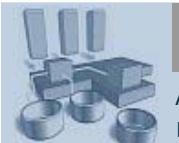


Training Selection for Tuning Entity Matching

Hanna Köpcke

Oberseminar Zingst 29.06.-03.07.2008



UNIVERSITÄT LEIPZIG

Abteilung Datenbanken
Institut für Informatik



Outline

- Entity matching
- Entity Matching Framework
- Training data
- Selection strategies
- Evaluation
- Summary and outlook



Entity matching

[The merge/purge problem for large databases - all 5 versions »](#)

①

MA Hernández, SJ Stolfo - [Proceedings of the 1995 ACM SIGMOD international conference](#) ..., 1995 - |

Page 1. **The Merge/Purge Problem for Large Databases** * Abstract Mauricio A. Hernández

Salvatore J. Stolfo {rauricio, sal}@cs.columbia.edu ...

[Cited by 342](#) - [Related Articles](#) - [Web Search](#)

[② \[CITATION\] The Merge/Purge Problem for Large Databases](#)

AH Mauricio JS Stolfo - [Proceedings of the 1995 ACM SIGMOD Conference on Management](#) ..., 1995

[Cited by 4](#) - [Related Articles](#) - [Web Search](#)

[③ \[CITATION\] The merge/purge problem for large databases](#)

①

MA Hernández SJ Stolfo - [Proceedings of the ACM SIGMOD International Conference on](#) ..., 1995

[Cited by 4](#) - [Related Articles](#) - [Web Search](#)

[\[CITATION\] The merge/purge problem for large datasets](#)

MA Hernandez, SJ Stolfo - [Proc. Of the SIGMOD](#), 1995

[Cited by 2](#) - [Related Articles](#) - [Web Search](#)

①

[④ \[CITATION\] andez, SJ Stolfo The merge/purge problem for large databases](#)

M Hern - [Proceedings of the 1995 ACM SIGMOD International Conference](#) ..., 1995

[Cited by 3](#) - [Related Articles](#) - [Web Search](#)

②

- ① Heterogeneous venue names
- ② Extraction errors
- ③ Typos (author name)
- ④ Missing authors

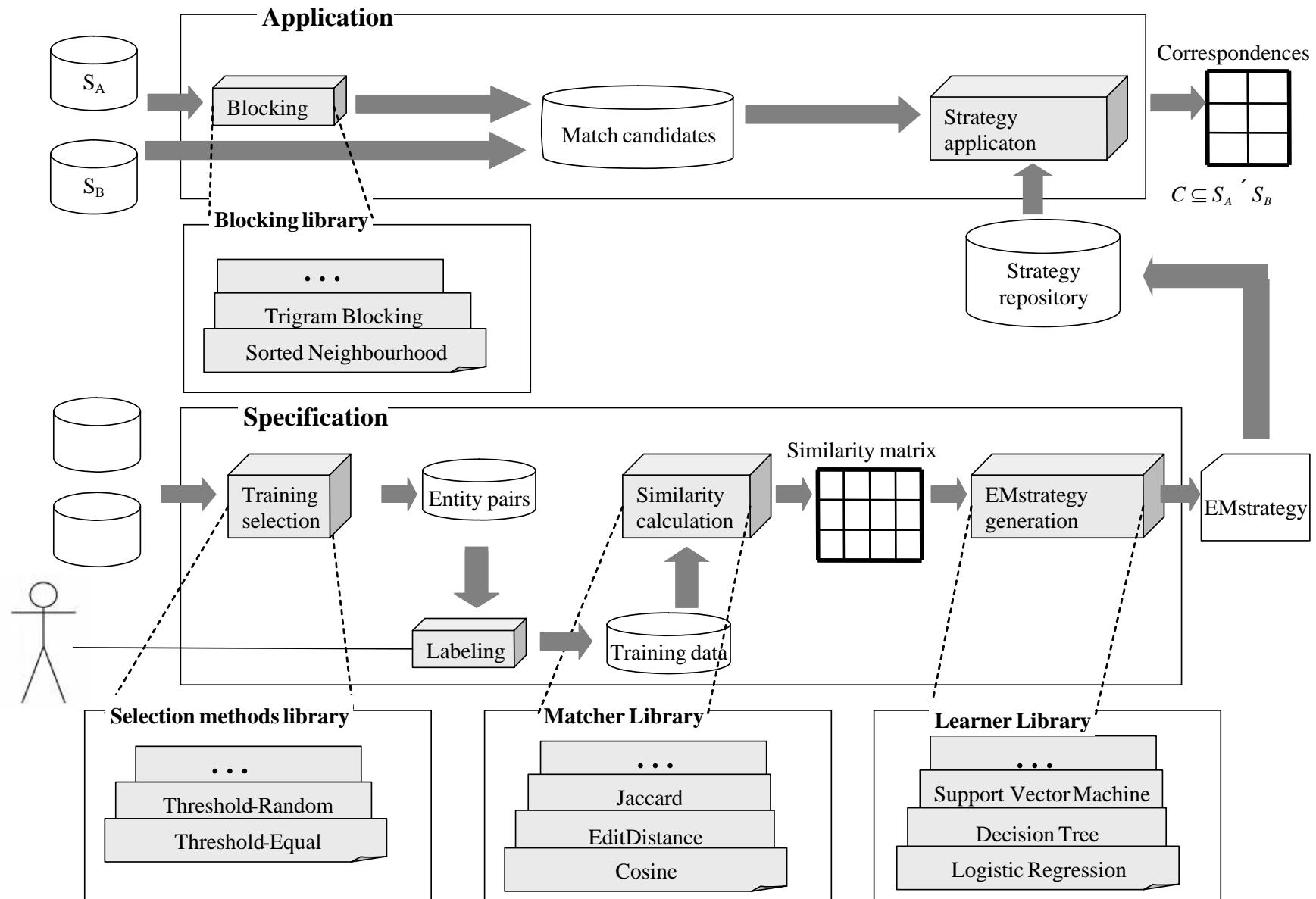


Entity matching

- Given two sets of entities $A \subseteq S_A$ and $B \subseteq S_B$ of a particular semantic entity type from data sources S_A and S_B , the entity matching (EM) problem is to identify all correspondences between entities in $A \times B$ representing the same real-world object.



Entity Matching Framework



Tuning approach

- Treat the objective of determining an EM strategy as a two-class (match or non-match) classification problem
- Employ supervised machine learning methods (learners)
- Requisite: training data



Training data

- Set of examples of matching and non-matching entity pairs

$$\left(\begin{array}{ccc} x_{11}, & \dots, & x_{1m}, y_1 \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ x_{n1}, & \dots, & x_{nm}, y_n \end{array} \right)$$

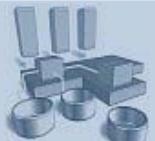


Training data selection

- The effectiveness of a learner critically depends on the size and quality of the available training data.
- Requirements:
 - Representative for the entities to be matched
 - Exhibit the variety and distribution of errors observed in practice
 - Observation of differences between the available matcher algorithms so that an effective combination of different algorithms can be learned
 - Little manual overhead for labeling

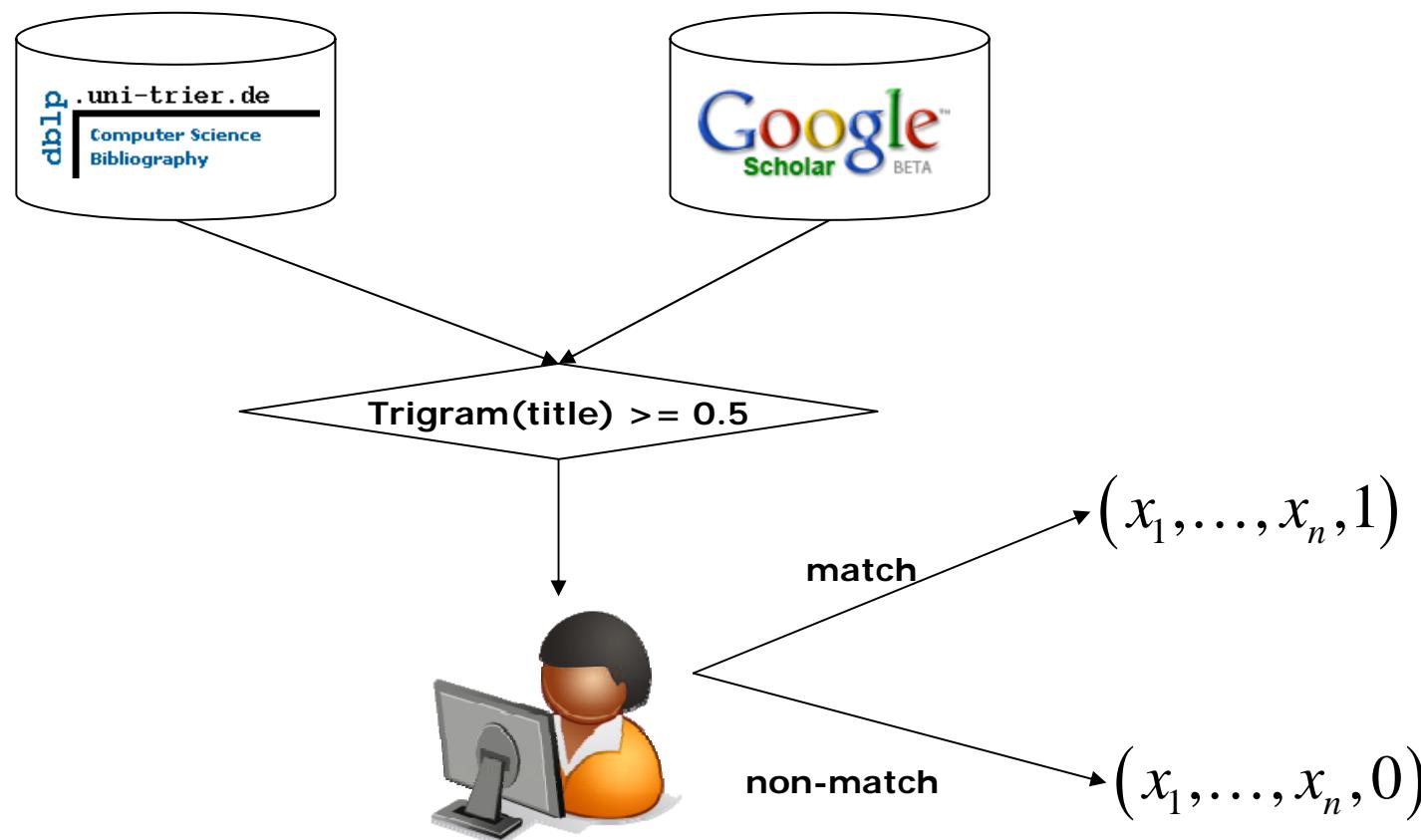


- Manual
- Semi-Automatic
 - Random
 - Threshold-Random
 - Active Learning
- Automatic
 - Nearest based



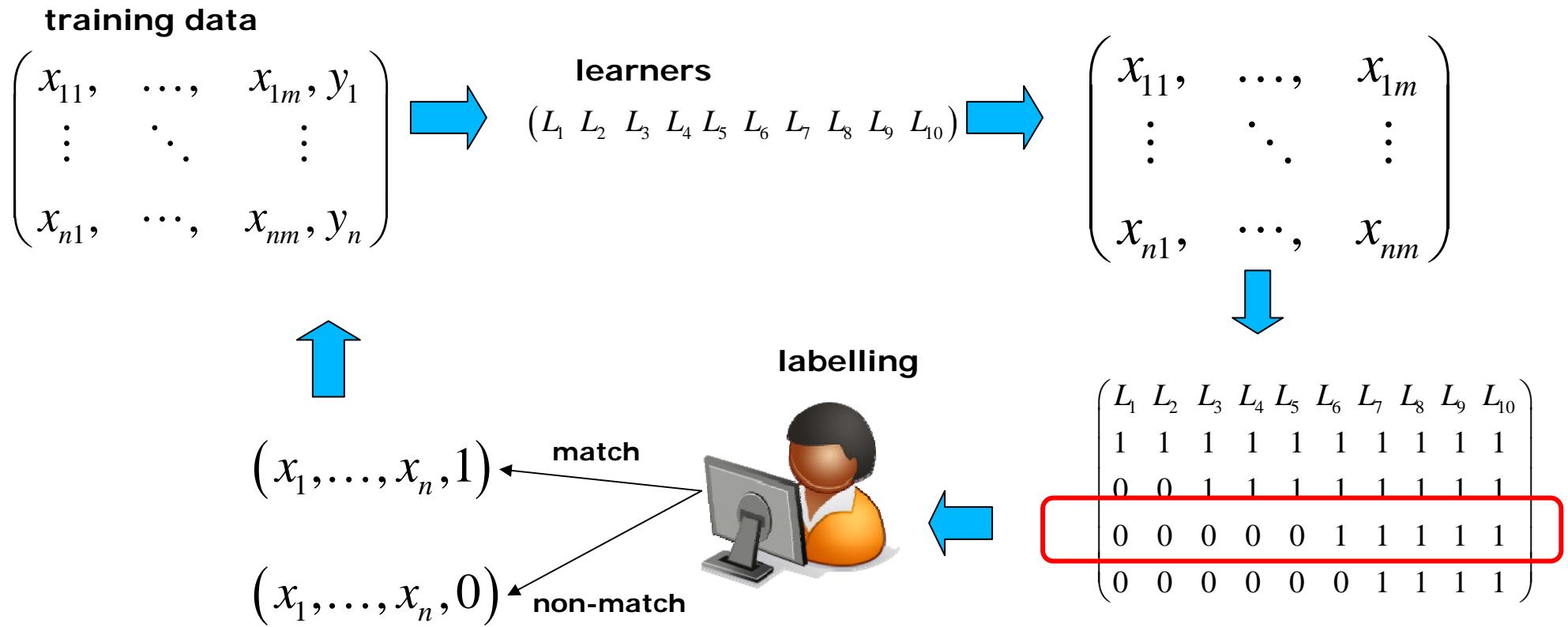
Threshold-Random

- *Threshold-Random* (n, m, t): n object pairs are randomly selected among the ones satisfying a given minimal threshold t applying a similarity measure m .

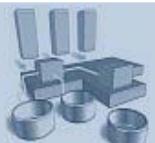


Active Learning (1)

- Attempts to iteratively identify those pairs leading to maximal performance improvements when added to the training set
- Committee of n learners



- Methods for creating committees:
 - Randomizing parameters
 - Partitioning training data
 - Attribute partition



Nearest based

- Proposed by Peter Christen et. al.*
- Selects entity pairs automatically, does not require manual labeling by a user
- The similarity vectors of the entity pairs are sorted according to their distances from the vectors containing only exact similarities and only total dissimilarities, respectively, and then selects the nearest entity pairs for training
- Distance measure: Manhattan distance $d(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i|$

$$\begin{pmatrix} 0.9, & 1.0, & 1.0, & 1.0, & 0.9 \\ 0.0, & 0.0, & 0.0, & 0.0, & 0.0 \\ 0.0, & 0.0, & 0.5, & 0.0, & 0.0 \\ 0.7, & 0.3, & 0.5, & 0.7, & 0.9 \end{pmatrix} \quad \begin{pmatrix} d(\mathbf{x}, \mathbf{1}) \\ 0.2 \\ 5 \\ 4.5 \\ 1.9 \end{pmatrix} \quad \begin{pmatrix} d(\mathbf{x}, \mathbf{0}) \\ 4.8 \\ 0 \\ 0.5 \\ 3.1 \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

*Peter Christen: Automatic Training Example Selection for Scalable Unsupervised Record Linkage, PAKDD, 2008.



Evaluation match tasks (1)

- Bibliographic domain
- Matching of publications



1 [EE] Mauricio A. Hernández, Salvatore J. Stolfo: The Merge/Purge Problem for Large Databases. SIGMOD 1995:127-138

1 [The merge/purge problem for large databases](#)
Mauricio A. Hernández, Salvatore J. Stolfo
May 1995 ACM SIGMOD Record, Volume 24 Issue 2
Publisher: ACM
Full text available: [pdf\(1.37 MB\)](#) Additional Information: [full citation](#), [abstract](#), [references](#), [cited by](#), [index terms](#)
Bibliometrics: Downloads (6 Weeks): 16, Downloads (12 Months): 191, Citation Count: 67



Many commercial organizations routinely gather large numbers of databases for various marketing and business analysis functions. The task is to correlate information from different databases by identifying distinct individuals that appear in a number ...

2 [The merge/purge problem for large databases](#)
Mauricio A. Hernández, Salvatore J. Stolfo
June 1995 SIGMOD '95: Proceedings of the 1995 ACM SIGMOD international conference on Management of data
Publisher: ACM
Full text available: [pdf\(1.37 MB\)](#) Additional Information: [full citation](#), [abstract](#), [references](#), [cited by](#), [index terms](#)
Bibliometrics: Downloads (6 Weeks): 16, Downloads (12 Months): 191, Citation Count: 67

Many commercial organizations routinely gather large numbers of databases for various marketing and business analysis functions. The task is to correlate information from different databases by identifying distinct individuals that appear in a number ...



Computer Science Bibliography



The [merge/purge problem for large databases](#) - all 6 versions »
MA Hernández, SJ Stolfo - Proceedings of the 1995 ACM SIGMOD international conference ..., 1995 - portal.acm.org
Page 1. **The Merge/Purge Problem for Large Databases** * Abstract Mauricio A. Hernández
Salvatore J. Stolfo {nrauricio, sal}@cs. columbi.a. edu ...
Cited by 369 - Related Articles - Web Search - BL Direct

[CITATION] The [merge/purge problem for large databases](#)
S Stolfo, M Hernandez - Proceedings of the ACM SIGMOD Conference on Management of ..., 1995
Cited by 5 - Related Articles - Web Search

[CITATION] The [merge/purge problem for large databases](#)
M Hernandez, S Stolfo - Proceedings of the ACM SIGMOD International Conference on ..., 1995
Cited by 6 - Related Articles - Web Search

[CITATION] The [Merge/Purge Problem for Large Databases](#)
AH Mauricio, JS Stolfo - Proceedings of the 1995 ACM SIGMOD Conference on Management ..., 1995
Cited by 4 - Related Articles - Web Search

[CITATION] andez, SJ Stolfo, The [merge/purge problem for large databases](#)
MA Hern - Proceedings of the 1995 ACM SIGMOD International Conference ..., 1995
Cited by 3 - Related Articles - Web Search

[CITATION] The [Merge/Purge Problem for Large Databases](#), proceedings of ACM SIGMOD
MA Hernandez, SJ Stolfo - PODS, San Jose, CA, 1995
Cited by 2 - Related Articles - Web Search



Evaluation match tasks (2)

- RIDDLE repository (<http://www.cs.utexas.edu/users/ml/riddle/>)
- Restaurant match problem

ZAGAT.
Eat, Drink, Stay, Play

Welcome, please [sign in](#)

New User? [Register](#) or [Subscribe](#) for Access | Help

Your Location
Philadelphia Metro Area Name, cuisine, texture, etc. in Restaurants
 Search By Name Only

Home Ratings & Reviews Add a Review Discussion Boards Zagat Buzz NEW! Video Mobile Store

Members SIGN IN

It's Even Better With Ratings!

Subscribe Today! Access over 30,000 ZAGAT Ratings & Reviews. Find the right place for every occasion.

Join for FREE! Vote in a Featured Survey

Get a FREE Guide! Vote in a Featured Survey

New Orleans Restaurants New Orleans Nightlife Europe's Top Restaurants See More Surveys

OUTTAKE OF THE WEEK "The waiter repeatedly called my aging parents 'Coach' and 'Darling.'"

Read more witty surveyor comments

CUSTOMIZED ZAGAT GUIDES

Philadelphia

Restaurants ALL | NEW | TOP LISTS

Nightlife ALL | NEW | TOP LISTS

Hotels ALL

Attractions ALL

New & Notable **Alfresco Dining**

Check out these new restaurants

Azul Cantina Chima Brazilian Steakhouse Kanella Misso Nicholas Sassafras International Cafe Table 31 10 Arts Tennessee's BBQ & Grill

Messo Photo: Kristina Scott

VISA-ZAGAT SURVEY'S DINING PARTNER VISA Advertisement

Your Opinion Matters.

Please participate in a brief survey.

Click here

Videos: New York City Video profiles of the hottest restaurants keep you up-to-date before you hit the town. Start Watching Now!

Walt Disney World Insider's Guide NEW! Video on ZAGAT.com Text Zagat Information Directly to Your Phone

| Restaurant Name ▾ | Neighborhood | Cuisine | Type | Food | Decor | Service | Cost |
|--|--------------|------------------|-------------|-------------------|-------|---------|------|
| Art's Deli | Studio City | Deli, Sandwiches | ZAGAT RATED | ? | ? | ? | ? |
| 12224 Ventura Blvd. (bet. Laurelgrove & Vantage Aves.) Studio City, CA Map 818-762-1221 | | | | Add Your Review ▶ | | | |

lacounty.gov County Directory of Information & Services | Public Alerts | Public Information | County Contact Information

Environmental Health County of Los Angeles Public Health Home | Program Home | About | FAQ | Comment | Contact | A-Z Index Font Size | A A A

Search Service Locator Select a Center

Facility Rating

Inspected retail food facilities receive a letter grade or a score according to their inspection score (The cities of Long Beach, Pasadena and Vernon inspect their own retail food facilities, so inspection results for those facilities do not appear on this site.)

Effective July 1, 2007, the California Uniform Retail Food Facilities Law (CURFFL) is being repealed in its entirety and replaced with the new California Retail Food Code commonly referred to as Cal Code. For a summary of changes go to California Retail Food Code.

As the result of the changes in the new inspection report, effective July 1, 2007, violation category descriptions and their numeric codes have changed. For a detailed explanation of the new Food Facility Rating System, go to the Revised Retail Food Inspection Guide.

Inspections conducted prior to July 1, 2007 will continue to display the violation sub-category numbers and descriptions from the previous version of the inspection report.

A detailed explanation of the Food Facility Rating system (prior to July 1, 2007), go to the Retail Food Inspection Guide.

You can look up the inspection results for restaurants, markets etc., in your area by using the search term below. Wholesale facilities such as Wholesale Food Markets and Food Warehouses are not rated/scored but full disclosure of violations are displayed. For a detailed explanation of Wholesale Food Inspections go to the Wholesale Food Inspection Guide.

To see information regarding which facilities have recently been closed, please refer to the Food Facility Closure List.

Search Help 

To limit your search to a particular Name, Zip Code, City, or Street Name, fill in one or more of the following:

Name:
Street Name:
...

| Facility Name | Address | City | Zip | Inspected | Score | Type | Info |
|-----------------------|-----------------------|----------------|-------|------------|-------|------------|----------------------|
| ART'S BURGERS | 11629 E VALLEY BLVD | EL MONTE | 91733 | 03/10/2008 | 93 | Restaurant | Info |
| ART'S CHILI DOG STAND | 1410 W FLORENCE AVE | LOS ANGELES | 90044 | 10/04/2007 | 98 | Restaurant | Info |
| ART'S COFFEE SHOP | 1917 ZONAL AVE | LOS ANGELES | 90033 | 04/02/2008 | 91 | Restaurant | Info |
| ART'S DELICATESSEN | 12224 VENTURA BLVD | STUDIO CITY | 91604 | 02/27/2008 | 91 | Restaurant | Info |
| ART'S SNACK BAR # 1 | 7601 E IMPERIAL HWY | DOWNEY | 90242 | 03/06/2007 | 100 | Restaurant | Info |
| ART'S SNACK BAR #2 | 7285 QUILL DR | DOWNEY | 90242 | 05/24/2007 | 96 | Restaurant | Info |
| ART'S SUBS | 20855 VENTURA BLVD #9 | WOODLAND HILLS | 91364 | 11/05/2007 | 99 | Restaurant | Info |
| ART'S WING AND THINGS | 4213 S CRENSHAW BLVD | LOS ANGELES | 90008 | 03/27/2008 | 93 | Restaurant | Info |



Evaluation datasets

| match task | # entities | | # attr. | # corresp. |
|---------------|------------|---------|---------|------------|
| | source1 | source2 | | |
| Scholar-DBLP | 64363 | 2616 | 4 | 5347 |
| ACM-DBLP | 2294 | 2616 | 4 | 2224 |
| Restaurant | 533 | 331 | 4 | 112 |



- Trigram similarity
- Trigrams are character substrings of length three.
- Example: **zingst**

Trigrams: {##z, #zi, zin, ing, ngs, gst, st\$, t\$\$}

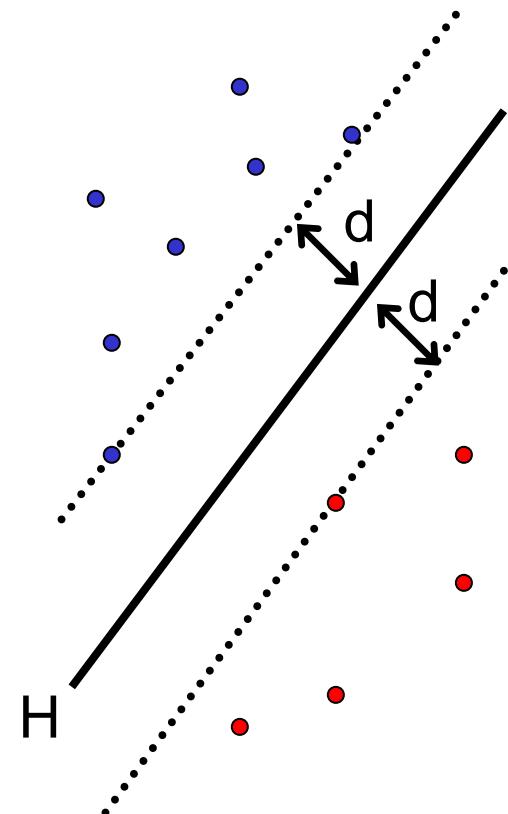
- String with length $l \rightarrow l + 3 - 1$ Trigrams
- Similar strings have many common Trigrams

$$\text{TrigramSim}(s_1, s_2) = \frac{2 * |\text{Trigrams}(s_1) \cap \text{Trigrams}(s_2)|}{|\text{Trigrams}(s_1)| + |\text{Trigrams}(s_2)|}$$



Support Vector Machine (SVM)

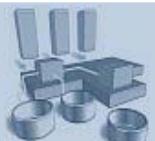
$$h(\vec{x}) = \text{sign} \left(b + \sum_{i=1}^n w_i x_i \right)$$



$$f_{\text{SVM}}(a, b) = \begin{cases} \text{match, } & \text{if } (0.8 \cdot \text{Trigram}(title_a, title_b) + 0.8 \cdot \text{Trigram}(authors_a, authors_b) - 1.1) > 0 \\ \text{non-match, } & \text{otherwise} \end{cases}$$



- Semi-automatic
 - Random
 - Threshold-Random(n , TrigramSimilarity, 0.5)
 - Active Learning
 - 20 initial training examples
 - disjoint partitioning to train SVM 10 times
 - strategies for intial training selection:
 - Random
 - Threshold-Random
 - Nearest based
- Automatic
 - Nearest based



Evaluation measures

- Training time
- Application time
- Performance: F-Measure



UNIVERSITÄT LEIPZIG

Abteilung Datenbanken
Institut für Informatik

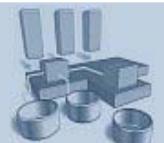
Training Selection for Tuning Entity Matching

Köpcke, Hanna Zingst, 01.07.2008

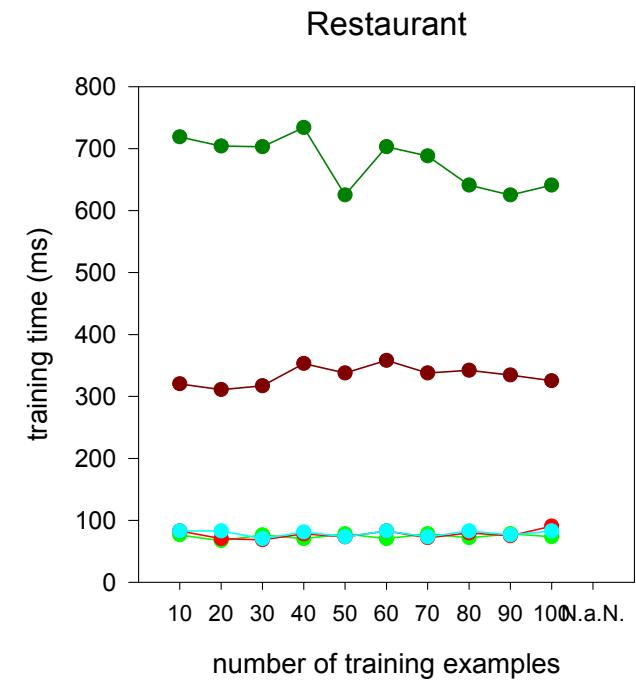
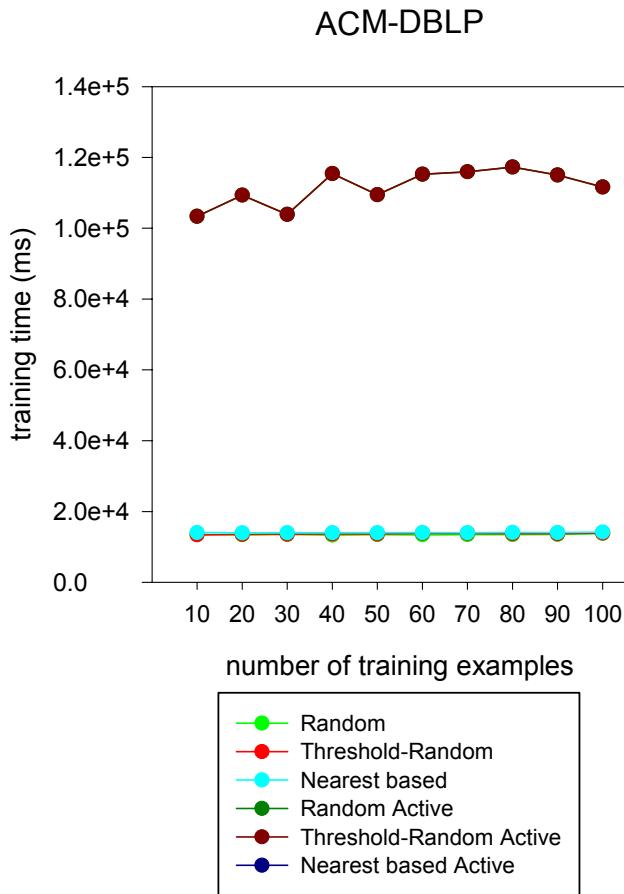
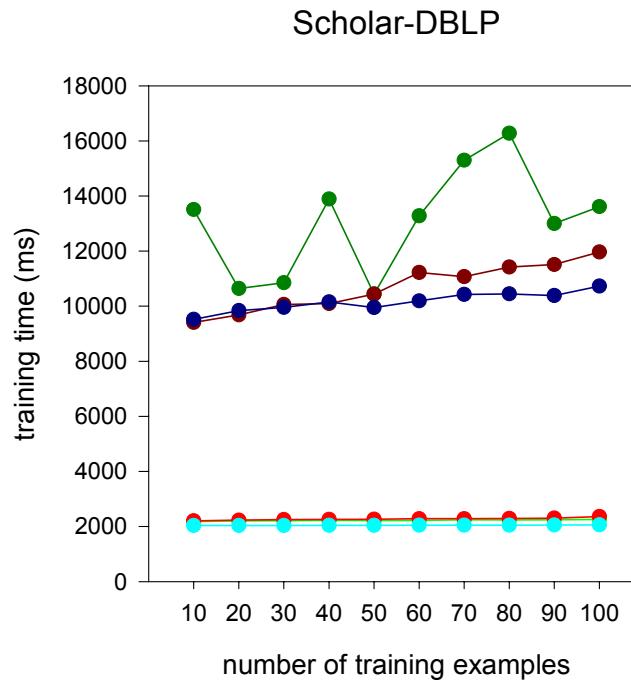
Folie 20



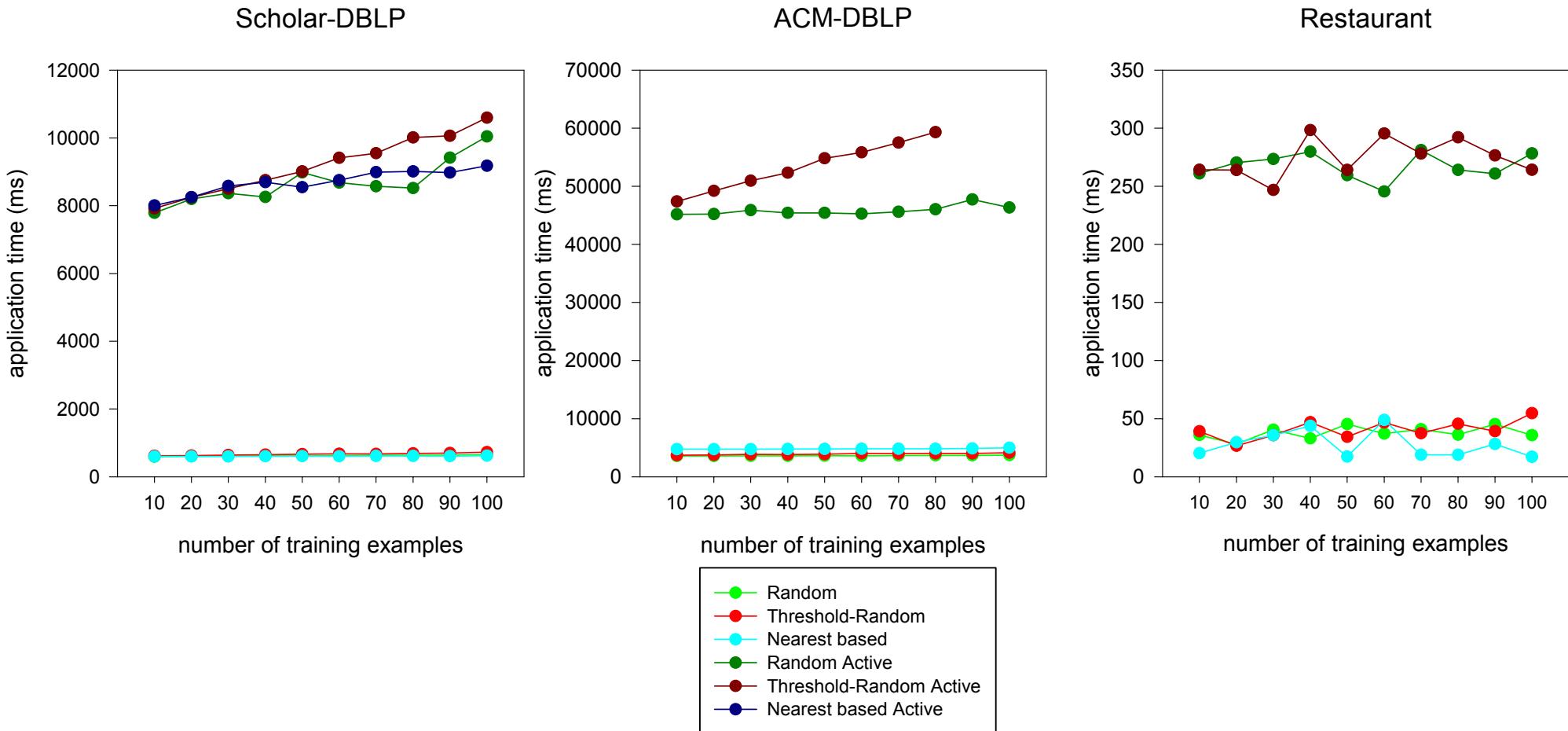
- Bibliographic match tasks:
 - trigram similarity on both title and authors with a threshold of 0. 5
 - Scholar-DBLP: 0.823 F-measure
 - ACM-DBLP: 0.914 F-measure
- Restaurant match task:
 - trigram similarity on name with threshold 0.8
 - 0.881 F-measure



Training time

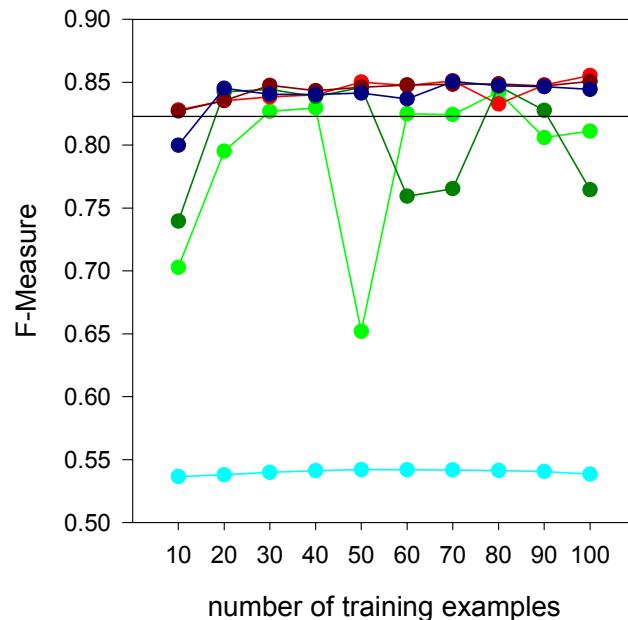


Application time

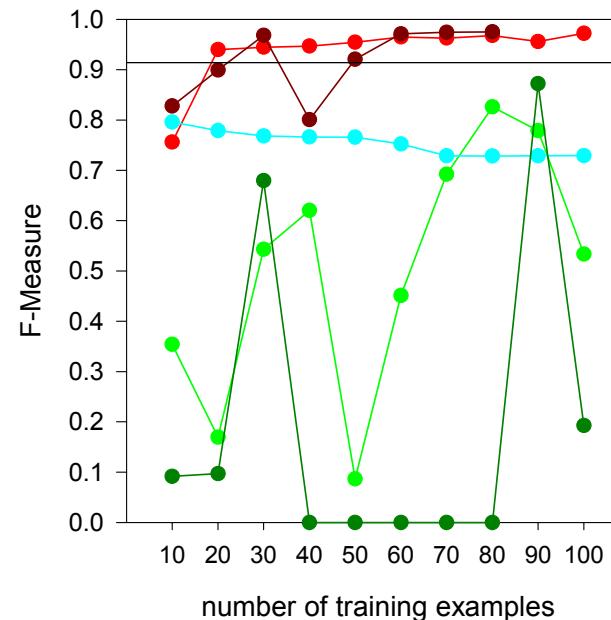


Performance

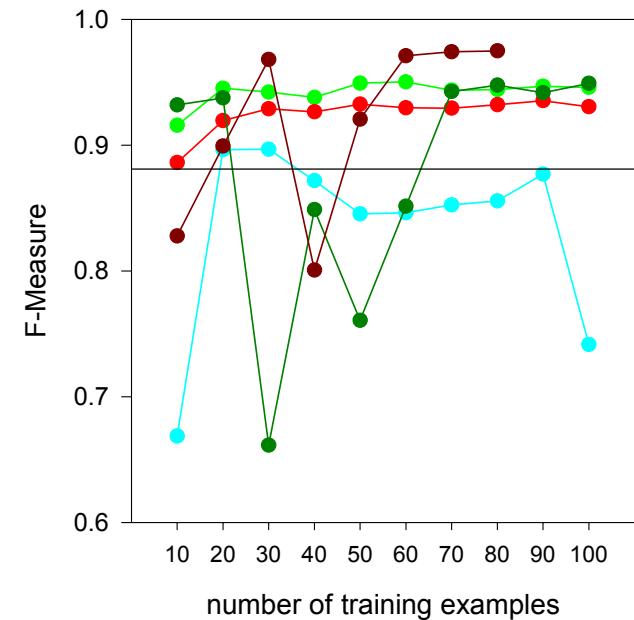
Scholar-DBLP



ACM-DBLP

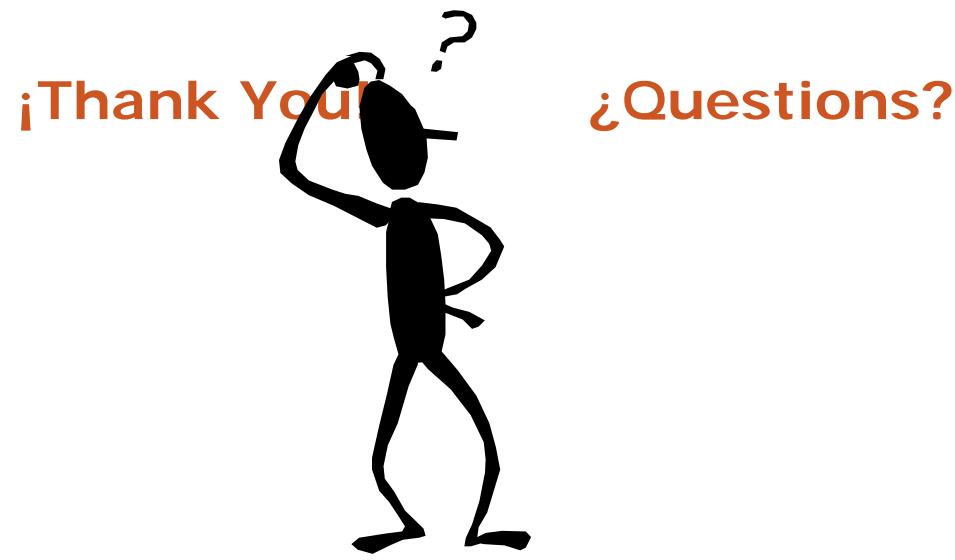


Restaurant



- Training selection
- Evaluation
- Ongoing work
 - Time for training selection
 - Efficient implementation of similarity measures
 - Blocking





¡Thank You!

¿Questions?

