Entity Resolution – azonosságfeloldás

"**Entity Resolution (ER)** is the process of identifying groups of records that refer to the same real-world entity."

„rejtett, való világbeli *entitások*hoz köthető *megfigyelések* csoportosítása az entitásonk köré"

ELTE, 2011.03.22.

Sidló Csaba - sidlo@sztaki.hu

Adatbányászat és Webes Keresés Kutatócsoport: http://datamining.sztaki.hu

MTA SZTAKI

datamining search

Entity Resolution – azonosságfeloldás

Témák mára:

- probléma leírása, példák, változatok
- megoldások:
  - attribútum-hasonlóság alapúak
  - kapcsolat alapúak (hálózati)
  - egzakt, szabály alapúak
- új eredmények:
  - megoldások minőségének mérése
  - több algoritmus kombinálása
  - megkötések szerepe
  - ...

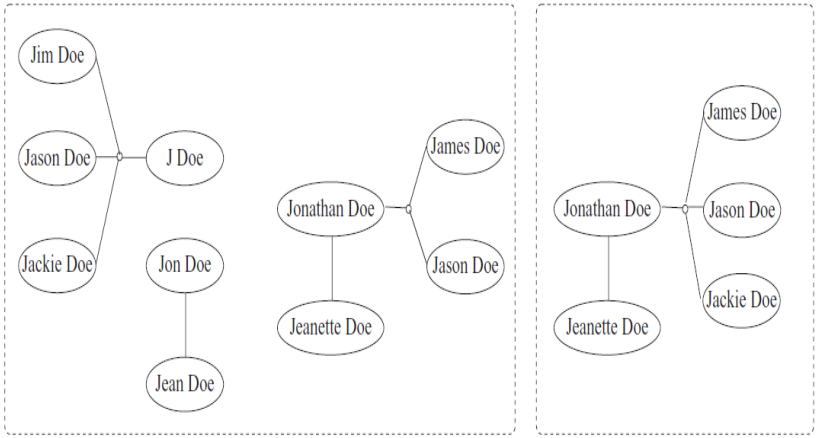
A témakör

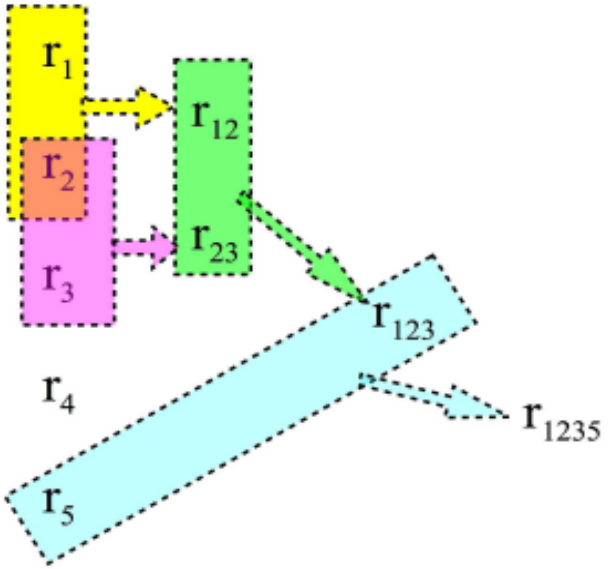
- elnevezések ~1960 óta hasonló problémákra
- (eltérő elnevezések → eltérő megközelítések):
  - **record linkage** (1960, Fellegi, Sunter)**,** duplicate detection, duplicate record detection, merge/purge; deduplikáció, duplikátum-keresés
  - **entity resolution** → “azonosságfeloldás”
  - instance identification, reference reconciliation, coreference resolution, database hardening, …
- kutatási terület, példa: Very Large Databases konferencia, 2010:
  - nagyságrendileg 80-90 cikkből
    - ~5 db entity resolution cikk,
    - ~10-15 szorosan kapcsolódó cikk

Entity Resolution (ER): a feladat

- megfogalmazási lehetőségek:
  - modell: rekordok halmaza / fa: XML / gráf
  - szemlélet:
    - match-merge → összevonás, reprezentatív elemmel
    - klaszterezés: csoportosítás, partíciókk
  - közelítő / egzakt megoldások
  - felügyelt tanulás (tanító halmaz) / nem felügyelt tanulás
- kapcsolódó területek:

  klaszterezés (adatbányászat), similarity join, string hasonlóságok, adatminőség, adattisztítás, adattárházak, adatintegráció, információ integráció, …





Példa: Google Places

sztaki

A [Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézete MTA **SZTAKI**](#) - more info »
1111 Budapest, Kende Street 13, Hungary
+36 1 209 5400

B [MTA Számítástechnikai és Automatizálási Kutató Intézet](#) - more info »
1111 Budapest, Kende Street 17, Hungary
+36 1 279 6000

C [MTA **SZTAKI** DSD (Department of Distributed Systems)](#) - more info »
1111, Lágymányosi Street 11., Hungary
+36 1 279 6212

D [MTA **SZTAKI** W3C Magyar Iroda](#) - more info »
1111 Budapest, Lágymányosi Street 11., Hungary
+36 1 279 6204

E [MTA **SZTAKI**](#) - more info »
Hungary
+36 1 279 6000
This is an unverified listing

F [Scope Meetings Ltd. (MTA **SZTAKI**)](#) - more info »
1111 Budapest, Kende Street 13, Hungary
+36 1 209 6001
This is an unverified listing
2 reviews
"Scope Meetings Ltd. was established in 1990 based on the experience and ..."

G [Data Mining and Web Search Group / Adatbányászati és Webes Keresés Kutatócsoport](#) - more info »
1111 Budapest, Lágymányosi Street 11, Hungary
+36 1 279 6172

példa forrása: Google Maps

Példa: ügyfelek

gyakori modell:
rekordok: {r_1, r_2, r_3, r_4, r_5, r_6…} → {<r_1>, <r_2, r_6>, <r_4>, <r_3, r_5>, ... }

előállítandó ennek egy particionálása (<> jelentése: klaszterek, mint halmaz):

Record	Name	Address(zip)	Email
r	John Doe	02139	jdoe@yahoo
s	John Doe	94305	
t	J. Foe	94305	jdoe@yahoo
u	Bobbie Brown	12345	bob@google
v	Bobbie Brown	12345	bob@google

gyakori problémák:

- heterogén adatforrások: redundáns, örökölt, átfedő stb. rendszerek
- heterogén formátum: különböző sémák, szabványok, szokások (pl. postai címek)
- adatminőség, lexikális heterogenitás: adatbeviteli hibák, hiányzó, kitöltetlen attribútumok, szabályok megkerülése (pl. default értékek, 11111-es azonosítók vagy 1970.01.01 dátum) stb.

Példa: ügyfelek

gyakori modell:
rekordok: {r_1, r_2, r_3, r_4, r_5, r_6…} → {<r_1>, <r_2, r_6>, <r_4>, <r_3, r_5>, ... }

előállítandó ennek egy particionálása (<> jelentése: klaszterek, mint halmaz):

Record	Name	Address(zip)	Email
r	John Doe	02139	jdoe@yahoo
s	John Doe	25	
t	J. Foe	9	jdoe@yahoo
u	Bobbie Brown	123	bob@google
v	Bobbie Brown	12345	b@google

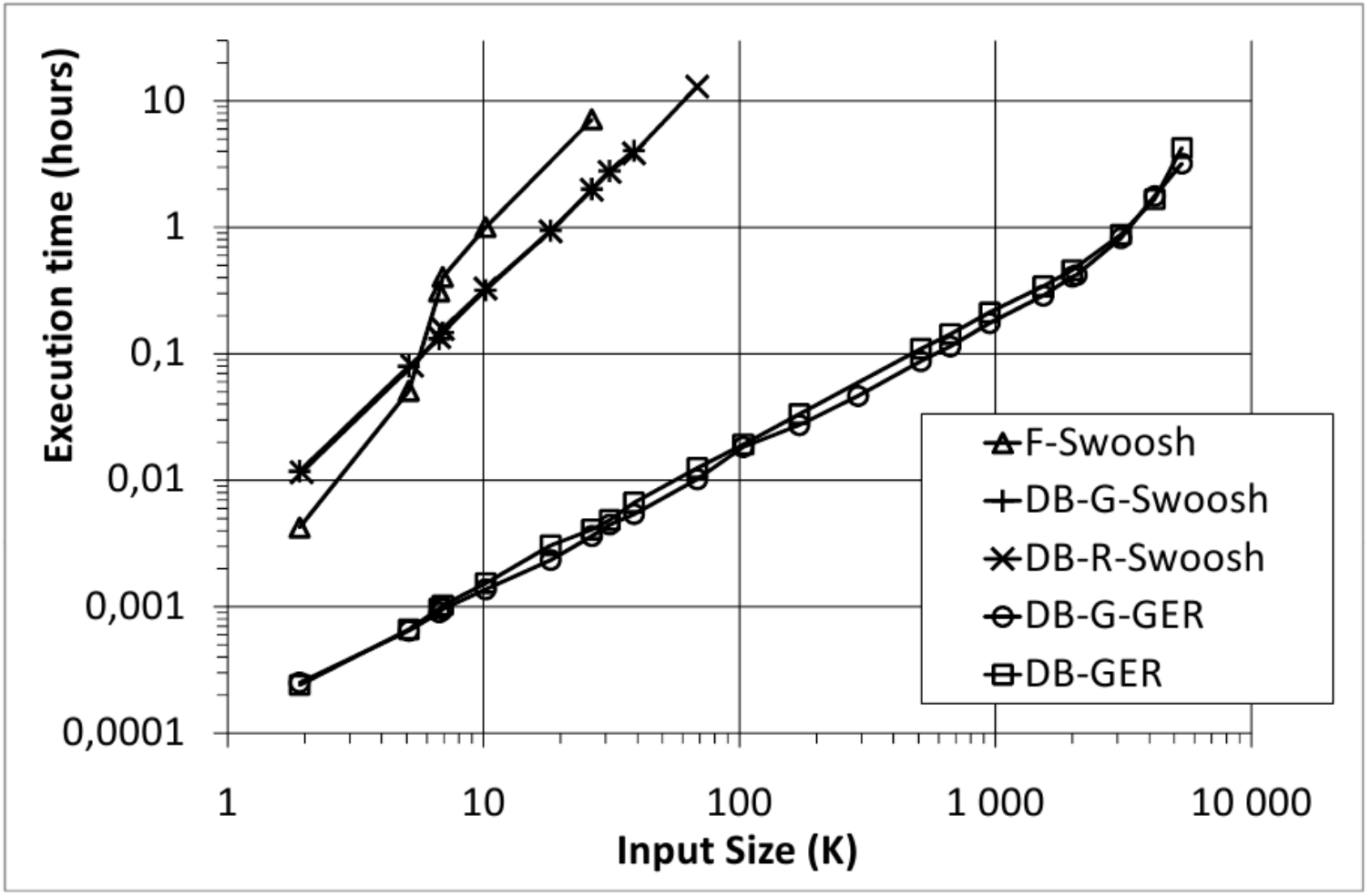
párok vizsgálata: **O (n^2) !**
A probléma nehéz.

gyakori problémák:

- heterogén adatforrások: redundáns, örökölt, átfed
- heterogén formátum: különböző sémák, szabványok, szokások (pl. postai címek)
- adatminőség, lexikális heterogenitás: adatbeviteli hibák, hiányzó, kitöltetlen attribútumok, szabályok megkerülése (pl. default értékek, 11111-es azonosítók vagy 1970.01.01 dátum) stb.

Példa: ügyfelek - futásidők

futásidő példa ügyfél adatbázison, ~15 M rekord, egy átlagosnál kicsit jobb asztali gépen, többféle hatékony algoritmus:



További alkalmazások

- klasszikus feladat: publikációs adatbázisok
  - kevés attribútum: írók nevei, esetleg munkahely
  - *kapcsolatok* entitások közt: közösen írt cikkek
- ügyfelek:
  - jellemzően sok attribútum: természetes + generált (id)
  - heterogén forrásrendszerek (különböző portfóliók, örökölt rendszerek, összeolvadások stb.); **hány ügyfelünk van igazából? kerestük már ajánlattal? szerződtünk már vele valaha?**
  - pl: Yahoo közel-keleti felvásárlás: **mennyi az új felhasználók száma valójában** – mennyit érdemes költeni?
- web:
  - weboldalak ('mirror detection'), termék-keresők termékei, entitások weboldalakon: személyek, dátumok, helyek stb.; **hány Facebook / IWIW felhasználó van igazából**?
- ...

kép forrása: http://www.flickr.com/photos/popilop/331357312/

Record Linkage model - 1969

- populations: A, B
- ebből M: matched, U: unmatched párok halmaza $A \times B = \{(a, b); a \epsilon A, b \epsilon B)$
- cél: positive link: $\mathbf{A}_1$: $(a, b) \in M$, positive non-link: $\mathbf{A}_3$: $(a, b) \in U$
  $\mathbf{A}_2$: possible link: nem tudjuk eldönteni
- A, B-ből egy-egy mintát alakítunk rekordokká → két file
- comparison vector: $\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], \cdots, \gamma^K[\alpha(a), \beta(b)]\}$
  - rekordra (α, β): jellemzők, állítások halmaza, pl. “name is the same”, “name is missing on one record”
  - → comparison space: minden lehetséges realizáció halmaza
- linkage rule: leképezés, “comparison space → random decision functions”

$$d(\gamma) = \{P(A_1 | \gamma), P(A_2 | \gamma), P(A_3 | \gamma)\}; \ \gamma \epsilon \Gamma \qquad \sum_{i=1}^{3} P(A_i | \gamma) = 1.$$

Ivan P. Fellegi, Alan B. Sunter: **A Theory for Record Linkage**, Journal of the American Statistical Association, 1969

Duplicate Record Detection

- hasonló alaphelyzet: A és B, két halmaz közti párokra $(\alpha, \beta) \in M$ vagy $(\alpha, \beta) \in U$
- lexikális heterogenitás feloldása a cél
- adatelőkészítés: parsing, data transformation, standardization (→ 'ETL')
- megoldások: vagy megtanulják a megoldást, vagy szakértői tudás kell
- ha van tanulóadat: M és U egy részhalmaza adott
  - statisztikai módszerek: $\underline{x}$: összehasonlítás vektor <α, β>-ra

$$\langle \alpha, \beta \rangle \in \begin{cases} M & \text{if } p(M|\underline{x}) \geq p(U|\underline{x}) \\ U & \text{otherwise.} \end{cases}$$

Bayes tétellel (l: 'likelihood ratio'):

$$\langle \alpha, \beta \rangle \in \begin{cases} M & \text{if } l(\underline{x}) = \frac{p(\underline{x}|M)}{p(\underline{x}|U)}) \geq \frac{p(U)}{p(M)} \\ U & \text{otherwise.} \end{cases}$$

naïve Bayes: attr. függetenség feltételezése,
cél a szükséges eloszlások becslése

$$p(\underline{x}|M) = \prod_{i=i}^{n} p(x_i|M)$$
$$p(\underline{x}|U) = \prod_{i=i}^{n} p(x_i|U).$$

Ahmed K. Elmagarmid and Panagiotis G. Ipeirotis and Vassilios S. Verykios: **Duplicate record detection: A survey**, IEEE TKDE, 2007

Duplicate Record Detection 2.

$P(x_i \mid M)$, $P(x_i \mid U)$: becslés a tanuló-halmaz alapján

enélkül is alkalmazható a modell: EM algoritmus

general expectation maximization (EM) működésének feltételei:

- nagy arányú egyezés (5 % fölött)
- a matchelők 'jól elkülönülnek'
- kis arányban hibás attribútumok
- függetlenségi feltételezés nagyjából igaz

ezeken lehet itt-ott gyengíteni.

Bayes következtetés: új megfigyelések - hipotézis

hibák (első- és másodrendű) súlyozása: 'Bayes decision rules for minimum cost'

- ellenőrzött tanulás:
  - CART, SVM, regresszió ...

Ahmed K. Elmagarmid and Panagiotis G. Ipeirotis and Vassilios S. Verykios: **Duplicate record detection: A survey**, IEEE TKDE, 2007

Duplicate Record Detection 3.

- active learning: emberi közreműködés, jól választott döntések meghozatala
  - többféle klasszifikátor egyidejűleg, majd a bizonytalan elemekre kérdezni
  - pl.: ALIAS rendszer
- ha nincs tanulóadat:
  - statisztikai módszerek (EM)
  - rekord szintű hasonlósági függvények: távolság alapú módszerek
    - távolság függvény + treshold-ok
    - speciálisan: szabály alapon; szakértők → szabályok → szabály alapú egyezések
  - unsupervised learning: klaszterező algoritmusok
    - speciális: sok, jellemzően kicsi klaszter

Ahmed K. Elmagarmid and Panagiotis G. Ipeirotis and Vassilios S. Verykios: **Duplicate record detection: A survey**, IEEE TKDE, 2007

Duplicate Record Detection: gyorsítás

eddig: A x B teljes összehasonlítás

- rekord párok számának csökkentése:
  - blocking: speciális hash függvény heurisztika alapján
  - sorted neighborhood:
    - kulcs készítése (attrib. konkatenáció pl.) → rendezés
    - fix méretű ablakban keressük a match-előket
  - tranzitivitás kihasználása (ha van)
  - canopies: átfedő blocking
  - set join
- rekord-rekord összehasonlítás gyorsítása:
  - dimenzionalitás csökkentése

Ahmed K. Elmagarmid and Panagiotis G. Ipeirotis and Vassilios S. Verykios: **Duplicate record detection: A survey**, IEEE TKDE, 2007

String matching / field similarity

- karakter-alapú hasonlóság metrikák:
  - edit distance /Levenstein-distance/: minimális szerkesztési távolság
    - karakter-műveletek: beszúrás, törlés, csere
    - dinamikus programozás megoldás: O(m*n) idő és tár
      n*m-es mátrix kitöltése
  - affine gap distance
    - edit distance bővítés, új műveletek: open gap / extend gap (kisebb súllyal)
  - Smith-Waterman: szavak elején és végén kisebb súlyú az eltérés
  - Jaro distance, Jaro-Winkler, ...
  - q-gram (karakter q-asok egyezése):
    - hash indexeléssel O(max {m, n}) (index az n-gramoknak)
    - positional q-gram: (i, q-gram)
  - kapcsolódó téma: DB string join

String matching / field similarity 2.

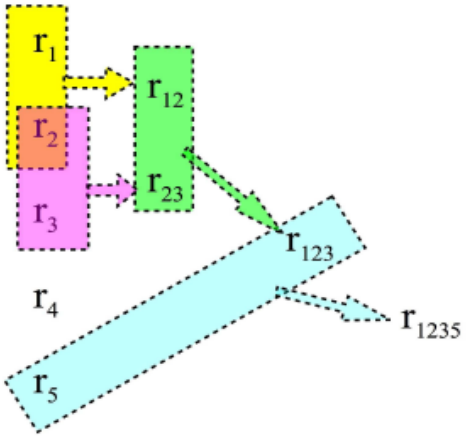
- token-alapú hasonlóság metrikák: “John Doe” vs. “Doe, John”
  - 'atomic strings': tokenekre bontás; egyezés alternatíva: prefixek esetén
  - TF-IDF (term frequency – inverse document frequency)

    term t_i document d_j: $\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $\mathrm{idf}_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$
  - WHIRL példa: szavak súlya:

    cosine sim. $v_\sigma(w) = \log(tf_w + 1) \cdot \log(idf_w)$ $sim(\sigma_1, \sigma_2) = \frac{\sum_{j=1}^{|D|} v_{\sigma_1}(j) \cdot v_{\sigma_2}(j)}{\|v_{\sigma_1}\|_2 \cdot \|v_{\sigma_2}\|_2}$
- fonetikus hasonlóság:
  - Soundex: egy-egy átkódolások (pl. D,T → 3; B, F, P, V → 1)
    - szaparátorok, magánhangzók: feladarabolás
    - állítólag létezik magyar soundex is
  - NYSIIS: bővítés – magánhangzók bevétele
    - betűk → fonetikusan hasonló betűk
  - ONCA: brit
  - Metaphone: soundex alternatíva

Generic Entity Resolution

- páronkénti döntés: rekord-párok összehasonlítása
- egzakt megoldás (nincs közelítés)
- nincsenek kapcsolatok: a rekord minden információt tartalmaz
- feature-ök: attribútum kombinációk
- fix séma (adott attribútumok)
- R: rekordok halmaza; „black box” függvények: match: RxR → {true, false}, merge: RxR → R parciális függvény (match-előkre értelmezett)
- merge lezárt: legkisebb új elemmel bővíthetetlen
- 'domination': rekordok rendezése – melyik 'jobb' leíró
- entity resolution (ER): legkisebb olyan lezárt, ami nem tartalmaz dominált elemeket



Benjelloun, Omar and Garcia-Molina, Hector and Menestrina, David and Su, Qi and Whang, Steven Euijong and Widom, Jennifer (2005): **Swoosh: A Generic Approach to Entity Resolution**. Technical Report. Stanford. → 2009: VLDB Journal

Generic Entity Resolution 2.

- ICAR tulajdonságok:

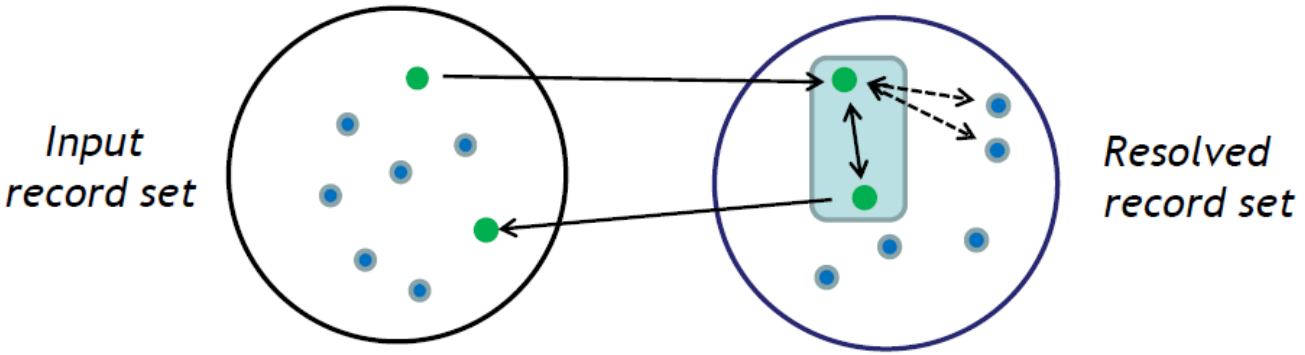
1. *Idempotence:* $\forall r,\ r \approx r$ and $\langle r, r \rangle = r$. A record always matches itself, and merging it with itself still yields the same record.
2. *Commutativity:* $\forall r_1, r_2,\ r_1 \approx r_2$ iff $r_2 \approx r_1$, and if $r_1 \approx r_2$, then $\langle r_1, r_2 \rangle = \langle r_2, r_1 \rangle$.
3. *Associativity:* $\forall r_1, r_2, r_3$ such that $\langle r_1, \langle r_2, r_3 \rangle \rangle$ and $\langle \langle r_1, r_2 \rangle, r_3 \rangle$ exist, $\langle r_1, \langle r_2, r_3 \rangle \rangle = \langle \langle r_1, r_2 \rangle, r_3 \rangle$.
4. *Representativity:* If $r_3 = \langle r_1, r_2 \rangle$ then for any r_4 such that $r_1 \approx r_4$, we also have $r_3 \approx r_4$.

- tranzitivitást nem teszünk fel
- ennek megfelelve ER véges lehet: a dominált rekordokat eldobhatjuk

Generic Entity Resolution 3.

Algoritmusok:

- brute-force: mindent mindennel összehasonlítunk és -vonunk, amíg ez lehetséges
- G-Swoosh:
  - egyesével bővíteni a lezártat merge-elt rekordokkal
  - ICAR nélkül is működik, de kell egy utólagos dominált-rekord eldobás
- R-Swoosh
  - ICAR esetére: összevonás esetén a forrás rekordok eldobhatók → csak domináns rekordok maradnak
- F-Swoosh
  - feature-based + feature indexeket használ (feature-feature párok, illetve negatív match feature értékek)



Generic Entity Resolution 4.

Swoosh tulajdonságok:

- inkrementális algoritmus
- helyes, optimális

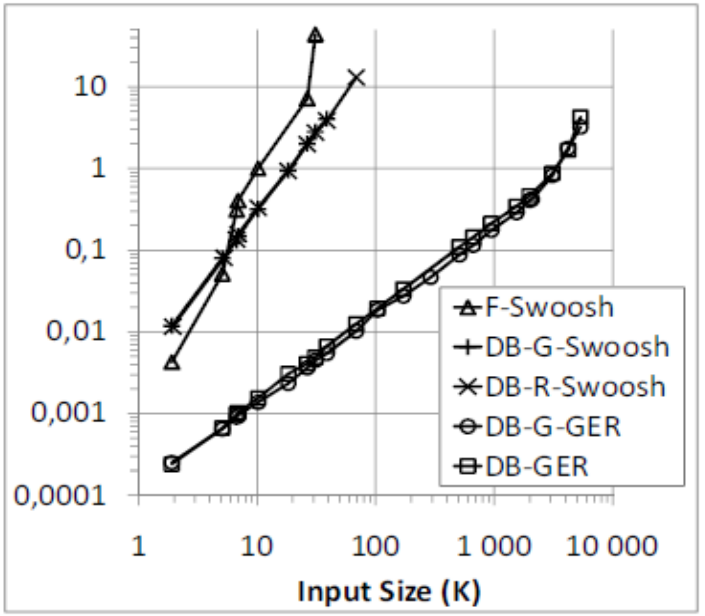
Bizonyosságok bevezetése: Koos algoritmus

- rekord: (r.c, r.A) pl. 0.8[név: John Doe, szül.dátum: 1965]
- feltevések: kommutativitás, idempotencia
- dominancia relációt szintén használnak

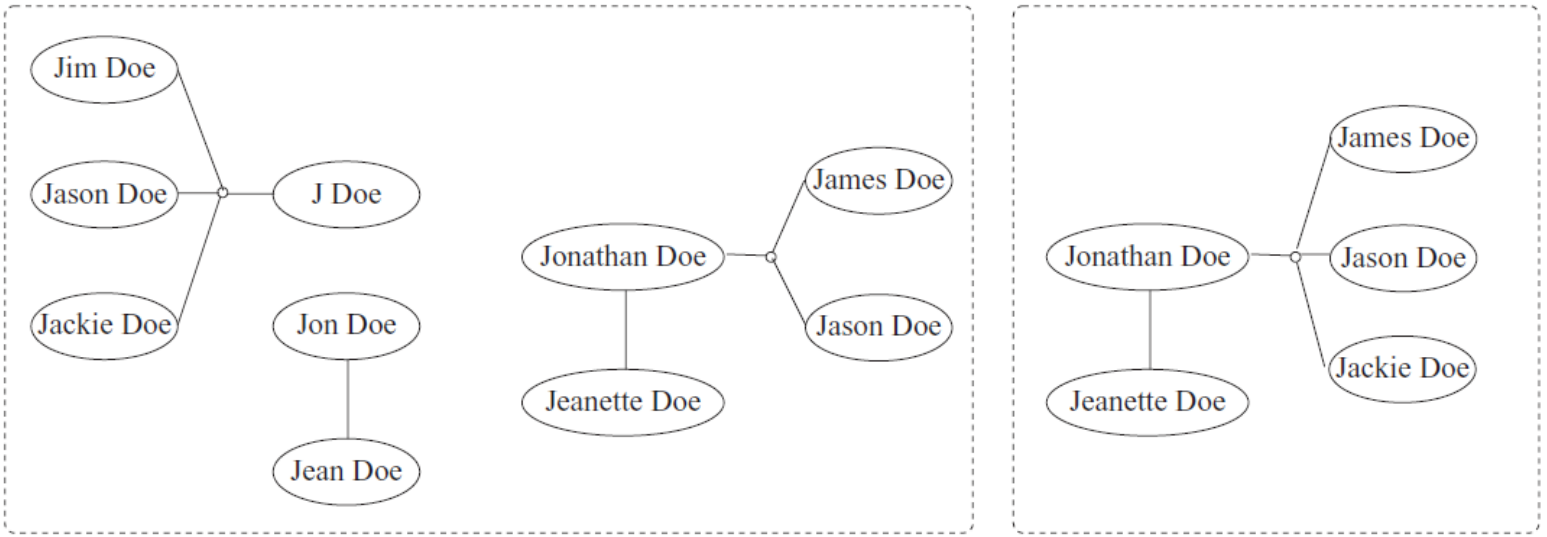
D-Swoosh, P-Swoosh: elosztott megoldások – nem tűntek túl használhatónak

Adatbázis generic ER:

- saját változat; rekord halmaz x rekord műveletek
- sql implementáció: kötegelt, optimalizálható lépések



Relational Clustering



- relációs = kapcsolati; cél: reference graph → resolved entity graph
- hipergráf; hiperélek: kapcsolatok
- “naïve relational”:

$$sim_{NR}(r_i, r_j) = (1 - \alpha) \times sim_A(r_i, r_j) + \alpha \times sim_H(r_i, r_j), \qquad 0 \leq \alpha \leq 1.$$

ahol (H helyett A): $sim_H(r, h_j) = max_{r' \in h_j} sim_H(r, r').$

Bhattacharya, Indrajit and Getoor, Lise: **Collective entity resolution in relational data**. ACM Trans. Knowl. Discov. Data, 2007

Relational Clustering 2.

- “collective relational”: speciális klaszterezés
- két klaszter távolsága:

$$sim(c_i, c_j) = (1-\alpha) \times sim_A(c_i, c_j) + \alpha \times sim_R(c_i, c_j), \qquad 0 \le \alpha \le 1$$

- egy klaszter szomszédsága:

$$Nbr(c) = \bigcup_{h \in c.H, r \in h.R} \{c_j \mid c_j = r.C\}$$

- hasonlóság mértékek:
  - közös szomszédok száma:

$$CommonNbrScore(c_i, c_j) = \frac{1}{K} \times \left| Nbr(c_i) \bigcap Nbr(c_j) \right|,$$

  - Jaccard-együttható:

$$JaccardCoeff(c_i, c_j) = \frac{|Nbr(c_i) \bigcap Nbr(c_j)|}{|Nbr(c_i) \bigcup Nbr(c_j)|}.$$

Relational Clustering 3.

- Adamic / Adar távolság:

  weboldalakra: $$similarity(X, Y) = \sum_{\text{shared feature } z} \frac{1}{\log(frequency(z))}.$$

  klaszterekre: $$Adar(c_i, c_j) = \frac{\sum_{c \in Nbr(c_i) \cap Nbr(c_j)} u(c)}{\sum_{c \in Nbr(c_i) \cup Nbr(c_j)} u(c)},$$

  ahol u(c) a klaszter label “egyedisége”, pl. $$u(c) = \frac{1}{\log(|Nbr(c)|)}$$

- szélesebb szomszédság is bevonható
- algoritmus:
  - mohó algoritmus: összevonni mindig a leghasonlóbbakat
  - bonyolultság: O(nf), ahol n az élek száma és egy él max. f klaszterhez rendelődik
  - javítás: blocking, bootstrapping (okos kezdeti klaszter kialakítás)

SIGMOD 2009

S. E. Whang, D. Menestrina, G. Koutrika, M. Theobald, H. Garcia-Molina: **Entity Resolution with Iterative Blocking**

- blocking techniques: legjobb blocking feltétel meghatározása a cél ; blokkokat nem szerveztük újra
- most: többszörös blokkokra osztás (akár átfedően), az új eredmények propagálása ezek közt
  - lefedési tulajdonság: a klaszter elemi rekordjai szerint döntök a blokkról
- keretrendszer: tetszőleges resolution algoritmus beilleszthető
  - általános core resolution alg. (CER): rekordok klaszterei → rekordok klaszterei
  - CER mindig csak összevon, sosem szed szét – dominancia: az eredmény mindig "jobban összevont"

Criterion	$b_{-,1}$	$b_{-,2}$	$b_{-,3}$
SC_1	r	s,t	u,v
SC_2	r,s	t	u,v

Criterion	$b_{-,1}$	$b_{-,2}$	$b_{-,3}$
SC_1	$r, \langle r,s\rangle$	$s,t, \langle r,s\rangle$	$\langle u,v\rangle$
SC_2	$\langle r,s\rangle$	t	$\langle u,v\rangle$

Criterion	$b_{-,1}$	$b_{-,2}$	$b_{-,3}$
SC_1	$\langle r,s\rangle, \langle r,s,t\rangle$	$\langle r,s,t\rangle$	$\langle u,v\rangle$
SC_2	$\langle r,s,t\rangle$	$\langle r,s,t\rangle$	$\langle u,v\rangle$

Criterion	$b_{-,1}$	$b_{-,2}$	$b_{-,3}$
SC_1	$\langle r,s,t\rangle$	$\langle r,s,t\rangle$	$\langle u,v\rangle$
SC_2	$\langle r,s,t\rangle$	$\langle r,s,t\rangle$	$\langle u,v\rangle$

SIGMOD 2009

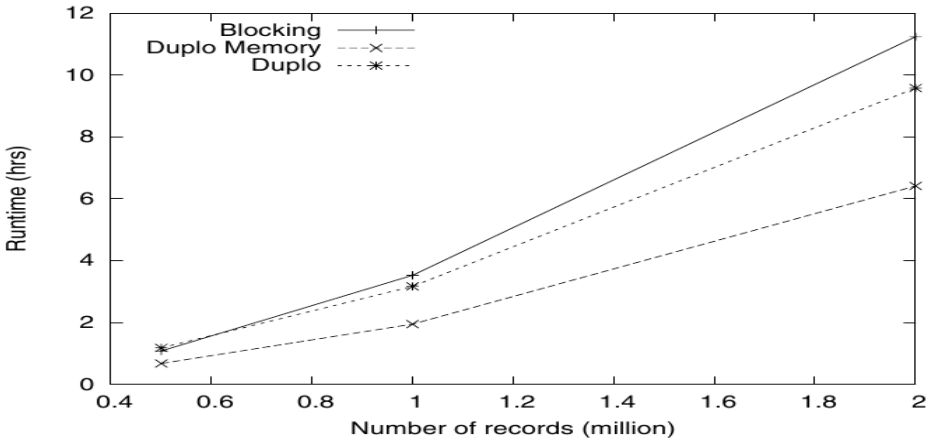
S. E. Whang, D. Menestrina, G. Koutrika, M. Theobald, H. Garcia-Molina: **Entity Resolution with Iterative Blocking (2)**

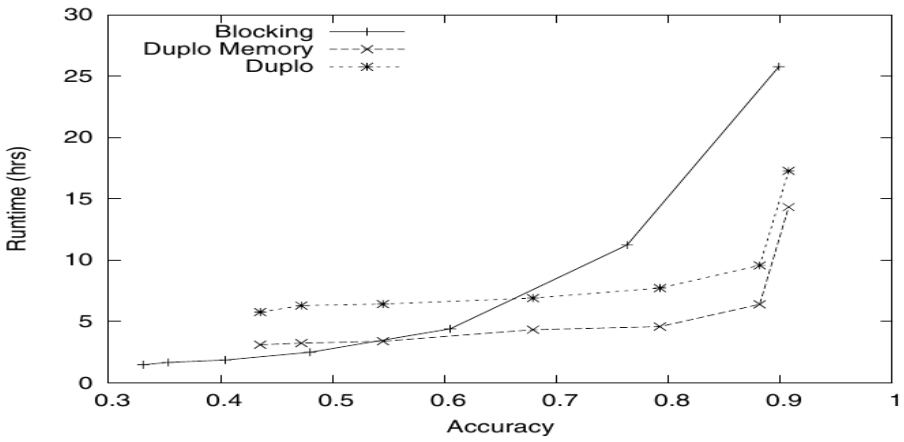
- algoritmusok:
  - elvárások eredményre: valid particionálás legyen, és ne maradjon benne match
  - 1. iterative blocking
    - konfliktusok: adott elem több klaszterbe is bekerül; pl. {<r, s>, <s, t>, <u, v>}
      - unmerge: {r, s, t, <u, v>}
      - connect: {<r, s, t>, <s, t>, <u, v>}
  - 2. Lego:
    - maximális rekordok cseréje: ha van <r,s> akkor r-et és s-et erre cseréljük; ehhez hash-elés
    - 'block que' bevezetése, módosul a blokk-sorrend: valószínűbb összevonások először

SIGMOD 2009

S. E. Whang, D. Menestrina, G. Koutrika, M. Theobald, H. Garcia-Molina: **Entity Resolution with Iterative Blocking** (3)

- 3. Duplo: disk-based, jól skálázódó
  - fix méretű szegmensek: memória méretű blokk-részek;
    - adott blocking feltételhez tartoznak, egyenletes véletlen rekord leosztással
    - feldolgozás: szegmensek → ezen belül a konkrét blokkok
    - bővítésre fenn kell tartani némi helyet
  - merge log: összevonások naplója; pl. r → <r, s>; s → <r, s>
    - maximális elemek hash táblája helyett
    - többnyire elég kicsi (kevés merge)
- mérések:
  - R-Swoosh CER és 'minhash signatures' (egyfajta Jaccard 3-gramokkal); minőség: blocking nélküli CER-rel összehasonlítva





VLDB 2010

Menestrina, David and Whang, Steven Euijong and Garcia-Molina, Hector:
Evaluating Entity Resolution Results

- “gold standard”: ember által kiértékelt eredmény; ritkán adott, enélküli kell

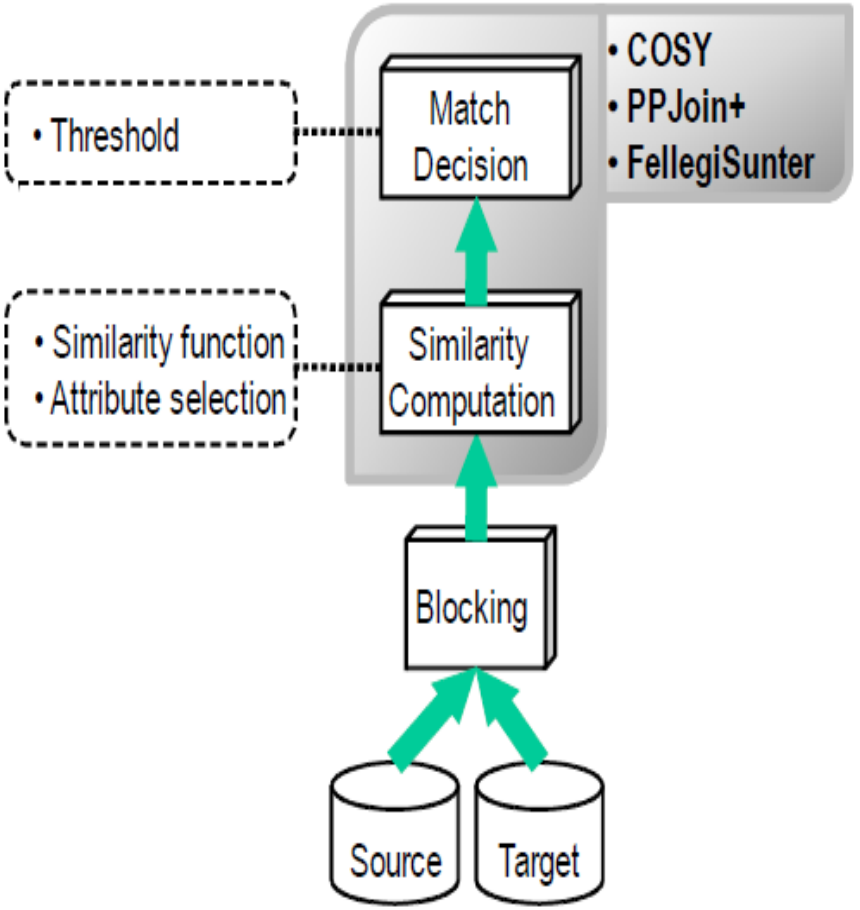
Set	ER Result
Gold Standard	$\{\langle a,b\rangle, \langle c,d\rangle, \langle e,f,g,h,i,j\rangle\}$
R_1	$\{\langle a\rangle, \langle b\rangle, \langle c\rangle, \langle d\rangle, \langle e,f,g,h,i,j\rangle\}$
R_2	$\{\langle a,b\rangle, \langle c,d\rangle, \langle e,f,g\rangle, \langle h,i,j\rangle\}$
R_3	$\{\langle a,b,c,d\rangle, \langle e,f,g,h,i,j\rangle\}$

- eddigi 'jóság' mértékek:
  - IR / AI, klaszterezés metrikák: klaszterek közti és klaszteren belüli hasonlóságok – nem elég speciálisak
  - precision, recall, F-measure; páronkénti F_1, klaszter F_1, VI: információ-vesztés mértéke
- generalized merge distance (GMD): edit-distance-szerű
  - műveletek: split, merge
  - mérték: legrövidebb műveleti út egyikből-másikba
  - súlyozás: súlyozhatóak a műveletek → legkisebb súlyú út
  - speciális esetei néhány előzőleg használt mérték
- 'slice' algoritmus: O(n) alg. ennek kiszámítására ??

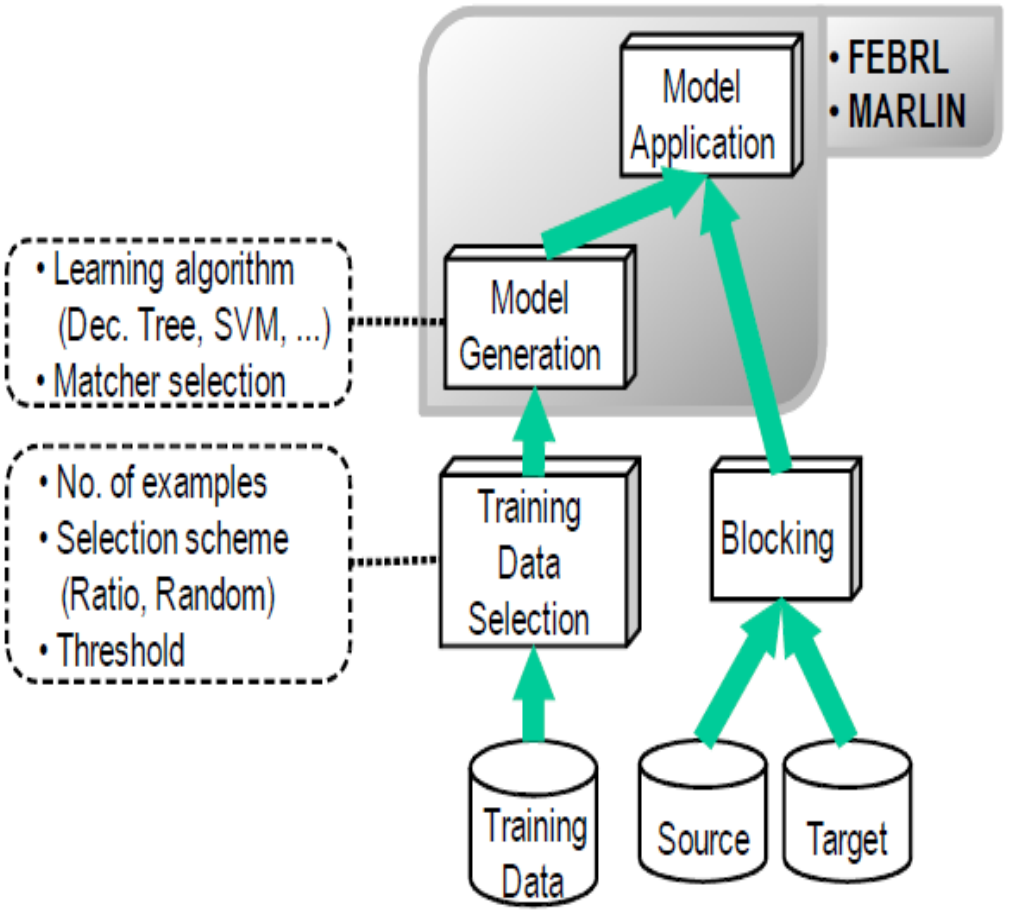
VLDB 2010 (demo: VLDB 2009)

Köpcke, H.; Thor, A.; Rahm, E.: **Evaluation of Entity Resolution Approaches on Real-world Match Problems**

- FEVER: új keretrendszer összehasonlításhoz és algoritmus-hangoláshoz



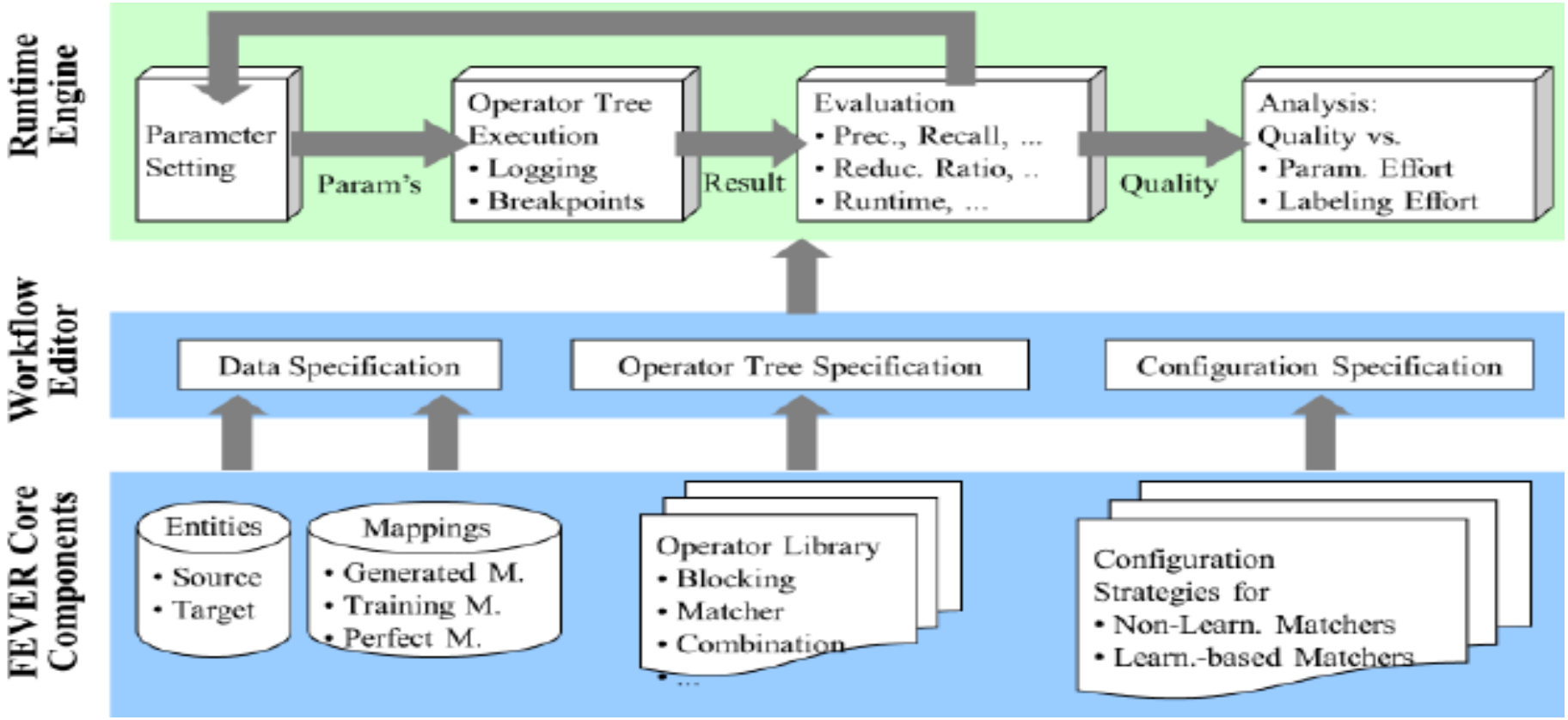
(a) Non-learning match approaches



(b) Learning-based match approaches

VLDB 2010 (demo: VLDB 2009)

Köpcke, H.; Thor, A.; Rahm, E.: **Evaluation of Entity Resolution Approaches on Real-world Match Problems (2)**



- algoritmusok:
  - PPJoin+, Fellegi-Sunter, ? commercial
  - keretrendszerek: FEBRL SVN-nel, MARLIN SVN-nel és többféle klasszifikátorral
  - sokminden más:
    http://dbs.uni-leipzig.de/de/research/projects/object_matching/fever/analysis_of_research_publications

VLDB 2010 (demo: VLDB 2009)

Köpcke, H.; Thor, A.; Rahm, E.: **Evaluation of Entity Resolution Approaches on Real-world Match Problems (3)**

- futásidő + minőségi jellemzők: precision, recall, F-measure
- adathalmazok (mind egyedileg crawl-olt):

Match task			Source size (#entities)		Mapping size (#correspondences)		
Domain	Attributes	Sources	Source 1	Source 2	Full input mapping (Cartesian product)	Reduced input mapping (blocking result)	perfect result
Bibliographic	- title - authors	DBLP-ACM	2,616	2,294	6 million	494,000	2,224
	- venue - year	DBLP-Scholar	2,616	64,263	168.1 million	607,000	5,347
E-commerce	- product name - description	Amazon-GoogleProducts	1,363	3,226	4.4 million	342,761	1,300
	- manufacturer - price	Abt-Buy	1,081	1,092	1.2 million	164,072	1,097

- megállapítások:
  - tanulóhalmaz hasznos,
  - kereskedelmi termék jó,
  - e-commerce feladatot nem tudták jól megoldani (→ kellenek a kapcsolatok)

VLDB 2010

S.Euijong, H.Garcia-Molina: **Entity Resolution with Evolving Rules**

- cél: ER eredmények frissítése, ha a szabályok változnak
  - inkrementális algoritmus a kívánatos
- szabályok: két rekord összehasonlításakor használt logika
  - Boolean match függvény / távolság függvény
- szabályok fejlődése:
  - szigorítás: kisebb probléma ; általános eset: ?
- megoldás:

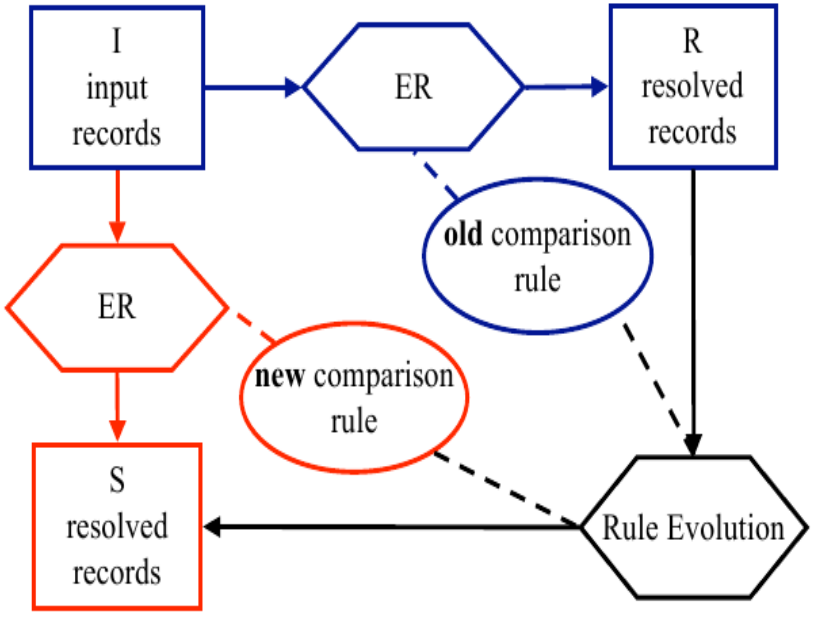
  1. algoritmus osztályok → hatékony szabály-változtató technikák

  2. részeredmények materializálása
     - párhuzam: materializált nézetek SQL lekérdezésekhez

Record	Name	Zip	Phone
r_1	*John*	54321	123-4567
r_2	*John*	54321	987-6543
r_3	*John*	11111	987-6543
r_4	*Bob*	*null*	121-1212

Fig. 1. Records to resolve

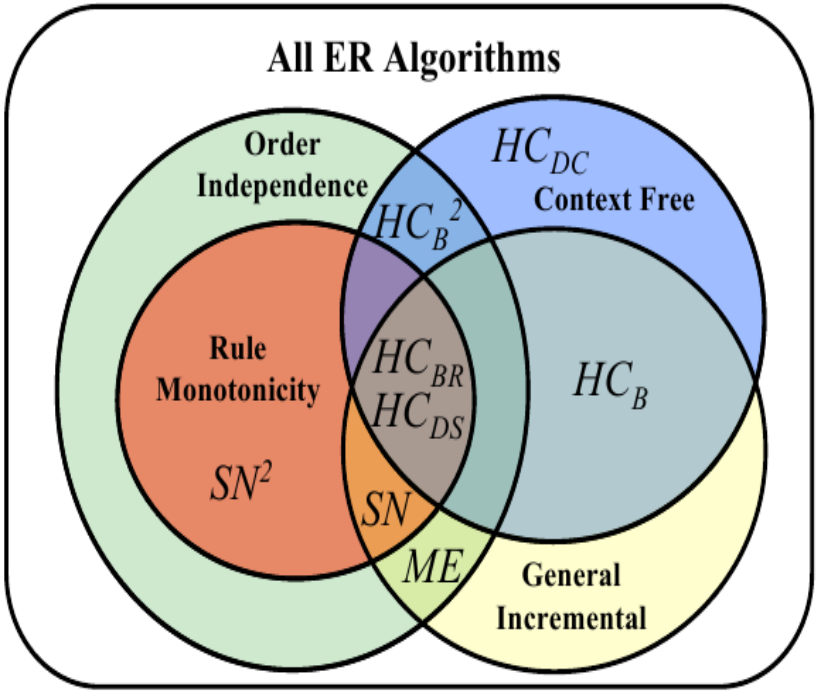
Comparison Rule	Definition
B_1	p_{name}
B_2	$p_{name} \wedge p_{zip}$
B_3	$p_{name} \wedge p_{phone}$



VLDB 2010

S.Euijong, H.Garcia-Molina: **Entity Resolution with Evolving Rules** (2)

- match szabály fejlesztés /hasonlóságra hasonló/:
  - B Boolean match fv., kommutatív
    - relatív szigorúság: B1 ≤ B2
  - P_i partícionálás→E alg.→$E(P_i, B)$ part.
  - partíció finomítás: $P_1 \leq P_2$ ha minden eleme részhalmaza a másik valamely elemének
- ER algoritmusok osztályai:
  - szabály monotonitás: szigorúbb szabály → finomabb partíció
  - környezetfüggetlenség: ha független (nem match-elő) halmazokat külön fel tud dolgozni
- szabály fejlesztő alg., RM + CF, RM esetére:
  - CNF részeihez eredményeket materializálunk → ezek 'találkozását' használjuk (a klaszterek metszeteit)



VLDB 2010

S.Euijong, H.Garcia-Molina: **Entity Resolution with Evolving Rules** **(3)**

- szabály alapú alg.:
  - sorted neighborhood
  - hierarchikus klaszterezés
  - Monge Elkan klaszterezés
- távolság alapú alg:
  - hierarchikus klaszterezés

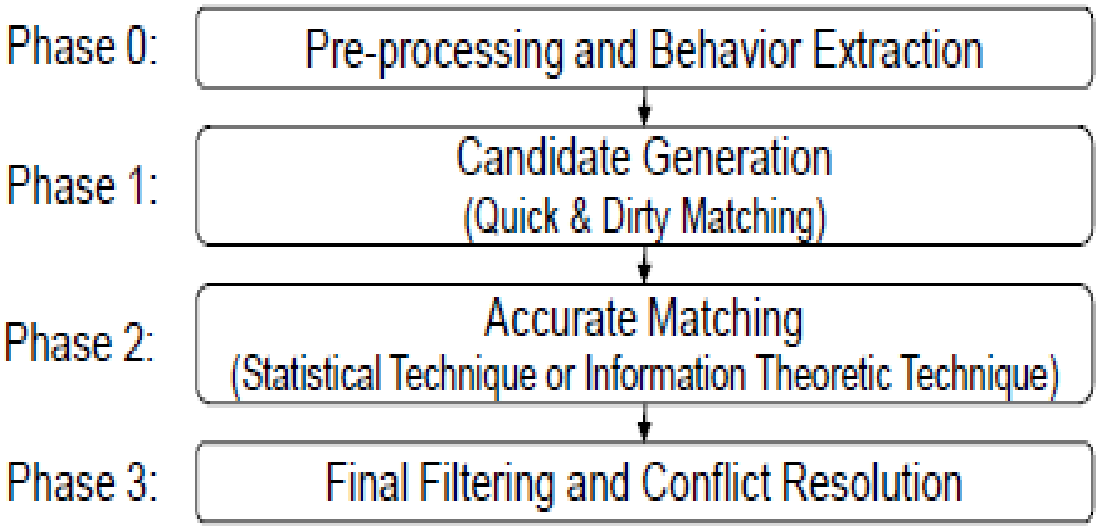
ER algorithm	Sho3K		Ho3K	
	Time O/H	Space O/H	Time O/H	Space O/H
SN	0.52 (0.02)	0.28	1.14 (0.27)	0.14
HC_B	0.87 (0.04)	0.14	3.18 (0.71)	0.1
HC_{BR}	11 (3E-6)	0.14	13.28 (1.06)	0.1
HC_{DS}	0.44	0.07	0.61	0.02

ER algorithm	Sh1K	Sh2K	Sh3K	Ho1K	Ho2K	Ho3K
ER algorithm runtime (seconds)						
SN	0.094	0.152	0.249	0.012	0.027	0.042
HC_B	1.85	7.59	17.43	0.386	2.317	5.933
HC_{BR}	3.56	19.37	48.72	0.322	1.632	4.264
HC_{DS}	8.33	40.38	111	5.482	27.96	73.59
Ratio of ER algorithm runtime to rule evolution runtime						
SN	4.09	4.22	4.45	1.2	1.93	2
HC_B	1.5	1.84	2.07	1.27	1.3	1.27
HC_{BR}	162	807	1218	36	136	237
HC_{DS}	298	708	918	322	499	545

VLDB 2010

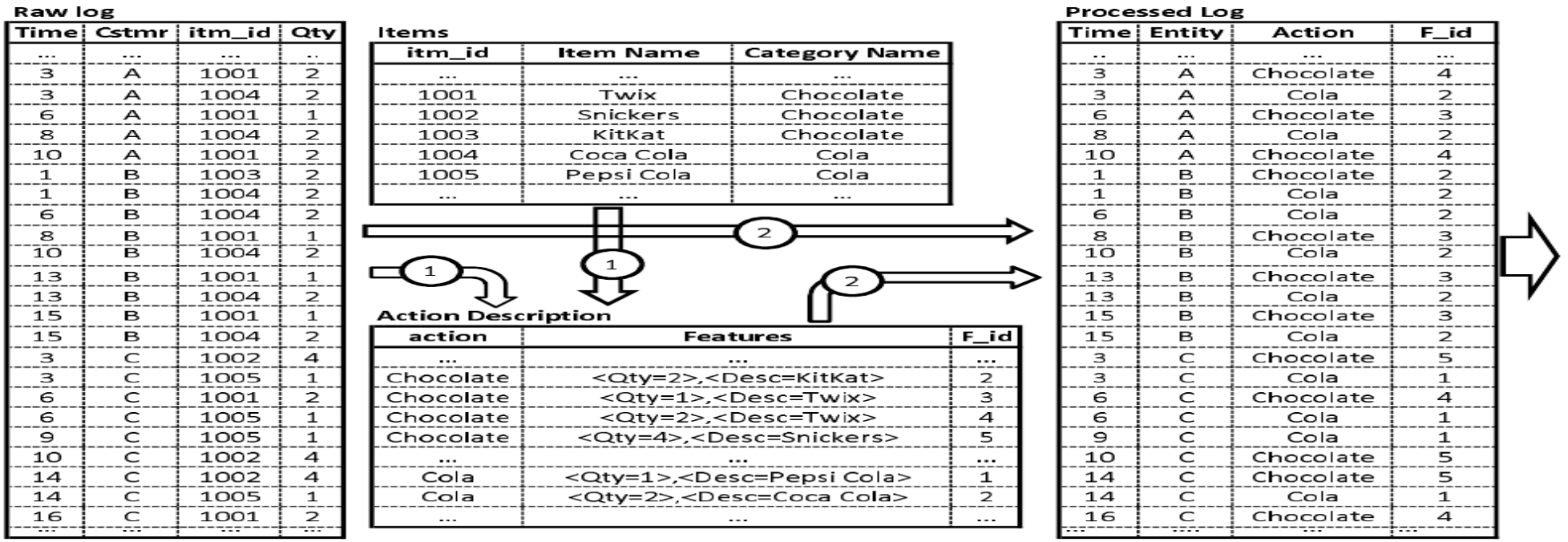
M.Yakout, A.K.Elmagarmid, H.Elmeleegy, M.Ouzzani, A.Qi: **Behavior Based Record Linkage**

- entitás viselkedése: tranzakciós log → események
  pl.: Yahoo – Maktoob felvásárlásnál a felhasználók
- viselkedési hasonlóság alapján: rosszul működik (kevéssé átfedő események)
- helyette: vonjuk össze páronként a viselkedési infót
  - ha “erősebb” mintákat kapunk → összevonhatóak
    - statisztikai modell: EM algoritmus
    - információ-elméleti modell: tömöríthetőség vizsgálata



VLDB 2010

M.Yakout, A.K.Elmagarmid, H.Elmeleegy, M.Ouzzani, A.Qi: **Behavior Based Record Linkage** **(2)**



Raw log

Time	Cstmr	itm_id	Qty
...	...	...	..
3	A	1001	2
3	A	1004	2
6	A	1001	1
8	A	1004	2
10	A	1001	2
1	B	1003	2
1	B	1004	2
6	B	1004	2
8	B	1001	1
10	B	1004	2
13	B	1001	1
13	B	1004	2
15	B	1001	1
15	B	1004	2
3	C	1002	4
3	C	1005	1
6	C	1001	2
6	C	1005	1
9	C	1005	1
10	C	1002	4
14	C	1002	4
14	C	1005	1
16	C	1001	2
...	...	...	...

Items

itm_id	Item Name	Category Name
...	...	...
1001	Twix	Chocolate
1002	Snickers	Chocolate
1003	KitKat	Chocolate
1004	Coca Cola	Cola
1005	Pepsi Cola	Cola
...	...	...

Action Description

action	Features	F_id
...	...	...
Chocolate	<Qty=2>,<Desc=KitKat>	2
Chocolate	<Qty=1>,<Desc=Twix>	3
Chocolate	<Qty=2>,<Desc=Twix>	4
Chocolate	<Qty=4>,<Desc=Snickers>	5
...	...	...
Cola	<Qty=1>,<Desc=Pepsi Cola>	1
Cola	<Qty=2>,<Desc=Coca Cola>	2
...	...	...

Processed Log

Time	Entity	Action	F_id
..	...	...	...
3	A	Chocolate	4
3	A	Cola	2
6	A	Chocolate	3
8	A	Cola	2
10	A	Chocolate	4
1	B	Chocolate	2
1	B	Cola	2
6	B	Cola	2
8	B	Chocolate	3
10	B	Cola	2
13	B	Chocolate	3
13	B	Cola	2
15	B	Chocolate	3
15	B	Cola	2
3	C	Chocolate	5
3	C	Cola	1
6	C	Chocolate	4
6	C	Cola	1
9	C	Cola	1
10	C	Chocolate	5
14	C	Chocolate	5
14	C	Cola	1
16	C	Chocolate	4
...	...	...	...

Time (Date)

		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
A	Chocolate	0	0	4	0	0	3	0	0	0	4	0	0	0	0	0	0
	Cola	0	0	2	0	0	0	0	2	0	0	0	0	0	0	0	0
B	Chocolate	3	0	0	0	0	0	0	3	0	0	0	0	3	0	3	0
	Cola	2	0	0	0	0	2	0	0	0	2	0	0	2	0	2	0
C	Chocolate	0	0	5	0	0	4	0	0	0	5	0	0	0	5	0	4
	Cola	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	1

viselkedési mátrix

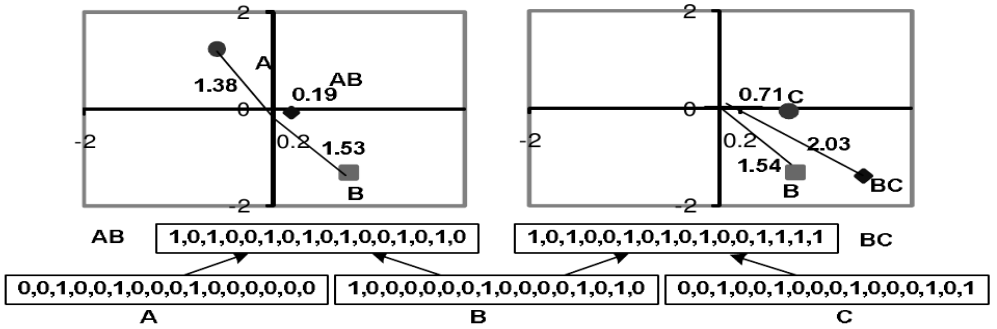
When merging A &B and then B &C

AB	Chocolate	3	0	4	0	0	3	0	3	0	4	0	0	3	0	3	0
	Cola	2	0	2	0	0	2	0	2	0	2	0	0	2	0	2	0
BC	Chocolate	3	0	5	0	0	4	0	3	0	5	0	0	3	5	3	4
	Cola	2	0	1	0	0	3	0	0	1	2	0	0	2	1	2	1

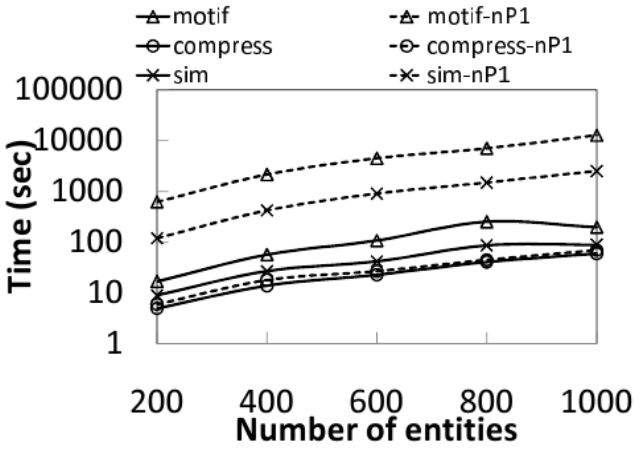
VLDB 2010

M.Yakout, A.K.Elmagarmid, H.Elmeleegy, M.Ouzzani, A.Qi: **Behavior Based Record Linkage** (3)

- viselkedés felismerési pontszám; célok:
  - konzisztens visszatérő események
  - feature jellemzők stabilitása: pl. 2 darab Twix-et vásárol mindig
  - cselekedetek közti asszociáció: pl. együtt vásárolt termékek
- jelöltállítás: két dimenzióba való leképezéssel, diszkrét Fourier transzf.; hasonló: Canopies



- kísérletek:
  - Wallmart adatok, művileg szétdobálva, sok RAM + MySQL
  - szöveges attribútumok hasonlóságának bevonása
- hasonló: webes ajánlások



VLDB 2010

S.Guo, X.L.Dong, D.Srivastava, R.Zajac: **Record Linkage with Uniqueness Constraints and Erroneous Values**

- hagyományosan két fázis:
  - record linkage: duplikátumok keresése
  - data fusion: duplikáltak összevonása, legjobb rekord előállítása
- problémák:
  - hibás rekordok megnehezítik az összevonást
  - egyediségi elvárások nem feltétlenül teljesülnek
  - lokális döntések globálisan rossznak bizonyulhatnak
- megoldás:
  - adatforrások, 'hard / soft uniqueness' megszorítások fogalmának bevezetése
  - a két lépés összevonása
  - hibás értékek megkeresése egyediségi megszorítások felhasználásával
  - visszavezetés: k-partite graph clustering problem

Soft Computing in XML Data Management (Springer, 2009)

An Overview of XML Duplicate Detection Algorithms

- algoritmusok: DogmatiX framework, XMLDup, SXNM

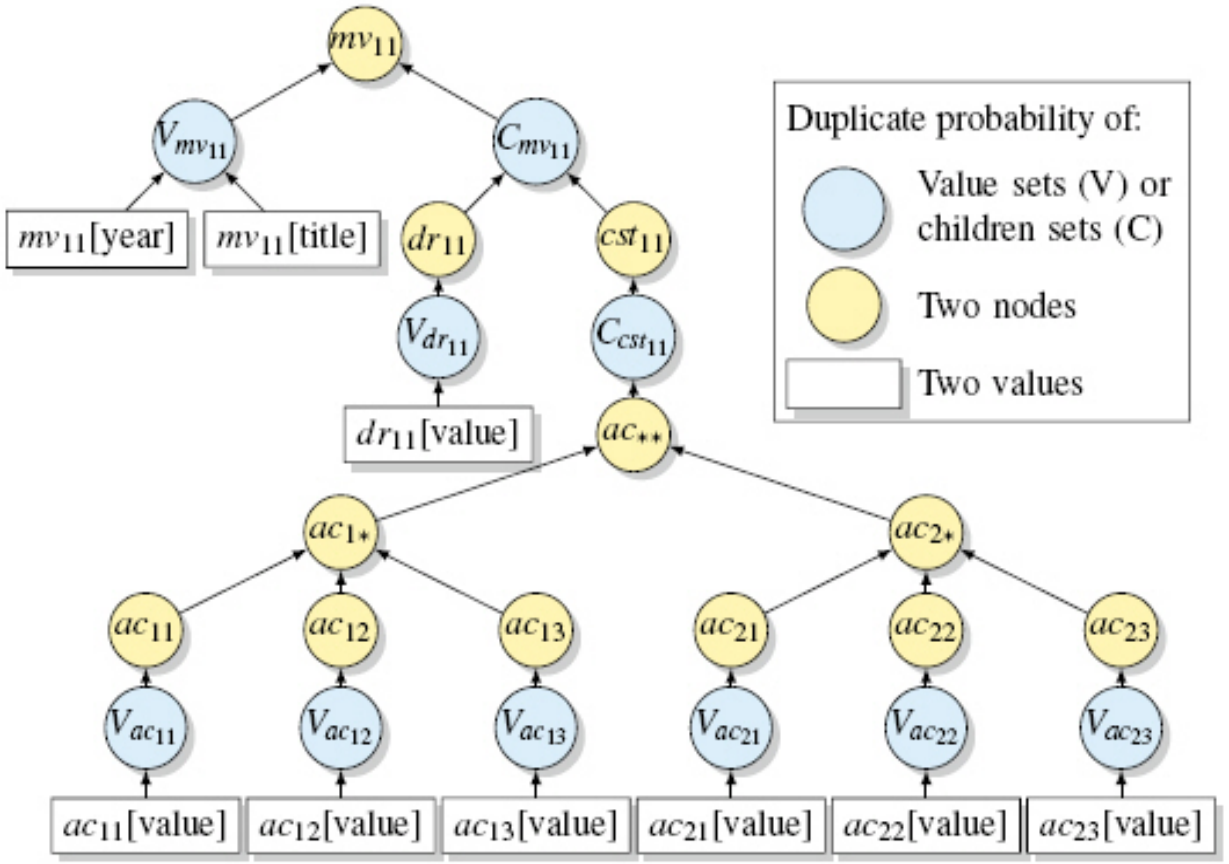


Fig. 7 Bayesian network to compute the similarity of the trees in Fig. 3, as shown in [27]

forrás: Molnár Miklós blogja,
http://liftinstinct.blogspot.com/2010/09/soft-computing-in-xml-data-management.html

Kísérleti adathalmazok

szabadon elérhető:

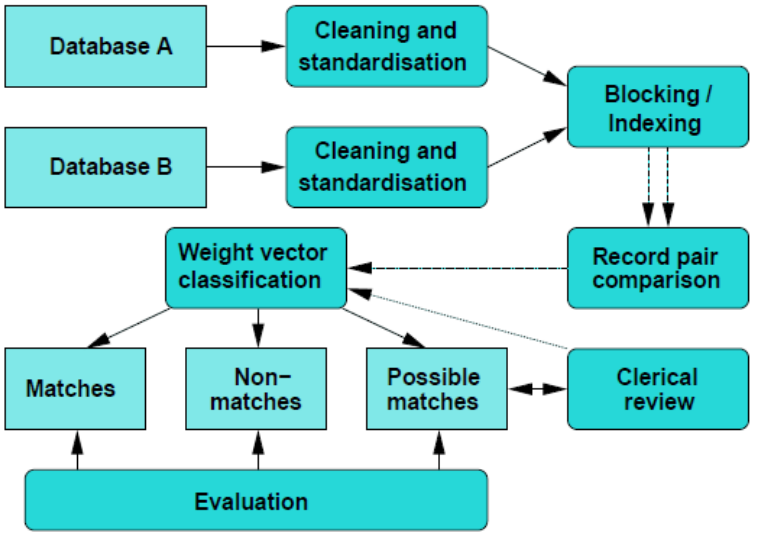
- CiteSeer, DBLP, Google Scholar, BioBase, ACM, Riddle db, ...: biográfiai adatbázisok
  - Scholar példa: lekérések cím és rendezvény szerint

zárt:

- biztosító, bank: ügyféltörzs
- összehasonlító oldalak termékadatai: Yahoo! Shopping
- hotel adathalmaz: Yahoo! Travel (gyűjtőoldal)
- yellowpages.com címadatok
- Walmart vásárlói és vásárlói aktivitás adat

Elérhető eszközök

- kutatási vagy free, open source:
  - FEBRL: Free Extensible Biomedical Recod Linkage; ausztrál



  - SERF: Stanford University, Hector Garcia-Molina – R-Swoosh implementáció,
    MTB: Duisburg
    DDUpe: Maryland
    MARLIN
  - String join algoritmusok (PPJoin+), klaszterező algoritmusok stb.
- fizetős: léteznek (Daurum, Infosolve OpenDQ, …); minőségük: ?

Data & Knowledge Engineering, 2010

Hanna Köpcke, Erhard Rahm: **Frameworks for entity matching: A comparison**

	BN [38]	MOMA [55]	SERF [5]	Active Atlas [53,54]	MARLIN [11,12]	Multiple Classifier System [62]	Operator Trees [13]	TAILOR [24]	FEBRL [18,17]	STEM [36]	Context Based Framework [16]
Training-based				Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Entity type	XML	Relational	Relational	Relational	Relational	Relational	Relational	Relational	Relational	Relational	Relational
Blocking	-	-	-								
Key definition				Manual	Manual	Manual	Manual	Manual	Manual	Manual	Manual
Partitioning											
Disjoint						?		Sorting, hashing	Sorting	Threshold	
Overlapping				Hashing	Canopy clustering	?	Canopy clustering	Sorted neighborhood	Sorted neighborhood, q-gram, canopy clustering	Sorted neighborhood	Canopy-like
Matchers	Attribute value, context	Attribute value, context	Attribute value	Attribute value	Attribute value		Attribute value	Attribute value	Attribute value	Attribute value	Attribute value, context
Matcher combination	Numerical	Workflow	Rules	Rules	Numerical, rules	Numerical, rules	Rules	Numerical, rules	Numerical	Numerical, rules	Numerical, rules
Learners				Decision tree	SVM, decision tree	7 (SVM, decision tree, etc.)	Operator Tree algorithm, SVM	Probabilistic, decision tree	SVM	SVM, decision tree, logistic regression, multiple learning	Diverse (SMOreg, Logistic, etc.)
Training selection				Manual, semi-automatic	Manual, semi-automatic	Manual	Manual	Manual	Manual, automatic	Manual, semi-automatic	Manual

Trendek, nyitott kérdések

- egyre szélesebb körű igények
- osztott rendszerek, jól skálázódó megoldások: hogyan?
- egyéb gyorsítási módszerek: blocking, iterative blocking, … ?
- duplikátum-azonosítás mint keresési probléma: indexelés
- viselkedés alapú azonosság-feloldás: tranzakciós logok
- időben változó adatok, időben változó szabályok, időben változó entitások
- felügyelt tanulás, „active learning”: hibák iránya fontos
- valószínűségi modellek vs. egzakt eredmények
- eredmény minősége: hogyan mérjük?
- entitások hierarchikus egymásra épülése
- kapcsolat alapú azonosság-feloldás, gráf-klaszterezés
- keretrendszerek, dobozos termékek: mennyire használhatóak? → iparági tudás beépítése

Főbb források

- Ivan P. Fellegi, Alan B. Sunter: **A Theory for Record Linkage**, Journal of the American Statistical Association, 1969
- Ahmed K. Elmagarmid and Panagiotis G. Ipeirotis and Vassilios S. Verykios: **Duplicate record detection: A survey**, IEEE TKDE, 2007
- Benjelloun, Omar and Garcia-Molina, Hector and Menestrina, David and Su, Qi and Whang, Steven Euijong and Widom, Jennifer (2005): **Swoosh: A Generic Approach to Entity Resolution**. Technical Report. Stanford. → 2009: VLDB Journal
- Bhattacharya, Indrajit and Getoor, Lise: **Collective entity resolution in relational data**. ACM Trans. Knowl. Discov. Data, 2007
- S. E. Whang, D. Menestrina, G. Koutrika, M. Theobald, H. Garcia-Molina: **Entity Resolution with Iterative Blocking,** SIGMOD 2009
- M.Yakout, A.K.Elmagarmid, H.Elmeleegy, M.Ouzzani, A.Qi: **Behavior Based Record Linkage**, VLDB 2010
- jó kiindulási pont: Stanford Entity Resolution Framework, http://infolab.stanford.edu/serf/
- XML könyvfejezet magyar bemutatása: http://liftinstinct.blogspot.com/2010/09/soft-computing-in-xml-data-management.html