Approximate Matching of Textual Domain Attributes for Information Source Integration

Andreas Koeller, Vinay Keelara

Montclair State University

Montclair, New Jersey

Motivation: (Heterogeneous) Information Source Integration

- Integrating information sources
- Heterogeneity in databases can be attributable to differences in
  - Data model (here: relational)
  - Schema
  - Data Domain (here: text)
  - Data Representation

Outline

- Background
- Similarity of Strings
- Similarity of Attributes
- Similarity of Relations
- Evaluation

Approaches to Database Integration

- Schema Integration: Identification of related attributes across two databases
  - Schema-based Matching: Identification of relationships based only on the schema of the data sources
  - Instance-based Matching: Identification of relationships based on the properties of the data contained in the database fields
- Data Integration: Translation of data in matching attributes into a common format

Deficiencies of Current Algorithms

- Schema-based Matching:
  - Dependence on the availability of either global or common domains
  - Dependence on the availability of accurate descriptions of the structure of the databases
- Instance-based Matching:
  - Consideration of only exact subsets in determining relationships between databases (e.g., FD and IND discovery)

Quality Issues in Source Integration

- Many instance-based approaches use concept of *dependencies* (functional dependencies, inclusion dependencies) → assumes exact inclusion of sets
- Real-world data sources are unlikely to be exactly included in one another, even if related: missing and extraneous tuples, etc.
- Also: Individual data *values* do not always match due to different descriptions of same items → string comparison at value level often not sufficient

Examination of Real World Data Sources: Variation in representation of real world entities

COMPANY 1
MICROSOFT
AMERICAN MARKETING
ATT
INTEL
ORACLE
SBC
IBM
BOEING

COMPANY 2
ATT TELECOM
MICROSOFT INC
INTEL RESEARCH LABS
ORACLE RESEARCH LABS
FOXPRO INTERNATIONAL INC
AMERICAN MARKETING RESEARCH
SBC TELECOM
TEXAS INSTRUMENTS LABS
BOEING CORP.
IBM INC
BOEING INC LABS
AMERICAN MARKETING RESEARCH
SBC INC

Outline

- Background
- Similarity of Strings
- Similarity of Attributes
- Similarity of Relations
- Evaluation

Relationship between data values

- Two values are said to be *co-referent* if they refer to the same real world entity.
- **Definition: *Similarity*** of two data values is a measure of confidence in the co-reference between them. We denote similarity of two value x and y by $\text{sim}(x,y)$ with $0 = \text{sim}(x,y) = 1$.
- A similarity of 1 means certain co-reference, whereas a similarity of 0 means unrelated semantics of x and y .

How to *measure* similarity?

- Treat values (“documents”) in textual domain attributes as sets of individual name constants
- Represent documents as vectors of real numbers, with each unique word in the database spanning one dimension of the vector space (Vector Space Model)
- The values of the coefficients of the vector are determined using the TF/IDF weighing scheme, then vector is normalized
- Similarity between documents is calculated by the dot product of the corresponding document vectors → angle between vectors

TF-IDF weighing scheme (one version)

- The magnitude of a vector coefficient is related to the importance of the term corresponding to the coefficient in the document considered

$$v_t = \begin{cases} (\log(TF_{v,t}) + 1) \cdot \log(IDF_T) & \text{if } TF_{v,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$TF_{v,t}$ is the term frequency

IDF_t is the inverse document frequency of t in document collection C, i.e.,

$$IDF_t = \frac{|C|}{|C_t|}$$

Similarity

- The dot product of the normalized TF-IDF vectors (the cosine of the angle between the vectors) used to calculate the similarity between tuples is given by

$$sim(\hat{x}, \hat{y}) = \sum_{t \in T} \hat{x}_t \cdot \hat{y}_t$$

With $\hat{x} = \frac{x}{|x|}$ and $\hat{y} = \frac{y}{|y|}$

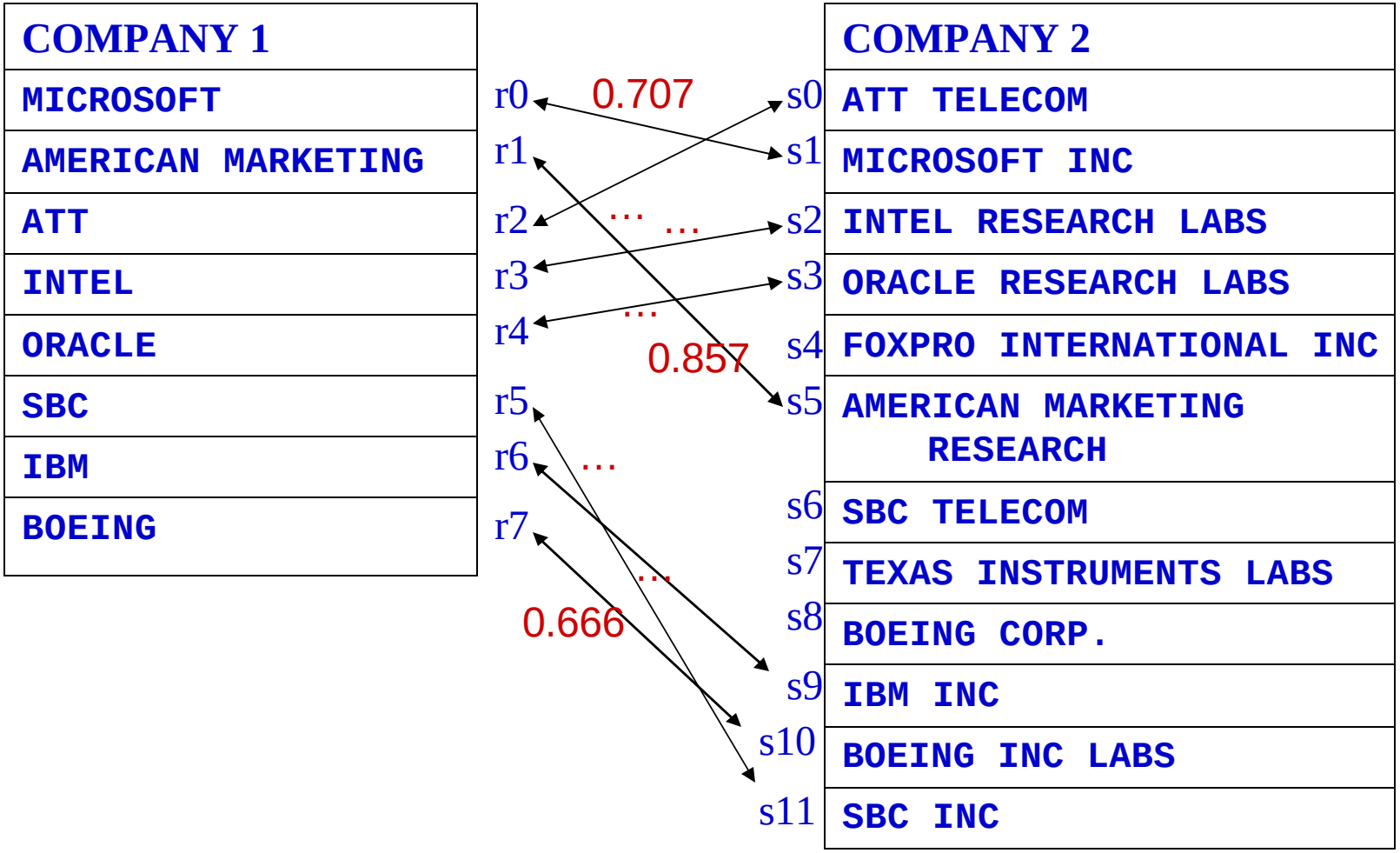
Outline

- Background
- Similarity of Strings
- Similarity of Attributes
- Similarity of Relations
- Evaluation

Similarity

- Two documents are similar if they share many important terms (i.e., terms with high TF-IDF values)
- Documents with no common terms are unrelated and are represented by orthogonal TF-IDF vectors
- How to extend value (document) similarity to sets of documents (i.e., attributes in a relation)?
- ➔ *Compute similarity of each value in first attribute to* ***best matching*** *value in second attribute.*

Matching values in attributes

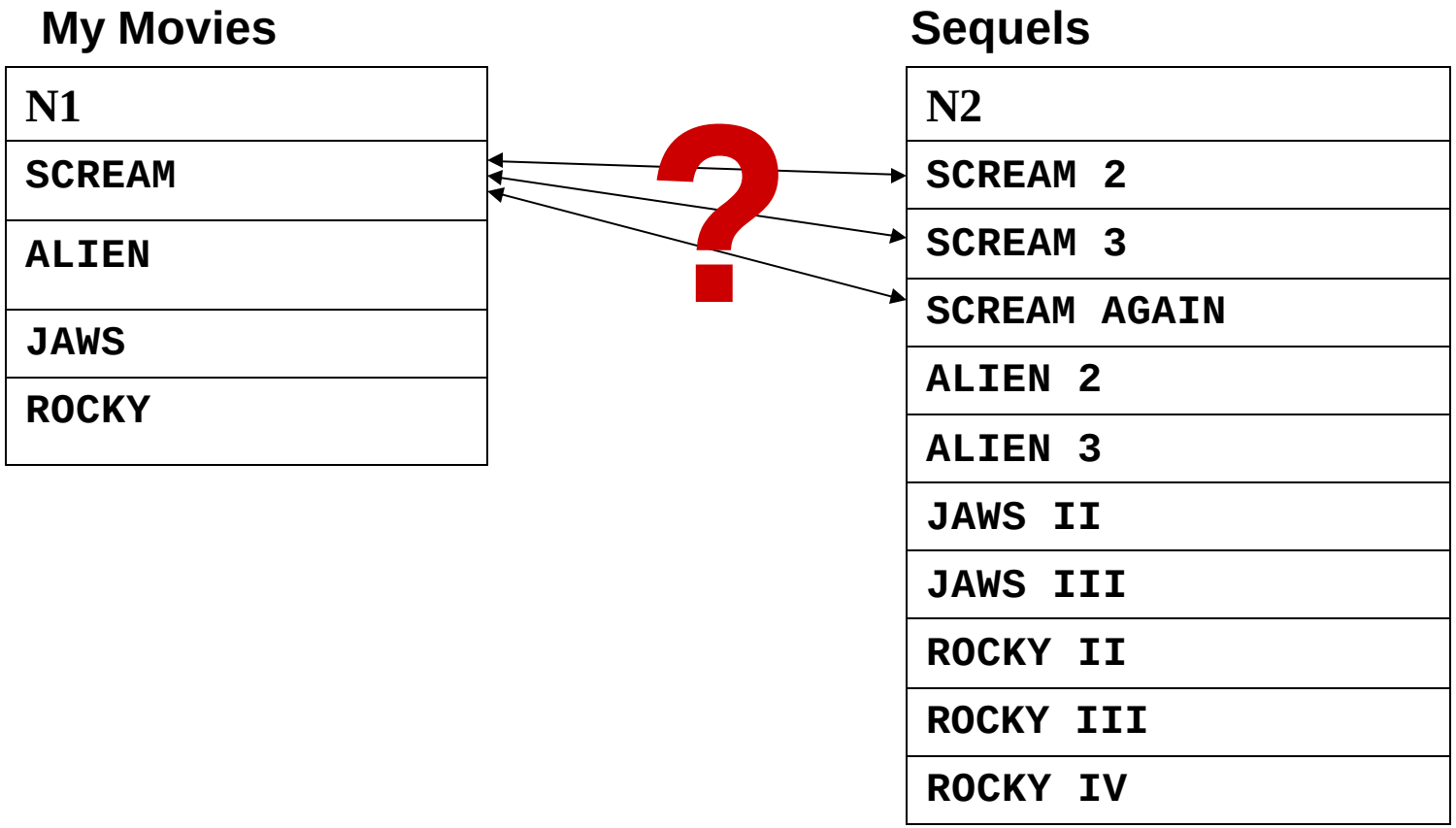


Attribute Similarity

- **Definition:** For each $x \in X$, the *attribute similarity vector* $V_{X,Y}$ contains the maximum similarity score with any tuple $y \in Y$.
- **Definition:** The *attribute similarity score* sim(X,Y) is the average of all components of the attribute similarity vector $V_{X,Y}$

sim(COMPANY1, COMPANY2)=.765

Attribute similarity in the presence of duplicates (non-key attributes)



Similarity of Non-key attributes

Definition:

Assume an attribute X, an attribute Y, and a tuple $x \in X$.
The *approximate multiplicity* of x in Y, denoted by $\tilde{m}(Y, x)$, is then defined as the number of tuples $y \in Y$ for which $\text{sim}(x, y) \geq c$, with c some constant between 0 and 1.

Useful values for c *(empirical):*

c=0.95 if X and Y are from different relations

c=0.995 if X and Y are from the same relation

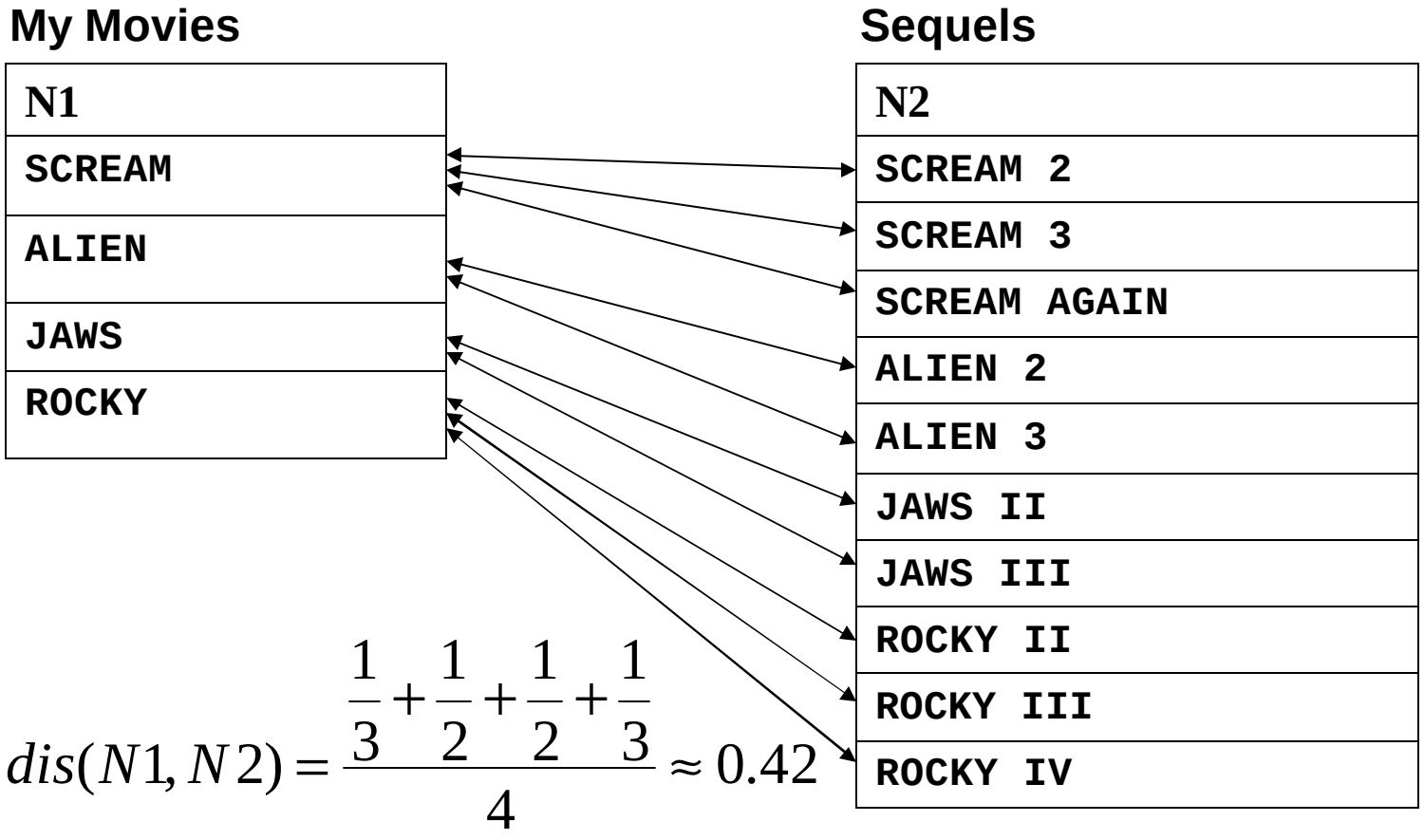
Similarity of Non-key attributes

Definition: Assume that |X|<|Y|. For each x ∈ X , the *attribute distribution vector* $D_{X,Y}$ contains the value distribution score d_x as follows:

$$d_x = \begin{cases} \dfrac{\tilde{m}(X,x)}{\tilde{m}(Y,x)} & \text{if } \tilde{m}(X,x) \le \tilde{m}(Y,x) \\ 0 & \text{otherwise} \end{cases}$$

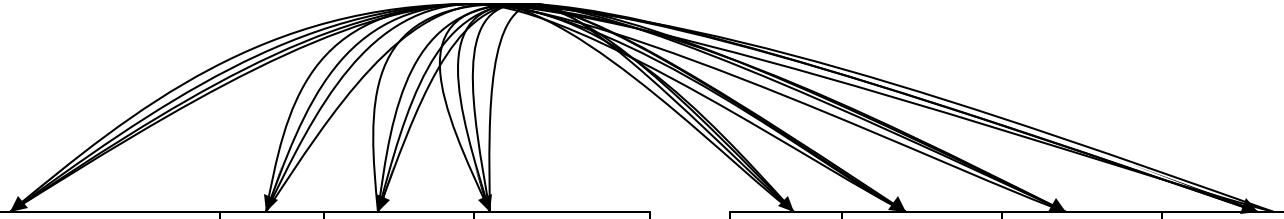
Definition: The *attribute distribution score* dis(*X*,*Y*) is the average of all components of the attribute distribution vector $D_{X,Y}$

Distribution Score



$$dis(N1, N2) = \frac{\frac{1}{3} + \frac{1}{2} + \frac{1}{2} + \frac{1}{3}}{4} \approx 0.42$$

Discovery of Relationships between attributes

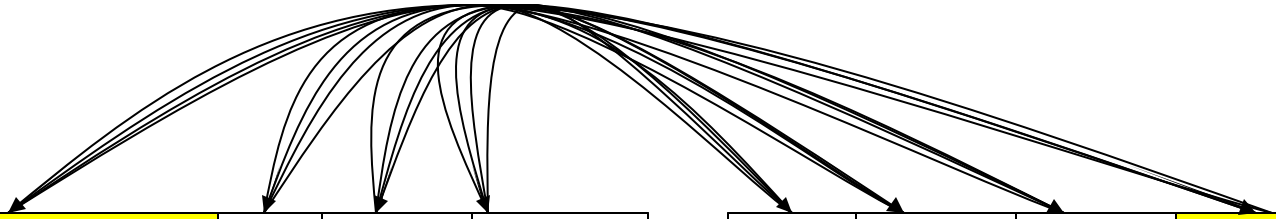


Name	Year	Director	Studio
Pygmalion	1938	Asquith	GB
Pygmalion	1934	Engel	PRC
The Ten Commandments	1956	DeMille	Paramount
The Ten Commandments	1923	DeMille	Paramount
Macbeth	1948	Schafer	Republic
Macbeth	1956	Polanski	Toho
Casablanca	1942	Mayo	Fox
Casablanca	1946	Curtiz	Warner
Casablanca	1959	Wald	Warner

Jahr	Studio	Regisseur	Titel
1937	Cinetone	Berger	Pygmalion
1938	GB	Asquith	Pygmalion
1934	PRC	Engel	Pygmalion
1956	Paramount	DeMille	The Ten Commandments
1923	Paramount	DeMille	The Ten Commandments
1948	Republic	Schafer	Macbeth
1956	Toho	Polanski	Macbeth
1942	Fox	Mayo	Casablanca
1946	Warner	Curtiz	Casablanca

Pairwise comparison, using distribution scores

Discovery of Relationships between attributes

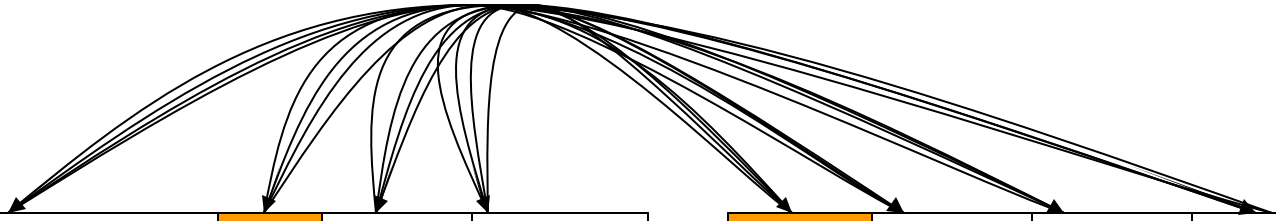


Name	Year	Director	Studio
Pygmalion	1938	Asquith	GB
Pygmalion	1934	Engel	PRC
The Ten Commandments	1956	DeMille	Paramount
The Ten Commandments	1923	DeMille	Paramount
Macbeth	1948	Schafer	Republic
Macbeth	1956	Polanski	Toho
Casablanca	1942	Mayo	Fox
Casablanca	1946	Curtiz	Warner
Casablanca	1959	Wald	Warner

Jahr	Studio	Regisseur	Titel
1937	Cinetone	Berger	Pygmalion
1938	GB	Asquith	Pygmalion
1934	PRC	Engel	Pygmalion
1956	Paramount	DeMille	The Ten Commandments
1923	Paramount	DeMille	The Ten Commandments
1948	Republic	Schafer	Macbeth
1956	Toho	Polanski	Macbeth
1942	Fox	Mayo	Casablanca
1946	Warner	Curtiz	Casablanca

The highest distribution score generated is .666

Discovery of Relationships between attributes

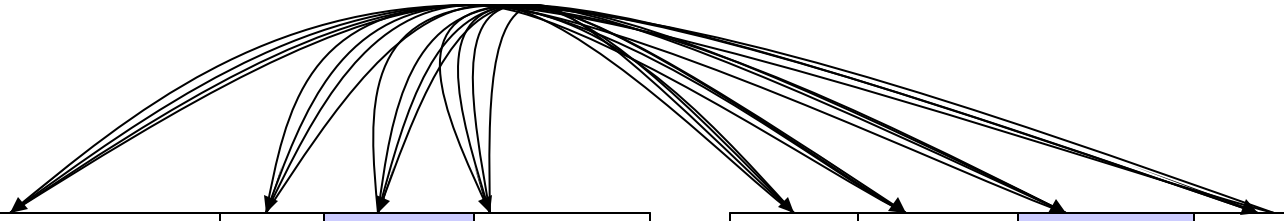


Name	Year	Director	Studio
Pygmalion	1938	Asquith	GB
Pygmalion	1934	Engel	PRC
The Ten Commandments	1956	DeMille	Paramount
The Ten Commandments	1923	DeMille	Paramount
Macbeth	1948	Schafer	Republic
Macbeth	1956	Polanski	Toho
Casablanca	1942	Mayo	Fox
Casablanca	1946	Curtiz	Warner
Casablanca	1959	Wald	Warner

Jahr	Studio	Regisseur	Titel
1937	Cinetone	Berger	Pygmalion
1938	GB	Asquith	Pygmalion
1934	PRC	Engel	Pygmalion
1956	Paramount	DeMille	The Ten Commandments
1923	Paramount	DeMille	The Ten Commandments
1948	Republic	Schafer	Macbeth
1956	Toho	Polanski	Macbeth
1942	Fox	Mayo	Casablanca
1946	Warner	Curtiz	Casablanca

The highest distribution score generated is .875

Discovery of Relationships between attributes

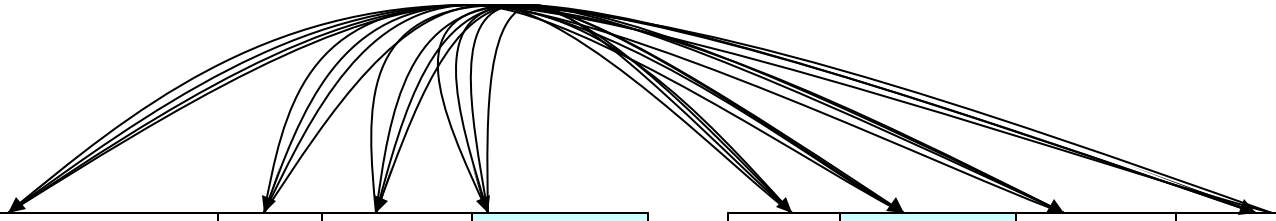


Name	Year	Director	Studio
Pygmalion	1938	Asquith	GB
Pygmalion	1934	Engel	PRC
The Ten Commandments	1956	DeMille	Paramount
The Ten Commandments	1923	DeMille	Paramount
Macbeth	1948	Schafer	Republic
Macbeth	1956	Polanski	Toho
Casablanca	1942	Mayo	Fox
Casablanca	1946	Curtiz	Warner
Casablanca	1959	Wald	Warner

Jahr	Studio	Regisseur	Titel
1937	Cinetone	Berger	Pygmalion
1938	GB	Asquith	Pygmalion
1934	PRC	Engel	Pygmalion
1956	Paramount	DeMille	The Ten Commandments
1923	Paramount	DeMille	The Ten Commandments
1948	Republic	Schafer	Macbeth
1956	Toho	Polanski	Macbeth
1942	Fox	Mayo	Casablanca
1946	Warner	Curtiz	Casablanca

The highest distribution score generated is .875

Discovery of Relationships between attributes

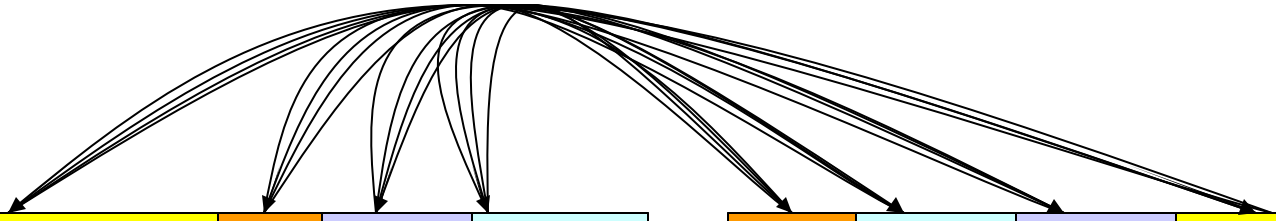


Name	Year	Director	Studio
Pygmalion	1938	Asquith	GB
Pygmalion	1934	Engel	PRC
The Ten Commandments	1956	DeMille	Paramount
The Ten Commandments	1923	DeMille	Paramount
Macbeth	1948	Schafer	Republic
Macbeth	1956	Polanski	Toho
Casablanca	1942	Mayo	Fox
Casablanca	1946	Curtiz	Warner
Casablanca	1959	Wald	Warner

Jahr	Studio	Regisseur	Titel
1937	Cinetone	Berger	Pygmalion
1938	GB	Asquith	Pygmalion
1934	PRC	Engel	Pygmalion
1956	Paramount	DeMille	The Ten Commandments
1923	Paramount	DeMille	The Ten Commandments
1948	Republic	Schafer	Macbeth
1956	Toho	Polanski	Macbeth
1942	Fox	Mayo	Casablanca
1946	Warner	Curtiz	Casablanca

The highest distribution score generated is .815

Discovery of Relationships between attributes



Name	Year	Director	Studio
Pygmalion	1938	Asquith	GB
Pygmalion	1934	Engel	PRC
The Ten Commandments	1956	DeMille	Paramount
The Ten Commandments	1923	DeMille	Paramount
Macbeth	1948	Schafer	Republic
Macbeth	1956	Polanski	Toho
Casablanca	1942	Mayo	Fox
Casablanca	1946	Curtiz	Warner
Casablanca	1959	Wald	Warner

Jahr	Studio	Regisseur	Titel
1937	Cinetone	Berger	Pygmalion
1938	GB	Asquith	Pygmalion
1934	PRC	Engel	Pygmalion
1956	Paramount	DeMille	The Ten Commandments
1923	Paramount	DeMille	The Ten Commandments
1948	Republic	Schafer	Macbeth
1956	Toho	Polanski	Macbeth
1942	Fox	Mayo	Casablanca
1946	Warner	Curtiz	Casablanca

Outline

- Background
- Similarity of Strings
- Similarity of Attributes
- Similarity of Relations
- Evaluation

Related attribute sets

- If *R[A1]* and *S[B1]* are related and *R[A2]* and *S[B2]* are related, does that mean that *R[A1,A2]* is also related to *S[B1,B2]*?
- Not really, since different *value pairs* might have contributed to distribution score
- ➔ the approximate version of the *IND inference problem (Inclusion Dependencies)*
- The exact problem is NP-hard

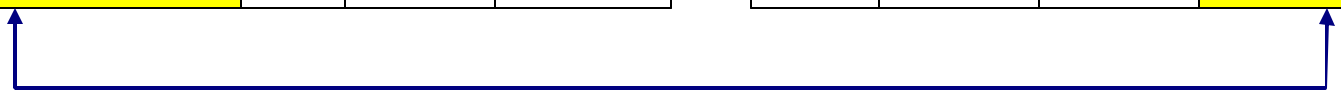
Discovery of Similarities between Relations

- Compute Distribution Scores for all pairs of attributes
- Merge best matching attribute pairs
- Compute *combined* distribution scores, until score falls below threshold
- Scores for each set will indicate how closely that particular subset of the two relations are related.

Discovery of Similarities between Databases

Name	Year	Director	Studio
Pygmalion	1938	Asquith	GB
Pygmalion	1934	Engel	PRC
The Ten Commandments	1956	DeMille	Paramount
The Ten Commandments	1923	DeMille	Paramount
Macbeth	1948	Schafer	Republic
Macbeth	1956	Polanski	Toho
Casablanca	1942	Mayo	Fox
Casablanca	1946	Curtiz	Warner
Casablanca	1959	Wald	Warner

Jahr	Studio	Regisseur	Titel
1937	Cinetone	Berger	Pygmalion
1938	GB	Asquith	Pygmalion
1934	PRC	Engel	Pygmalion
1956	Paramount	DeMille	The Ten Commandments
1923	Paramount	DeMille	The Ten Commandments
1948	Republic	Schafer	Macbeth
1956	Toho	Polanski	Macbeth
1942	Fox	Mayo	Casablanca
1946	Warner	Curtiz	Casablanca



Distribution score: 0.666

Discovery of Similarities between Databases

The exact problem (IND discovery) has exponential complexity in # of attributes.

Name	Year	Director	Studio
Pygmalion	1938	Asquith	GB
Pygmalion	1934	Engel	PRC
The Ten Commandments	1956	DeMille	Paramount
The Ten Commandments	1923	DeMille	Paramount
Macbeth	1948	Schafer	Republic
Macbeth	1956	Polanski	Toho
Casablanca	1942	Mayo	Fox
Casablanca	1946	Curtiz	Warner
Casablanca	1959	Wald	Warner

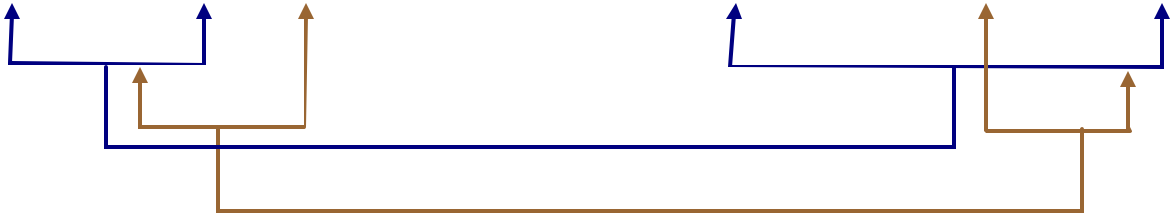
an	cinema	nom	film
1937	Cinetone	Berger	Pygmalion
1938	GB	Asquith	Pygmalion
1934	PRC	Engel	Pygmalion
1956	Paramount	DeMille	The Ten Commandments
1923	Paramount	DeMille	The Ten Commandments
1948	Republic	Schafer	Macbeth
1956	Toho	Polanski	Macbeth
1942	Fox	Mayo	Casablanca
1946	Warner	Curtiz	Casablanca

Distribution score: 0.777

Discovery of Similarities between Databases

Name	Year	Director	Studio
Pygmalion	1938	Asquith	GB
Pygmalion	1934	Engel	PRC
The Ten Commandments	1956	DeMille	Paramount
The Ten Commandments	1923	DeMille	Paramount
Macbeth	1948	Schafer	Republic
Macbeth	1956	Polanski	Toho
Casablanca	1942	Mayo	Fox
Casablanca	1946	Curtiz	Warner
Casablanca	1959	Wald	Warner

an	cinema	nom	film
1937	Cinetone	Berger	Pygmalion
1938	GB	Asquith	Pygmalion
1934	PRC	Engel	Pygmalion
1956	Paramount	DeMille	The Ten Commandments
1923	Paramount	DeMille	The Ten Commandments
1948	Republic	Schafer	Macbeth
1956	Toho	Polanski	Macbeth
1942	Fox	Mayo	Casablanca
1946	Warner	Curtiz	Casablanca

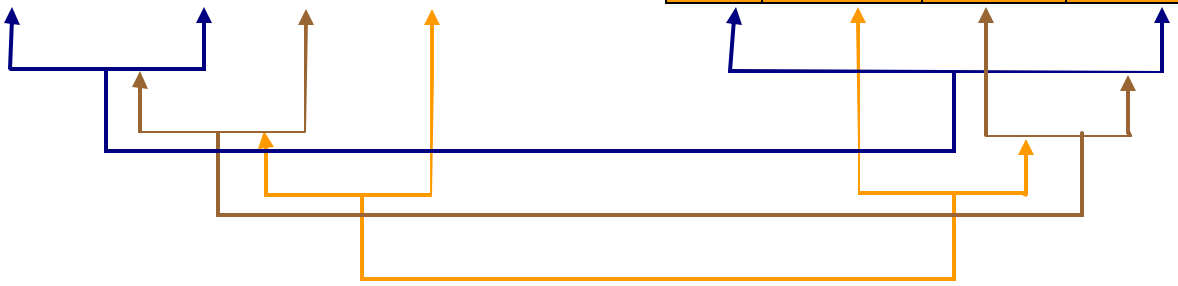


Distribution score: 0.888

Discovery of Similarities between Databases

Name	Year	Director	Studio
Pygmalion	1938	Asquith	GB
Pygmalion	1934	Engel	PRC
The Ten Commandments	1956	DeMille	Paramount
The Ten Commandments	1923	DeMille	Paramount
Macbeth	1948	Schafer	Republic
Macbeth	1956	Polanski	Toho
Casablanca	1942	Mayo	Fox
Casablanca	1946	Curtiz	Warner
Casablanca	1959	Wald	Warner

an	cinema	nom	film
1937	Cinetone	Berger	Pygmalion
1938	GB	Asquith	Pygmalion
1934	PRC	Engel	Pygmalion
1956	Paramount	DeMille	The Ten Commandments
1923	Paramount	DeMille	The Ten Commandments
1948	Republic	Schafer	Macbeth
1956	Toho	Polanski	Macbeth
1942	Fox	Mayo	Casablanca
1946	Warner	Curtiz	Casablanca



Distribution score: 0.888

Outline

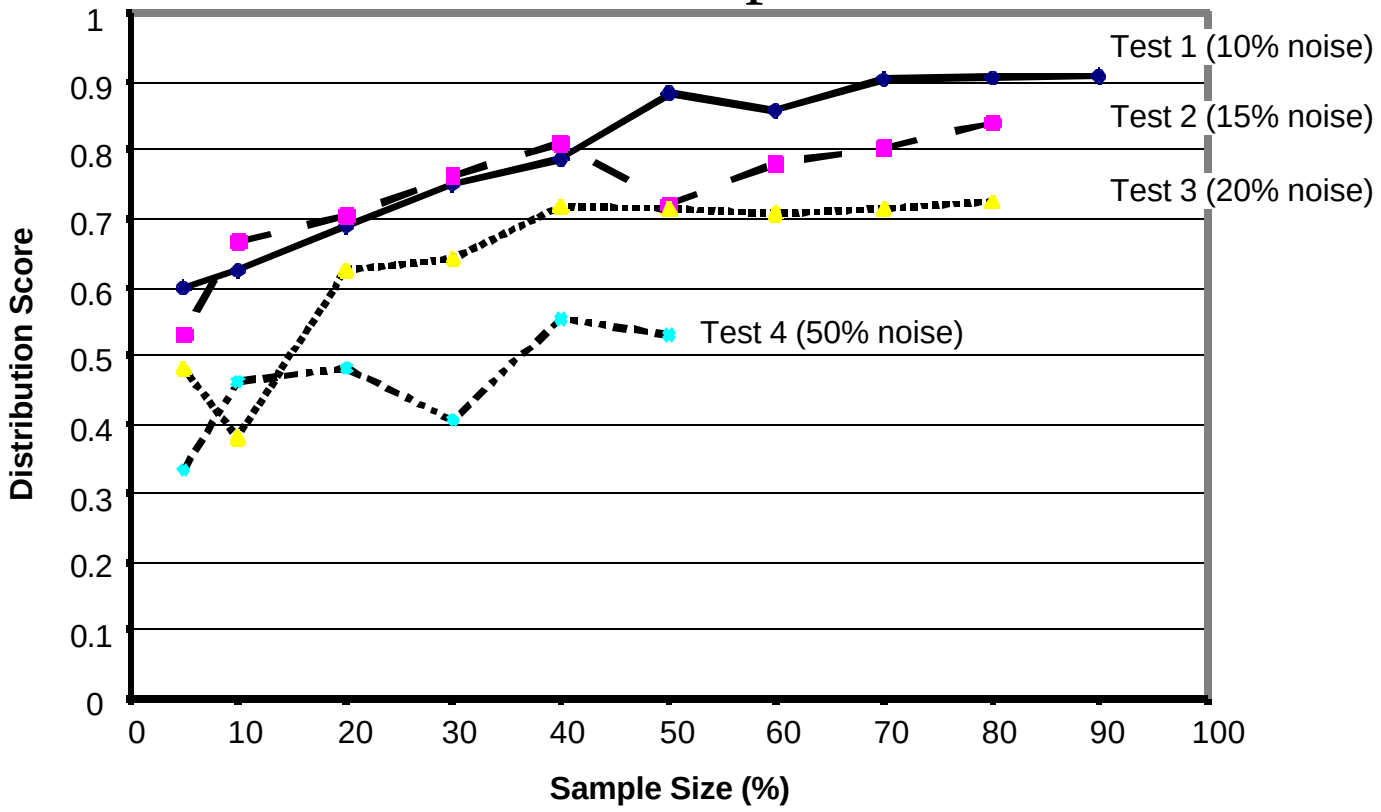
- Background
- Similarity of Strings
- Similarity of Attributes
- Similarity of Relations
- Evaluation

Experimental Evaluation

- Suitability of Similarity Score and Distribution Score for determining “relatedness” of relations
- Behavior of distribution score under various scenarios of distinct and non-distinct values in attribute
- Performance as function of # of tuples and # of attributes in source relations

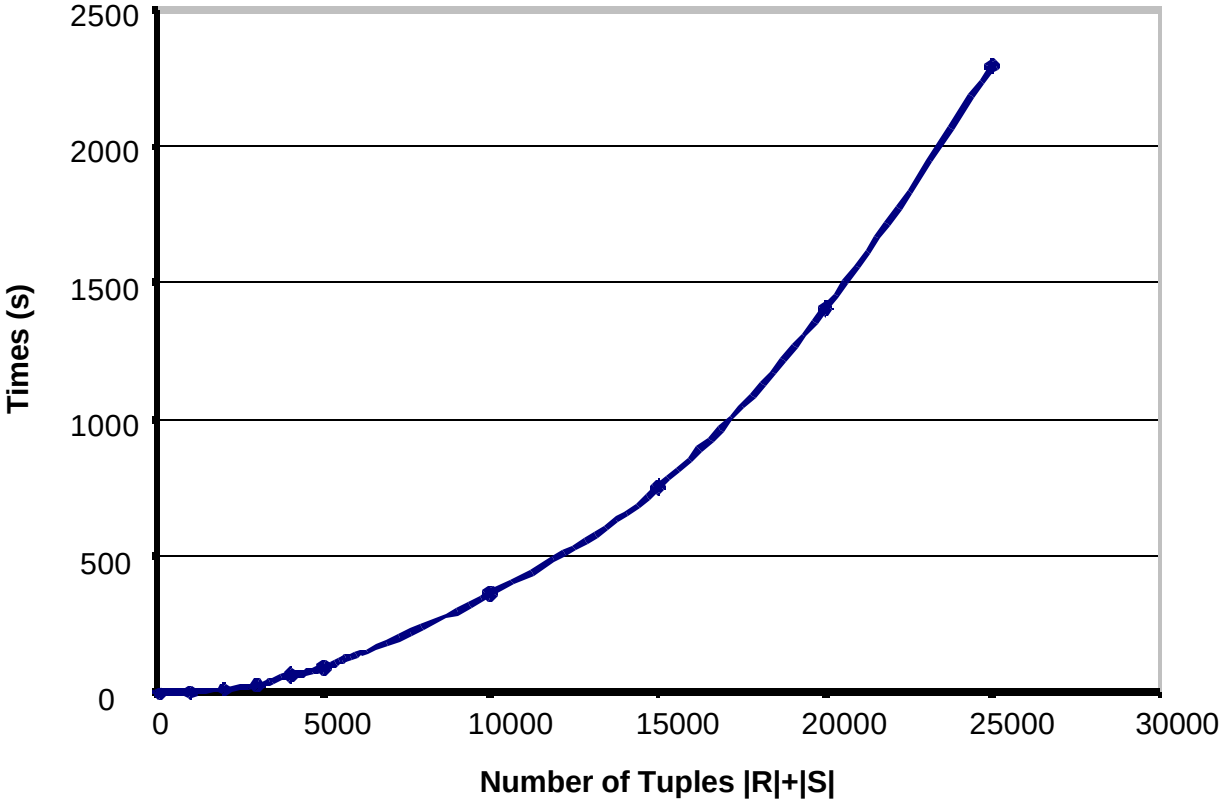
Experiment

- Presence of Extraneous Tuples in the Relations



Experiment

- Run Time v/s Number of Tuples in the Relations



Conclusion

- Addresses the problem of variations in representation of real world entities
- Detect relationships among relations with partial overlaps
- New algorithm provides a score that reflects degree of confidence that two relations can be integrated
- Retains real world relevance, because of low polynomial complexity.

- Additional Slides

Related Work

- Inclusion dependency discovery: de Marchi & Petit (2003) and Koeller & Rundensteiner (2003).
- Both papers propose algorithms for the detection of inclusion dependency patterns across databases. However, neither paper deals with noise or non-matching values.
- Other instance-based techniques with exact value matching, e.g., Kang/Naughton (*SIGMOD 2003*)

Related Work

- Variants of the SQL operators for similarity predicates, e.g., Schallehn, Sattler & Saake (*DKE 2004*)
- AI solutions (neural networks, machine learning), e.g., Savnik/Flach (*AI Communications 1999*), Bilenko/Mooney (*IJCAI-03*)
- TF-IDF for similarity joins: “WHIRL”, Cohen (*SIGMOD Record 1998*)
- FD discovery with degrees of satisfaction, Wei/Chen (*Intelligent Systems 2004*)