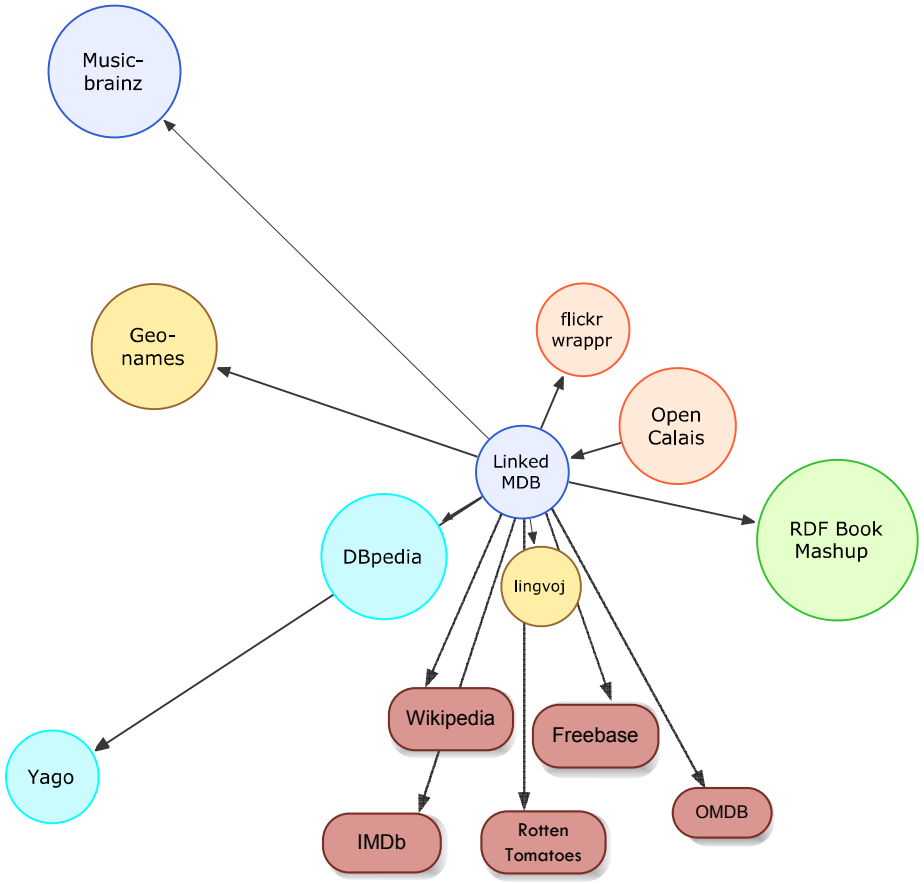
- The first linked data source dedicated to movies and movie-related information
- Currently published by D2R Server
- Contains ~4 million triples and ~0.5 million links to other sources and web pages
- Won the first prize at *Triplification Challenge*
  - I-Semantics 2008 Conference

LinkedMDB in LOD cloud



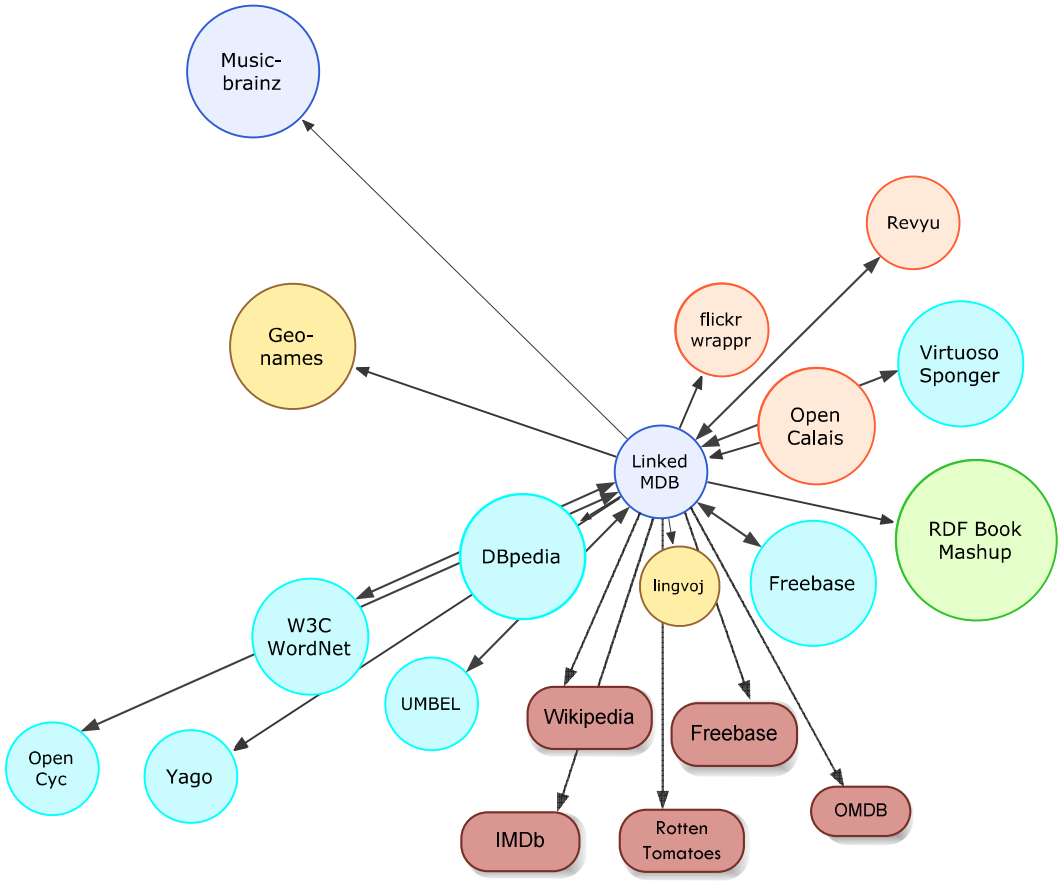
UofT: DB Group

LinkedMDB in LOD cloud



- Linked Data Sources
  - DBpedia/YAGO
  - Lingvoj
  - Musicbrainz
  - Geonames
  - FlickrWrappr
  - RDF Book Mashup
- Movie Web Pages
  - Freebase
  - Wikipedia
  - IMDb
  - Rotten Tomatoes
  - OMDB

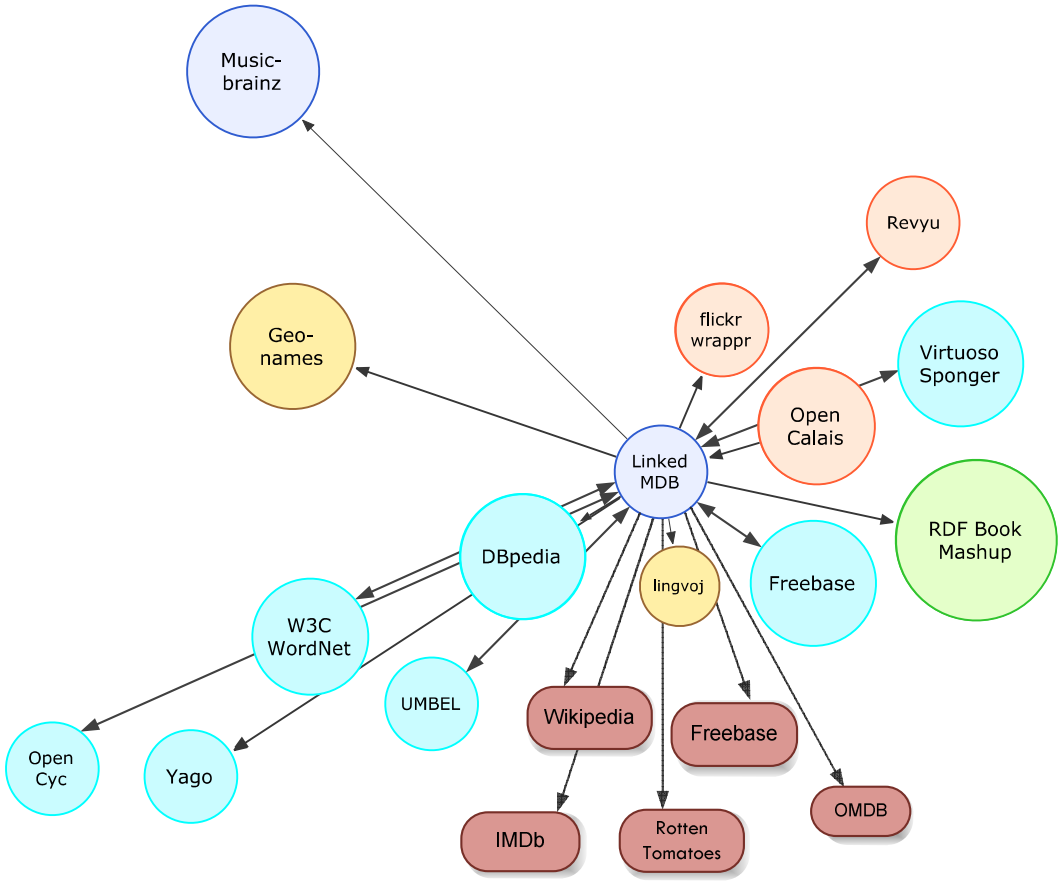
LinkedMDB in LOD cloud



- Linked Data Sources
  - DBpedia/YAGO
  - Lingvoj
  - Musicbrainz
  - Geonames
  - FlickrWrappr
  - RDF Book Mashup
- Other possible links
  - Revyu
  - Freebase
  - Virtuoso Sponger
  - UMBEL
  - OpenCyc
  - Wordnet

UofT: DB Group

LinkedMDB in LOD cloud



- Linked Data Sources
  - DBpedia/YAGO hard
  - Lingvoj easy
  - Musicbrainz hard
  - Geonames easy
  - FlickrWrappr easy
  - RDF Book Mashup hard
- Movie Web Pages
  - Freebase easy
  - Wikipedia hard
  - IMDb hard
  - Rotten Tomatoes hard
  - OMDB hard

Linkage Challenges – A Few Examples

- Different Names for the same movie
  - Alternative titles
    - "High School Musical 3" – "HSM3", "Batman Returns 2" and "The Dark Knight"
  - Different Styles
    - "A Thousand and One Nights" and "1001 Nights"
    - "The Shining" and "The Shining (film)"
  - Non-English titles
    - "Adu Puli Attam" – "Aadu_Puli_Aattam" or "Sacco and Vanzetti" "Sacco_e_Vanzetti"
- Same title for different movies
  - "Chicago" (1927 movie) - "Chicago" (2002 movie)
- Similar names for different movies
  - "Spiderman 1" and "Spiderman 2"
  - "Face to Face" and "Face to Fate"

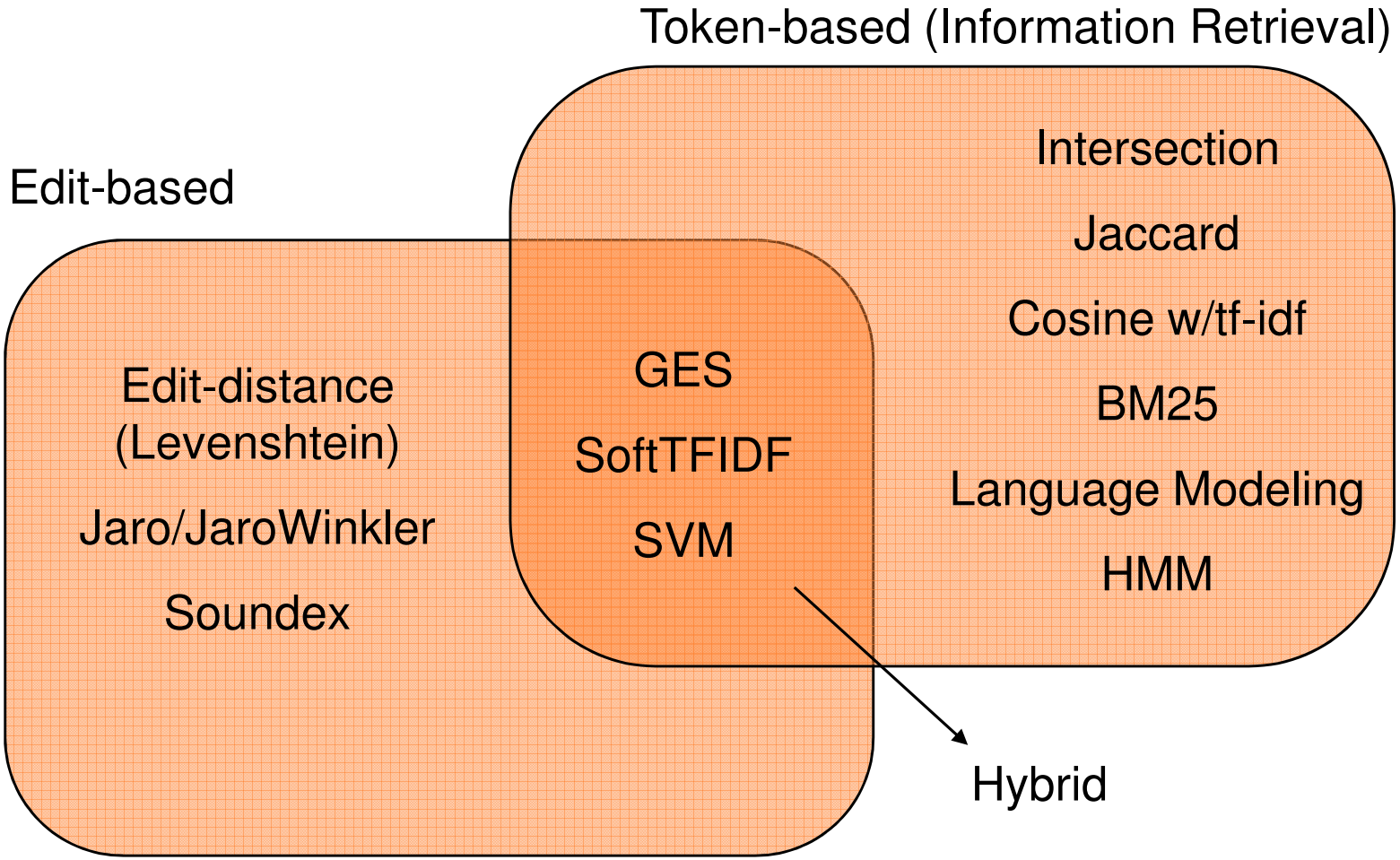
Linkage Methodology

- Effectiveness
  - Approximate String Matching/Joins
    - Automatic and domain independent
  - Semantic Matching
  - User-defined Transformation Rules
- Efficiency
  - Efficient indexing, hashing and blocking techniques
    - Avoid full similarity computation
    - Minimize preprocessing on the data

Leverage Existing Techniques

- A huge amount of work on approximate string joins [Tutorial-VLDB'05, Tutorial-SIGMOD'06]
- Many string similarity measures proposed
  - Survey for duplicate detection in [DDSurvey-TKDE'07]
  - A comparison for name-matching in [NameMatching-IJCAI'03] by Cohen et al.
  - Benchmarked for declarative approximate selection in [D.App.σ-SIGMOD'07]

String Similarity Measures

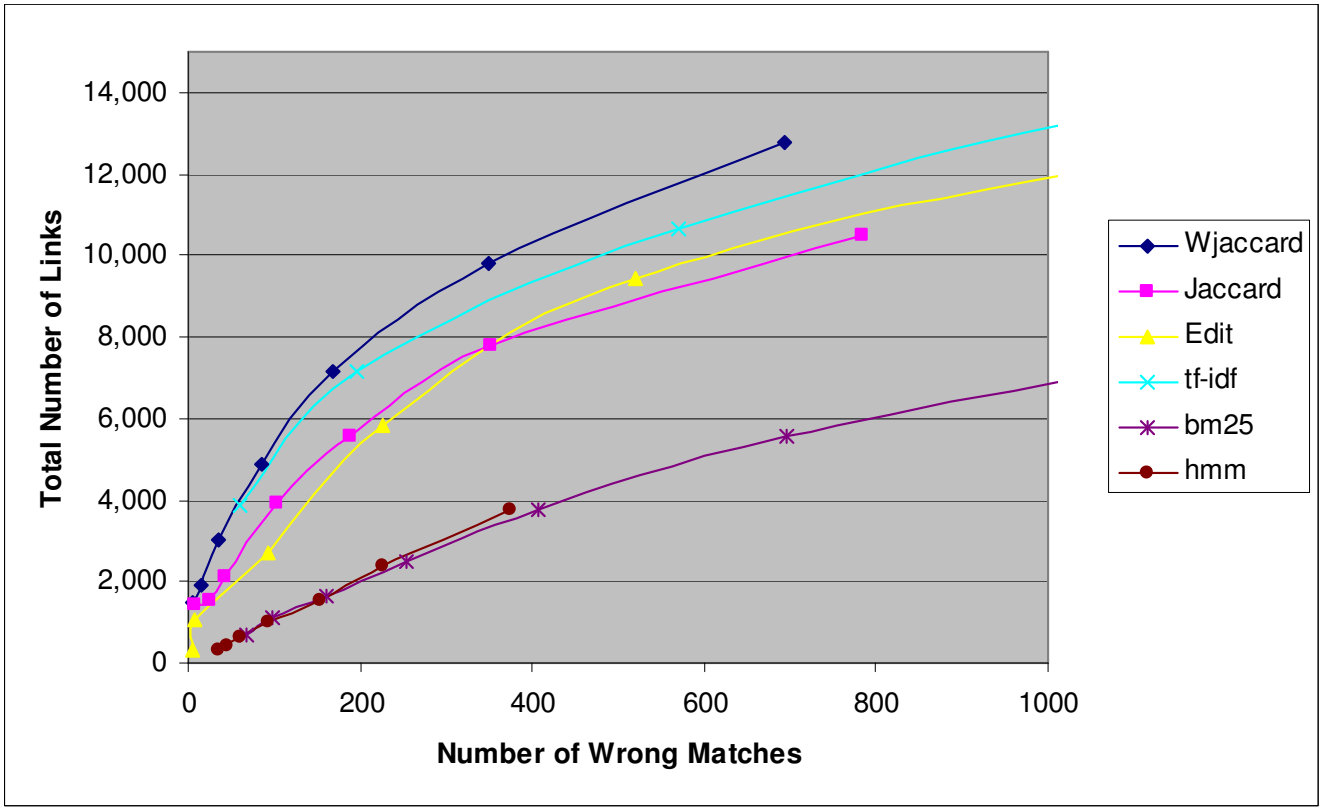


Implementation

- Algorithms based on q-grams
  - Efficient set similarity joins techniques can be used
  - Techniques from Information Retrieval can be used
    - To enhance accuracy
      - Similarity measures from IR
    - To enhance efficiency
      - By IR indexing techniques
    - Idea: treat *q-grams* in strings like *words* in documents
- Implementing the similarity measures using vanilla SQL
  - Why?
    - Ease of implementation
    - Keeping the data in the data source
      - A major concern for enterprises
    - Rely on the existing techniques for efficiency

Accuracy Evaluation

- Goal
  - Maximum number of links (proxy for maximizing recall)
  - Minimum number of incorrect links (maximizing precision)



The measure of choice in this scenario: Weighted Jaccard

Threshold Selection and Accuracy

(Example using links to movie titles in DBpedia)

Measure	Threshold	#Total	#Wrong	Accuracy	Measure	Threshold	#Total	#Wrong	Accuracy
Weighted Jaccard	0.6	12,756	693	94.57%	Cosine w/tf-idf	0.6	24,549	7,189	70.72%
	0.65	9,823	350	96.44%		0.65	21,068	4,546	78.42%
	0.7	7,130	169	97.63%		0.7	17,623	2,541	85.58%
	0.75	4,874	86	98.24%		0.75	14,082	1,243	91.17%
	0.8	3,018	36	98.81%		0.8	10,671	571	94.65%
	0.85	1,913	15	99.22%		0.85	7,169	197	97.25%
	0.9	1,505	6	99.60%		0.9	3,886	61	98.43%
Jaccard	0.6	10,476	785	92.51%	BM25	0.6	7,889	1,258	84.05%
	0.65	7,798	353	95.47%		0.65	5,565	695	87.51%
	0.7	5,545	189	96.59%		0.7	3,760	407	89.18%
	0.75	3,909	104	97.34%		0.75	2,485	254	89.78%
	0.8	2,117	43	97.97%		0.8	1,658	160	90.35%
	0.85	1,531	25	98.37%		0.85	1,092	98	91.03%
	0.9	1,432	7	99.51%		0.9	715	67	90.63%
Edit Similarity	0.6	16,137	3,260	79.80%	HMM	0.6	3,737	374	89.99%
	0.65	12,550	1,219	90.29%		0.65	2,396	226	90.57%
	0.7	9,423	519	94.49%		0.7	1,534	154	89.96%
	0.75	5,848	227	96.12%		0.75	992	92	90.73%
	0.8	2,719	93	96.58%		0.8	646	61	90.56%
	0.85	1,043	7	99.33%		0.85	447	45	89.93%
	0.9	334	4	98.80%		0.9	306	35	88.56%

Threshold Selection and Accuracy

(Example using links to movie titles in DBpedia)

Measure	Threshold	#Total	#Wrong	Accuracy	Measure	Threshold	#Total	#Wrong	Accuracy
Weighted Jaccard	0.6	12,756	693	94.57%	Cosine w/tf-idf	0.6	24,549	7,189	70.72%
	0.65	9,823	350	96.44%		0.65	21,068	4,546	78.42%
	0.7	7,130	169	97.63%		0.7	17,623	2,541	85.58%
	0.75	4,874	86	98.24%		0.75	14,082	1,243	91.17%
	0.8	3,018	36	98.81%		0.8	10,671	571	94.65%
	0.85	1,913	15	99.22%		0.85	7,169	197	97.25%
	0.9	1,505	6	99.60%		0.9	3,886	61	98.43%
Jaccard	0.6	10,476	785	92.51%	BM25	0.6	7,889	1,258	84.05%
	0.65	7,798	353	95.47%		0.65	5,565	695	87.51%
	0.7	5,545	189	96.59%		0.7	3,760	407	89.18%
	0.75	3,909	104	97.34%		0.75	2,485	254	89.78%
	0.8	2,117	43	97.97%		0.8	1,658	160	90.35%
	0.85	1,531	25	98.37%		0.85	1,092	98	91.03%
	0.9	1,432	7	99.51%		0.9	715	67	90.63%
Edit Similarity	0.6	16,137	3,260	79.80%	HMM	0.6	3,737	374	89.99%
	0.65	12,550	1,219	90.29%		0.65	2,396	226	90.57%
	0.7	9,423	519	94.49%		0.7	1,534	154	89.96%
	0.75	5,848	227	96.12%		0.75	992	92	90.73%
	0.8	2,719	93	96.58%		0.8	646	61	90.56%
	0.85	1,043	7	99.33%		0.85	447	45	89.93%
	0.9	334	4	98.80%		0.9	306	35	88.56%

Linkage Metadata

- Metadata about the source of the links
  - Where do they come from?
  - Can we trust them?
  - What technique is used for linkage?
  - If the linkage is based on record matching, what is the similarity/confidence score?
- Allows the user to
  - Get information about a link
  - Set user-specific requirements for the links

Linkage Metadata

- Two linkage entities
  - oddinker:interlink

Property	Value
rdfs:label	1036 (Interlink)
oddlinker:link_source	<http://data.linkedmdb.org/resource/film/2014>
oddlinker:link_target	<http://dbpedia.org/resource/The_Shining_%28film%29>
oddlinker:link_type	owl:sameAs
oddlinker:linkage_run	<http://data.linkedmdb.org/resource/linkage_run/1>
oddlinker:linkage_score	0.567848166224181
movie:linkid	1036 (xsd:int)
rdf:type	oddlinker:interlink

  - oddlinker:linkage_run

Property	Value
oddlinker:linkage_date	7-7-2008
oddlinker:linkage_method	WeightedJaccard
is oddlinker:linkage_run of	<http://data.linkedmdb.org/resource/interlink/1>
is oddlinker:linkage_run of	<http://data.linkedmdb.org/resource/interlink/1036>
is oddlinker:linkage_run of	<http://data.linkedmdb.org/resource/interlink/100>

UofT: DB Group

Conclusion

- Need for automatic linkage tools
  - Functionality
    - Ease of use
    - Domain Independence
  - Scalability
    - Web-scale link discovery
- Approximate string matching as the first step in linkage
  - More complex techniques are often required
    - Using structural information [DBTune-LDOW'08], [CollER-DEngBul'06]
- Need for linkage metadata
  - Where do the links come from?
- Need for linkage evaluation interfaces
  - Users can easily increase the quality

The End

Questions ?

http://linkedmdb.org

References

- [Tutorial-VLDB05] Approximate Joins: Concepts and Techniques.
  N. Koudas and D. Srivastava. VLDB'05 Tutorial
- [Tutorial-SIGMOD'06] Record linkage: similarity measures and algorithms.
  N. Koudas, S. Sarawagi and D. Srivastava. SIGMOD'06 Tutorial
- [DDSurvey-TKDE'07] Duplicate Record Detection: A Survey
  Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis and Vassilios S. Verykios. In IEEE Transactions on Knowledge and Data Engineering
- [NameMatching-IJCAI'03] A Comparison of String Distance Metrics for Name-Matching Tasks
  William W. Cohen, Pradeep Ravikumar, Stephen E. Fienberg. In IJCAI-03 Workshop on Information Integration on the Web
- [D.App.σ-SIGMOD'07] Benchmarking declarative approximate selection predicates.
  A. Chandel, O. Hassanzadeh, N.Koudas, M. Sadoghi, and D. Srivastava. In SIGMOD'07.

References

- [DBTune-LDOW'08] Automatic interlinking of music datasets on the semantic web.
  Y. Raimond, C. Sutton, and M. Sandler. In LDOW'08.
- [CollER-DEngBul'06] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data.
  IEEE Data Eng. Bull, 29(2):4–12, 2006.
- [LSH, STOC'97] Locality-preserving hashing in multidimensional spaces. Indyk, Motwani, Raghavan, and Vempala. In STOC'97.
- [MP-LSH-VLDB'07] Multi-Probe LSH: Efficient Indexing for HighDimensional Similarity Search
  Qin Lv, William Josephson, Zhe Wang, Moses Charikar and Kai Li In VLDB'07
- [ExactSSJoin-VLDB'06] Efficient Exact Set Similarity Joins
  A. Arasu, V. Ganti, R. Kausshik, In VLDB'06