

# ***CIKM 2010***

19<sup>th</sup> ACM International Conference on  
Information and Knowledge Management

Toronto, Kanada, 26.-30. října, 2010

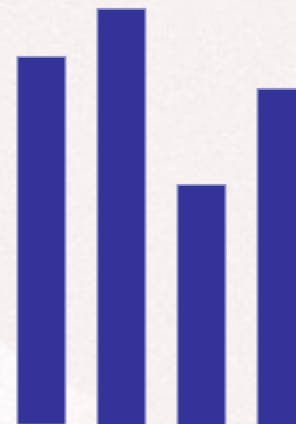
# Schema Extraction

## Divesh Srivastava

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



FK



PK

# *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>

» Personal workstations

» Affordable power

» Compact power

» Extreme power



Enhanced performance in an affordable package.

» HP Z400 Workstation **NEW!**  
Starting at: \$ 929.00\*  
As low as \$27/mo.\*\*

- Up to 16 GB of system memory
- Up to 4.5 TB of internal storage
- Up to NVIDIA Quadro FX4800 or dual NVIDIA Qu graphics

» HP xw4600 Work  
Starting at: \$ 679  
As low as \$21



Eight core performance in a compact footprint.

» HP Z600 Workstation **NEW!**  
Starting at: \$ 1,589.00\*  
As low as \$46/mo.\*\*

- Up to 24 GB of system memory
- Up to 4.5 TB of internal storage
- Up to eight 2D displays



Ultimate performance with extreme expandability.

» HP Z800 Workstation **NEW!**  
Starting at: \$ 1,839.00\*  
As low as \$53/mo.\*\*

- Up to 192 GB of system memory
- Up to 7.5 TB of internal storage

Eastside Mothers & More Message Board | Chapter News and Events | ... | Microsoft Library

Pages: [1] 2 3 ... 7

| Subject   | Views | Last post  |
|---|-------|--|
| 2009 Book Club selections   | 5     | 462 July 14, 2009, 10:54:29 PM by Cathy Smith          |
| 2010 Book Club Sugg...  | 8     | 78 July 12, 2009, 05:49:19 PM by Meredith Schwietzer   |
| Five Reasons to Jo...   | 8     | 78 July 12, 2009, 05:49:19 PM by Meredith Schwietzer   |
| 7/22 Book Club: ...   | 3     | 215 February 28, 2007, 10:37:16 AM by Sarah Mangold    |
| The Guernsey Lit Society <b>new</b>                                     | 3     | 215 February 28, 2007, 10:37:16 AM by Sarah Mangold    |
| 6/17 Book Club  | 12    | 112 <b>Today at 07:26:10 PM</b> by Cathy Smith         |
| The Girls from A...   | 12    | 112 <b>Today at 07:26:10 PM</b> by Cathy Smith         |
| Anita Blake series  | 6     | 70 July 13, 2009, 01:50:00 PM by Gabrielle Nonast      |
| Gatsby  | 6     | 70 July 13, 2009, 01:50:00 PM by Gabrielle Nonast      |
| 5/20 Book Club: Self-M...   | 41    | 404 June 18, 2009, 09:43:29 PM by Michelle Jek...      |
| The Glass Castle: Interview   | 41    | 404 June 18, 2009, 09:43:29 PM by Michelle Jek...      |
| April 22 7PM: Merle's Door: Lesson: Freethinking Dog = 1 2 3 4 =        | 49    | 563 April 19, 2009, 02:08:02 PM by Kristi Ray          |
| April 30: The Beautiful Things that Heaven Bears <b>new</b> = 1 2 3 4 = | 36    | 507 April 16, 2009, 07:53:30 AM by Trina Sooy          |
| Favorite Children's Books? <b>new</b> = 1 2 3 =                         | 72    | 685 March 20, 2009, 11:46:23 AM by Jaime Teevan        |
| 3/18 Book Club: Never Let Me Go = 1 2 3 4 5 =                           | 70    | 890 March 14, 2009, 01:59:50 PM by Katie White         |
| January 30, 2008 Book Club: Eat, Pray, Love = 1 2 3 4 5 =               | 0     | 29 March 14, 2009, 10:46:49 AM by Della Dawn Fallah    |
| Elizabeth Gaskell   | 0     | 29 March 14, 2009, 10:46:49 AM by Della Dawn Fallah    |
| 2/25 Book Club: Extremely Loud & Incredibly Close = 1 2 3 4 5 =         | 69    | 763 March 02, 2009, 04:49:36 PM by Meredith Schwietzer |

Workshops, Collaborations and Papers:

- Co-Chair: SIGIR Doctoral Consortium, Singapore, July 20, 2008.
- Co-Organizer: NSF Workshop on Personal Information Management, Seattle WA, Jan 27-29, 2005.
- Co-Organizer: SIGIR Workshop on Empirical Measures of User Interests and Preferences, Toronto CA, Aug 1, 2003.
- Collaborator: "Collaborative Information Retrieval", a multidisciplinary research project to understand the social aspects of information retrieval in a variety of workplace settings. In collaboration with Raya Finkel and Harry Bruce (U Washington School), Steve Patrock (Boeing), A.M. Polsteren (Rao National Laboratory), and Jonathan Grudin (Microsoft Research).
- Collaborator: "Keeping Forward Things Forward", a research project to understand the ways in which people manage information for subsequent re-access. In collaboration with William Jones, Harry Bruce and Mike Eisenberg (U Washington School).
- S.T. Dumais (2010). Temporal dynamics and information retrieval. *CIKM 2010, Keynote Talk (upcoming)*.
- J. Teevan and S.T. Dumais (in press). Web retrieval, ranking and personalization. To appear in *Interactive Information Seeking and Retrieval*, L. Rasmussen and D. Kelly (Eds.).
- R.W. White, P. Bennett and S.T. Dumais (2010). Predicting short-term interests using activity-based search context. To appear in *Proceedings of CIKM 2010*.
- S.T. Dumais (2010). Understanding and supporting people in dynamic information environments. (Slides.) *ECDL 2010, Keynote Talk*.
- D. Liebling, D. Ramage, S. T. Dumais and S. Drucker (2010). Interactively exploring Twitter with topic models. In *Proceedings of KDD 2010 (Demos)*.
- S.T. Dumais, G. Bascher and E. Catrol (2010). Individual differences in user patterns for Web search. In *Proceedings of WWW 2010*.
- S.T. Dumais (2010). Staff I've Seen: Retrospective and prospective. *SIGIR 2010 Desktop Search Workshop, Keynote Talk*.

Web Images Videos Shopping News Maps More | MSN HTML

bing

jaime teevan

Web Images

RELATED SEARCHES  
Susan Dumais

**Jaime Teevan, Ph.D.**  
Jaime Teevan, Ph.D. Researcher studying information retrieval and human computer interaction at Microsoft Research.  
research.microsoft.com/en-us/um/people/teevan - Cached page - Mark as spam

**Jaime Teevan, Work**  
Jaime Teevan: Doctoral candidate at Massachusetts Institute of Technology. Research in information retrieval and information architecture.  
people.csail.mit.edu/teevan/work - Cached page - Mark as spam

**Jaime Teevan**  
Jaime Teevan: Doctoral candidate at Massachusetts Institute of Technology. Research in information retrieval and information architecture.  
people.csail.mit.edu/teevan - Cached page - Mark as spam

**DBLP: Jaime Teevan**  
2010: 34 - Jaime Teevan, Susan T. Dumais, Daniel J. Liebling: A longitudinal study of how highlighting web content change affects people's web interactions.  
www.informatik.uni-trier.de/~ley/db/indicies/a/tree/Teevan-Jaime.html - Cached page - Mark as spam

**Microsoft Academic Search: The Re-Search Engine: Helping People Return**  
Christine Alvarado, Jaime Teevan, Mark S. Ackerman, David Karger. ... Jaime B. Teevan.  
academic.research.microsoft.com/Paper/6025483.aspx?query=ethernet - Cached page - Mark as spam

**Microsoft Academic Search: The Re-Search Engine: Helping People Return**  
The Re-Search Engine: Helping People Return to Information in Dynamic Information ... Jaime B. Teevan.  
academic.research.microsoft.com/Paper/6025483.aspx?query=gs - Cached page - Mark as spam

**Technology Review TR35: Jaime Teevan 32**  
From MIT, information on Emerging Technologies & impact on business & society  
www.technologyreview.com/TR35/Profile.aspx?cand=T&TRID=778 - Cached page - Mark as spam

# *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 equivalent if their minimized forms are isomorphic
  - XML Schema Computations: Schema Compatibility Testing and Subschema Extraction

# *Tutorial – Applied TextMining Ronen Feldman*

- TextRunner

<http://www.cs.washington.edu/research/text>

- 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)

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

- 600 mil uživatelů, 3. největší “.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