



# O pen Source Text Data

&

# how can we work with it

Robert McArthur

CSIRO ICT

***“Federal Attorney-General Philip Ruddock has rejected criticism about Dr Haneef's detention and says that with 31,000 documents to consider, police need time to cover every base.”***



## Interested/motivated humans are good

**A time-honored Washington practice of trying to extinguish, preempt, or redirect news coverage by dumping stacks of previously secret government documents on the press may be in for some changes after a headlong collision with hundreds of liberal Web loggers in the wee hours of yesterday morning.**

**On Monday night, the Justice Department delivered to Congress more than 3,000 pages of e-mails, memos, and other records about the firing of eight U.S. attorneys. The handover came so late that many news organizations had to scramble to try to skim a few headlines from the files before late night deadlines.**

**Despite the late hour, readers of a liberal Web site, XXX tackled the task with gusto. They quickly began grabbing 50-page chunks of the scanned documents from a House of Representatives Internet server, analyzing them and excerpting them. The first post about the Department of Justice records hit the left-leaning news and commentary site at 1:04 a.m. Within half an hour, there were 50 summaries posted by readers gleaning the documents. By 4:30 a.m., more than 220 postings were up detailing various aspects of the files.**



**We've already started scouring newly-released documents relating to the misuse of National Security Letters to collect Americans' private information. But don't let us have all fun — you, too, can dive into the docs and help uncover the truth about the FBI's abuse of power.**

**All 1138 pages are freely downloadable (with searchable text) from EFF's website, and we'll be posting a new batch every month. We've had over 8000 downloads so far, and the blogosphere is starting to light up with feedback and analysis of the documents, which were disclosed after EFF sued the government under the Freedom of Information Act (FOIA) earlier this year. Over at Wired, Threat Level reports that much of the mischief at the FBI seems to be emanating from a mysterious "Room 4944", and this anonymous blogger is asking questions about who knew what when.**

## Where's the money?

***Rather than running product listings in trade publications and newspapers, media insiders say tech companies prefer to buy keyword ads so they can send buyers straight to the gear they want. "Search is what ignited everything," says Geoff Ramsey, Chief Executive of eMarketer, a firm which aggregates and analyzes online marketing statistics.***

***When Prime Minister John Howard lost his YouTube virginity yesterday morning, he probably didn't expect a search of his name to rank his two minute long piece behind videos entitled "John Howard is a farting fossil fuel" and "John Howard downloads some p-rn".***



# Large textual utterance collections

**Mailing lists**

**Discussion groups**

**Blogs**

**No standard collection,  
sometimes as part of other  
places, e.g. TREC/INEX tracks  
W3C, CLIR**

- TREC Blog track 2006
- Old WWW workshop 2005, 2006 ↩
- New conference ICWSM

**The blogosphere is the most explosive social network you'll never see. Recent studies [ha!] suggest that nearly 60 million blogs exist online, and about 175,000 more crop up daily (that's about 2 every second).**

## IM

- No known collection(s)
- One in four workers admits to spending more than two hours a day chatting with their significant other over instant messaging. ('05)
- 1b/day on MS IM (mid '06)
- One third of South Korean teenagers average 100 text messages per day
- half of teenagers regularly wake up to incoming text messages from their friends. Regularly wake up.
- The average e-mail is opened within 48 hours.

## Email

- Enron – only organisational, not complete
- TREC/INEX – smallish, niche
- E-mail traffic per user, per day increased by 33% between 2005 and 2006.
- The average size of e-mail messages is also increasing quickly. The average size of messages without attachments jumped by 30% between 2005 and 2006, and the average size of messages with attachments increased by 34% in this time period.
- The combination of increased e-mail traffic and larger messages had a dramatic impact on bandwidth storage requirements in 2006. The average bandwidth storage requirement per user, per day jumped by 61% between 2005 and 2006.

**QA, citation analysis, biomedical, legal...**



# Open source textual utterance data

## Utterance text data

- mailing lists, discussion groups, blogs, email, IM
- is where *lots* of individual's time is being spent
- real-life
- unsanitised – not “final” docs, quickly sent
- utterances – from the mouth

## Important for search – lots of interest, few resources

### But...

- ungrammatical, misspellings, odd phrases, omissions, repetition
  - opinion
  - mixed interest / accidental
- human
- often expertise,

# What to do with the data?

**What can we find out from what people say?**

**What? – wrong question!**

**How can we *help* people by knowing something about what *they* have said, or *about* them?**

- Context for improved retrieval?
- Personal information for better delivery?
- Deep personal knowledge about their psyche/beliefs

**How can *other people* be assisted by knowing things someone else knows, or about someone else, or that someone else is in a similar situation?**

**Déjà vu - "already seen" fr**

**Assisted serendipity (chance discovery)**

**Tacit knowledge**

**Abduction – C.S.Peirce**

- method of reasoning in which one chooses the hypothesis that would, if true, best explain the relevant evidence
- Abductive reasoning starts from a set of accepted facts and infers to their most likely, or best, explanations
- humans have an innate ability to infer correctly; possessing this ability is explained by the evolutionary advantage it gives

**Experience management, mining & sharing**

- yahoo answers is great for extroverts

## Socio-cognitive basis

- How do people represent knowledge internally
- How do people externalise their knowledge in utterances
- How do people take in other people's utterances and make inferences
- How can computers assist with any of these
- Cognitive science + sociology + philosophy + linguistics + computing
- Practical, real-life examples

**How can information retrieval, with the help of linguistic and cognitive knowledge, deal with these problems and retrieve appropriate results (often in the context of exploratory search)?**

**If a person can understand the text, how can we get closer to effective automatic systems acting correctly?**

**Examples of such queries, from real life, can be investigated using socio-cognitively motivated techniques for extracting knowledge.**



Tech Boom, Media Bust

Brian Caulfield, 07.16.07, 11:15 AM ET

It was a slow Friday at *Red Herring* magazine. The receptionist at the Silicon Valley tech title had stepped away from her desk. So a messenger strolls in from the summer sunshine, finds a 20-something reporter on her first real job and hits her with an eviction notice. *Red Herring* has three days to pay the rent or get out. Word got around, fast. Then someone looked outside. There, driving up in a rented silver Mazda minivan is a correspondent with gossip blog Valleywag. [Aaaaaand she's got a camera.](#)

Silicon Valley is booming again. But if you work in tech media, there's blood on the floor. Take *Red Herring*. It hung onto its offices after getting the eviction notice earlier this month. But gossip site Valleywag is breaking story after story not just on its beat--but about its woes. Meanwhile, bigger publications are hurting too: Time Warner's *Business 2.0* saw ad pages drop 21.8% through March from the same period a year ago; *PC Magazine's* editor in chief walked out the door after ad pages fell 38.8% over the same period; and one-time online powerhouse CNET is reporting growing losses even as the companies it covers flourish. It may be happening in tech first, but there's no reason the same thing won't happen, eventually, in every media niche.

Things couldn't be much more different than the last boom. While online upstarts such as HotWired struggled to make money--they had to *invent* the banner ad--print titles flourished. *The Industry Standard*, founded in 1997, set ad sales records. *Business 2.0* came out of nowhere to scoop up gobs of ads against articles detailing how to succeed in the new economy. And one-time venture capital bible *Red Herring* ballooned to hundreds of pages. Then the tech downturn hit. *The Industry Standard* closed. The assets of *Red Herring* and *Business 2.0* were sold to new owners.

But while the good times are back--the tech-heavy Nasdaq hit a six-and-a-half-year high last week--tech trade and new-economy publications have not bounced back. The first problem: online keyword advertising. Media insiders say search engines such as Google have snarfed up the product-driven ads. Rather than running product listings in trade publications and newspapers, media insiders say tech companies prefer to buy keyword ads so they can send buyers straight to the gear they want. "Search is what ignited everything," says Geoff Ramsey, Chief Executive of eMarketer, a firm which aggregates and analyzes online marketing statistics.

Meanwhile, *Industry Standard* founder John Battelle is keeping the bonfire of the print titles burning. His Federated Media Publishing is selling ads on more than 100 blogs, giving ad buyers the ability to spend big money on a collection of highly specialized sites--many of them focused on tech--that suit their needs. "If Cisco has to spend, I don't know, a couple of million dollars on a trade campaign, they are not spending it with *Red Herring* or *Business 2.0*. They are spending it with Federated Media, with bloggers who cover the sector," says Rafat Ali, editor and publisher of online media tracker PaidContent.org.

And while blog networks are quickly gaining scale, even their most coveted offerings are cost-competitive. To make a back-of-the-napkin comparison based on rate cards: A start-up looking to get attention will grab a third-of-a-page color ad in a magazine with a rate base of 600,000 and might pay \$27,300; or it can pay \$21,000 for 600,000 impressions for its ads on TechCrunch--a site covering start-ups represented by Battelle's Federated Media--assuming they take the priciest ad slot on one of tech's hottest sites.

That's no surprise, given that it takes fewer resources for blogs to crank out content than it does print titles. Web sites such as GigaOm, TechCrunch and Valleywag--with a few laptops, a web server and some hustle--are crowding into beats once dominated by trade publications and enthusiast magazines who rely on printing presses and full-time writers and editors. Bottom line: A successful blog can simply grab more readers, per employee, than more traditional media.

Talk to blogger Matt Marshall. He walked away from covering venture capital at one of California's biggest newspapers, the *San Jose Mercury News*, to run a venture capital Web site from the second bedroom of his Fremont, Calif., home. He has no employees. Federated Media handles the ad sales for a 40% cut. And Marshall says he now makes more than he did as a reporter. Meanwhile, the *Mercury News* laid off 31 of his former colleagues this month. "Where they can actually succeed is by taking a particular vertical and absolutely nailing it," eMarketer's Ramsey says of bloggers like Marshall.

Of course, blogging is not the express lane to riches its more exuberant backers would have you believe. The anonymous satirist who runs "The Secret Diary of Fake Steve Jobs" started hitting up his readers for money-making ideas just weeks after being named to *Business 2.0's* list of "50 Who Matter Now," even while, in character as Apple Chief Steve Jobs, he boasted about Apple's huge stock gains. And while Marshall says he's making a living, he's still living lean: he says he works until 3 a.m. many nights. "I can go under any day, and that's what brings the passion to this," Marshall says.

The truth is, the vast majority of bloggers will never garner more than a few dozen readers. Then again, most of today's print-heavy news outlets are scaling back in the face of the relentless online competition. Marshall's father, Tyler Marshall, walked away from journalism after winning a Pulitzer Prize at *The Los Angeles Times*, bought out in a round of downsizing at the venerable newspaper. When Marshall told his father about his plan to launch his own publication, the older Marshall didn't discourage him. After all, what did he have to lose?

**But first, let's recall the three generations according to Broder:**

**First generation: search engines are using almost only on-page data such as text and formatting information to compute result ranking (1995-1997, cf. Alta Vista, Excite, etc...).**

**Second generation: search engines are using off-page, web-related data such as link analysis, anchor-texts, and click-through data (1998-..., cf. Google).**

**Third generation: search engines try to blend data from multiple, heterogeneous sources trying to answer 'the need behind the query'. The computed results are customized according to the user's information needs, taking into account the user's personal data background, context, and intention (now? - ...).**

**Broder distinguishes different sorts of web search queries:**

**Navigational: intended use is to reach a particular web page (similar to 'known item' search in classical information retrieval). Therefore, navigational queries usually do have only one 'right' result.**

**Informational: intended use is to acquire information assumed to be present on one or more web pages (as in classical information retrieval).**

**Transactional: intended use is to find a web page, where further transactions (e.g. shopping) will take place.**

**If we take social bookmarking services, navigational queries can be computed simply by using the user's *personomy* (i.e. the set of all tags used by a distinct user). If the goal is to find a web page, which has been already accessed in the past, the page might be found quickly, if the user has registered the page within the bookmarking service (which comes to '*Finding*' information). But, the query might also be resolved by using other people's tags, if somebody has tagged the page (with objectively descriptive tags).**

**Social bookmarking services are also useful for the other two purposes. In addition, if the page is found, the social networking information can be utilized for 'discovering' new, previously unknown, but related (similar) information (which comes to '*Discovering*' information). Hotho et al. present an adaptive ranking algorithm (FolkRank) for social bookmarking systems and discuss the problems that arise for tag-based search engines [3].**

**But, to answer the 'need behind the query' as Broder states in his definition of 'Third generation search engines', further personalization is mandatory. Only, if the search engine is able to find out the context of a query w.r.t. a given user and a given situation (i.e. even the same user might have different information needs in different situations), then it is possible to grasp the actual context of the query, and thus, also the 'need behind the query'...**