CIKM 2010

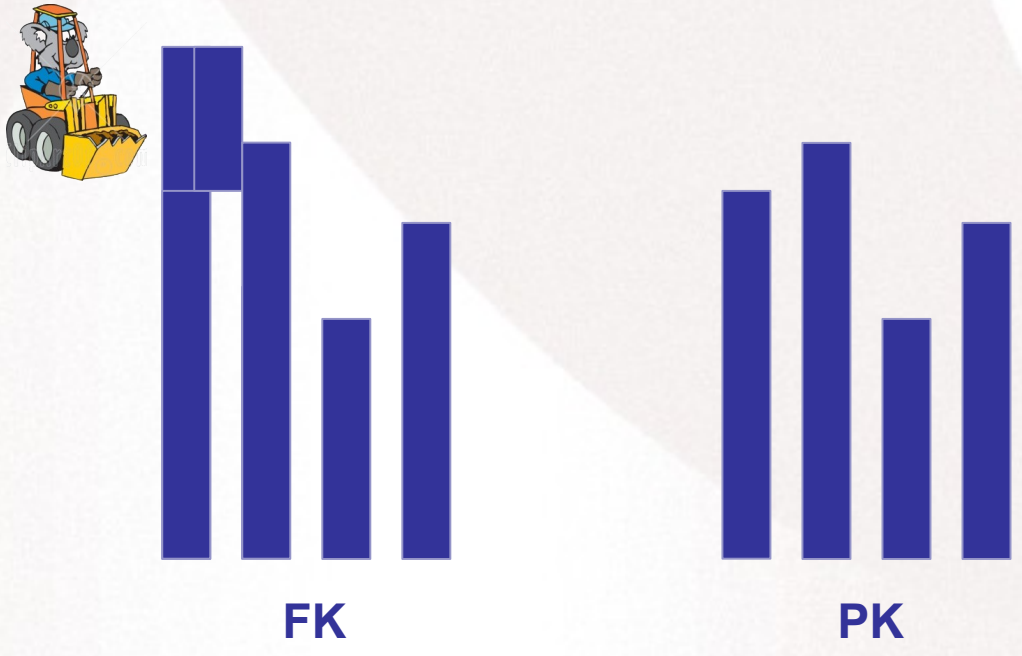
19th ACM International Conference on Information and Knowledge Management

Toronto, Kanada, 26.-30. října, 2010

Tomáš Kliegr, 18. října, KEG 2010

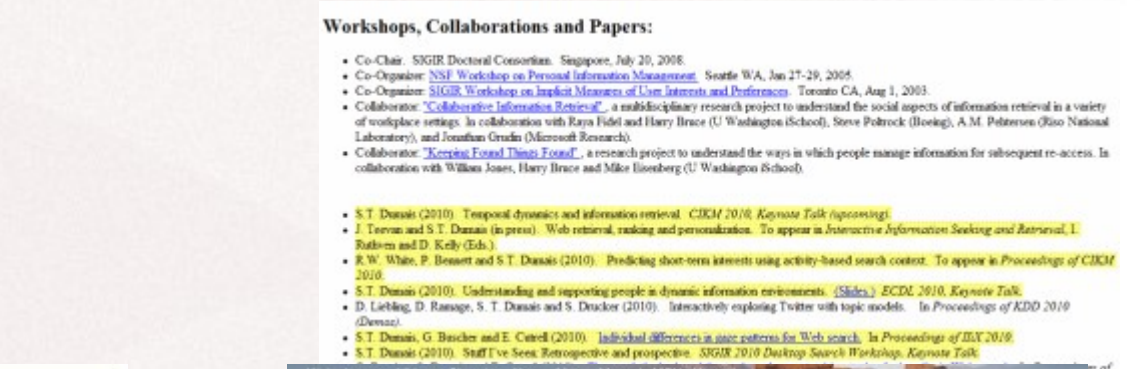
Schema Extraction
Divesh Srivastava

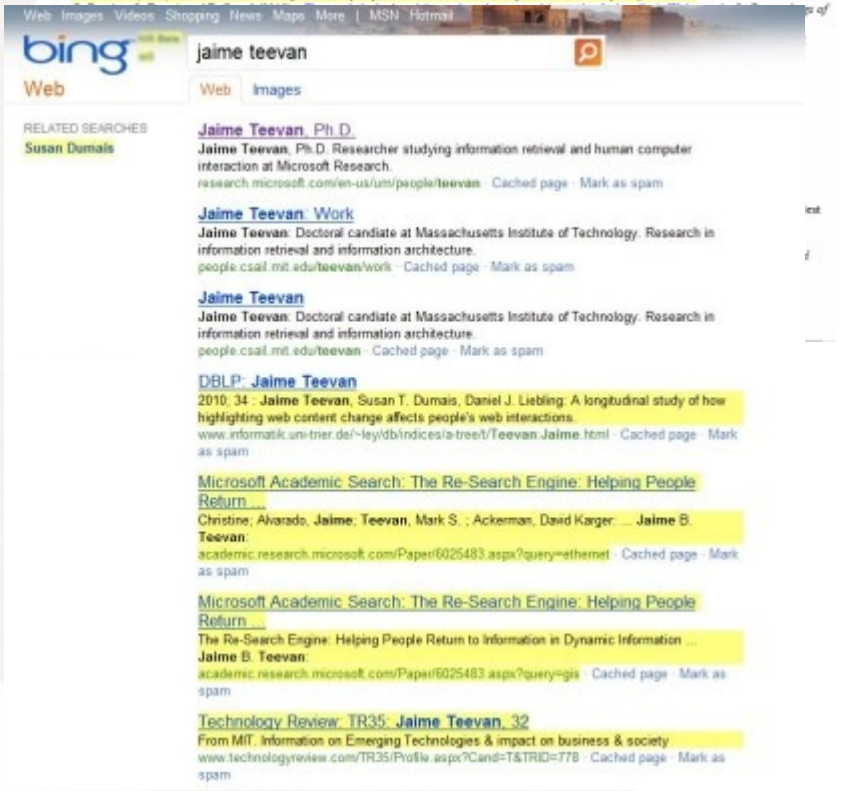
- Discover foreign/primary-key relationships between tables
- Velmi kladné zkušenosti s metrikou Earth Mover Distance



Temporal Dynamics and IR Susan Dumais

- Vyčerpávající statistiky týkající se změn na webových stránkách
- Shrnuty na http://www.searchenginecaffe.com/2010/11/susan-dumais-cikm-2010-keynote-temporal.html
- Browser plugin pro zvýraznění změn:
- http://research.microsoft.com/en-us/projects/diffie/default.aspx





http://research.microsoft.com/en-us/projects/diffie/examples.aspx

Regulární výrazy

- Převod textu (recenzí) do sémantických dimenzí pomocí regulárních výrazů
  - Urbanizer.com
- Určení kompatibility XML schémat pomocí regulárních výrazů (XML výraz je převeden na regulární výraz)
  - Schema Minimalization
  - Two Sas are equivallent if their minimized forms are isomorhpic
  - XML Schema Computations: Schema Compatibility Testing and Subschema Extraction

Tutorial – Applied TextMining Ronen Feldman

- TextRunner
  http://www.cs.washington.edu/research/tex
- KnowItAll

  The input to KnowItAll is a set of entity classes to be extracted, such as “city”, “scientist”,“movie”, etc., and the output is a list of entities extracted from the Web.

  JAPE-like patterns: NP2 "was acquired by" NP1
- SRES - (Self-Supervised Relation Extraction

www.cis.upenn.edu/~ungar/KDD/KDD_tutorial.pdf

SRES

Vstup: pozitivní a negativní příklady

- Acquisition: PeopleSoft – Oracle, Sun – Oracle,...
- Toward this end, <Arg1> in July acquired <Arg2>
- “Earlier this year, <Arg1> acquired <Arg2>”
- <ARG1> * acquired <ARG2>

  - use linguistic annotations to replace non-important word with a star

  – Každý vzor musí obsahovat alespoň jedno slovo blízké predikátu (zde acquire), to se zjišťuje z Wordnetu

-

Vyhodnocení

- Pravidlový přístup (KnowItAll)
  - Vysoká redundance
  - Jednoduché tvary
- Učení (SRES)
  - Instance s nízkou frekvencí
  - Expresivnější reprezentace pravidel
  -

The Shift from Search to Matching Vasco Pedro

Original Tags

Japan Tokyo Kumamoto business subway

Clean Tags

Japan Tokyo Kumamoto Prefecture Business Subway

Content Categories

places-and-geography

Semantic Tags

Japan
Country Location Tag Military Combatant

Subway
Company Restaurant Employer Juggle

Business
Domain Domain Profile

Kumamoto Prefecture
Location Administrative Division Japanese prefecture Dated location

Tokyo
City/Town Location Administrative Division Top Architectural City

Original Tags

Japan Tokyo Kumamoto chopsticks

Clean Tags

Japan Tokyo Kumamoto Prefecture Chopsticks

Semantic Tags

Japan
Country Location Tag Military Combatant

Chopsticks
Ontology Instance

Kumamoto Prefecture
Location Administrative Division Japanese prefecture Dated location

Tokyo
City/Town Location Administrative Division Top Architectural City

Wikipedia

- TAGME: on-the-fly annotation of short text fragments by Wikipedia entities
  - http://tagme.di.unipi.it/
- Extracting Structured Information from Wikipedia Articles to populated Infoboxes
- Efficient Wikipedia-Based Semantic Interpreted by Exploiting Top-k processing
- Elusive Vandalis Detection at Wikipedia: A text stability approach Wikipedia as

QQ

- 600 mil uživatelů, 3. nějvětší “.com” společnost (Tencent)
- 100 MLD USD obrat
  - 77 % z prodeje virtuálního zboží
  - Většina stahuje hry a sport
- 22% VŠ
- Průměrný denní login time 94 minut FB, 10 minut QQ
- “Nadprůměrná úroveň cenzury” Gordon Sun, Chief Scientist, Tencent