Using the Fuzzy Match algorithm for data cleaning

Ing. David Pejčoch, DiS.

DIKE, University of Economics, Prague
Kooperativa, pojišťovna a.s., VIG
dpejcoch@koop.cz

KEG seminar, 17.4.2008

Outline of presentation

- Introduction
- DQ dimensions
- CDI DQ process
- Match and Merge techniques
- Exact join failure
- Fuzzy Matching definition
- The K-Fuzzy Match problem
- Examples of measures
- Optimization of Fuzzy Match
- Open problems and challenges
- Conclusion

Data Quality dimensions

Accuracy	Closeness between v and v' => Wrong values, duplicites in DB table
Completness	=> Missing values
Currency	How promptly data are updated
Volatility	Frequency with which data vary in time
Timeliness	How current data are for the task at hand
Consistency	=> Data Integration (CDI, PIM, ...)
...	

CDI Data Quality process [7]

- **Define** - Understand the data required to answer business needs
- **Locate** - Locate and validate the correct data sources
- **Profile** - Analyze, characterize, and compare the content
- **Standardize** - Spelling: Bob -> Robert, Consistency of coding (i.e. YYYYMMDD), Parsing: {Paul, Anthony, Samuelson}
- **Match and Merge** - Reconcile and combine data
- **Deploy** - Put records to the Customer Hub
- **Permanent monitoring**

- Typical problem: matching of records against the reference table

Match and Merge techniques

- **Exact Join** (Deterministic Approach)
- **Probabilistic Approach**
  - **Machine learning methods**
  - **Approximate Joins** (Set Joins) using HAVING condition
  - **Fuzzy Match** = Approximate (fuzzy) join using string matching techniques
    - **Token based measures** (Inverse document frequency, Jaccard Coefficient, Probabilistic models like Kullback-Liber Divergence, etc.)
    - **Edit based measures** (Levenshtein/ED, Jaro, Jaro/Winkler, etc.)
    - **Hybrid measures** (Fuzzy Match Similarity)

Exact join failure

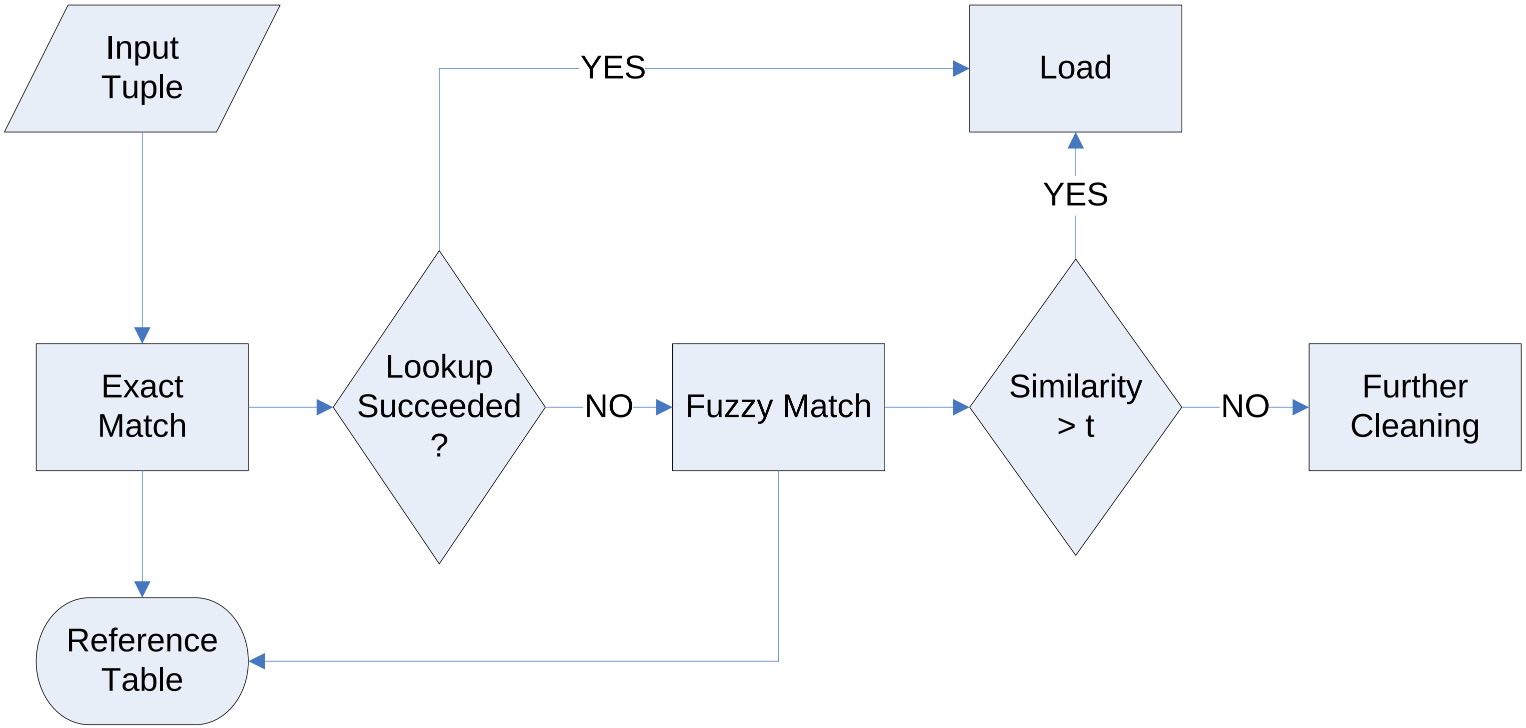
- Typical usage: joining PK with FK
- Problem: Multi-columns join (different syntax of atributes)
- Nicknames (Robert / Bob, Mirek / Miroslav), shortcuts (Road, Rd., nám., Nám., n., Náměstí, bratří, bří), order of tokens in attributes (Chuck Patridge, Patridge Chuck)
- LIKE, CONTAINS can't manage with misspellings

External list	**Reference record from database**
Ing. David Pejčoch, DiS. Kooperativa, a.s. Fiktivní n. 172 110 00 Praha	Pejčoch, David Kooperativa, pojišťovna a.s., VIG Fiktivní nám. 172 Praha 1 110 01

Fuzzy Matching definition

- [1]: Fuzzy matching = matching of incoming record against the reference table.
- Fuzzy Match = opposite of Exact Matching (Deterministic Record Linkage using exact match key)

A Template for using Fuzzy Match [1]



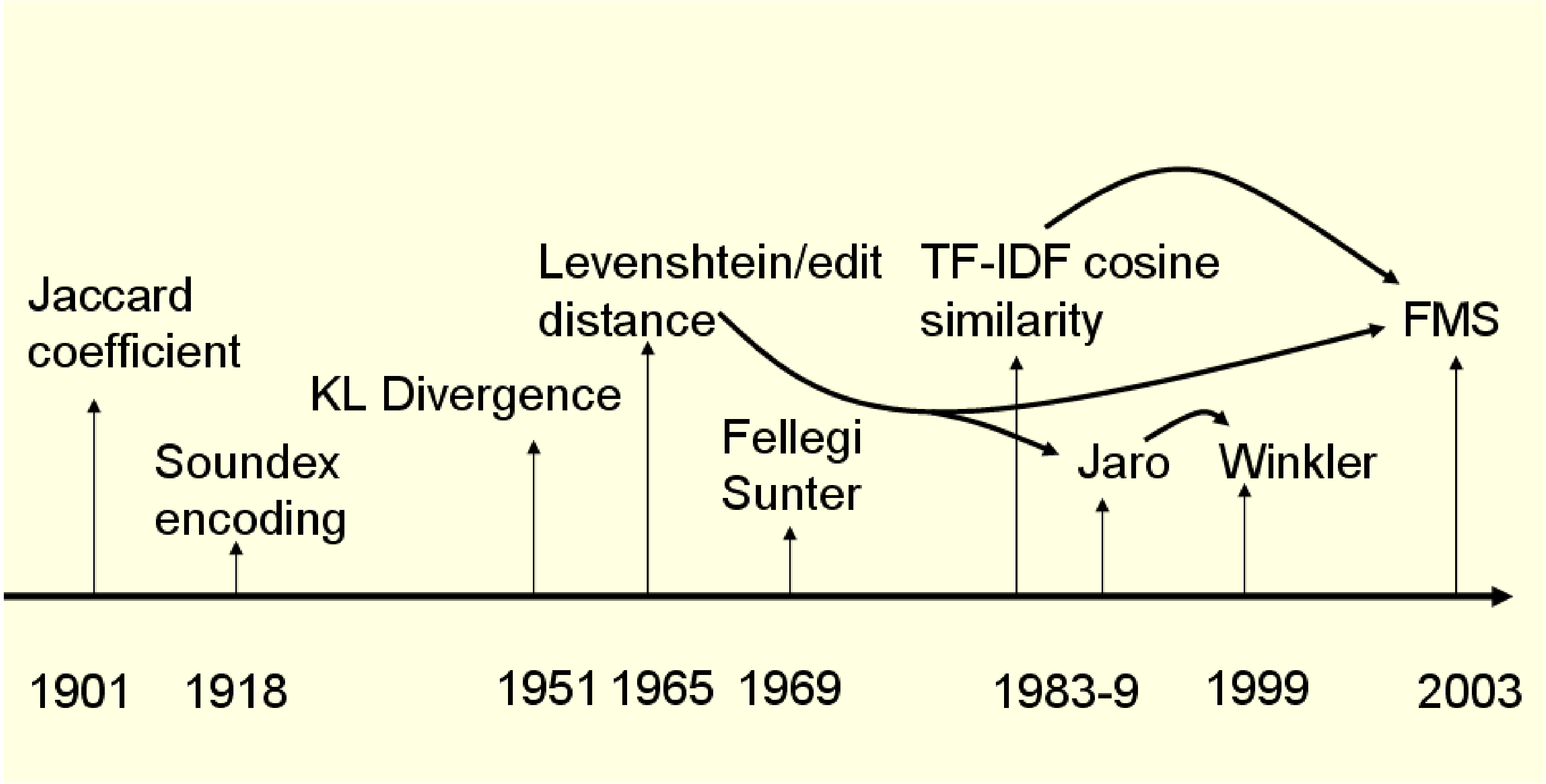
The K-Fuzzy Match Problem [1]

- Given a reference relation R, minimum similarity treshold c $\in$ <0;1>, the similarity function f, and an imput tuple u, find the set FM(u) of fuzzy matches of at most K tuples from R such that
  - f(u,v) >= c, for all v in FM(u)
  - f(u,v) >= f(u,v´) for any v in FM(u) and v´in R – FM(u).

ID	Company Name	City	State	Zip
I1	Beoing Company	Seattle	WA	98004
I2	Beoing Co.	Seattle	WA	98004
I3	Boeing Corporation	Seattle	WA	98004
I4	Company Beoing	Seattle	NULL	98004

ID	Company Name	City	State	Zip
R1	Boeing Company	Seattle	WA	98004
R2	Bon Corporation	Seattle	WA	98014
R3	Companions	Seattle	WA	98024

Historical Timeline [10]



Token based measures: Jaccard coefficient

- v = R[R1, Boeing Company, Seattle, WA, 98004]
- u = R[I1, Beoing Company, Seattle, WA, 98004]
- S = tok(v[1]) = {boeing, company}
- T = tok(u[1]) = {beoing, company}
- Jaccard(S,T) = $|S \cap T| / |S \cup T|$ = 7 / (6 + 6 + 7) = 0,37
- Jaccard(S,T) = 1 => identical

Edit based measures: Levenshtein / Edit distance

- ed(s1,s2) = min. count of insert, delete, replace operation needed to transformation of s1 to s2, normalized by max(d(s1),d(s2)).
- Levenshtein is not normalized.
- Better results: application to q-grams made of tokens.
- ed(s1,s2) = 0 => identical

c	**o**	**m**	**p**			**a**				**n**	**y**
c	**o**	**r**	**p**	**o**	**r**	**a**	**t**	**i**	**o**	**n**	
0	0	1	0	1	1	0	1	1	1	0	1

ed(s1,s2) = 7/11 = 0,64

b	**o**			**n**	
b	**o**	**e**	**i**	**n**	**g**
0	0	1	1	0	1

ed(s1,s2) = 3/6 = 0,5

"Boeing Corporation" = "Bon Corporation" instead of "Boeing Company"

Hybrid measures: The Fuzzy Match Similarity

- Consider string as a sequence of tokens => eliminate insufficiency of ed
- Reflect different importance of tokens (using IDF) - frequence of token in reference relation R
- Domain independent measure
- Can manage with incorrect records

Calculation of weights: $$w(t,i) = IDF(t,i) = \log \frac{|R|}{freq(t,i)}$$

IDF = Inverse Document Frequency
R = total number of records
freq(t,i) = frequency of token in attribute

Fuzzy Match Similarity: Costs of transformation

- **Token replacing costs** = ed(t1,t2) * IDF of replaced token
- **Token deleting costs** = IDF of deleted token
- **Token inserting costs** = inserting factor $c_{ins} \in \langle 0;1 \rangle$ * IDF of inserted token

u[Beoing Corporation, Seattle, WA, 98004]
v[Boeing Company, Seattle, WA, 98004]

Replacing "beoing" for "boeing" and "corporation" for "company"
ed("beoing" ,"boeing") = 0,33
ed("corporation","company") = 0,64
w("beoing",1) = log (4/1) = 0,602
w("corporation",1) = log (4/3) = 0,125
tc(u,v) = 0,33 * 0,602 + 0,64 * 0,125 = 0,278

Fuzzy Match Similarity

$$fms(u,v) = 1 - \min\left(\frac{tc(u,v)}{w(u)}, 1\right)$$

u[Beoing Corporation, Seattle, WA, 98004]

v[Boeing Company, Seattle, WA, 98004]

tc(u,v) = 0,278

w(u) = 0,125 + 0,64 + 0 + 0,125 + 0,125 = 1,015

fms(u,v) = 1 - min(0,278 / 1,015; 1) = 0,726

fms(u,v) = 1 => similar

FMS Approximation

- Consider different order of tokens in input tuple and reference relation => possible to compare tokens among each other
- fms_{apx} is upper bound of fms
- Records which differ only in order of tokens are evaluated as identical
- Application of fms on subset of q-grams called min-hash signature. For q = 3, s = "corporation" set of q-grams QG3('corporation') = {cor, orp, rpo, por, ora, rat, ati, tio, ion}

$$fms^{apx}(u,v) = \frac{1}{w(u)} \sum_{i} \sum_{t \in tok(u[i])} w(t) \max_{r \in tok(v[i])} \left(\frac{2}{q} sim_{mh}(QG(t), QG(r)) + d_q \right)$$

FMS$_{apx}$ Example

- q = 3
- max number of q-gams H = 2
- v[Company Beoing, Seattle, NULL, 98004]
- u[Boeing Company, Seattle, WA, 98004]
- min-hash signature u = [eoi, ing], [com, pan], [sea, ttl], [980, 004]
- min-hash signature v = [oei, ing], [com, pan], [sea, ttl], [wa], [980, 004]
- w(u) = 0,25 + 0,5 + 1 + 2 = 3,75
- Fms$_{apx}$ ignore inserting costs of 'WA' !!!!
- Fms$_{apx}$(u,v) = 1 / w(u) * w(beoing) * (2/3 * 0,5 + 1 - 1/3) = 3,75 / 3,75 * 1 = 1 => similar
- For comparision: fms(u,v) = 0,726 + inserting costs of 'WA'

Optimization

- w(t) = IDF * subjective weights of atribute
- Drop vowels
- Hashing => Phonetic scheme SOUNDEX, NYSIIS (The New York State Identification and Intelligence System)
- Blocking = using another atribute to reduce search space (i.e. ZIP code)
- Pruning = deleting records that cannot match
- q-grams + index

Soundex

- Phonetic scheme for encoding names
- Algorithm:
  - retain first letter
  - delete a, e, i, o, u, y, w, h
  - encode remaining consonants
  - delete adjacent letters with the same code
  - syntax of code must be letter and three digits => add zeroes
- Many different names have the same Soundex code
- Some names that are closely related are coded differently
- [24]: Best on European last name

Soundex - Example

- Smith = Smythi = S530
- Lee = Liu = L000
- Rogers = R262, Rodgers = R326
- Ševc = Švec = S120
- Srp = Srb = S610
- Šemberiová = S516, Zsemberiová = Z251

Letters	**Code**
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

Indexing

- **Naïve algorithm** - Compare each tuple with others
- **M-tree index** - Often used in multimedia databases queries. Makes partitions of objects based on distance. Not implemented in DWH.
- **Error Tolerant Index** (using by fms_{apx}) => temporary table contaning minhash q-grams with B-tree index

Q-gram	Coordinate	Column	Freq	Tid list
oei	1	1	1	{R1}
ing	2	1	1	{R1}
com	1	1	2	{R1,R3}
pan	2	1	2	{R1,R3}
bon	1	1	1	{R2}

Open problems and challenges

- **Absence of standard benchmark** for similarity measures (i.e. [10]) => collection of ~30 measures, SAS code of measures + collecting metrics for benchmark (precision, false negative percentage, ...)
- **Combination** of similarity measures with methods of machine learning
- **Full automation** of domain independent solutions vs. involving of domain knowledge
- **Improving the performance** of algorithms without loosing accuracy
- **Combining** incoming and reference records
- Multi-table joins
- Improving indexing
- Improving hashing

Conclusion

- FMS was implemented as FUZZY LOOKUP + FUZZY GROUPING components in MS SQL Server 2005, 2008 (SQL Server Integration Services)
- Edit distance and Jaro-Winkler distance were implemented in Match-Merge Operator in Oracle Warehouse Builder 10g
- Domain specific solutions: Trillium, DataFlux, FirstLogic ...

Reference

- [1] Chaudhuri S., Ganjam K., Ganti V., and Motwani R.: Robust and efficient fuzzy match for online data cleaning. In *Proceedings of the ACM SIGMOD*, June 2003.
- [2] Wang R. Y., Ziad M., Lee Y. W.: Data Quality, The Kluwer International Series on Advances in Database Systems Volume 23, Springer, Berlin, 2001.
- [3] Rubens N. O.: The Applicationn of Fuzzy Logic to The Construction of The Ranking Function of Information Retrieval Systems, Computer Modelling and New Technologies, 2006, Vol. 10., No. 1, str. 20-27.
- [4] Dasu T., Johnson T.: Exploratory Data Mining and Data Cleaning John Wiley 2003
- [5] Gravano L., Ipeirotis P. G., Jagadish H. V., Koudas N., Muthukrishnan S., Srivastava D.: Approximate string joins in a database (almost) for free. In Proceedings of VLDB, Roma, Italy. September 11-14 2001.
- [6] Navarro G., Baeza-Yates R., Sutinen E., Tarhio J.: Indexing methods for approximate string matching. IEEE Data Engineering Bulletin, 24(4):19—24,2001.
- [7] Dyché J., Levy E: Customer Data Integration – Reaching a Single Version of Truth, John Wiley & Sons, Inc., 2006.

Reference

- [8] English L.: Improving Data Warehouse and Business Information Quality – Methods for Reducing Costs and Increasing Profits, John Wiley & Sons, 1999.
- [9] Patridge Ch.: The Fuzzy Feeling SAS Provides: Electronic Matching of Records without Common Keys, Observations – the technical journal for SAS Software Users, 1998, SAS Institute Inc., Cary.
- [10] Srivastava D.: Record Linkage: A Database Approach, AT&T Labs-Research: http://www.research.att.com/~divesh/
- [11] http://www.levenshtein.net/
- [12] Broder A.: On the resemblance and containment of documents. In Compression and Complexity of Sequences (SEQUENCES ,97), 1998.
- [13] Cohen E.: Size estimation framework with applications to transitive closure and reachability. Journal of Computer and System Sciences, 1997.
- [14] Cohen W.: Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity. In Proceedings of ACM SIGMOD, Seattle, WA, June 1998.
- [15] Ananthakrishna R., Chaudhuri S., and Ganti V.: Eliminating fuzzy duplicates in data warehouses. In *Proceedings of the 28th international conference on very large databases (VLDB)*, pages 586–597, Hong Kong, August 20-23 2002.
- [16] Ciaccia P., Patella M., Zezula P.: M-Tree: An efficient access method for similarity search in metric spaces. VLDB 1997.

Reference

- [17] Winkler W.: The state of record linkage and current research. *Statistics of Income Division, Internal Revenue Service Publication R99/04* http://www.census.gov/srd/papers/pdf/rr99-04.pdf*.*
- [18] Winkler W. E., Thibaudeau Y.: An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U. S. Decennial Census, U. S. Burreau fo Census, 1991.
- [19] Gravano L., Ipeirotis P. G., Jagadish H. V., Koudas N., Muthukrishnan S, Srivastava D.: Approximate string joins in a database (almost) free. In Proceedings of VLDB, Roma, Italy. September 11-14 2001.
- [20] Baillie M., Azzopardi L, Crestani F.: Towards Better Measures: Evaluation of Estimated Resource Description Quality For Distributed IR.
- [21] Cohen W. W., Ravikumar P., Fienberg S. E.: A Comparison of String Distance Metrics for Name-Matching Tasks. IIWeb 2003: 73-78
- [22] http://download.oracle.com/docs/cd/B31080_01/doc/owb.102/b28223/ref_dataquality.htm
- [23] MSDN - SQL Server Developer Center: http://msdn2.microsoft.com/en-us/library/ms137786(SQL.100).aspx
- [24] Herzog T. N., Scheuren F. J., Winkler W. E.: Data Quality and Record Linkage Techniques, Springer, 2007.

Thanks for your attention

Questions?