

Des Aiguilles dans une Botte de Foin

Une Étude sur

La Visibilité de L'information sur

Internet

Needles in the Haystack
A Study on
The Visibility of Information on the
Internet

Outlines

-Internet at huge

How large is it

How fast is it growing

Can we perceive the size?

-How much pattern information we have

-What search methods we use

Text lookup, (not a viable option anymore)

Semantics Web, (not our job)

(semantic web, semantic indexing, smart spiders, ...)

Pattern communities

A better option for dissemination?

Our Approach

-Internet is huge, it is impractical to come up with accurate numbers about it, therefore our study follows the following rules:

-Use reliable sources only

-Cross-check numbers for added reliability

-Use more conservative estimations when possible

according to estimation contexts, we select either lower or upper range of estimations numbers in such a way to avoid “bloated results”

-For extremely large numbers, we only use “orders of magnitude” to present a general idea about the situations. In this case exact estimated numbers are less important

Pew Internet and American Life Project:

72 million persons in USA alone go online everyday (5/2002)

Moreover: 29% use a search engine to find information everyday,

Reality #1:

We do our math :

$$29\% * 72 = 20.88$$

(million persons in USA looking for information online every day)

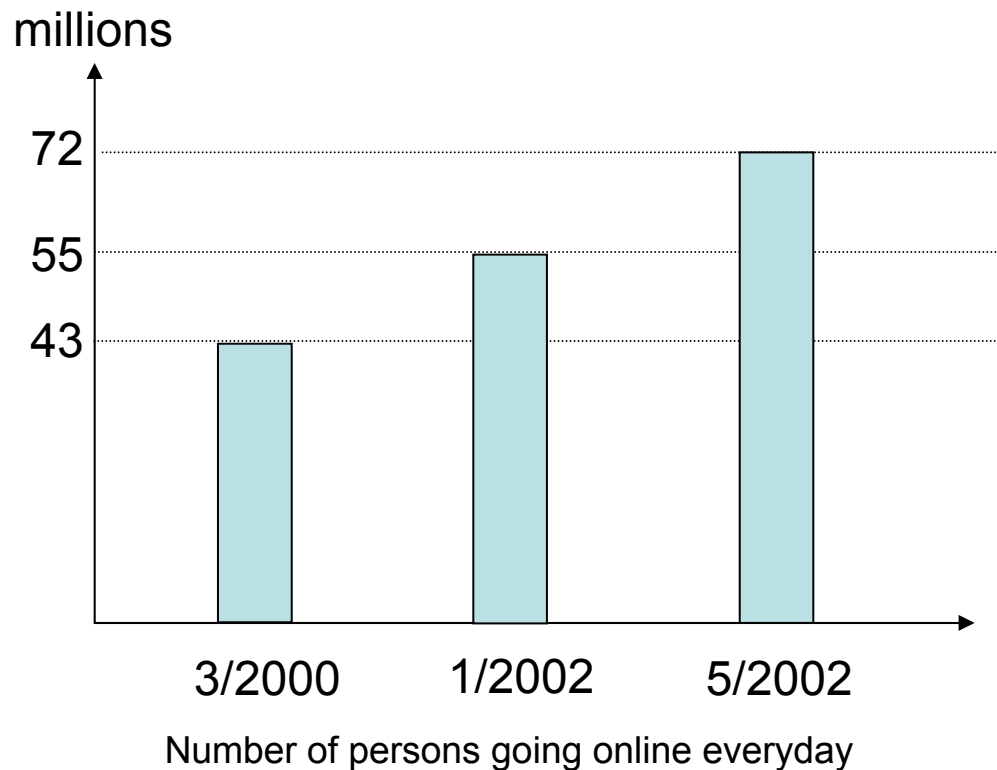
Sources:

1- Report 1: Getting serious online,

March 3rd, 2002

2- Report 2: One year later,

September 5th, 2002



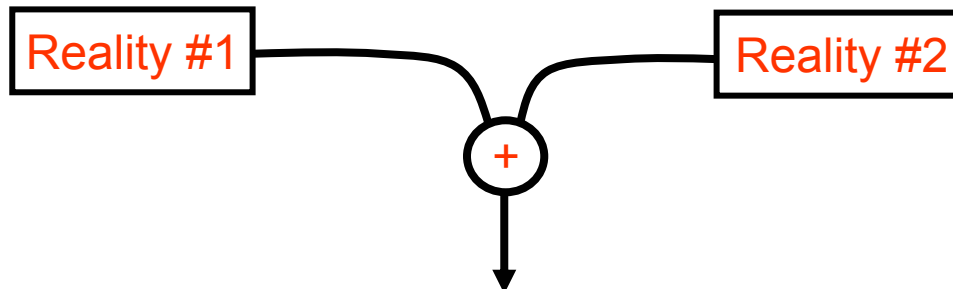
Internet is Getting Bigger

Reality #2: Find answers to virtually any question
Find articles about virtually any **"thing"**
Ubiquitous connectivity for everyone

Enterprises

Researchers

People



Logical conclusion:

I will put my patterns on the Internet for people to find and use them

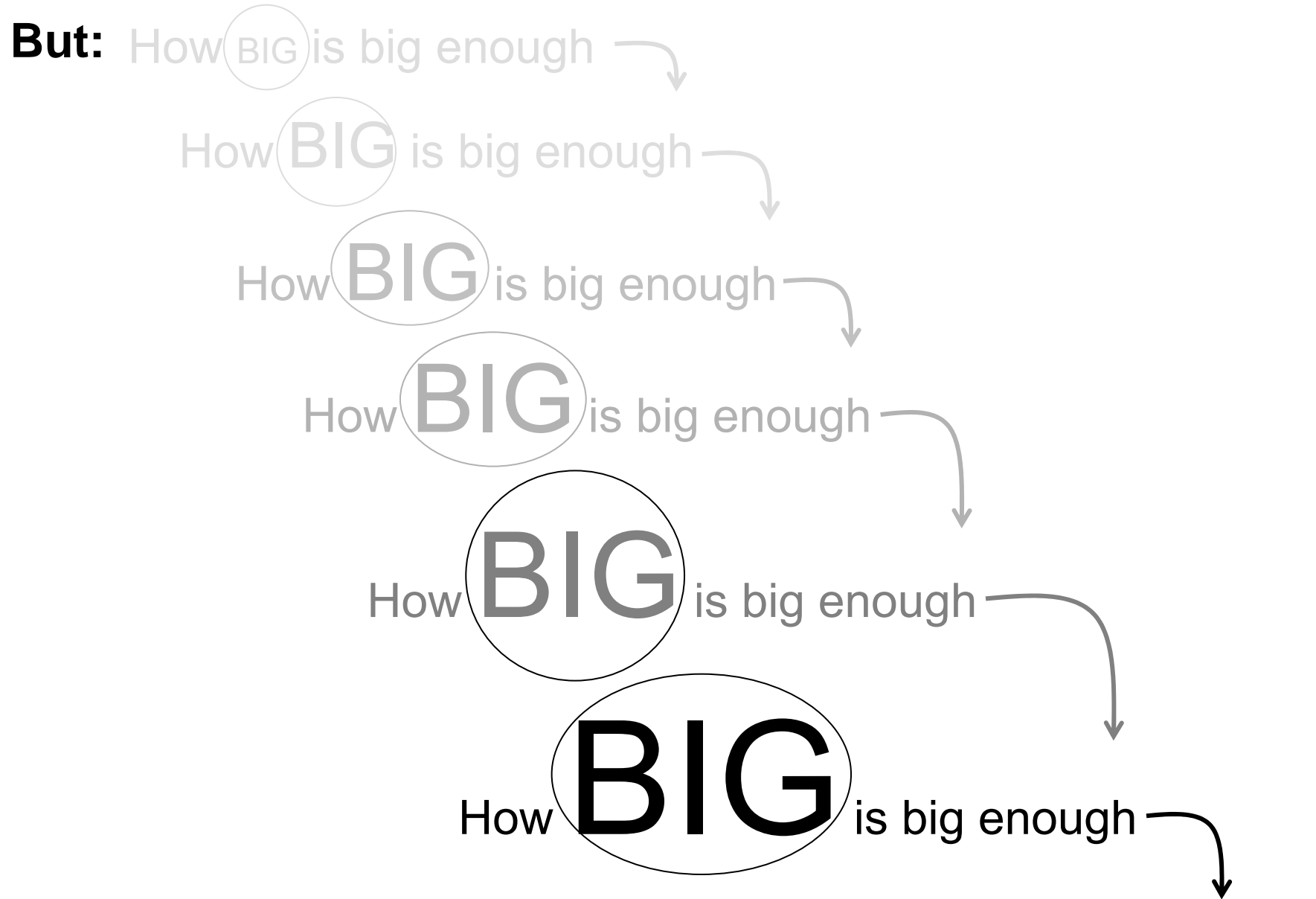
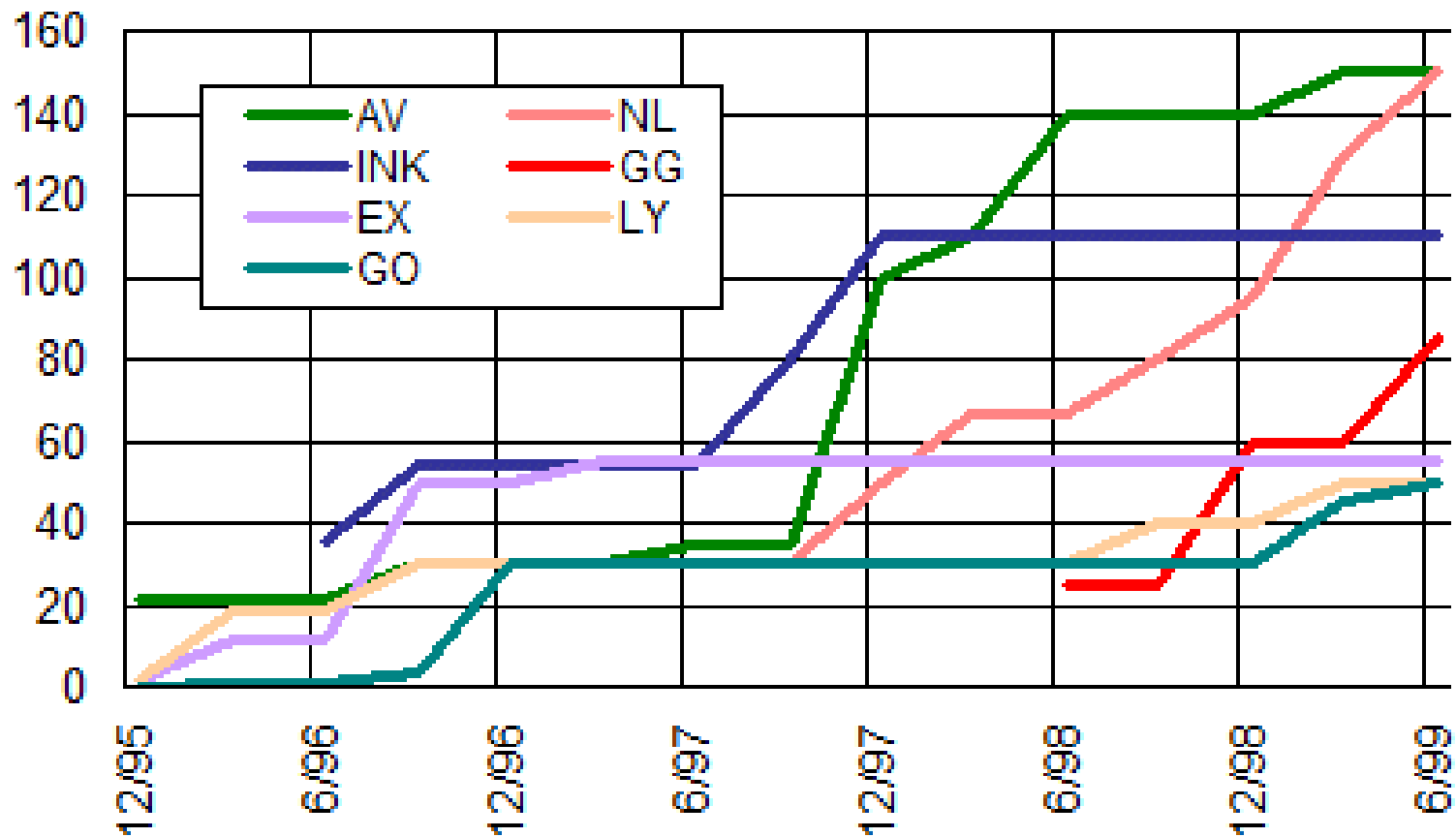


Chart #1



MILLIONS Of Textual Documents Indexed
December 1995-June 1999

SearchEngineWatch.com

Chart #2

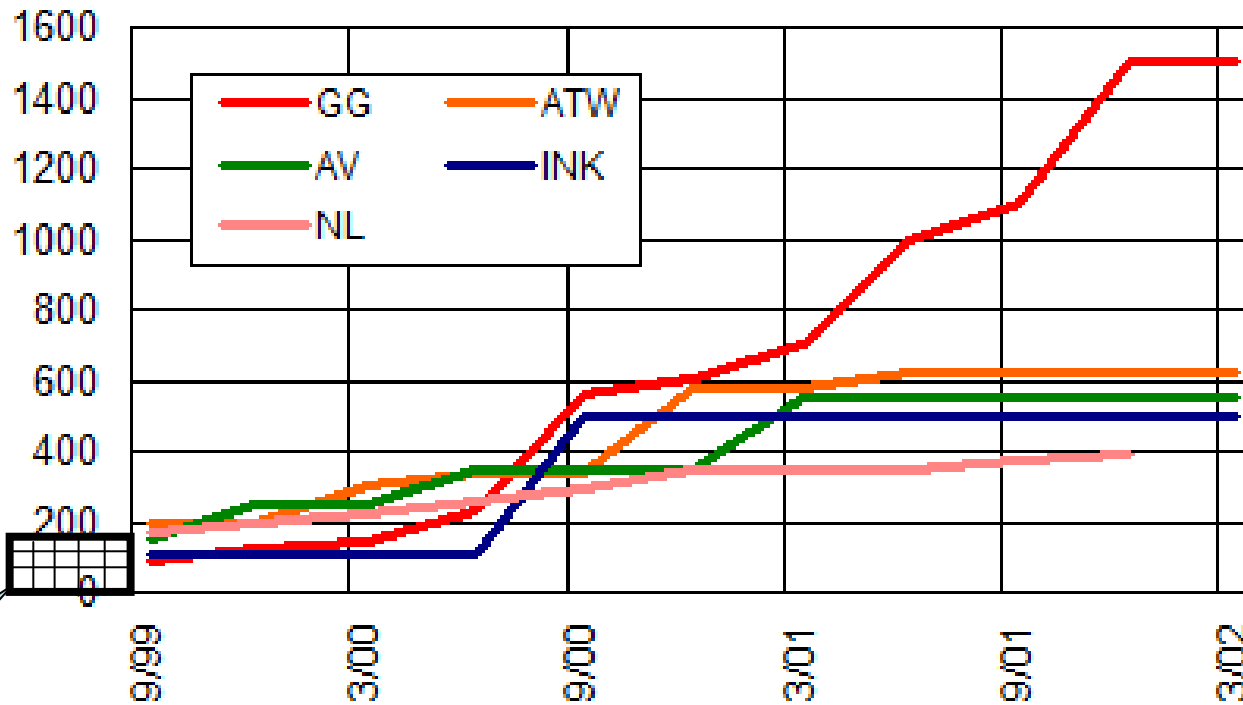


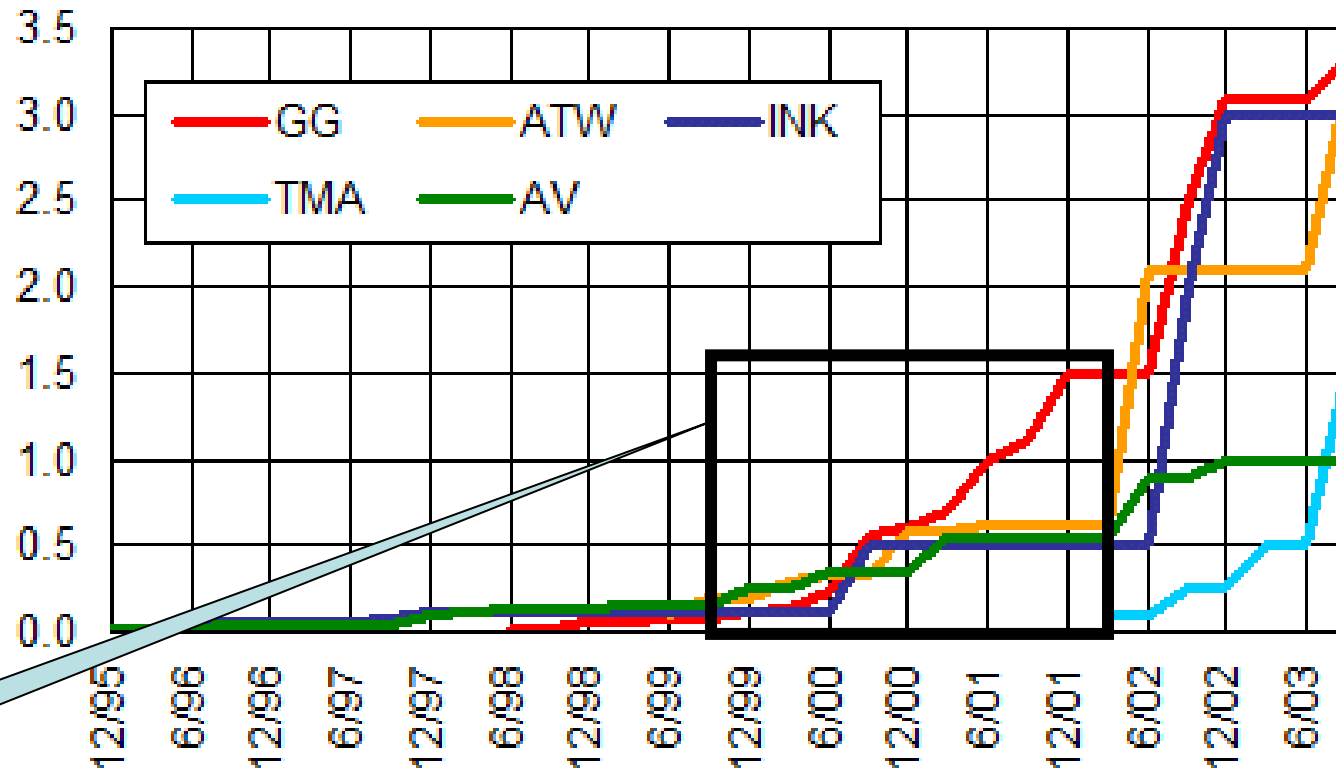
Chart #1 is here

MILLIONS Of Textual Documents Indexed
September 1999-March 2002
SearchEngineWatch.com

Google Announces Largest Index

The Search Engine Report, July 5, 2000
Google breaks the 500 million page mark

Chart #3



Billions Of Textual Documents Indexed

December 1995-September 2003

SearchEngineWatch.com

Google to Overture: Mine's Bigger

Search Day: Aug. 27, 2003

And the War Goes on:

FAST Announces Largest Search Engine

The Search Engine Report, Aug. 2, 1999

Who's The Biggest Of Them All?

The Search Engine Report, Nov. 1, 1999

FAST Gets Bigger, Partners With Lycos

The Search Engine Report, Feb. 3, 2000

Numbers, Numbers -- But What Do They Mean?

The Search Engine Report, March 3, 2000

Inktomi Reenters Battle For Biggest

The Search Engine Report, June 2, 2000

Google Announces Largest Index

The Search Engine Report, July 5, 2000

Google Fires New Salvo in Search Engine Size Wars

SearchDay, Dec. 11, 2001

One more page ?

Where are We Today:

Search Engine	Reported Size	Page Depth
Google	8.1 billion	101K
MSN	5.0 billion	150K
Yahoo	4.2 billion (estimate)	500K
Ask Jeeves	2.5 billion	101K+

Our Study:

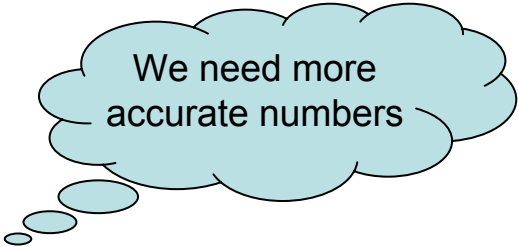
We are not in for size and quality improvements of global indexing
(which has a lot of ongoing research, for example see

Latent Semantic Indexing, University of Tennessee,
<http://www.cs.utk.edu/~lsi/>)

**We are in for the effect of Internet on pattern visibility,
and how to improve it.**

What can we do with these numbers:

- They give an idea about how large the Internet is
- They show how fast it is growing



We need more accurate numbers

What should we **Not** do with these numbers

- We can not estimate the size of information on the Web

Reported numbers are those of selected web pages only, many pages are ignored

Number of pages give no idea about their size or contents

Page depth is the **max** amount of text copied from **indexed** pages only, It is just a company's policy:

Number of Pages X Page Depth = meaningless number: Too many unknowns

- We can not predict how these numbers will be next year

Numbers can remain fixed for some time if no new "size war" erupted

→The rate of growth of these numbers is insignificant to our study

How Large is the Information on the Internet?

How Much Information 2003?

A major study at University of California at Berkeley, some findings:

- *We produced 5 exabytes of new information in 2002
- *Our annual production rate grows 30% every year
- *92% of new information is found on hard discs
- *Only 0.01 % of new information is found on paper
- *WWW contains 170 terabytes on its surface (static sites, HTML)
- ***WWW contains 92,000 terabytes of deep pages (CMS)**

Content Management Systems (CMS):

Don't save pages as static HTML code. Clear out the formatting details and save info contents only on the server. Use processing logic to add any formatting style upon rendering. Why?

Deep ≠ dynamic, but deep generates dynamic

Annual production rate grows about 30% every year:

2 Exabytes new information in 1999

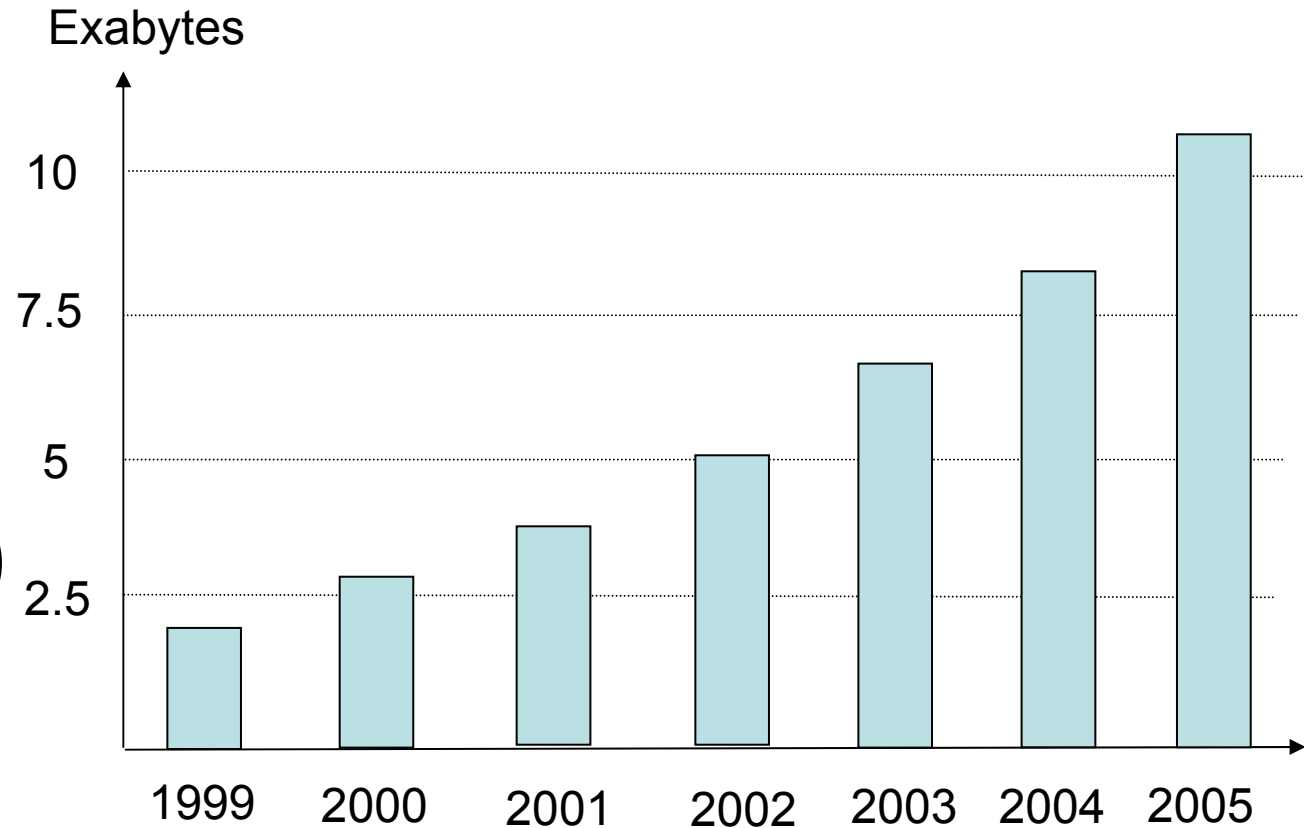
5 Exabytes new information in 2002

Estimations:

2003: 6.5 Exabytes

2004: 8.45 Exabytes

**2005: 11 Exabytes
Of new information**



What is Exactly an Exabyte :

International Electrotechnical Commission (IEC):

Byte,

X1000 = Kilobytes (10³ Bytes),

X1000 = Megabytes (10⁶ Bytes),

X1000 = Gigabytes (10⁹ Bytes),

X1000 = Terabytes (10¹² Bytes),

X1000 = Petabytes (10¹⁵ Bytes),

X1000 = Exabytes (10¹⁸ Bytes)

Moreover, we will be producing:

A new 0.5 Zettabytes in 6 years, another new 1 Zettabytes 3 years later, then ...

What is a Zettabyte anyways? We will need to worry about it only in few years.

Let's first perceive the size of an Exabyte

How Big is What?

A binary decision: 1 bit

Pickup truck full of books: 1 Gigabytes

A library floor full of academic journals: 100 Gigabytes

11 exabytes = one million buildings of 110 floors full of academic journals

We produce this amount every year

The library of congress (17 million books): 136 terabytes

11 exabytes= 81 000 times the books inside the library of congress

Next year, we will produce another 14 exabytes of new information

Did we loose perception of information size in the new era of computers

Today, one DVD disk (25 grams) can hold contents of 7 pickup trucks full of books (5 tons)

One hard disk (500 grams) can hold contents of two library floors full of books (250 tons)

Did we increase our capacity to use information at the same rate! Or even close!

Or it will just land in “the archives” while we are busy producing new information

Size matters,

but it is becoming our enemy

Can we say:

-Bigger Internet means more “quality, reliable” information!

or just more information

-Bigger Internet means better search!

or the opposite is true

-Bigger Internet means bigger indexes!

and less index updates

-Bigger Internet means better visibility!

or the opposite is true

Study on visibility of patterns on the Internet

Part I: Estimating the quality of search by keyword

Keyword Indicators

How Much Information 2003?

The report uses a “keyword indicator” to estimate the functionality of a page

- 30% of pages contained the word “search”, indicating they are complex sites
- 7.7% contained the word “password” or “login”, indicating protected pages

Is “Pattern” a good keyword indicator

Keyword Indicators

(cont.)

If a page has the word “pattern”, does it indicate that a page is really about patterns!

“Pattern” is a popular, general purpose word, used very frequently, everywhere

-When we write text, we use “patters” in everyday language as frequently as we use

“will”, “object” and “normal” (Source: Collins Cobuild Bank of English, 3rd edition, 2001)

Some synonyms that we replace by the word “pattern” in all texts

arrangement	decoration	device	diagram	figure
guide	impression	instruction	marking	mold
motive	original	ornament	patterning	plan
stencil	template	trim	order	copy
constellation	kind	method	orderliness	sequence
shape	sort	style	system	type
variety	archetype	beau-ideal	criterion	cynosure
ensemble	exemplar	guide	mirror	norm
paradigm	paragon	prototype	sample	specimen
standard	copy	imitation	(Source: Roget’s 21st century thesaurus)	

But we still expect pattern users to find patterns on the Internet using keyword search

Part II: Estimating The size of pattern information

Building Pattern Corpus: An extended empirical study

- 7 groups asked to trace and collect patterns on the Web
- No group was able to built a “Universal Corpus”
- The results < 30 Megabytes
- We estimated a cap of 50 Megabytes of all patterns available
- The actual number is less, but we stay on the conservative side

The estimation of books: 50 books x 5 Megabytes = 250 Megabytes

- most books are less than 5 Meg, but we stay on the conservative side
- We consider books that offer patterns in them, not books about pattern concepts

The total pattern information \approx 300Megabytes

- This refers to actual patterns

Part III: Estimating the size of patterns compared to the Web

Using the mathematical model of the UCB report (slide 14):

The **cumulative body of information** we have today

(Current body of information on hand):

47 Exabytes, **55% in the last 3 years alone** (most of it on hard disks)

The cumulative body of information on the **Web**: (optimistic estimation)

785 Petabytes

The size of patterns compared to the size of information:

1- Total Pattern information to total information: 11 orders of magnitude

1 to 100 000 000 000 (10^{11})

2- **Pattern on the Web** to **total web** Information: 10 orders of magnitude

1 to 10 000 000 000 (10^{10})

→ A more accurate comparison

Part IV: Searching for Patterns on the Internet

Pattern books, Amazon.com

8,633	matches (02/2002)
11,852	matches (02/2003)
114,101	matches (02/2004)
22,711	matches (02/2005)

Remember: Paper represents only 0.01% of new information

Possible factors of the ~~increase~~ “explosion” of 2004:

- The inclusion of used books and items,
- Dumb search criteria (early versions)
- All our advanced search refinements in 2004 ranged between 25 000 and 114 000 (refer to previous talk of UPA, 2004)

Possible factors of the “back on track” of 2005:

- Smarter indexing and improved search criteria
- Several contact attempts with amazon.com

BUT ...the trend is growing

Part IV: Searching for Patterns on the Internet (cont.)

Pattern web pages in Google

29,700,000	matches in 0.73 seconds (02/2003)
35,300,000	matches in 0.28 seconds (02/2004)
43,500,000	matches in 0.19 seconds (02/2005)

Observation:

We are getting more results, faster

Hardware speed is growing, prices are falling

Information size is growing, its electronic share is growing

What are we missing? See the box on slide 17

Remarks:

-Who can go through 43 million pages looking for anything

-How many pages can we practically go through

-Is it better to “ask a friend” or hope to “be lucky” in this new era

If we need to use the Internet effectively, we need to do things differently

**Remember:
Electronic media
(hard disks) contain
92% of new
information**

IEEE: Similar Observation

IEEE Journal "Spectrum", Tomorrow's Technology Today:

(November 2004)

Does Google Like Me?

"I assume that Google's ranking has directed me to the most relevant and informative Web pages"

But:

- A search for author's own web page returns about 1.8 million hits in 0.2 seconds
- "I am impressed with the SPEED, but not with the RESULTS"
- The first few hundred hits were useless; the actual web page possibly "somewhere down past the million mark in the list". Several keyword searches did not give better results
- "... , this is partly a problem in semantics"

**Don't we all share the same perception?
Do we have other choices?**

This is only half the problem

(but 92% of the slides)

Conclusion

We need to effectively find patterns and use pattern information

-Pattern communities on the Internet already exist, but they are:

A mini WEB of **crisscross reference pages** about pattern + pattern “stuff”

One page has links to 75 pattern websites, each loaded with many more links, most pointing to each other, some back to original pages: A maze

Lots of textual information

cognitive load, scalability problem, reuse problem

Current state-of-the-”ART”

A lot of people are writing [about] patterns,

but few people are actually using them

Stay away from “searching the web”?

Stay away from “patterns”?

Conclusion (cont)

~~Who~~ what can go through 43 million pages looking for anything

A Software, not a human

- It is not how big-, but how **effective** information is:

-**Separate** “patterns” from “pattern literature”

-**Identify** pattern semantics

Programmability, tool integration, code generation

-**Write** “smart patterns”: programmable objects, not text format

-**Store** in XMLDB (eXist)

robustness, extensibility, scalability

-**Add** tools to automatically manipulate them

-We are working on it

References:

Robust Hyperlinks Cost Just Five Words Each

UC Berkeley, January 2000

<http://www.cs.berkeley.edu/~phelps/Robust/papers/robust-hyperlinks.html>

Technical paper on how web pages could be assigned a lexical code to make it easier to locate them.

September 1998 Search Engine Coverage Update

NEC Research Institute, September 1998

<http://www.neci.nj.nec.com/homepages/lawrence/websize98.html>

An update to the findings reported in Science magazine in April 1998, by its authors. It found that coverage was getting worse since the original study.

Robert W. Lucky, *Does Google Like Me?* IEEE Spectrum: Tomorrow's Technology Today, November 2004, pp 136

Gaffar, A., and Seffah, A. *An XML Multi-tier Pattern Dissemination System*, Encyclopedia of Database Technologies and Applications, Idea Group Publication Inc., USA April 2005, ISBN 1-59140-560-2.