

➔ Highlighting the Information

Collective and Emerging Ways

Gwendal Simon¹

Department of Computer Sciences
GET - ENST-Bretagne
janvier 2007

¹gwendal.simon@enst-bretagne.fr

➔ A New Era Of Information

A *massive* amount of **accessible** information

- 106 millions of websites
- 3.6 millions of *french* bloggers (+53.7% in one year)
- 4.5 millions of *french* web-surfers produce comments
- ...

... without any explicit coordination

... while humans have a limited capacity of attention

⇒ The Economy of **Attention**

➔ The Challenge of Highlighting Information

Many shifts are occurring :

- decentralizing the editors
- decentralizing the archivists
- decentralizing the distributors
- decentralizing the moderators

→ Trends and Challenges

Part I

Communication Schemes

➔ A New Era Of Information

XXth century provide massive access

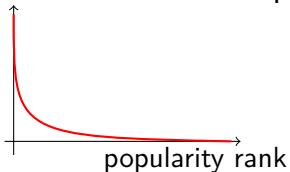
- **but only two main communication schemes :**
 - few-to-all
 - 1.5 billion TV sets in use in the world
 - few authorized incumbents (superbowl : ~ 140 millions viewers)
 - one-to-one
 - 2.7 billion mobile phones in use
 - 42% of Americans are active in SMS

New perspectives in XXIth century :

- **new schemes of communication**
 - many-to-many : forum, communities...
 - all humans are broadcasters : blogs...

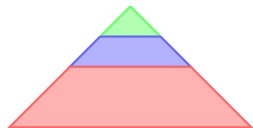
➔ Few apparent changes

nb of visitors



Few-to-all *zipfian* communication :

- a few pages are very popular
- the *long tail*



Even in open forums :

- 1% main contributors
- 9% little contributors
- 90% lurkers

➔ Web2.0 Communication

Several types of **blogs**

- *diary blogs* :
 - "message in a bottle" communication
 - few fleeting virtual links
- *community blogs* :
 - dense multimedia communication
 - based on existing real links
- *affinity blogs* :
 - communication on a precise topic
 - new strong virtual links
- *media blogs* :
 - expressing convictions
 - building a worldwide public

➔ Network Utility

Network's value :

- *Sarnoff law* : broadcast network
- *Metcalfe law* : phone network
- *Reed law* : community network

net class	for one user	global	net fusion
broadcast	1	n	$n + m$
phone	n	n^2	$n^2 + m^2 + n * m$
community	2^n	2^n	$2^n * 2^m$

Part II

Information Ranking

➔ Slashdot Moderation

Slashdot : collective moderation to guarantee "*stuff that matters*"

▪ Moderation Principles :

- anybody can post
- anybody can become a moderator
- each user is characterized by a *karma*
- each comment has a score from -1 to +5
- initial score depends on karma from -1 to +2
- comments are displayed upon a threshold (3 by default)
- moderators can increase or decrease the score
- moderators have a limited number of actions by day

▪ Meta-Moderation Principles :

- anybody can become a meta-moderator
- a meta-moderator rates a moderation action ("fair" or "unfair")

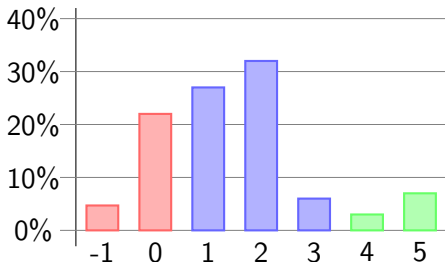
➔ Slashdot Moderation - Study

25,000 distinct moderators

293,000 moderations

490,000 comments

- moderated comments : 28%
- positive moderation : 79%
- relative consensus :
 - 85% same moderation
 - 92% agree meta-mod.
 - 2% "often" disagree



➔ Slashdot Moderation : Initial Influences

Moderation reversal : positive fair cancel negative unfair

- only 37% of "unfair" moderations are reversed

→ a first unfair moderation can affect the final score

		Ending Score						
		-1	0	1	2	3	4	5
Starting Score	-1	93%	4%	1%	0.6%	0.4%	0.2%	0.4%
	0	13%	76%	6%	2%	0.8%	0.6%	1%
	1	2%	3%	73%	11%	4%	2%	5%
	2	0%	0%	2%	71%	11%	5%	11%

➔ Digg Voting System

Main differences with Slashdot :

- not only on comments but also on stories
- only positive votes
- unlimited number of votes
- . . . easy-to-use eye-candy system

An old debate :

- the wisdom of crowd
- the mass domination

➔ Google PageRank

Some intuitions :

- v links to $u \implies v$ votes for u
- highly linked pages are more "important" than pages with few links
- the *random surfer model* : clic - clic - clic and "get bored"

A recursive algorithm based on web hyperlinks

- a page is good if the pages that point to it are good

$$R(u) = (1 - d) + d * \left(\sum_{v \rightarrow u} \frac{R(v)}{\|v \rightarrow x\|} \right)$$

- the *relevance* is merged with the *popularity*

➔ Authorities and Hubs

On a peculiar topic, distinction between

- **authorities**
 - the most prominent sources of primary content
 - many links from pages related to this topic
- **hubs**
 - high-quality resource lists that direct users
 - many links whatever the topic (including this one)

A mutually reinforcing relationship :

- a *good* hub is a page that points to many good authorities
- a *good* authority is a page that is pointed to by many good hubs

➔ Web2.0 Trends

New approach to hyperlinks :

- blogrolls, trackback and hyperlinks in posts : more reciprocal
- focus on data, not on pages : less influenced
- hubs are fueled by anonymous : more serendipitous and open

website	authority	audience	Forbes
engadget.com	1	466	23
blog.hexun.com	2	268	-
boingboing.net	3	1,666	15
techcrunch.com	5	477	10
postsecret.com	9	16,954	14
perezhilton.com	35	744	2
buzzmachine.com	210	40,620	5

➔ Marketing Issues

Investing the *viral marketing* :

- a conventional way : corporate blogs
 - advertisement and buzz
- a risky way : anonymous posts and comments
 - focus group and recommendation network
- a bottom-up way : stimulate user-generated contents
 - branding and network externality

Part III

Collaborative Classification

➔ Tagging Bookmarks

del.icio.us collective bookmarking

- a bookmark : page + user + some tags
- a page may be bookmarked by several users
- a **semantic** emerges from tags
 - absolute freedom : "*as close as possible to no rules at all*"
 - no central indexing nor pre-conceived ontology
 - more illuminating as the number of taggers grows
 - allow time and task related tags

⇒ a large-scale adaptive indexing system

➔ Collaborative Tagging - Study

Studying a casual *del.icio.us* feed (64 popular pages) :

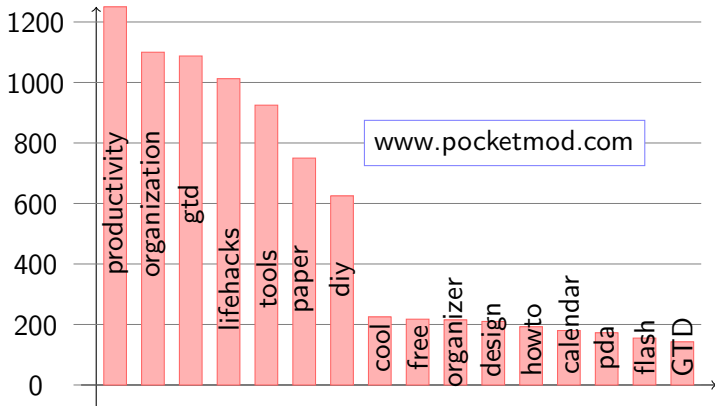
- number of posts $\in [53 \dots 5,172]$
- number of tags $\in [49 \dots 13,809]$
- number of *unique* tags $\in [23 \dots 1,252]$
- 30% of tags are used only once (typo?)

User practices :

- only 6% of users do not tag
- but 65% assign less than 3 tags to a bookmark

➔ A Cluster of Tags

A consensus on a **cluster** of tags (≈ 7)



➔ Emerging Conventions

Still no consensus on conventions :

- **several classes of users** :
 - synonymy, capital vs. lower case, singular vs. plural, acronyms. . .
- **a mix of communal and individual tags**
 - most popular tags are rarely used together
 - user-related task and time related tags (toread, gtd, . . .)

Two types of folksnomies :

- ***broad folksnomies* (del.icio.us)** :
 - a lot of people describing one item
 - give a sense to an item
- ***narrow folksnomies* (flickr)** :
 - a lot of items described by one person
 - give a sense to a tag

Part IV

Conclusion

➔ An Ongoing Research Topic

An abrupt growth of "the web" :

- more contents produced on a higher frequency
- more hyperlinks exhibiting a more horizontal space
- more ways to access data through stronger decentralized relays

The paradox of the Web2.0 :

- producing hyperlinked information is democratizing
- surfing *efficiently* on the web requires a good practice

➔ References I



D. Cardon and H. Delaunay-Teterel

La production de soi comme technique relationnelle. Un essai de typologie des blogs par leurs publics

Rezeaux, Hermès-Lavoisier, vol 24(138), 2006



C. Lampe and P. Resnick

Distributed Moderation in a Large Online Conversation Space
in ACM Computer Human Interaction, 2004



S. Brin and L. Page

The Anatomy of a Large-Scale Hypertextual Web Search Engine.
Computer Networks and ISDN Systems, vol. 30, 1998

➔ References II



J. Kleinberg

Authoritative Sources in a Hyperlinked Environment.

Journal of the ACM, vol. 46(5), 1999



M.E.I. Kipp and D. Grant Campbell

Patterns and Inconsistencies in Collaborative Tagging Systems : an Examination of Tagging Practices

in American Society for Information Science and Technology, 2006