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"

MIBE konferencia, 2011.06.16.

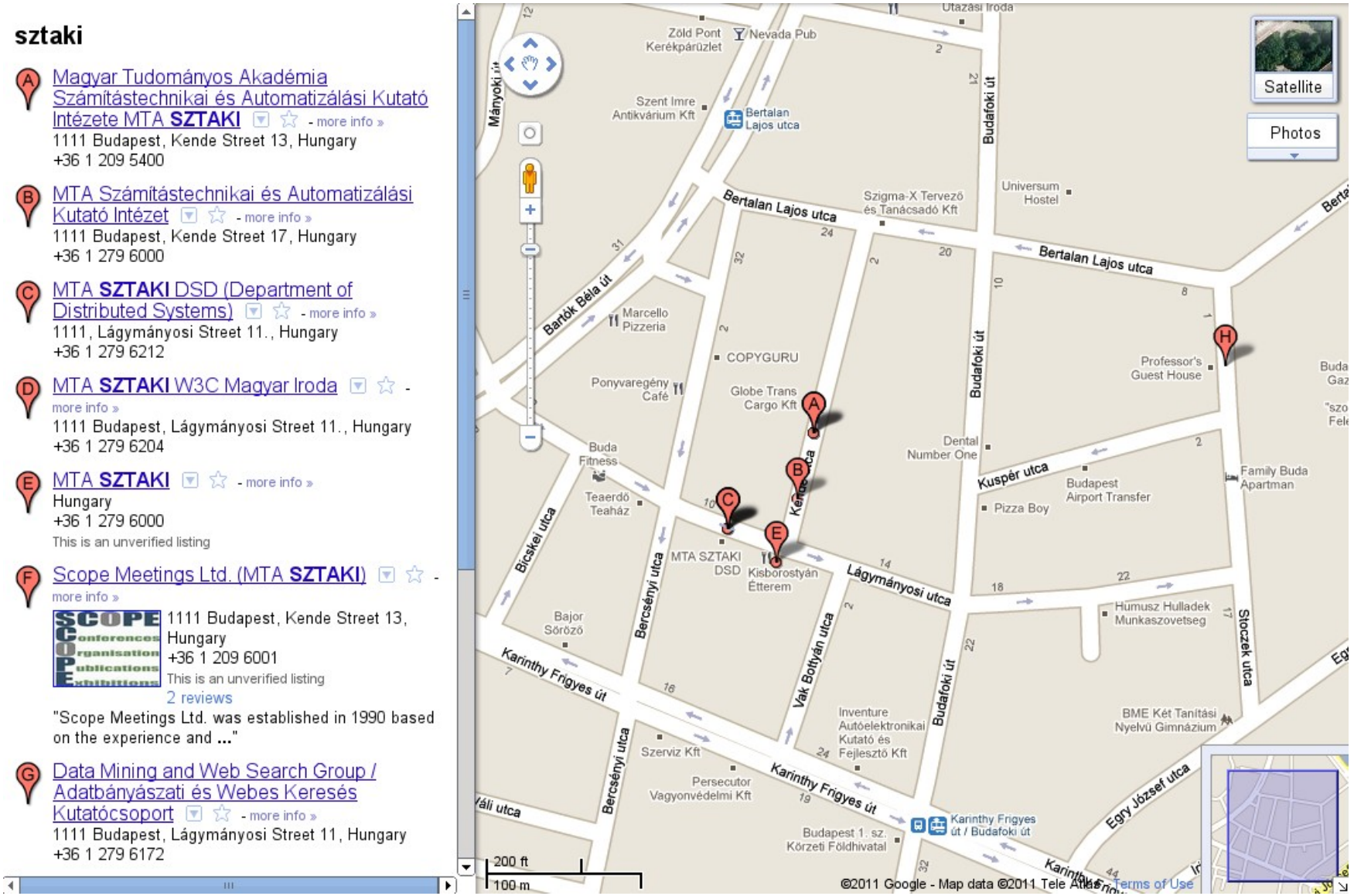
Sidló Csaba - sidlo@ilab.sztaki.hu

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

MTA SZTAKI

datamining search

Példa: Google Places

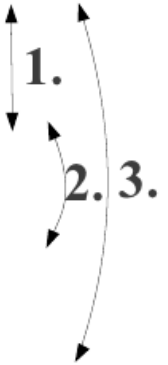


példa forrása: Google Maps keresés: „MTA SZTAKI”

Példa: ügyfelek

gyakori modell: előállítandó rekordok egy particionálása, csoportosítása

név	e-mail	ID
Kovács Mária	marcsi@mail-1.com	50071
Nagy Istvánné	nagyne.marcsi@mail-2.hu	50071
Nagy Istvánné K. Mária	nagyne.marcsi@mail-2.hu	79216
Kovács M.	marcsi@mail-1.com	34302



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), változó attribútumok (név, cím, telefonszám stb.)

Példa: 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? ...**
- felhasználás:
  - CRM, CDI (Customer Data Integration)
  - marketing (kampányokhoz: vásárolt címadatok),
  - döntéstámogatás, adattárházak, adatintegráció,
  - törzsadat építés (MDM: Master Data Management), ...
- bonyolultabb igények:
  - háztartások azonosítása
  - kapcsolatok felhasználása: házastárs, telefonhívás, emailküldés, szerződő, kedvezményezett …
  - hasonlóság felhasználása: névhasonlóság, címhasonlóság stb.
  - tanuló rendszerek: szabályok, minták felismerése, ...

További alkalmazások

- klasszikus feladat: publikációs adatbázisok
  - kevés attribútum: írók nevei, esetleg munkahely
  - *kapcsolatok* entitások közt: közösen írt cikkek
- ügyfelek:
  - pl: Yahoo közel-keleti felvásárlás: **mennyi az új felhasználók száma valójában** – mennyit érdemes költeni?
- web:
  - weboldalak ('mirror detection'); entitások weboldalakon: személyek, dátumok, helyek stb.; **hány Facebook / IWIW felhasználó van igazából? mutasd a keresett személy összes regisztrációját!**
  - termék-keresők termékei; **mutasd az összes ajánlatot a keresett termékre**
- ...

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

Az azonosságfeloldás témaköre

- elnevezések (eltérő elnevezések → eltérő megközelítések):
  - **record linkage** (1960)**,** duplicate detection, duplicate record detection, merge/purge; deduplikáció („dedup”), duplikátum-keresés
  - **entity resolution** → “azonosságfeloldás”
  - instance identification, reference reconciliation, coreference resolution, database hardening, …

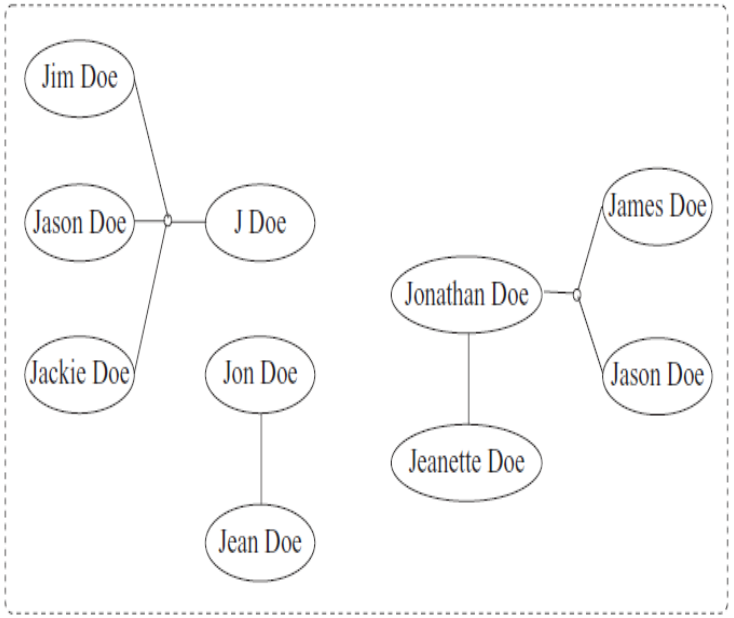
- kapcsolódó területek:

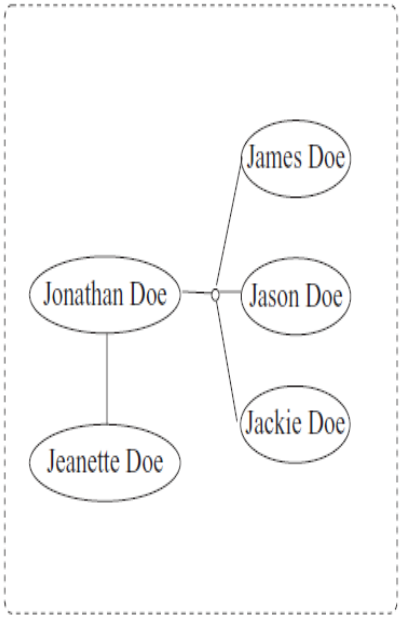
  klaszterezés (adatbányászat), similarity join, string hasonlóságok, adatminőség, adattisztítás, adattárházak, adatintegráció, információ integráció, …

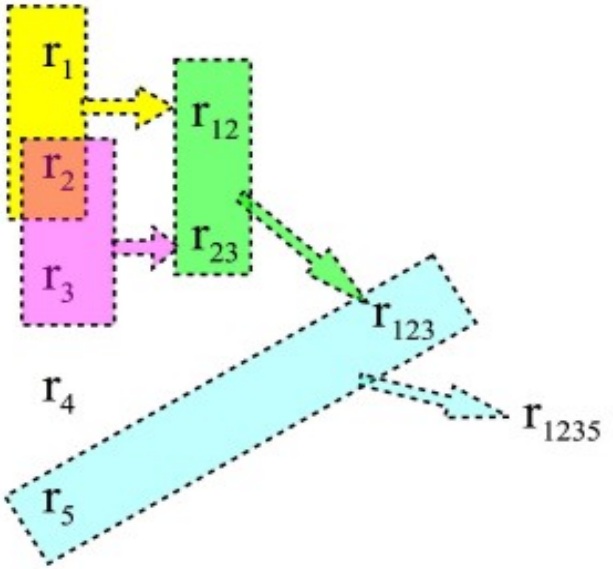
- aktív kutatási terület, példa: Very Large Databases konferencia, 2010:
  - nagyságrendileg 80-90 cikkből
    - ~5 db entity resolution cikk,
    - ~10-15 szorosan kapcsolódó cikk

Az azonosságfeloldás feladatának megfogalmazása

- modell: rekordok halmaza / fa: XML (szemantikus web, ontológiák) / gráf
- match-merge → összevonás, reprezentatív elemmel / klaszterezés: csoportosítás, partíciókkal
- közelítő (valószínűségi, fuzzy) / egzakt (szabályalapú) megoldások
- felügyelt tanulás (tanító halmaz) / nem felügyelt tanulás

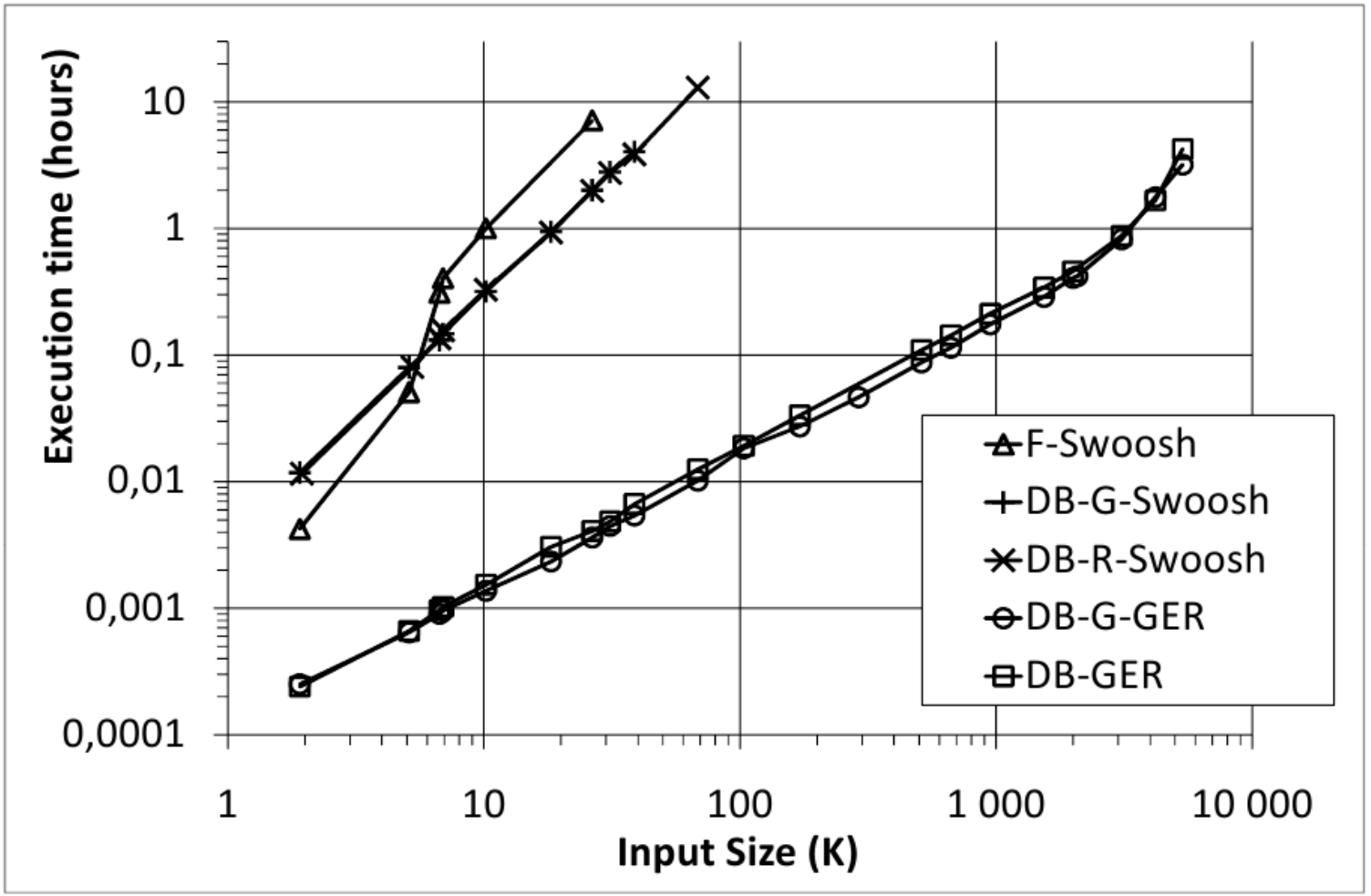






A fő nehézség: számítási bonyolultság

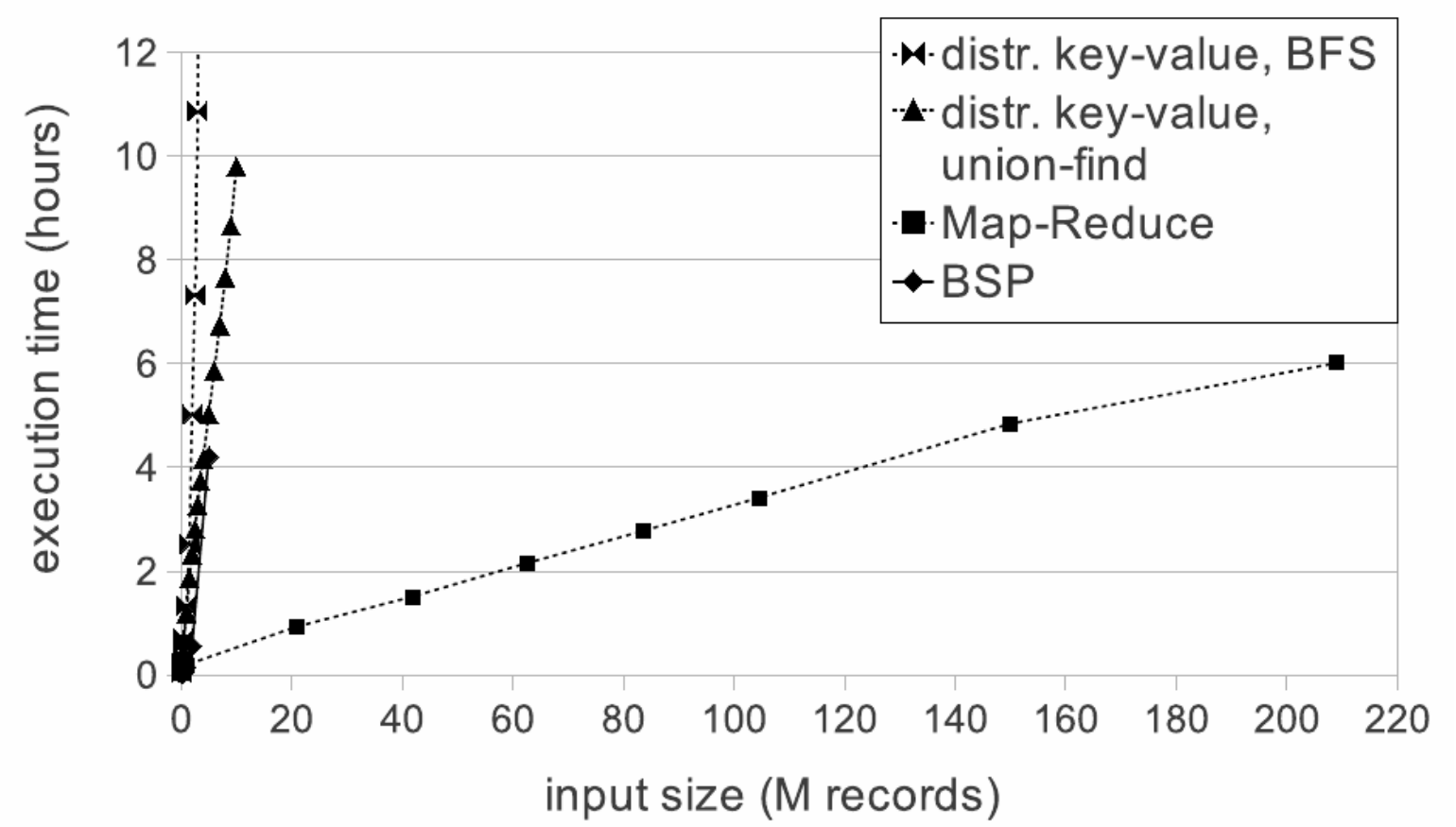
futásidő példa ügyfél adatbázison, ~15 M rekord, egy átlagosnál kicsit jobb asztali gépen, többféle hatékony algoritmus (saját implementációk):



forrás: Cs. I. Sidló: Generic Entity Resolution in Relational Databases, ADBIS 2009

Egy megoldás: elosztott algoritmusok

futásidő példa ügyfél adatbázison, ~200 M rekord, gyengébb 15 gép, különböző elosztott algoritmusok



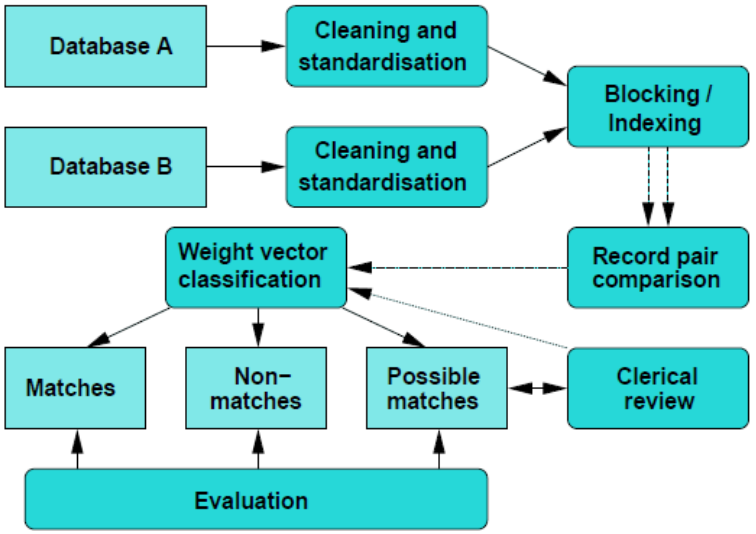
saját osztott algoritmus változatok, ügyfél adatokon

További nehézség: bonyolult logika

- hasonlóságok:
  - karakterlánc hasonlóságok: szerkesztési, kiejtés szerinti stb.,
  - dokumentumok hasonlósága,
  - képek hasonlósága, stb.
- átfedő csoportok, nem egyértelmű csoportosítások (pl. találjuk meg a háztartásokat):
  - fuzzy csoportosítás (egyszerre több entitáshoz tartozás)
  - valószínűségi modellek (bizonyos valószínűség melletti összetartozás)
- gépi tanulás: következtessünk ismert (pl. kézzel összerendelt) entitásokból
- minőség mérése: nehéz összehasonlítási alaphoz jutni
- események, tranzakciók felhasználása: aktivitás alapú azonosságfeloldás
- hibák elkülönítése: hibás összevonás büntetése

Elérhető eszközök

- egyedi megoldások:
  adatbázis lekérdezések, saját heurisztikák, egyedi algoritmusok
- kutatási vagy free, open source:
  - FEBRL: ausztrál
    SERF: Stanford University
    MTB: Duisburg
    DDUpe: Maryland
    MARLIN
  - String join algoritmusok (PPJoin+), klaszterező algoritmusok stb.



- kereskedelmi: különböző megközelítések, csomagok, környezetek, teljesítmény, ár
  - Infosolve OpenDQ, IBM QualityStage, Eobjects DataCleaner, Sparsity Technologies Daurum, Infoglide Identity Resolution, The Link King (SAS-hoz), ...

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