Introduction

IN4325 – Information Retrieval, 2012

Organizational matters

Course overview

- 2 lectures a week for 7 weeks
- 4 programming assignments
  - **Groups** of 2-3 students (find your own, email me your group name and members)
  - Assignments are handed out on Wednesdays; they are due the first following Wednesday with a classs
- Final assignment
  - **Individual** work
  - Write a paper about an IR research topic
  - Hand-in intermediate results (topic, outline)

Assignments 1-4

- Group work
- The Wednesday lecture will provide you with the background knowledge to do these assignments

Final assignment

- Individually
- Write an **8-page paper** about an information retrieval research topic of your choice
  - Can be a lecture topic
  - Start thinking about a potential topic early on in the course
- Intermediate deadlines will provide you with an opportunity to develop the topic and paper outline
  - Handing in intermediate results is **voluntary but strongly encouraged**
  - I will provide feedback until the middle of April

Assignments timeline

	Handed out	Due
Assignment 0	8/2/2012	----
Assignment 1	15/2/2012	22/2/2012
Assignment 2	22/2/2012	7/3/2012
Assignment 3	7/3/2012	14/3/2012
Assignment 4	14/3/2012	21/3/2012
Assignment 5	21/3/2012	First week of Q4
Assignment 5a	21/3/2012	28/3/2012
Assignment 5b	28/3/2012	15/4/2012

Grading

- Course grade: 0.5 * [A1/2/3/4] + 0.5 * A5
- Assignments 0, 5a and 5b are voluntary and do not count towards your grade
- To pass
  - an overall grade of 6 or higher is required
  - AND all required assignments are handed in on time
  - AND all required assignments are passed with a score of 6 or higher

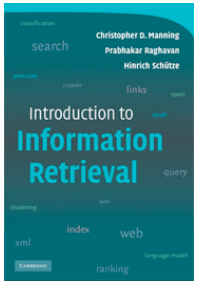
Questions?

- Are always welcome!
- Always per email c.hauff@tudelft.nl
  - Emails with **[IN4325]** in the subject line are read first!!
- Questions & answers may be posted on blackboard if they are relevant for others as well!

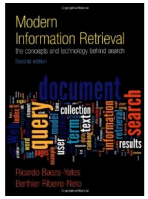
Reading material (lectures)

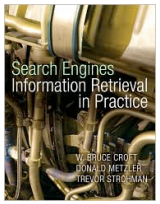
Additional papers will be announced on blackboard if necessary.

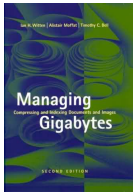
- *Introduction to Information R*etrieval by Manning, Raghavan and Schütze, University Press, 2008.
  - Book used in this course
  - Available online: http://nlp.stanford.edu/IR-book/



- Additional suggestions
  - *Modern Information Retrieval* by Baeza-Yates and Ribeiro-Neto, Addison-Wesley Professional; 2nd edition, 2011.
  - *Search Engines: Information Retrieval in Practice* by Croft, Metzler and Strohman, Addison-Wesley, 2009.
  - *Managing Gigabytes: Compressing and Indexing Documents and Images* by Witten, Moffat and Bell, Morgan Kaufmann, 1999.

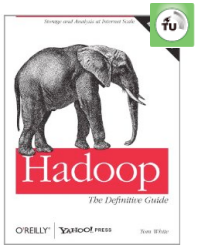






Reading material (assignments)

- The assignments are focused on MapReduce & Hadoop (more about this in the next lecture)



- Hadoop programming
  - *Hadoop: The Definitive Guide* by Tom White, O'Reilly Media, 2011.
- MapReduce algorithm design
  - *Data-Intensive Text Processing with MapReduce* by Lin, Dyer and Hirst, Morgan and Claypool Publishers, 2010
    - Available online: http://www.umiacs.umd.edu/~jimmylin/book.html

Information retrieval

Information retrieval in industry

Information retrieval definitions

- "The goal of a machine method of information retrieval is purely and simply that of being able to find and to recover at will information stored in a collection of documents. […] It is oriented completely towards actual use of the information, and to the convenience of the user."

  Calvin N. Mooers, Scientific information retrieval systems for machine operation; case studies in design. XIIth International Congress of Pure and Applied Chemistry, 1951

Information retrieval definitions

- "The goal of a machine method of information retrieval is purely and simply that of being able to find and to recover at will information stored in a collection of documents. […] It is oriented completely towards actual use of the information, and to the convenience of the user."

  Calvin N. Mooers, Scientific information retrieval systems for machine operation; case studies in design. XIIth International Congress of Pure and Applied Chemistry, 1951

- "An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request."

  F.W. Lancaster. Information Retrieval Systems: Characteristics, Testing and Evaluation, Wiley, New York, 1968.

Information retrieval definitions II

- "Information retrieval studies the retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a user information need usually expressed in natural language."

Baeza-Yates and Ribeiro-Neto. Modern Information Retrieval, 1999

Information retrieval definitions II

- "Information retrieval studies the retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a user information need usually expressed in natural language."

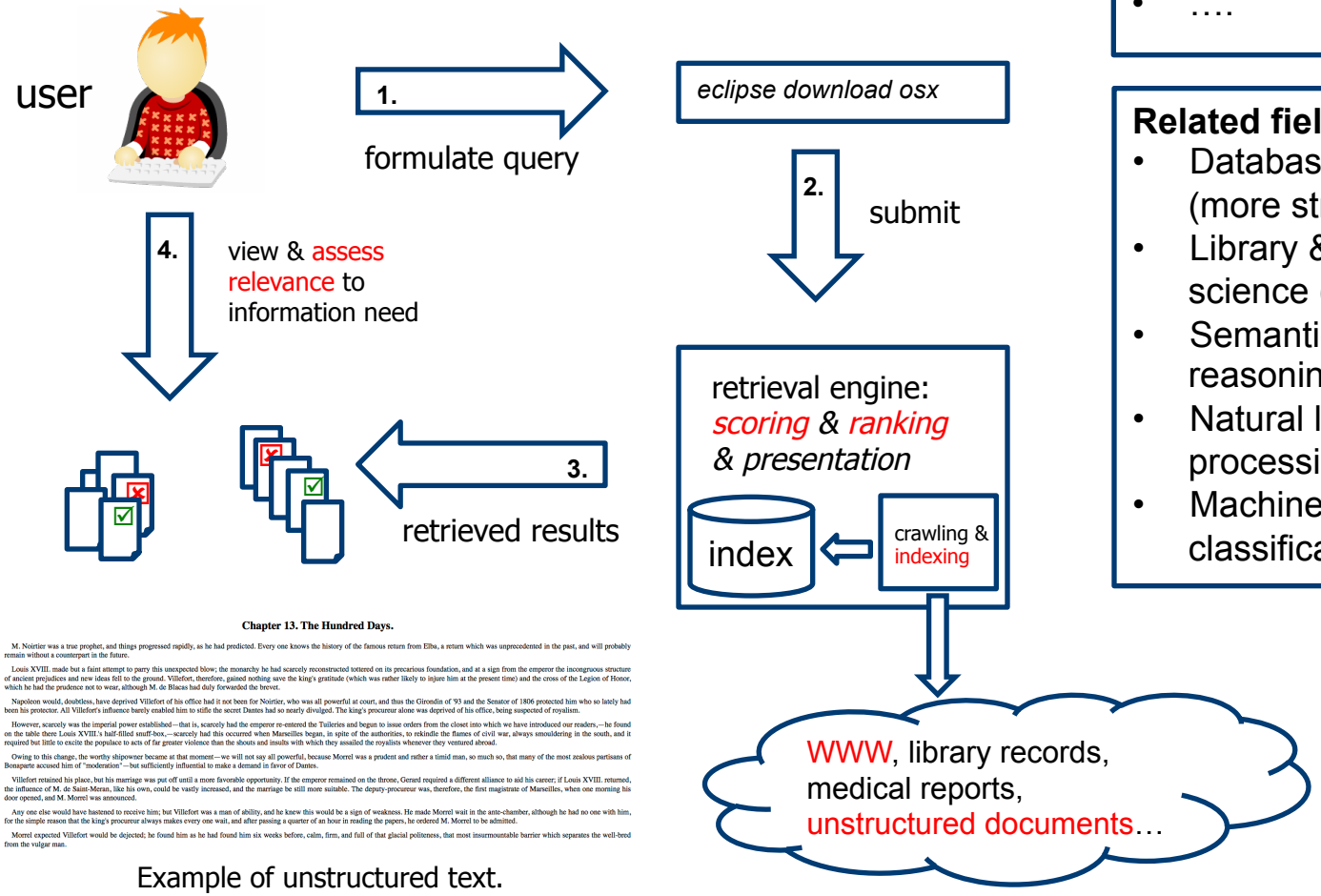
  Baeza-Yates and Ribeiro-Neto. Modern Information Retrieval, 1999

- "IR has traditionally been concerned primarily with the process of representation of texts and queries, and a comparison of these representations. […] it is becoming increasingly evident that IR is an inherently interactive process […] This means, in particular, that supporting and taking advantage of the interaction of the user with the other components of the IR system is crucial for effective IR system design."

  N.J. Belkin and C. Cooll, Cases, Scripts, and Information-Seeking Strategies: On the Design of Interactive Information Retrieval Systems. Expert Systems With Applications, Vol. 9, No. 3, 1995

This course in short (red)

information need: *I am supposed to use Eclipse for the assignments. Where can I download a version for Mac OS X?*



Chapter 13. The Hundred Days.

M. Noirtier was a true prophet, and things progressed rapidly, as he had predicted. Every one knows the history of the famous return from Elba, a return which was unprecedented in the past, and will probably remain without a counterpart in the future.

Louis XVIII. made but a faint attempt to parry this unexpected blow; the monarchy he had scarcely reconstructed tottered on its precarious foundation, and at a sign from the emperor the incongruous structure of ancient prejudices and new ideas fell to the ground. Villefort, therefore, gained nothing save the king's gratitude (which was rather likely to injure him at the present time) and the cross of the Legion of Honor, which he had the prudence not to wear, although M. de Blacas had duly forwarded the brevet.

Napoleon would, doubtless, have deprived Villefort of his office had it not been for Noirtier, who was all powerful at court, and thus the Girondin of '93 and the Senator of 1806 protected him who so lately had been his protector. All Villefort's influence barely enabled him to stifle the secret Dantes had so nearly divulged. The king's procureur alone was deprived of his office, being suspected of royalism.

However, scarcely was the imperial power established—that is, scarcely had the emperor re-entered the Tuileries and begun to issue orders from the closet into which we have introduced our readers,—he found on the table there Louis XVIII.'s half-filled snuff-box,—scarcely had this occurred when Marseilles began, in spite of the authorities, to rekindle the flames of civil war, always smouldering in the south, and it required but little to excite the populace to acts of far greater violence than the shouts and insults with which they assailed the royalists whenever they ventured abroad.

Owing to this change, the worthy shipowner became at that moment—we will not say all powerful, because Morrel was a prudent and rather a timid man, so much so, that many of the most zealous partisans of Bonaparte accused him of "moderation"—but sufficiently influential to make a demand in favor of Dantes.

Villefort retained his place, but his marriage was put off until a more favorable opportunity. If the emperor remained on the throne, Gerard required a different alliance to aid his career; if Louis XVIII. returned, the influence of M. de Saint-Meran, like his own, could be vastly increased, and the marriage be still more suitable. The deputy-procureur was, therefore, the first magistrate of Marseilles, when one morning his door opened, and M. Morrel was announced.

Any one else would have hastened to receive him; but Villefort was a man of ability, and he knew this would be a sign of weakness. He made Morrel wait in the ante-chamber, although he had no one with him, for the simple reason that the king's procureur always makes every one wait, and after passing a quarter of an hour in reading the papers, he ordered M. Morrel to be admitted.

Morrel expected Villefort would be dejected; he found him as he had found him six weeks before, calm, firm, and full of that glacial politeness, that most insurmountable barrier which separates the well-bred from the vulgar man.

Example of unstructured text.

Not covered [for now]:

- result presentation (clustering vs. ranked list, snippets, interfaces)
- crawling
- cross-lingual IR
- ….

Related fields:

- Database management (more structured data)
- Library & information science (categorization)
- Semantic web (logic, reasoning)
- Natural language processing
- Machine learning (text classification & mining)

Why so complicated?

- Searching for the lines in the book *Count of Monte Christo* that contain the terms *Dantes* AND *prison* but NOT *Albert*

- Naïve solution
  - Grep all lines that contain *Dantes*, then grep those containing *prison* and finally strip out lines containing *Albert*

```
more countOfMonteChristo.txt|grep Dantes|grep prison|grep -v Villefort
```

- Problems
  - Proximity operations not easy to implement (e.g. *Dantes* within max. 3 terms of *prison*)
  - *Set of matching results (yes/no decision)*
  - What about approximate/semantic matches (Edmond instead of Dantes, cell instead of prison. etc.)

Why so complicated?

- Searching for the lines in the book *Count of Monte Christo* that contain the terms *Dantes* AND *prison* but NOT *Albert*

- Naïve solution
  - Grep all lines that contain *Dantes*, then grep those containing *prison* and finally strip out lines containing *Albert*

```
more countOfMonteChristo.txt|grep Dantes|grep prison|grep -v Villefort
```

Elaborate queries require the user to anticipate possibly used terms:
(*Edmond* OR *Dantes* OR Monte-Cristo) AND (*prison* OR *cell* OR *imprisoned*) NOT *Albert*

nent (e.g. *Dantes* within

n)

  - What about approximate/semantic matches (Edmond instead of Dantes, cell instead of prison. etc.)

Why so complicated? II

- What about using a *term-document* matrix?
  - Here: a line is a document

	L1	L2	L3	L4	L5	L6	L7	L8	L9
Edmond	0	0	0	1	1	0	1	0	1
Dantes	1	0	1	0	1	0	1	0	0
Monte-Cristo	0	1	0	0	0	1	0	0	0
prison	1	0	0	1	0	0	1	1	0
cell	0	1	1	0	0	0	1	0	0
imprisoned	1	0	0	0	0	0	1	0	0
Albert	1	1	0	0	1	0	0	1	0

1 if *Edmond* occurs in line L9; 0 otherwise

Only feasible for extremely small corpora. The matrix gets too large too quickly.
Still no ranked retrieval.

Dantes AND *prison* ⇨ bitwise AND

```
1 0 1 0 1 0 1 0 0
1 0 0 1 0 0 1 1 0
-----------------
1 0 0 0 0 0 1 0 0
```

⇨ L1 L7

Dantes NOT *Albert* ⇨ bitwise AND complement

```
1 0 1 0 1 0 1 0 0
0 0 1 1 0 1 1 0 1
-----------------
0 0 1 0 0 0 1 0 0
```

⇨ L3 L7

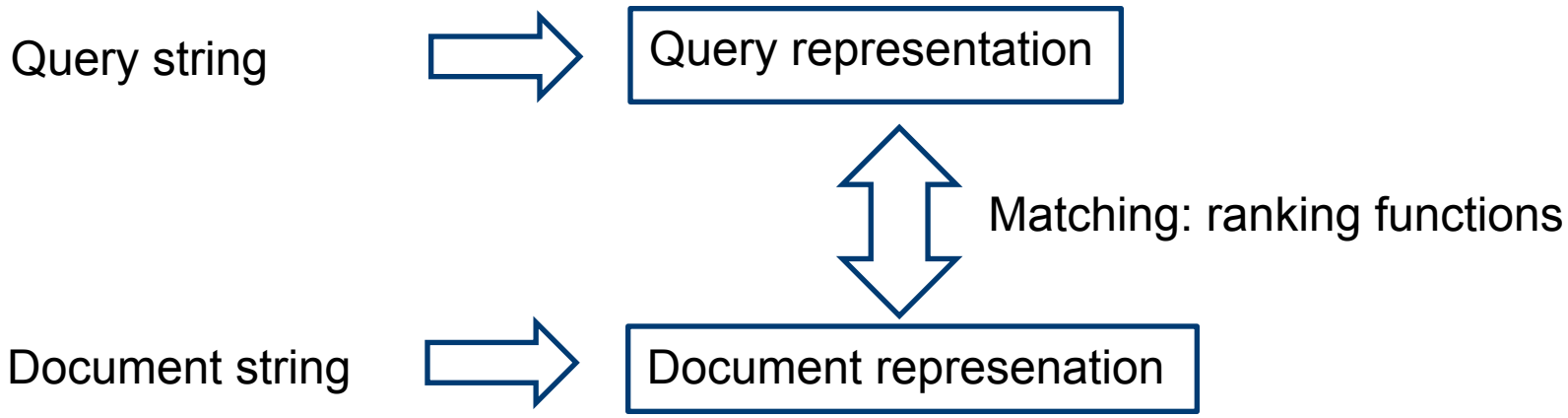
Too many vs. too few

~50,000 lines of text in the book

Boolean queries	#Lines retrieved
Edmond OR *Dantes* OR *Monte-Cristo*	1995
Edmond OR *Dantes* OR *Monte-Cristo* NOT *Albert*	1976
prison OR *cell* OR *imprisoned*	587
prison	240
(*Edmond* OR *Dantes* OR *Monte-Cristo*) AND (*prison* OR *cell* OR *imprisoned*)	22
Edmond AND *prison*	3

Why so complicated? III

Since simple string matching is not enough, we model ...

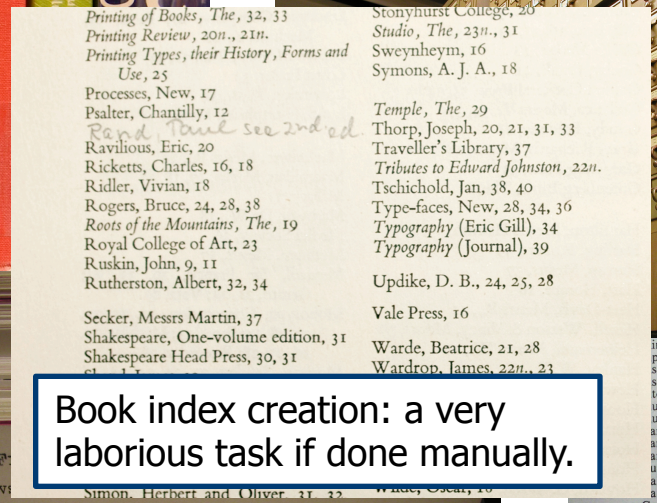


A little bit of IR history

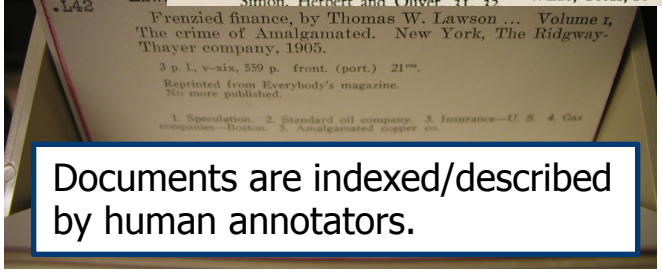
Information retrieval roots

Library classification models and card catalogues (1950's)

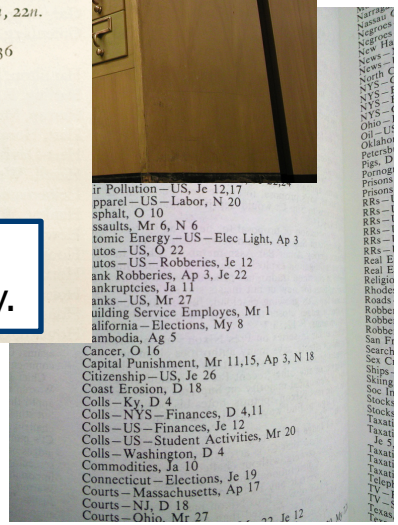
Library of Congress Classification



Book index creation: a very laborious task if done manually.



Documents are indexed/described by human annotators.



NY Times Print Index

The goal of classification schemes (many exist) is to remove the ambiguity of natural language.

Universal decimal classification

expand all | **collapse all**

- **0** SCIENCE AND KNOWLEDGE. ORGANIZATION. COMPUTER SCIENCE. INFO
  - **00** Prolegomena. Fundamentals of knowledge and culture. Propaedeutics
    - **001** Science and knowledge in general. Organization of intellectual w
      - **001.1** Concepts of science and knowledge
        - **001.18** Future of knowledge
      - **001.32** Learned, scientific societies. Academies
      - **001.8** Methodology
        - **001.89** Organization of science and scientific work
      - **001.9** Dissemination of ideas
    - **002** Documentation. Books. Writings. Authorship
    - **003** Writing systems and scripts
    - **004** Computer science and technology. Computing. Data processing
      - **004.01/.08** Special auxiliary numbers for computing
      - **004.2** Computer architecture
      - **004.3** Computer hardware
      - **004.4** Software
      - **004.5** Human-computer interaction. Man-machine interface. User
      - **004.6** Data
      - **004.7** Computer communication. Computer networks
      - **004.8** Artificial intelligence
      - **004.9** Application-oriented computer-based techniques
    - **005** Management
    - **006** Standardization of products, operations, weights, measures and t
    - **007** Activity and organizing. Communication and control theory gener
    - **008** Civilization. Culture. Progress
  - **01** Bibliography and bibliographies. Catalogues
  - **02** Librarianship

004	Computer science and technology. Computing. Data processing
004.2	**Computer architecture**
004.22	Data representation
004.23	Instruction set architecture
004.25	Memory system
004.27	Advanced architectures. Non-Von Neumann architectures

http://www.udcc.org

Vannevar Bush (1890-1974)

Early visions of the Web in 1945

"Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, "**memex**" will do. […] a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. **It is an enlarged intimate supplement to his memory**. […] It affords an immediate step, however, to associative indexing, the basic idea of which is a provision whereby **any item may be caused at will to select immediately and automatically another**. Thereafter, at any time, when one of these items is in view, the other can be instantly recalled merely by tapping a button below the corresponding code space. […] It is exactly as though the physical items had been gathered together from widely separated sources and bound together to form a new book. It is more than this, for **any item can be joined into numerous trails**. […] **Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them**. […] There is a new profession of **trail blazers**, those who find delight in the task of establishing useful trails through the enormous mass of the common record."

hyperlinks

Wikipedians

Medlars

Launched in 1964

- Medical Literature Analysis and Retrieval Service: computer-based retrieval system for medical research literature
  - Searching based on human-assigned indexing terms
  - Boolean logic queries (*cancer AND mouse*)
  - Search process:
    1. User (e.g. medical researcher) corresponds with **expert searcher** via mail, phone or face-to-face [delegated search vs. end-user search]
    2. Expert searcher formulates query
    3. Batch run of queries overnight on dedicated machine
    4. Result printouts posted to the user

In general: librarians, paralegals, patent officers, etc.

Querying the system could take **days or weeks**.

NCBI Resources How To My NCBI Sign In

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | Newcastle | Search

Help

- newcastle
- newcastle disease
- newcastle disease virus
- newcastle-ottawa
- newcastle-ottawa scale
- newcastle university
- newcastle disease virus cancer
- newcastle disease vaccine
- newcastle vaccine

Turn off

Display Settings: Abstract

Send to:

IJMR Free Full Text

Indian J Med Res. 2009 Nov;130(5):507-13.

Newcastle disease virus as an on

Ravindra PV, Tiwari AK, Sharma B, Chauhan RS.
Molecular Biology Laboratory, Division of Animal Biotechno

Abstract

Cancer is a major cause of deaths in humans. The ities of current chemo- and radiotherapies have provided ses for the treatment of cancer termed, oncolytic virotherapy has recently emerged. Newcastle disease virus (NDV) is one such virus with an inherent oncolytic property. NDV causes a highly infectious disease in poultry worldwide. In humans it is reported to have oncolytic and immuno-stimulatory effects. It specifically replicates in tumour cells while sparing normal cells and cause oncolysis. For many years different strains of the NDV have been investigated for treatment of various human cancers. Recent advances in reverse genetics provided investigators the tools to produce recombinant NDV with improved oncolytic property.

PMID: 20090097 [PubMed - indexed for MEDLINE] Free full text

Publication Types, MeSH Terms

Publication Types

Research Support, Non-U.S. Gov't
Review

MeSH Terms

Animals
Apoptosis
Humans
Neoplasms/pathology
Neoplasms/therapy*
Newcastle disease virus/genetics
Newcastle disease virus/physiology*
Oncolytic Virotherapy/methods*
Oncolytic Viruses/genetics
Oncolytic Viruses/physiology

Manually added by experts! Even today.

LinkOut - more resources

Related citations

Review [Progress in using Newcastle disease virus for [Sheng Wu Gong Cheng Xue Bao. 2010]

Type I interferon-sensitive recombinant newcastle disease virus for oncolyti [J Virol. 2010]

Generation of a recombinant oncolytic Newcastle disease virus and expression c [Gene Ther. 2008]

Analysis of three properties of Newcastle disease virus for fighting cancer: [Methods Mol Biol. 2012]

Review Newcastle disease virus (NDV): brief history of its oncolytic strains. [J Clin Virol. 2000]

See reviews...

See all...

Cited by 1 PubMed Central article

Cytolytic replication of echoviruses in colon cancer cell lines. [Virol J. 2011]

Related information

Related Citations

Cited in PMC

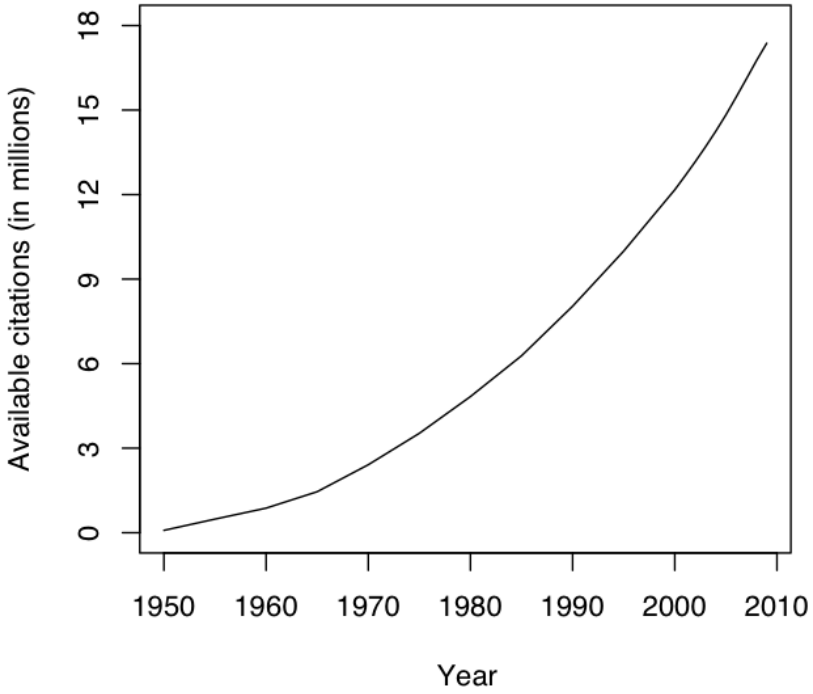
Recent activity

Turn Off Clear

- Newcastle disease virus as an oncolytic agent. PubMed
- Clinical Evidence for the Regression of Liver Fibrosis. PubMed
- Non-drug therapies for lower limb muscle cramps. PubMed
- Variability among the neuraminidase, non-structural 1 and PB1-F2 proteins i PubMed

They like to publish ...

- MEDLINE: part of PubMed
  - Bibliographic database
  - Exponential growth
  - 2010 statistics
    - 18 million citations
    - 5,516 journals
    - 700,000 additions



Slide provided by Dolf Trieschnigg, University of Twente.

SMART project

1960s-1990s

- Gerard Salton (1927-1995)
  - "the man most responsible for the establishment, survival, and recognition of Information Retrieval" [8]
- Developed pioneering solutions that are still in use today
  - Fully automatic indexing of document texts
  - Scoring functions to determine how well the query and document match
  - The concept of result ranking (vs. result set)
  - Relevance feedback (exploit known relevant documents)
- Scoring & ranking were added to commercial systems in the late 1980s and became ubiquitous with web search engines in the 1990s

SMART project

1960s-1990s

Computing power & costs: in **1973** one 'run' on the **1400-document** Cranfield collection took **11.2** minutes in processing time and it cost **$86.22** [1].

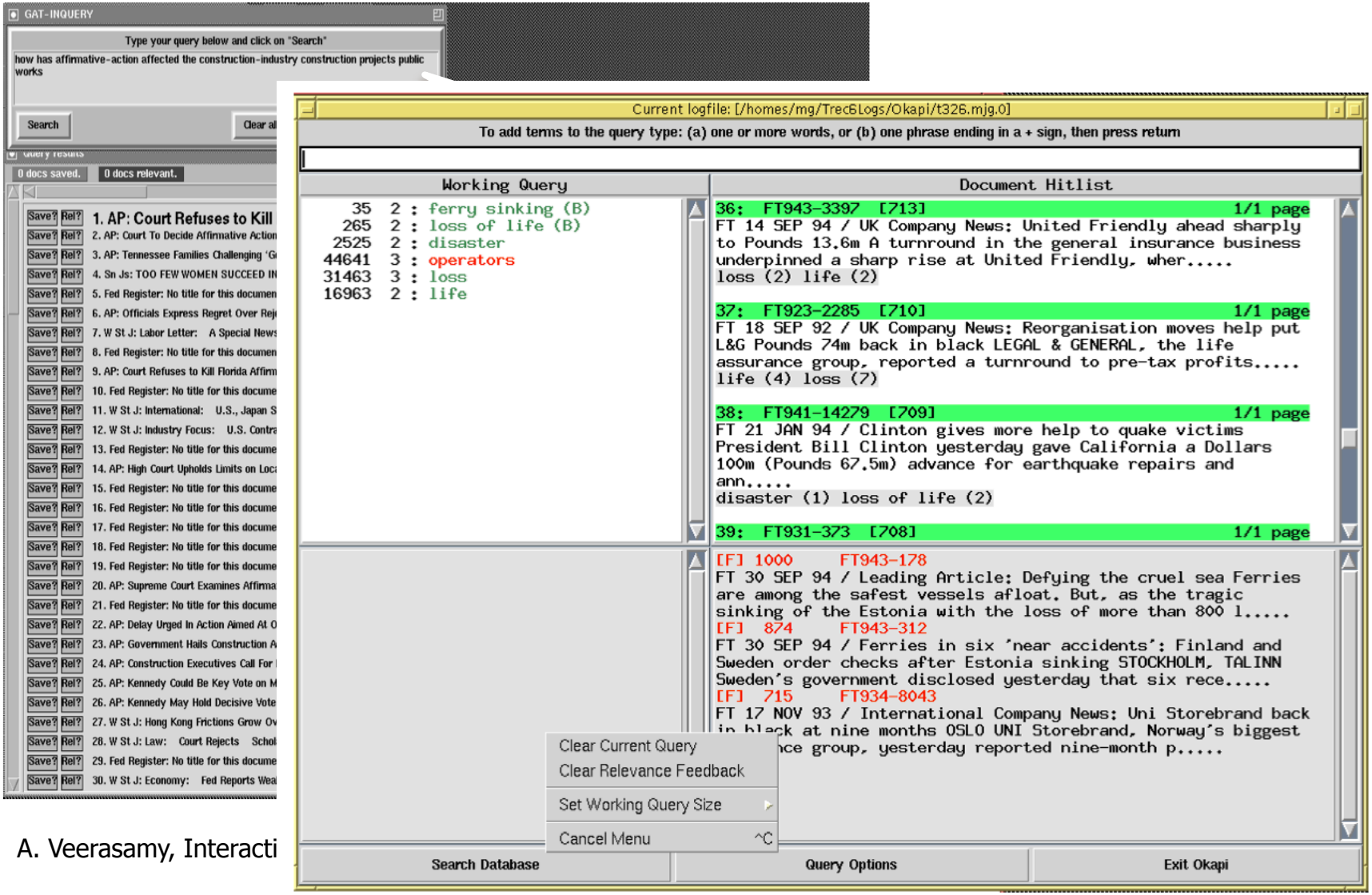
- Gerard Salton (1927-1995)
  - "the man most responsible for the establishment, survival, and recognition of Information Retrieval" [8]
- Developed pioneering solutions that are still in use today
  - Fully automatic indexing of document texts
  - Scoring functions to determine how well the query and document match
  - The concept of result ranking (vs. result set)
  - Relevance feedback (exploit known relevant documents)
- Scoring & ranking were 1980s and became ubic 1990s

"… it is nice to be able, in England, to search a file in California of **five hundred thousand documents** using an arbitrarily complex search specification, and get answers back in **six seconds**."
(*Karen* Spärck *Jones*, **1979**)

Okapi, BM25 and Okapi BM25

1990s-2000s

- Stephen Robertson & collaborators (London City University)
- Okapi is a retrieval system with ranking function BM25
- Focus on weighting terms, exploiting user feedback and user-system interaction



A. Veerasamy, Interacti

M.M. Beaulieu and M.J. Gatford, Interactive Okapi at TREC-6, 1997

TREC

Text REtrieval Conference (1992-*)

http://trec.nist.gov/ http://trecvid.nist.gov/

- Conducted by the US National Institute of Standards and Technology, co-sponsored by DARPA
- Several "tracks" per year (a good way to learn about current work)

Ad-hoc retrieval task (1992)
New queries over the same data.

Routing task (1992)
The same queries over new data.

Interactive track (1995)

Multilingual track (1995)
Non-English texts.

Database merging track (1995)

Confusion track (1995)
OCR simulation

Cross-Language track (1997)
Documents of languages A & B are searched with queries of language C.

Spoken document track (1997)

Question Answering track (1999)

Web track (1999)

Video track (2001)
Now split from TREC: TRECVID.

Novelty track (2002)
Find the relevant and novel docs.

Genomics track (2003)

Terabyte track (2004)

Enterprise track (2005)
Enterprise data (intranet, email).

Spam track (2005)

Blog track (2006)

Legal track (2006)
Business records.

Million query track (2007)

Chemical IR track (2009)
Search for chemical patents and research articles.

Entity track (2009)
Search for oganizations, people, products.

Microblog track (2011)

Benchmarks are very important to IR

The largest ongoing benchmarks apart from TREC/TRECVID

- CLEF
  - Conference and Labs of the Evaluation Forum
  - http://www.clef-initiative.eu/
- MediaEval
  - Benchmarking Initiative for Multimedia Evaluation
  - http://www.multimediaeval.org/
- NTCIR
  - NII Test Collection for IR Systems
  - http://research.nii.ac.jp/ntcir/index-en.html
- FIRE
  - Forum for Information Retrieval Evaluation
  - http://www.isical.ac.in/~clia/

Current systems and technologies

Current systems and technologies I

Auto-completion

Google

dam square amsterdam

dam square amsterdam
dam square amsterdam **hotels**
dam square amsterdam **address**
dam square amsterdam **the netherlands**

Search

Everything
Images
Maps
Videos
News
Shopping
More

All results
Sites with images
More search tools

Dam Square
www.dutchsouvenirs.com/
3 Google reviews - Write a review

Dam Square 17 1012 JS Amsterdam, Netherlands
020 6203432

Dam Square - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Dam_Square**
Dam Square, or simply the Dam (Dutch: de Dam) is a town square in **Amsterdam**, the capital of the Netherlands. Its notable buildings and frequent events make it ...
Location and description - History - Dam Square shooting, 1945 - Present

Amsterdam.info - **Dam Square Amsterdam** hotels
www.amsterdam.info › Sights › Squares
The **Dam square** is the very central square of **Amsterdam**.

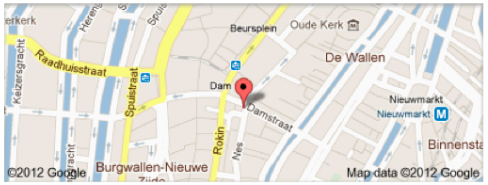
Dam Square, Amsterdam - Things to Do - VirtualTourist
www.virtualtourist.com › ... › Amsterdam › Things to Do
Dam Square reviews and photos from real travelers and locals in **Amsterdam**, Netherlands.

Dam Square - Amsterdam - Reviews of Dam Square - TripAdvisor
www.tripadvisor.com/Attraction_Review-g188590-d189381-Review...
84 reviews
Dam Square, **Amsterdam**: See 84 reviews, articles, and 110 photos of Dam Square, ranked No.87 on TripAdvisor among 458 attractions in Amsterdam.

Dam Square Amsterdam, Netherlands
www.360cities.net › The World › Europe › Netherlands › Amsterdam
10 Jan 2006 – 360° panoramic photography by Jook Leung | 360VR Images. Visit us to see more amazing panoramas from **Amsterdam** and thousands of ...

Query intent

Dam Square - Directions



Details

Transit: Dam

More reviews

virtualtourist.com (38)
flipkey.com (39)
tripadvisor.co.uk (40)
tripadvisor.com (11)

Where is the spam?

Ads adaptation (none shown)

Current systems and technologies II

Google

flight amsterdam dublin

Search

About 9,440,000 results (0.35 seconds)

~9 million results: ranking is of utmost importance

Everything
Images
Maps
Videos
News
Shopping
More

Show search tools

Ads - Why these ads?

Amsterdam - Dublin? - Goedkoop Vliegen naar Dublin?
www.cheaptickets.nl/**Dublin**
Boek Nu Online: **Amsterdam - Dublin**!
Vertrouwd boeken - Allerlaagste tarieven - Goedkope Citybreaks

Greatest Offers on - **Amsterdam-Dublin Flights** | eDreams.com
www.edreams.com/**Amsterdam_Dublin**
Book Now -Limited Seats Available!
► Early 2012 Flight Specials - Amsterdam-London from €20

Flight Amsterdam Dublin - Op 1 website alle vliegtickets
www.worldticketcenter.nl/AMS-DUB
Goedkoop en snel geregeld

Flights from **Amsterdam, The Netherlands** (AMS) to **Dublin, Ireland**...
Non-stop flights: 5 per day, 1h 40m duration
Airlines: Aer Lingus
+ Schedule of non-stop flights

Cheap **Flights** - To and from **Dublin**, Europe, UK & USA - Aer Lingus
www.aerlingus.com/
Book cheap **flights** online today with Aer Lingus. Fly to Ireland, the UK, Europe and N. America including Canada with us as well as find hotels and more.
Book Flights - Web-Check-in - Manage Booking - Travel Information

Cheap **flights** from **Dublin** to **Amsterdam** (Netherlands)
www.cheap**flights**.co.uk/**flights**/**Amsterdam**/**Dublin**/
Dublin to **Amsterdam flights**. Search and compare cheap **flights** from **Dublin** (Ireland) to **Amsterdam** (Netherlands) to find the latest deals from all major airlines ...

Ads - Why these ads?

Find Cheap Airfares Fast
www.kayak.com/**Flights**
kayak.com is rated ★★★★½
100s Of **Flights** On One Simple Site.
Compare Many Flying Options At Once

Flight Amsterdam Dublin
www.schipholtickets.nl/Ierland
Laatste stoelen met schérpe prijzen
Vergelijk & Bespaar snel online!

Egypt Airlines
www.egyptair.com/Egypt
Discover Luxury in the clouds
And enjoy our superior service

Amsterdam Dublin?
www.ebookers.nl/**Dublin**
Vliegticket Schiphol - **Dublin**?
Géén verborgen kosten. Boek nu!

Van **Amsterdam** naar **Dublin**
amsterdam-dublin.vlucht24.nl
Boek nu snel **Amsterdam** - **Dublin**.
Vliegtickets Vinden op Vlucht 24!

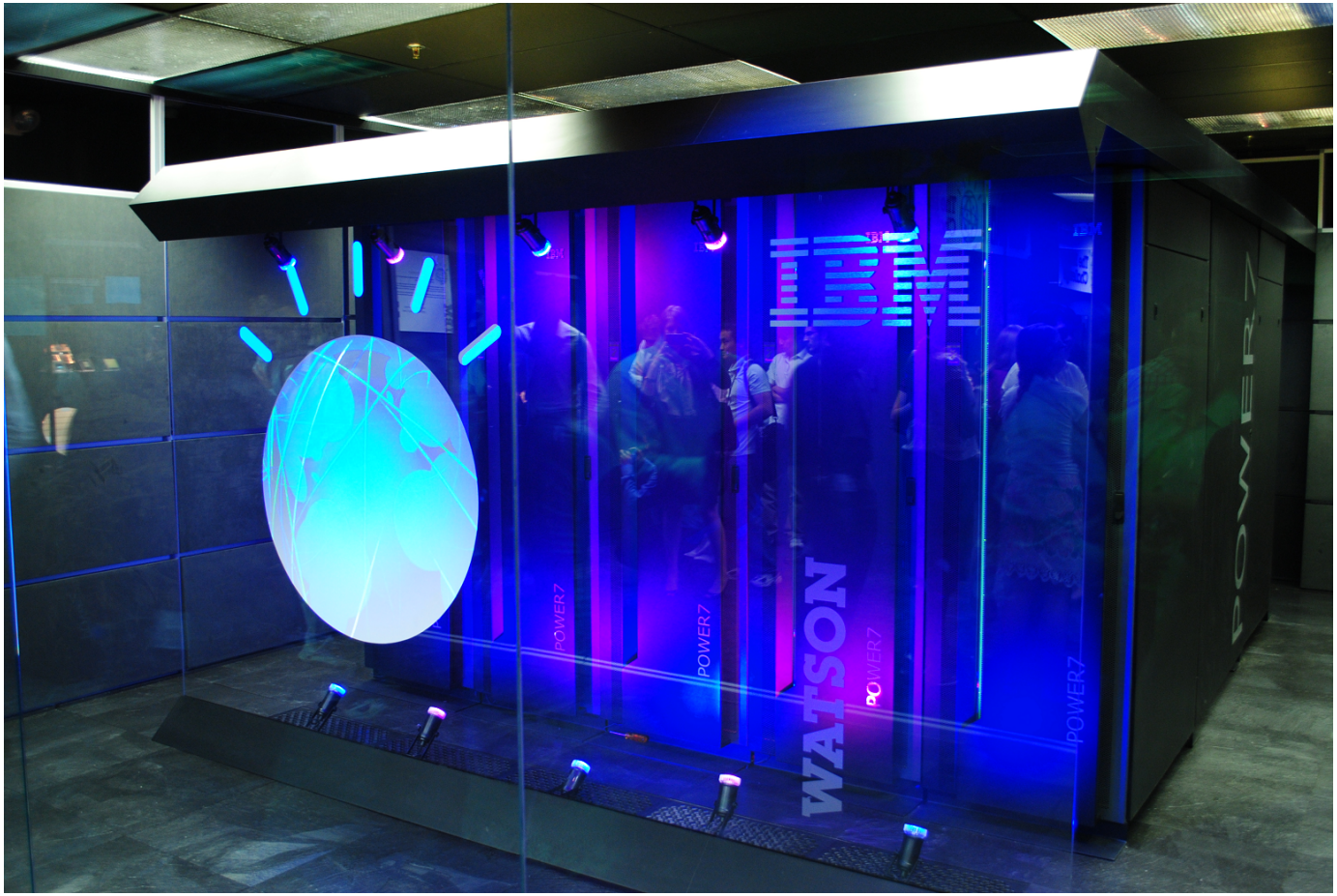
Flights to **Amsterdam**
www.vliegwinkel.nl/**Amsterdam**
All **flights** & airlines to **Amsterdam**
Compare, book and save money now!

Query intent & vertical search

Ads adaptation

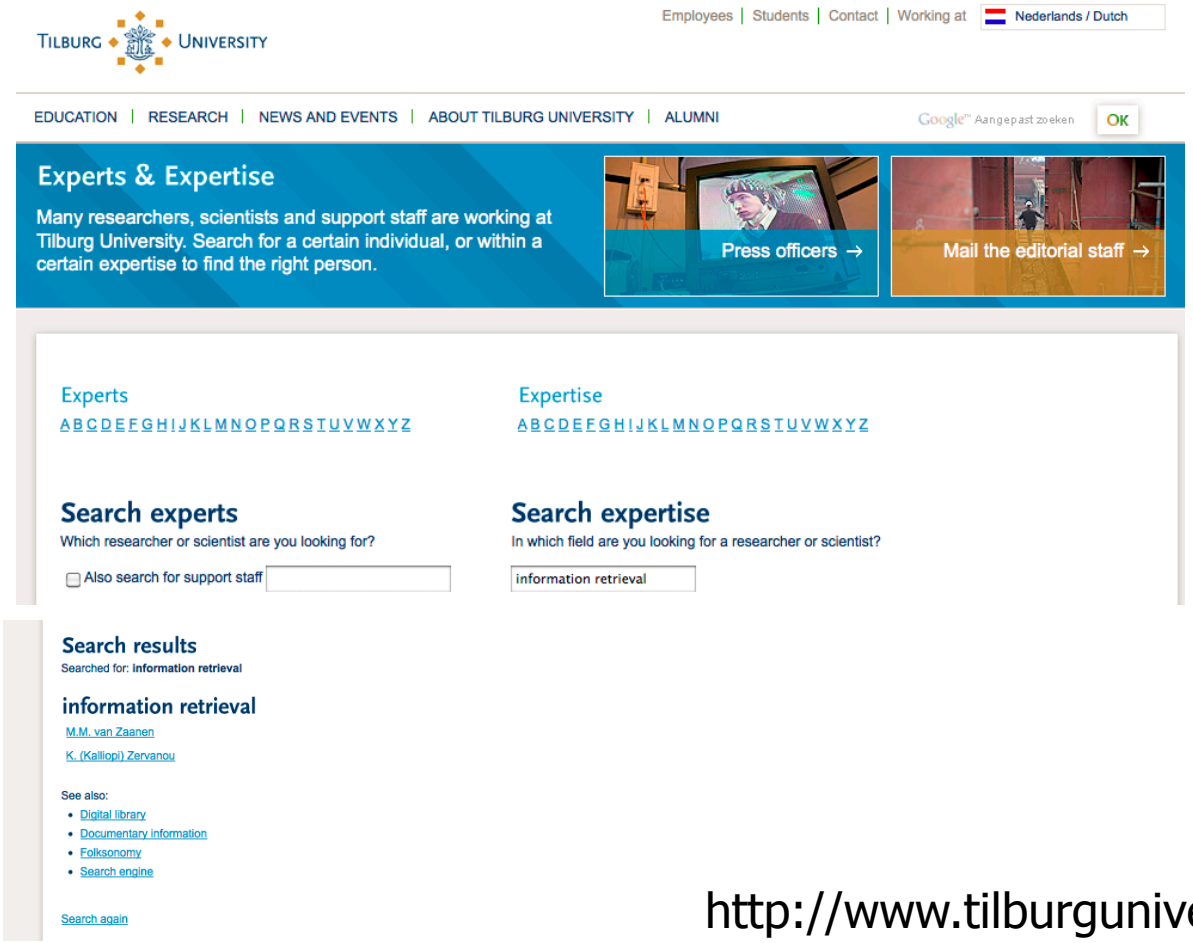
Current systems and technologies III

Searching for answers (not documents): IBM Watson @ Jeopardy!



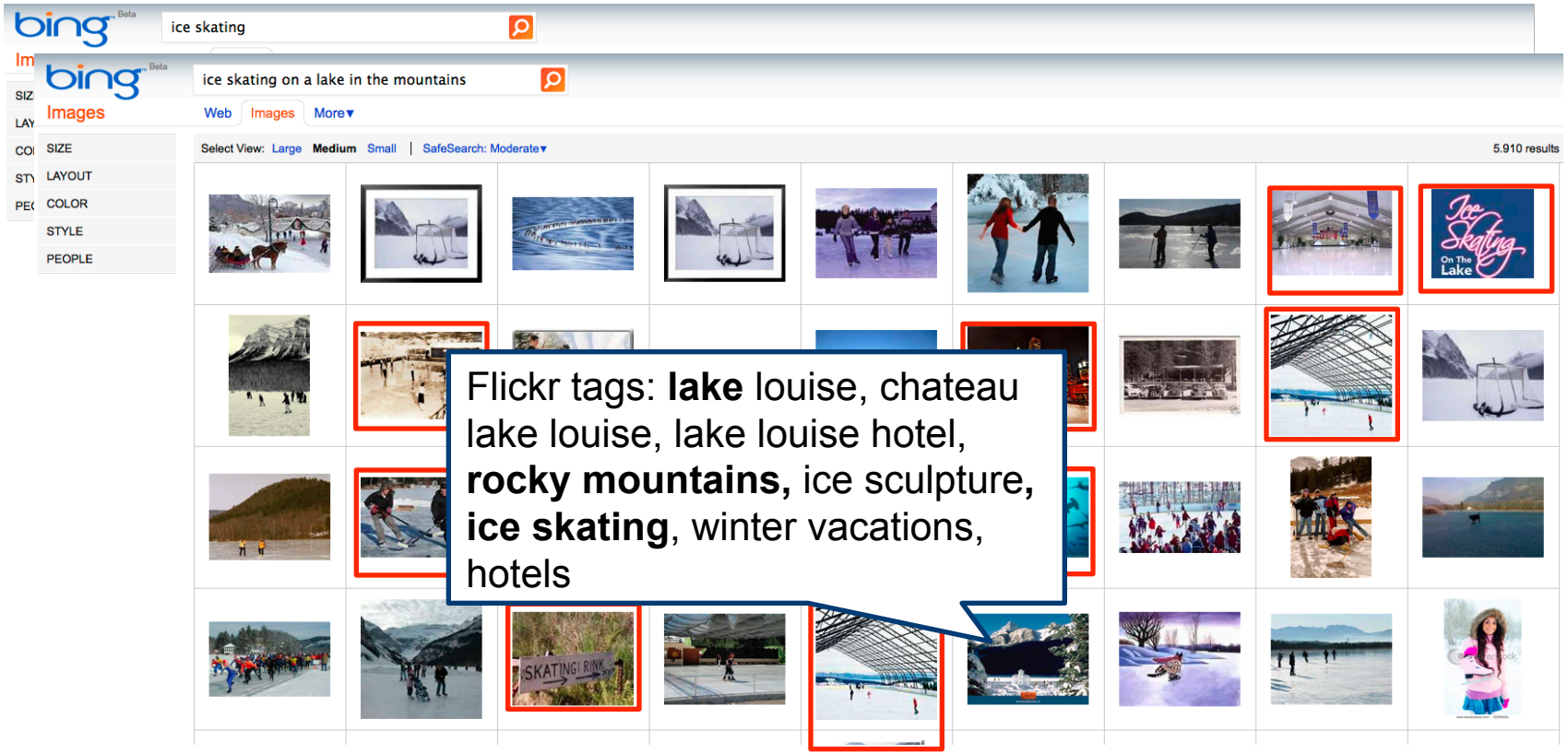
Current systems and technologies IV

Searching for expertise (not documents)



http://www.tilburguniversity.edu/webwijs/

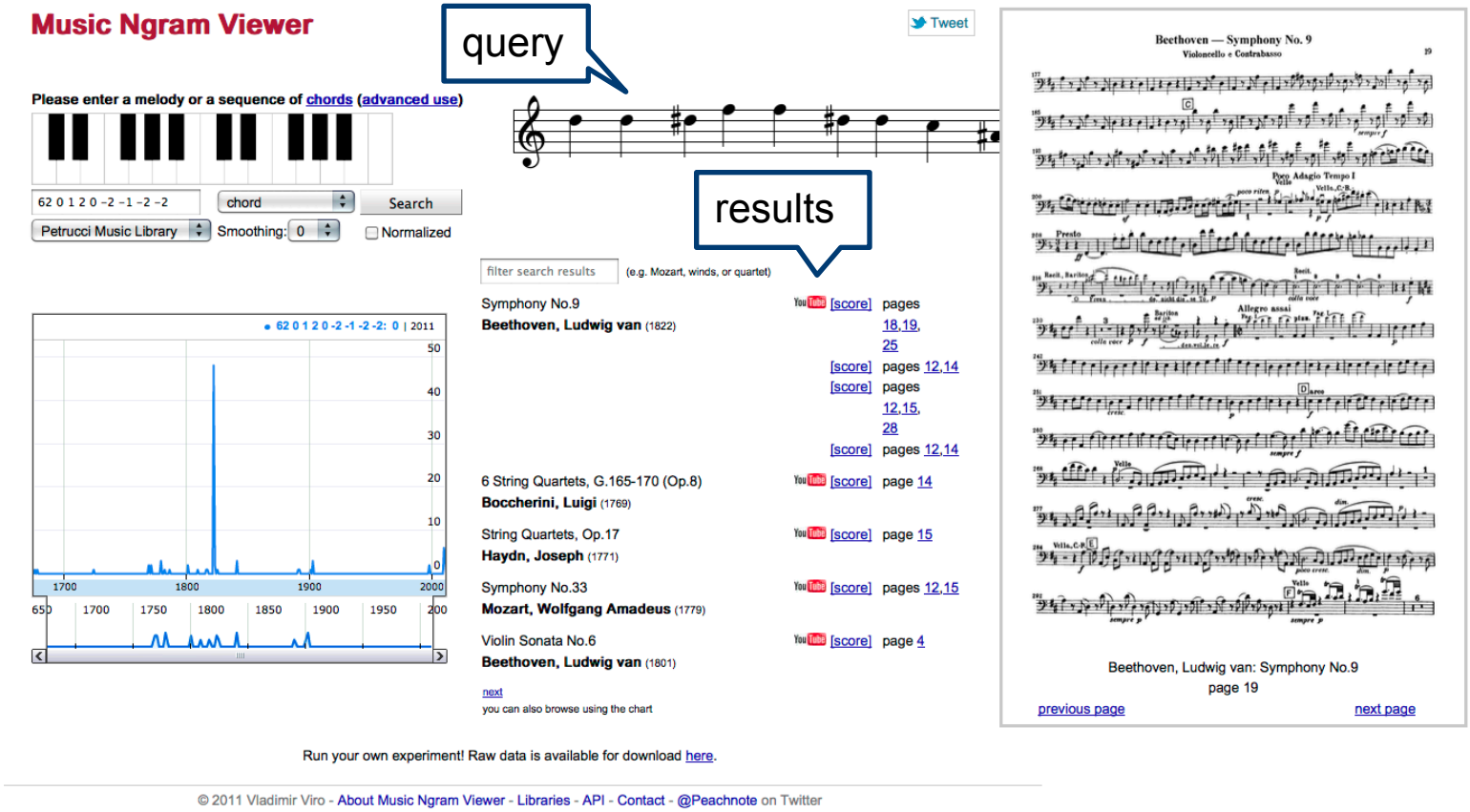
Current systems and technologies V



Flickr tags: **lake** louise, chateau lake louise, lake louise hotel, **rocky mountains,** ice sculpture**, ice skating**, winter vacations, hotels

Current systems and technologies VI

Music retrieval



http://www.peachnote.com

A few questions for you

- Who has programming experience in **Java**?
- Who is familiar with **functional programming**?
- Who has worked with **Hadoop**?
- Who has worked with **Amazon Web Services**?

Big data

What is 'Big data'?

- "Big data refers to enormous amounts of unstructured data produced by high-performance applications" [1]
  - Scientific computing applications
  - Social networks
  - E-government applications
  - Medical information systems

- Issues
  - Scalability
  - Heterogeneity
  - Data analysis

How big is big?

- Data repositories will only grow bigger with time.
- More data usually translates into more effective algorithms.

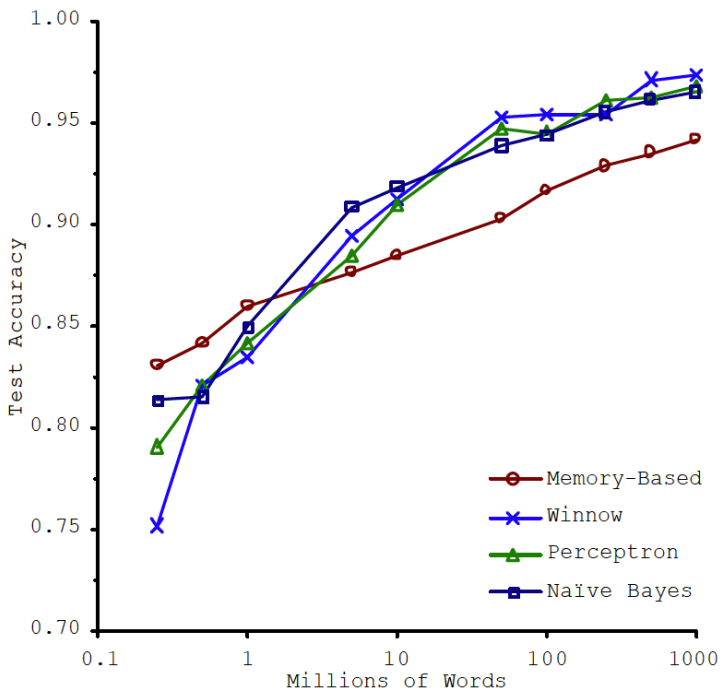
- YouTube: 4 billion views a day, one hour of video upload per second
- Facebook: 483 million daily active users (December 2011)
- Twitter: 140 million tweets on average per day (March 2011)
- Google: >1 billion searches per day (March 2011)

- Google processed 100 TB of data per day in 2004 and 20 PB data per day in 2008
- Large Hadron Collider at CERN: when fully functional, it will generate 15 petabytes of data per year
- Internet Archive: contains 2 petabytes of data, grows 20 terabytes per month (2011)

The more data, the better

The unreasonable effectiveness of data. A. Halevy, P. Norvig and F. Pereira, 2009.

"So, follow the data."



Confusion set disambiguation

- *then* vs. *than*
- *to* vs. *two* vs. *too*
- ....

Scaling to very very large corpora for natural language disambiguation. M. Banko and E. Brill, 2001.

Cloud computing

- "Anything running inside a browser that gathers and stores user-generated content" (Jimmy Lin)
- Utility computing
  - A computing resour
  - A "cloud user" buys "cloud provider" (pa
    - Virtual machine ir
  - IaaS: infrastructure
  - Amazon Web Servic dominant provider

Region: EU (Ireland)

	Linux/UNIX Usage	Windows Usage
Standard On-Demand Instances		
Small (Default)	$0.095 per hour	$0.12 per hour
Large	$0.38 per hour	$0.48 per hour
Extra Large	$0.76 per hour	$0.96 per hour
Micro On-Demand Instances		
Micro	$0.025 per hour	$0.035 per hour
Hi-Memory On-Demand Instances		
Extra Large	$0.57 per hour	$0.62 per hour
Double Extra Large	$1.14 per hour	$1.24 per hour
Quadruple Extra Large	$2.28 per hour	$2.48 per hour
Hi-CPU On-Demand Instances		
Medium	$0.19 per hour	$0.29 per hour
Extra Large	$0.76 per hour	$1.16 per hour

Amazon EC2 Pricing, 6th February 2012

Cloud computing

- "Anything running inside a browser that gathers and stores user-generated content" (Jimmy Lin)

- Utility computing
  - A computing resource as a metered service
  - A "cloud user" buys any amount of computing power from a "cloud provider" (pay-per-use)
    - Virtual machine instances
  - IaaS: infrastructure as a service
  - Amazon Web Services (EC2: elastic compute cloud) is the dominant provider

MapReduce

① Programming model for distributed computations on **large-scal**e data, inspired by the functional programming paradigm
② Execution framework for clusters of commodity hardware

- Developed by researchers at Google in 2003
  - Built on principles in parallel and distributed processing

- "MapReduce is used for the generation of data for Google's production web search service, for sorting, for data mining, for machine learning and many other systems" [12]

- Designed for **batch** processing over large data sets

Ideas behind MapReduce I

- Scale "out", not "up"
  - Many commodity servers are more cost effective than few high-end servers
- Assume failures are common
  - A 10,000-server cluster with a mean-time between failures of 1000 days experiences on average 10 failures a day.
- Move processes to the data
  - Moving the data around is expensive
- Process data sequentially and avoid random access
  - Data sets do not fit in memory, disk-based access (slow)

Ideas behind MapReduce II

- Hide system-level details from the application developer
  - Frees the developer to think about the task at hand only (no need to worry about deadlocks, …)
  - MapReduce takes care of the system-level details (separation of what and how to compute)

To be continued in the next lecture ....

Sources

- http://www.flickr.com/photos/annarbor/4349874305
- http://www.flickr.com/photos/annarbor/4349874915
- http://www.flickr.com/photos/ccacnorthlib/4775120660
- http://www.flickr.com/photos/awhitis/4266617452/
- http://www.flickr.com/photos/karenhorton/4893479758
- http://en.wikipedia.org/wiki/File:IBM_Watson.PNG

① On the history of evaluation in IR. Stephen Robertson. 2008
② Information retrieval and digital libraries: lessons of research. Karen Spaerck Jones. 2006.
③ Information retrieval viewed as temporal signaling. Calvin Mooers. 1950.
④ Making information retrieval pay. American Chemical Society 118th Meeting, abstracts of papers
⑤ Information retrieval. Keith van Rijsbergen
⑥ Experimentation as a way of life: Okapi at TREC. S. Robertson, S. Walker and M. Beaulieu, 2000.
⑦ http://www.cs.cornell.edu/Info/Department/Annual96/Beginning/salton.html
⑧ As we may think. Vannevar Bush, Atlantic Monthly, 1945. http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/
⑨ Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. B. Sparrow, J. Liu and M. Wegner, Science, 2011. http://www.sciencemag.org/content/early/2011/07/13/science.1207745