- The following material is the result of a curriculum development effort to provide a set of courses to support bioinformatics efforts involving students from the biological sciences, computer science, and mathematics departments. They have been developed as a part of the NIH funded project "Assisting Bioinformatics Efforts at Minority Schools" (2T36 GM008789). The people involved with the curriculum development effort include:

- Dr. Hugh B. Nicholas, Dr. Troy Wymore, Mr. Alexander Ropelewski and Dr. David Deerfield II, National Resource for Biomedical Supercomputing, Pittsburgh Supercomputing Center, Carnegie Mellon University.
- Dr. Ricardo Gonzalez-Mendez, University of Puerto Rico Medical Sciences Campus.
- Dr. Alade Tokuta, North Carolina Central University.
- Dr. Jaime Seguel and Dr. Bienvenido Velez, University of Puerto Rico at Mayaguez.
- Dr. Satish Bhalla, Johnson C. Smith University.

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

- This material is targeted towards students with a general background in Biology. It was developed to introduce biology students to the computational mathematical and biological issues surrounding bioinformatics. This specific lesson deals with the following fundamental topics:
  - Computing for biologists
  - Computer Science track

  - This material has been developed by:

    Dr. Hugh B. Nicholas, Jr.
    National Center for Biomedical Supercomputing
    Pittsburgh Supercomputing Center
    Carnegie Mellon University

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Bioinformatics Data Management

Lecture 2

Unstructured Data Repositories

Bienvenido Vélez

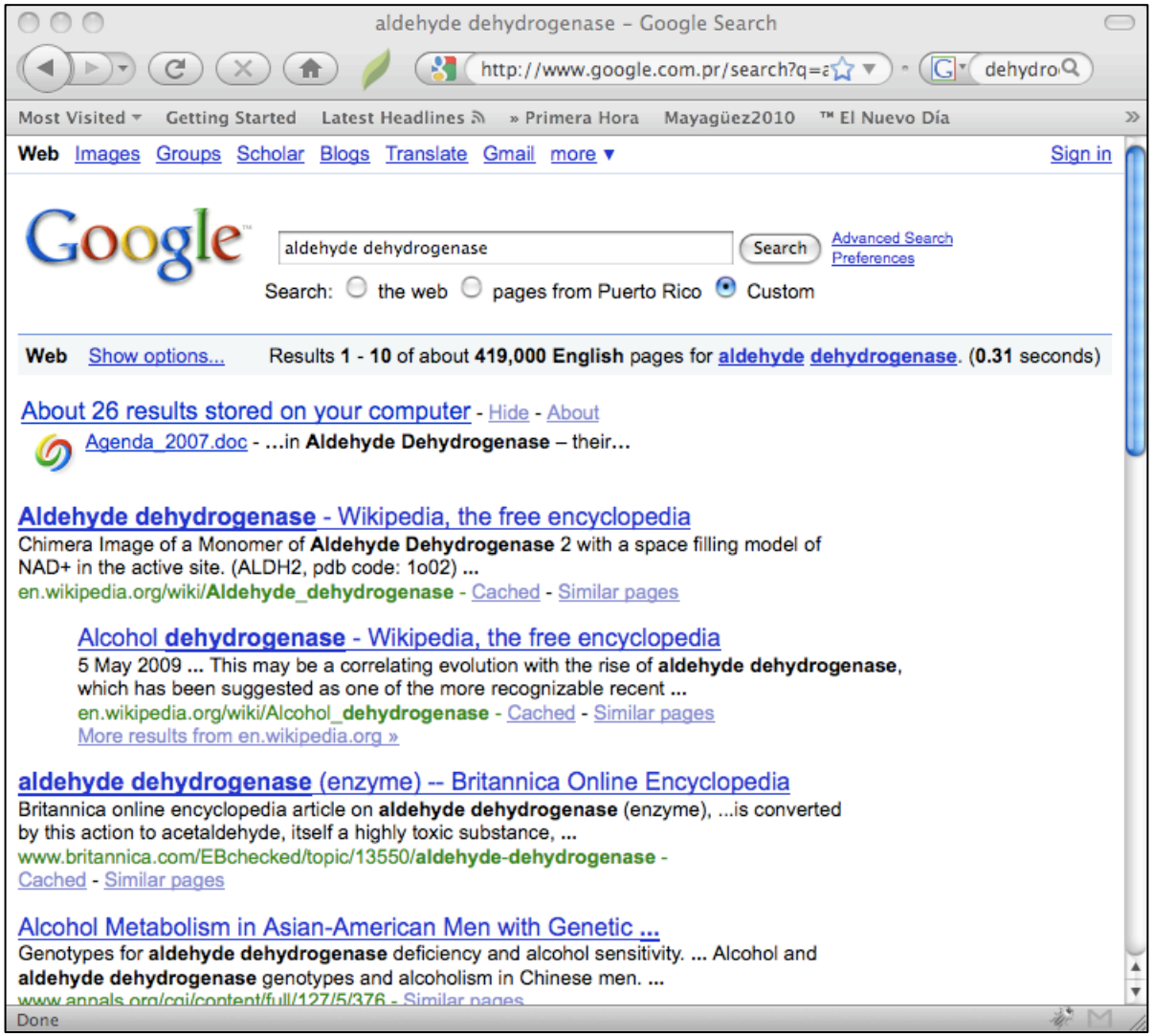
UPR Mayaguez

Reference: BioInformatics for Dummies

Unstructured Data Repositories: Outline

- Introduction and Examples
- Query Models
- Implementation Issues
- References

Introduction and Examples



Basic Paradigm

Find all documents containing some terms and rank them according to expected relevance

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Characteristics of Unstructured Databases

- Contain a collection of generic "documents" (many formats)
- Each document has an associated set of terms
- Queries also consists of sets of terms
- Documents are ranked according to their similarity to the query
- There are lots of ways of measuring query/document similarity

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

PubMed a Domain Specific Unstructured Database

aldehyde dehydrogenase - PubMed Results

http://www.ncbi.nlm.nih.gov/sites/entrez

Most Visited | Getting Started | Latest Headlines | » Primera Hora | Mayagüez2010 | ™ El Nuevo Día | Últimas Noticias | El Vocero | DigiZen: Un blogf... | CaribbeanBusinessP... | KnowledgeTreeLiv...

NCBI

PubMed www.pubmed.gov

A service of the U.S. National Library of Medicine and the National Institutes of Health

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search PubMed for aldehyde dehydrogenase Go Clear Advanced Search Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort By Send to

All: 7331 | Review: 529

Items 1 - 20 of 7331

Page 1 of 367 Next

1: Novel pharmacogenetic markers for treatment outcome in azathioprine-treated inflammatory bowel disease.
Smith M, Marinaki A, Arenas M, Shobowale-Bakre M, Lewis C, Ansari A, Duley J, Sanderson J.
Aliment Pharmacol Ther. 2009 Jun 3. [Epub ahead of print]
PMID: 19500084 [PubMed - as supplied by publisher]
Related Articles

2: Suppressing Glioblastoma Stem Cell Function by Aldehyde Dehydrogenase Inhibition with Chloramphenicol or Disulfiram as a New Treatment Adjunct: An Hypothesis.
Kast RE, Belda-Iniesta C.
Curr Stem Cell Res Ther. 2009 Dec 1. [Epub ahead of print]
PMID: 19500061 [PubMed - as supplied by publisher]
Related Articles

3: The role of aryl hydrocarbon receptor in regulation of enzymes involved in metabolic activation of polycyclic aromatic hydrocarbons in a model of rat liver progenitor cells.
Vondrácek J, Krcmár P, Procházková J, Trilecová L, Gavelová M, Skálová L, Szotáková B, Buncek M, Radilová H, Kozubík A, Machala M.
Chem Biol Interact. 2009 Jul 15;180(2):226-37. Epub 2009 Mar 27.
PMID: 19497421 [PubMed - in process]
Related Articles

4: Small-Molecule Targeting of the Mitochondrial Compartment with an Endogenously Cleaved Reversible Tag.
Ripcke J, Zarse K, Ristow M, Birringer M.
Chembiochem. 2009 Jun 2. [Epub ahead of print]
PMID: 19492396 [PubMed - as supplied by publisher]
Related Articles

Also try:
- aldehyde dehydrogenase 2
- mitochondrial aldehyde dehydrogenase
- betaine aldehyde dehydrogenase
- aldehyde dehydrogenase cancer
- aldehyde dehydrogenase 1

Titles with your search terms
- High aldehyde dehydrogenase and expression of cancer stem cell markers selects for breast [J Cell Mol Med. 2008]
- Aldehyde dehydrogenase discriminates the CD133 liver cancer stem cell populations. [Mol Cancer Res. 2008]
- Activation of aldehyde dehydrogenase-2 reduces ischemic damage to the heart. [Science. 2008]

» See more...

Recent Activity

Turn Off Clear

aldehyde dehydrogenase (7331) PubMed

Done

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Unstructured Data Repositories: Outline

- Introduction and Examples
- Query Models
- Implementation Issues
- References

Query Models Subtopics

- Boolean Query Model
- Vector Space Query Model
- Practical Query Models: The Case of Google

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Boolean Query Model: Definition

- A Boolean query consists of one of the following:
  - A single term
  - The negation (NOT) of a single Boolean query
  - A conjunction (AND) of two Boolean queries
  - A disjunction (OR) of two Boolean queries

Examples of Boolean queries	
aldehyde	Single term
(or aldehyde dehydrogenase)	Disjunction
(and aldehyde dehydrogenase)	Conjunction
(and (and aldehyde dehydrogenase) (not isocitrate))	Compound

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Boolean Query Model: Definition

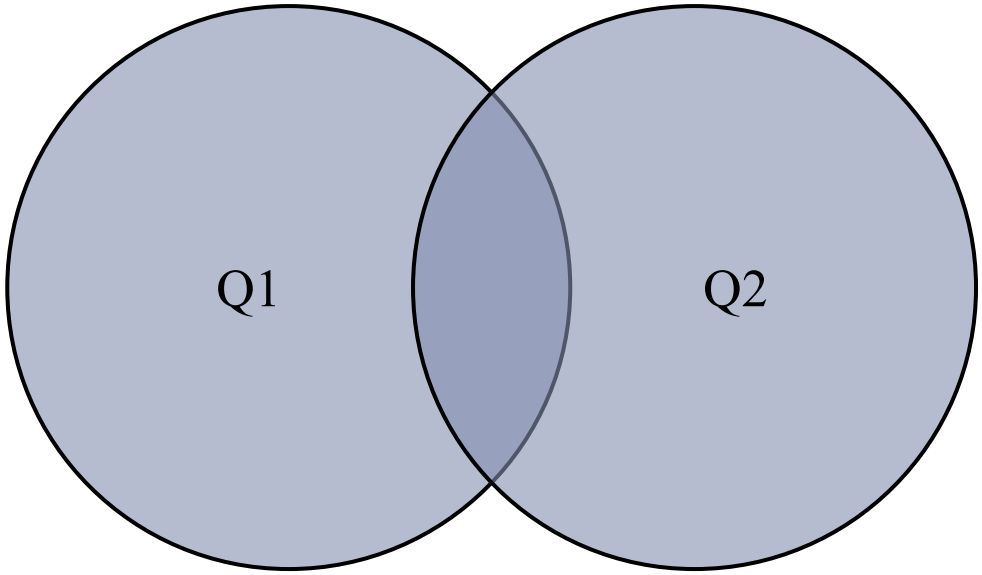
Query consists of:	Document Matches When:
t	Document contains at least one occurrence of term t
(not q)	Document does not match query q
(and q1 q2)	Document matches boths queries q1 and q2
(or q1 q2)	Document matches either query q1 or q2

A strictly Boolean query model does not include a mechanism for ranking matching documents

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Computing Boolean Queries

All documents



and Q1, Q2

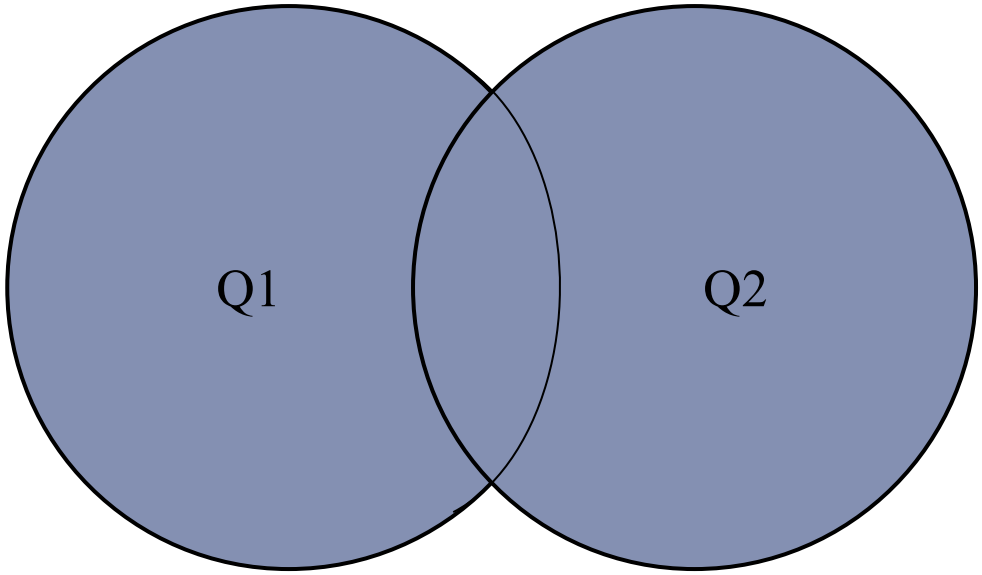
selected

Not selected

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Computing Boolean Queries

All documents



or Q1, Q2

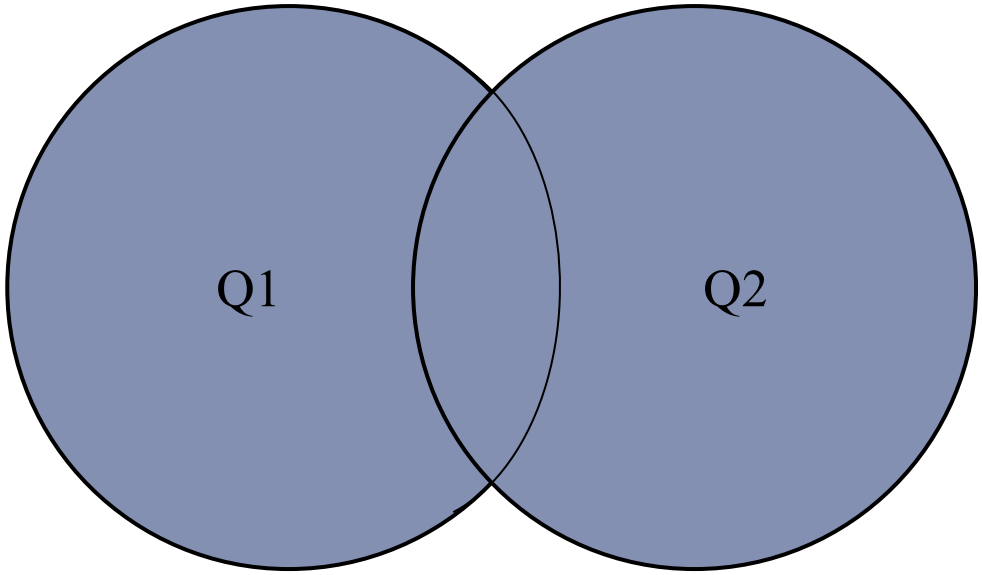
selected

Not selected

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Computing Boolean Queries

All documents



or Q1, Q2

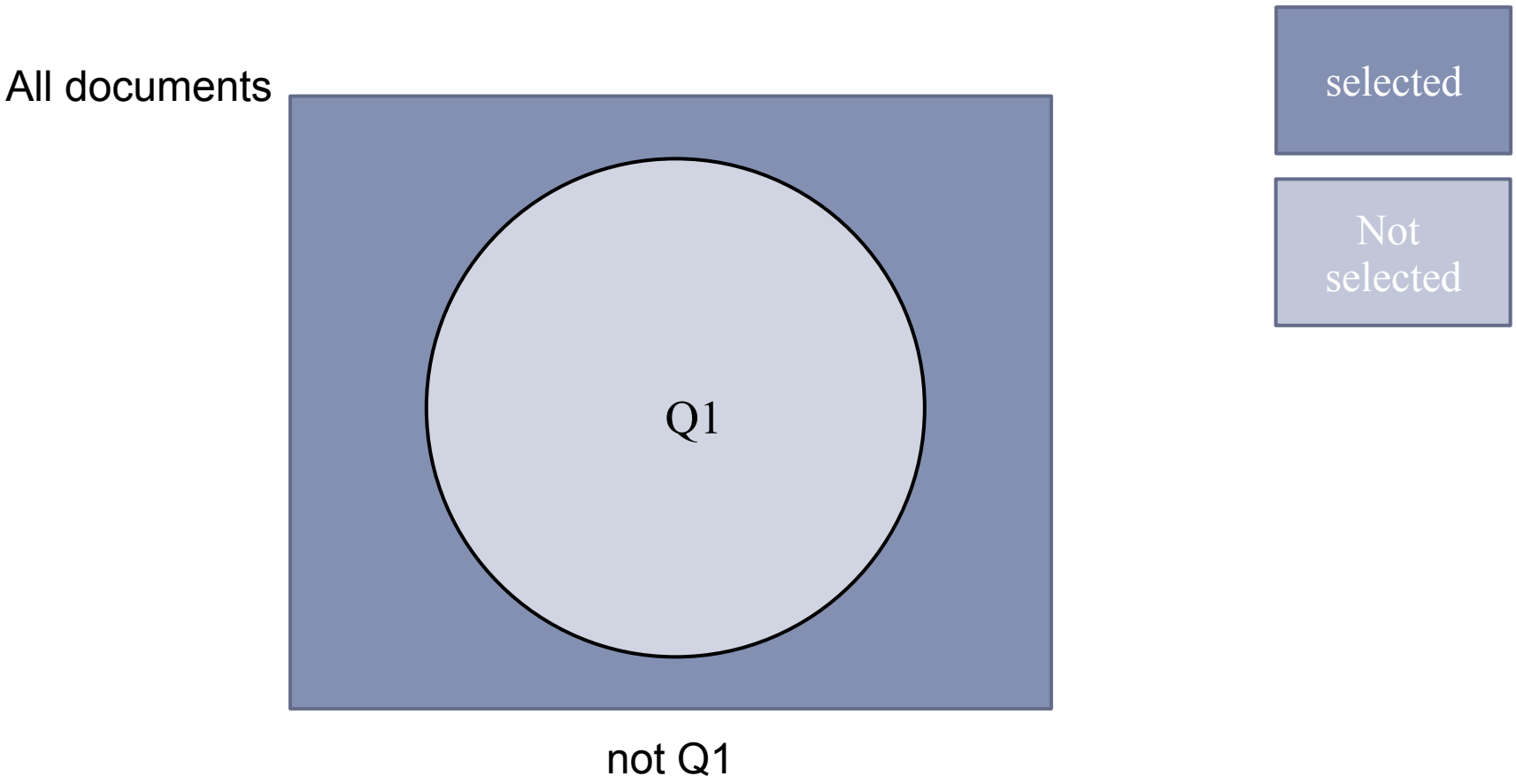
selected

Not selected

SLIDE HIDDEN

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Computing Boolean Queries



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

What Makes a Good Boolean Query?

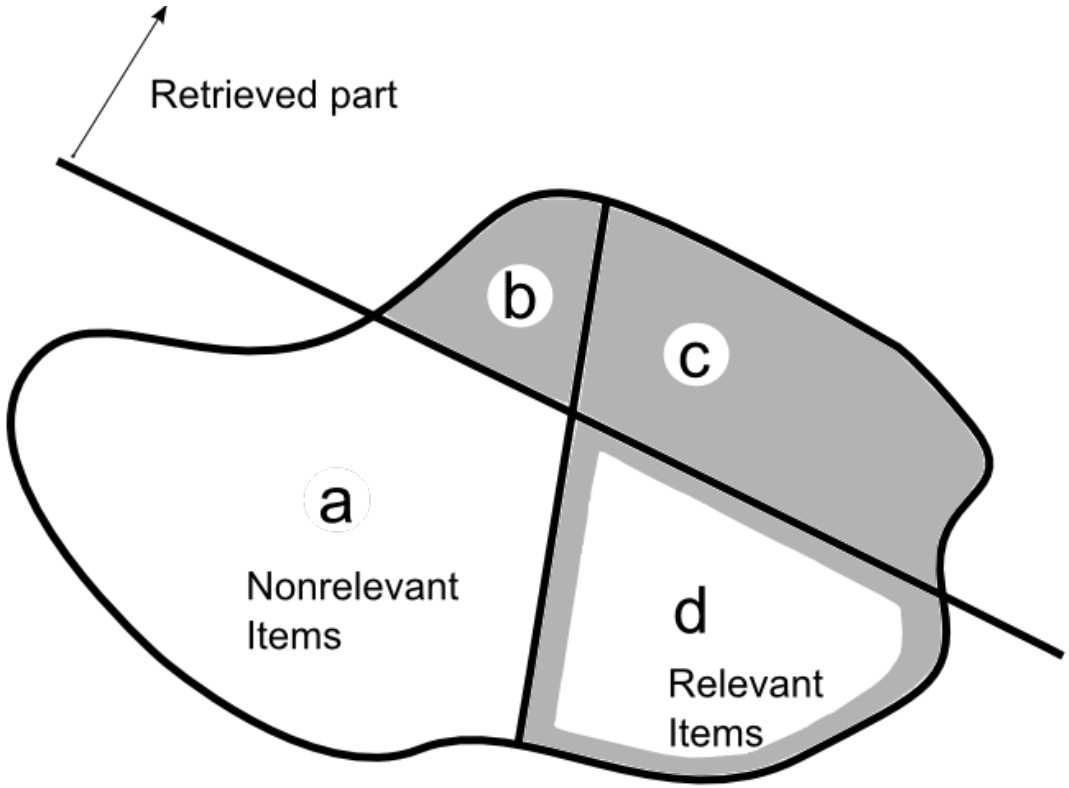
- It should match all "relevant" documents
- It should not match any " non-relevant" documents

  - Question 1: When is a document relevant?
    - Answer: When it fulfils our information need

  - Question 2: Are all matching documents relevant?
    - Answer: Not really, although we design the query with this goal in mind.

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Measuring Boolean Query effectiveness



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Measuring Boolean Query Effectiveness

- Precision

$$P = \frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}} = \frac{|c|}{|b| + |c|}$$

- Recall

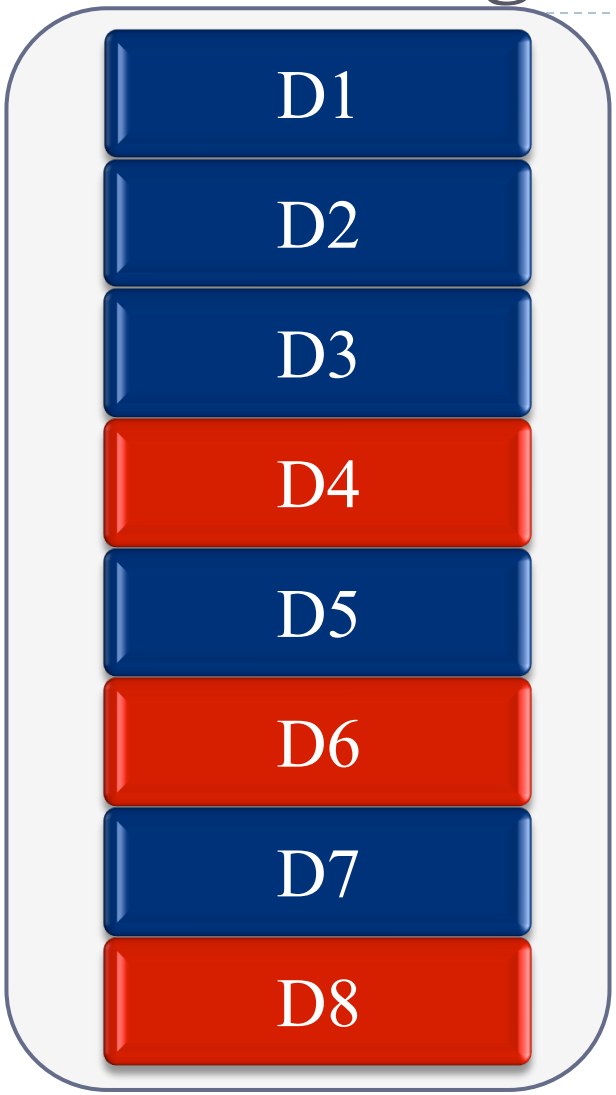
$$R = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items in collection}} = \frac{|c|}{|c| + |d|}$$

Question: Can we really measure the effectiveness of any query? Why?

Formula 9.1 and 9.2
From: Gerald Salton, Automatic Text Processing

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Measuring Boolean Query Effectiveness

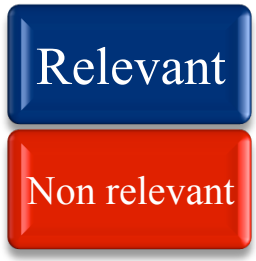


Not specific order

Retrieval documents (result set)

Total relevant retrieved = 5
Total retrieved = 8
Total relevant = 20 (assumed)

Precision = 5/8 = 52.5%
Recall = 5/20 = 25%



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Query Models Subtopics

- Boolean Query Model
- Vector Space Query Model
- Practical Query Models: The Case of Google

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Vector Space Query Model: Definition

- A Vector Space query consists of a list of terms with corresponding numeric weights
- Usually implicit default weights are assigned when needed
- Examples of vector space queries
  - aldehyde
  - (aldehyde:10 dehydrogenase:25)
  - (isocitrate:1 aldehyde:3 dehydrogenase:10)

SLIDE HIDDEN

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Vector Space Query Model: Definition

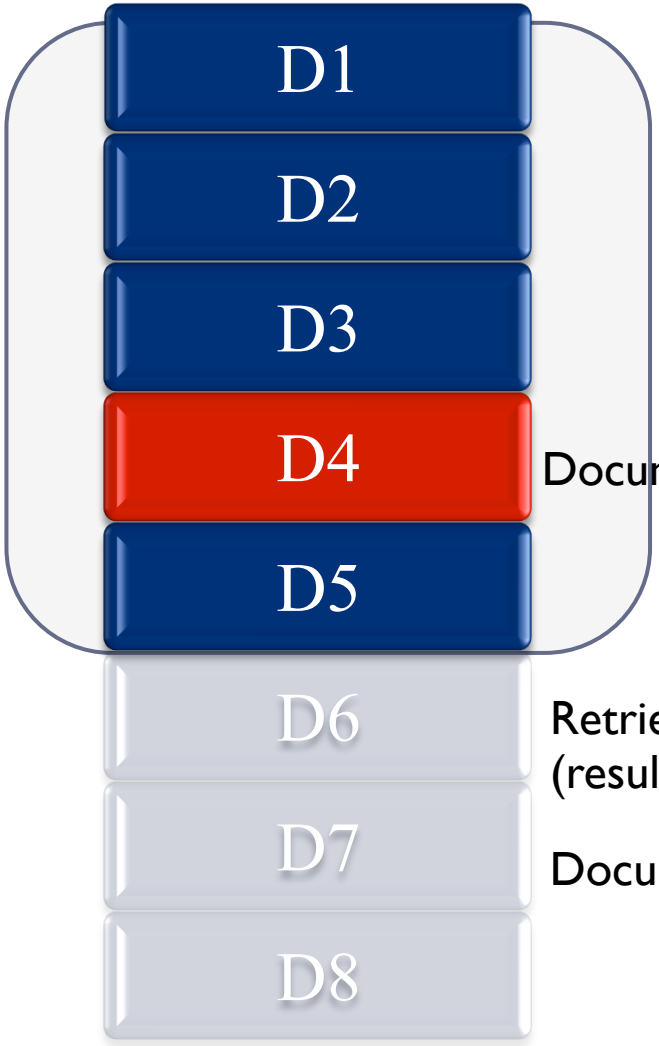
- A Vector Space query consists of a list of terms with corresponding numeric weights
- Usually implicit default weights are assigned when needed

Query consists of:	Document Matches When:
aldehyde	Document contains term aldehyde
(aldehyde:10 dehydrogenase:25)	Document that contains terms aldehyde or dehydrogenase. Documents with dehydrogenase ranked higher.
(isocitrate:-3 aldehyde:3 dehydrogenase:10)	Document may contains terms aldehyde, dehydrogenase or isocitrate. Documents with dehydrogenase ranked higher.

A vector query model can support multiple mechanisms for ranking matching documents according to expected similarity

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Measuring Vector Space Query Effectiveness



Documents analyzed

Retrieval documents (result set)

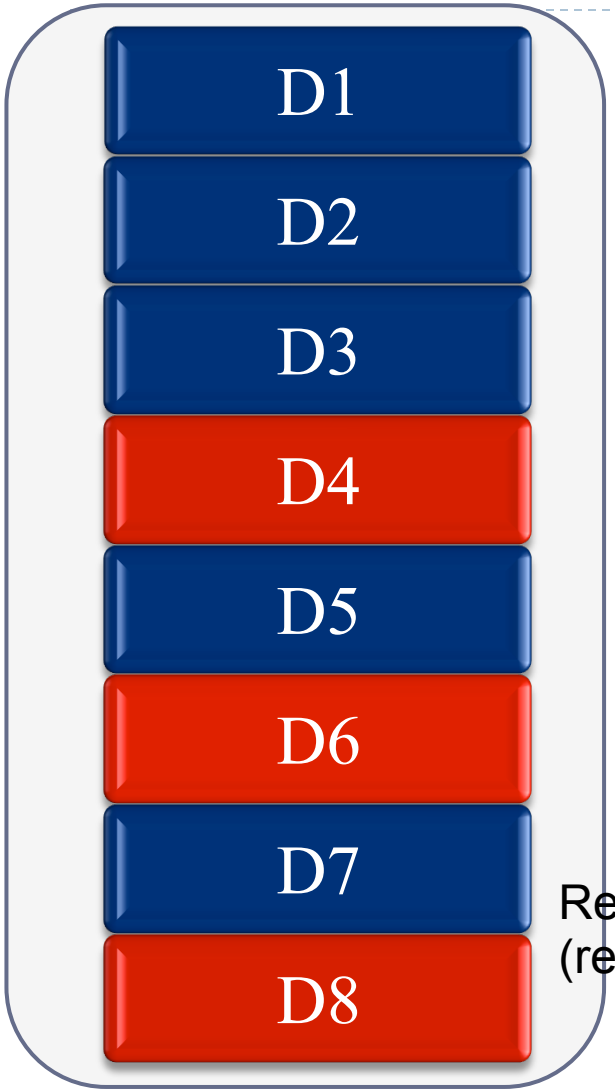
Documents D6-D8 not analyzed yet

Total relevant retrieved = 4
Total retrieved = 5
Total relevant = 10

Precision = 4/5 = 80%
Recall = 4/10 = 40%

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Measuring Vector Space Query Effectiveness



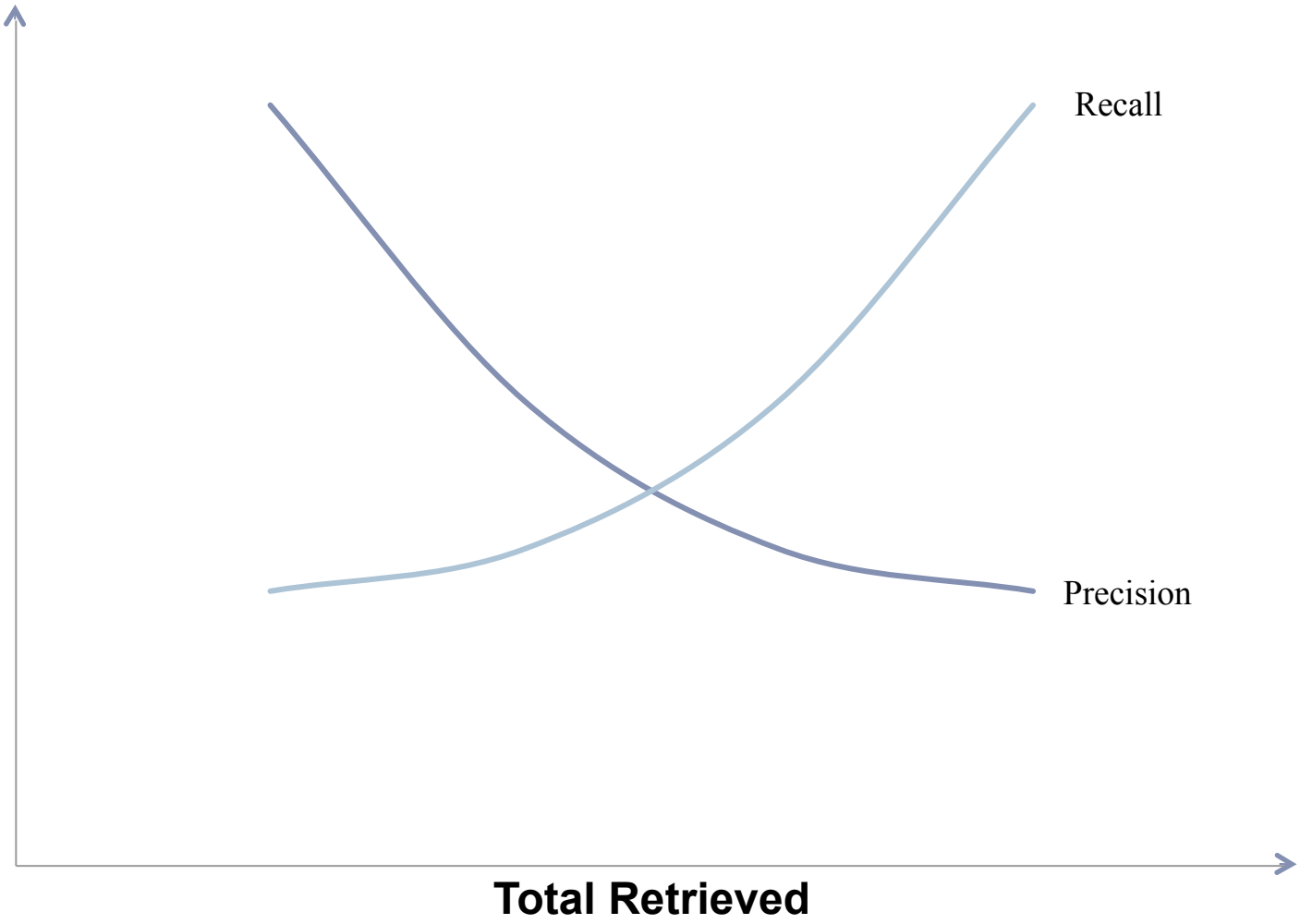
Retrieval documents (result set)

Total relevant retrieved = 5
Total retrieved = 8
Total relevant = 10 (assumed)

Precision = 5/8 = 62.5%
Recall = 4/10 = 50%

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

In Vector Space Query Model Precision and Recall constitute a tradeoff



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Examples of Similarity Measures

Similarity Measure sim(X,Y)	Evaluation for Binary Term Vectors	Evaluation for Weighted Term Vectors
Inner product	$\lvert X \cap Y \rvert$	$\sum_{i=1}^{t} x_i \bullet y_i$
Dice coefficient	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2\sum_{i=1}^{t} x_i \bullet y_i}{\sum_{i=1}^{t} x^2{}_i + \sum_{i=1}^{t} y^2{}_i}$
Cosine coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{1/2} \bullet \lvert Y \rvert^{1/2}}$	$\frac{\sum_{i=1}^{t} x_i \bullet y_i}{\sqrt{\sum_{i=1}^{t} x^2{}_i} \bullet \sum_{i=1}^{t} y^2{}_i}$
Jaccard coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum_{i=1}^{t} x_i \bullet y_i}{\sum_{i=1}^{t} x^2{}_i + \sum_{i=1}^{t} y^2{}_i - \sum_{i=1}^{t} x_i \bullet y_i}$

Legend:
$X=(x_1,x_2,\ldots,x_t)$
$\lvert X \rvert$ = number of terms in X
$\lvert X \cap Y \rvert$ =number of terms appearing jointly in X and Y

SLIDE HIDDEN

Table from:
Gerald Salton,
Automatic Text Processing
Page 318

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Examples of Similarity Measures

Similarity Measure sim(X,Y)	Evaluation for Binary Term Vectors	Evaluation for Weighted Term Vectors
Inner product	$\lvert Q \cap D \rvert$	$\sum_{i=1}^{t} q_i \bullet d_i$
Dice coefficient	$2\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert + D}$	$\frac{2\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2 i + \sum_{i=1}^{t} d^2 i}$
Cosine coefficient	$\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert^{1/2} \bullet \lvert D \rvert^{1/2}}$	$\frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sqrt{\sum_{i=1}^{t} q^2 i} \bullet \sum_{i=1}^{t} d^2 i}$
Jaccard coefficient	$\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert + \lvert D \rvert - \lvert Q \cap D \rvert}$	$\frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2 i + \sum_{i=1}^{t} d^2 i - \sum_{i=1}^{t} q_i \bullet d_i}$

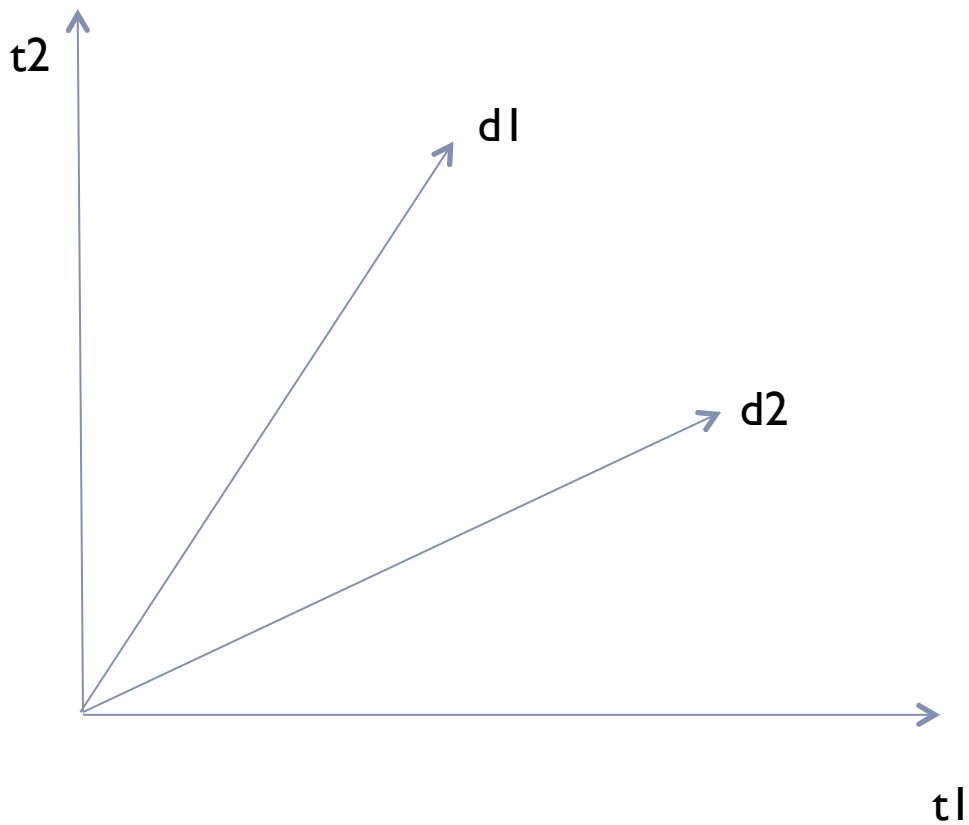
Legend:
$X=(x_1,x_2,\ldots,x_t)$
$\lvert X \rvert$ = number of terms in X
$\lvert X \cap Y \rvert$ =number of terms appearing jointly in X and Y

Table from:
Gerald Salton,
Automatic Text Processing
Page 318

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Diced coefficient: A Normalized Similarity Measure

2D example



$0 \leq \cos \alpha \leq 1$

$\cos \alpha = 1$

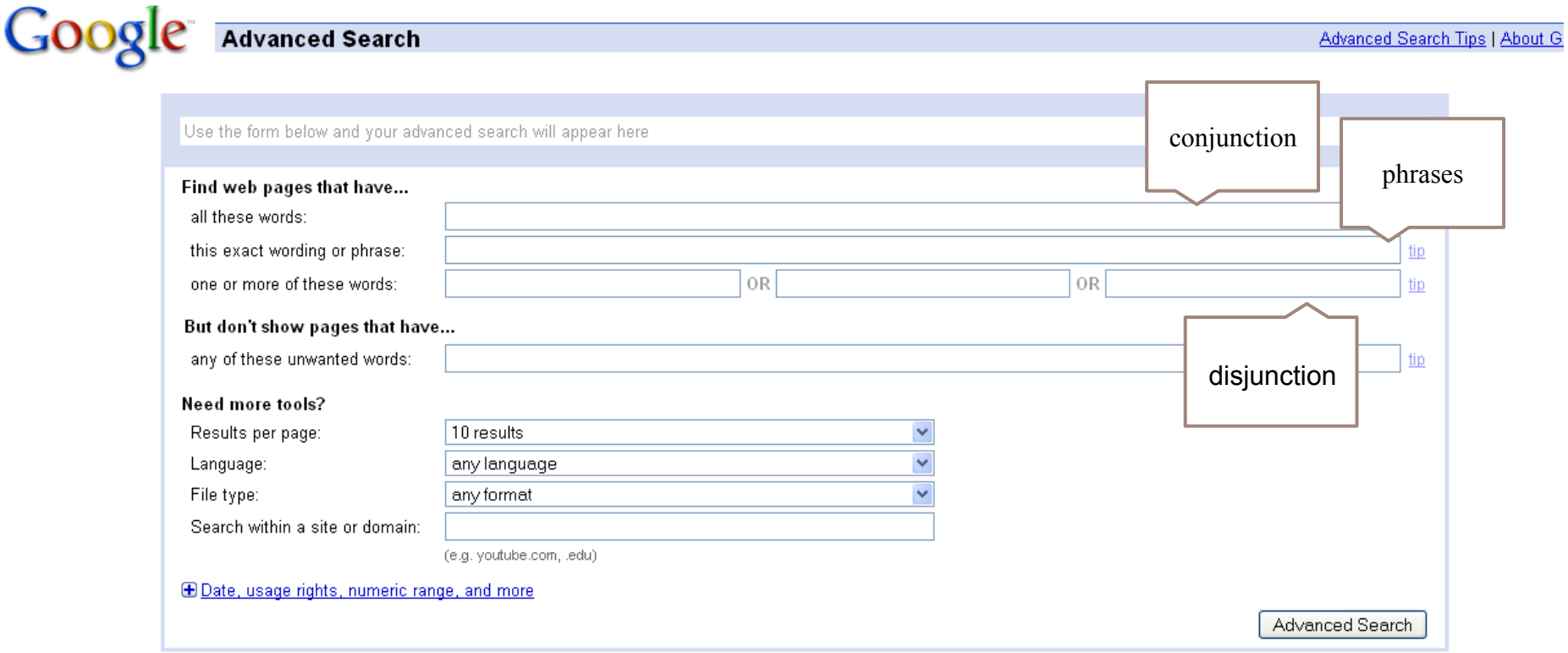
These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Query Models Subtopics

- Boolean Query Model
- Vector Space Query Model
- Practical Query Models: The Case of Google

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Google's Query Model



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Google's Query Model

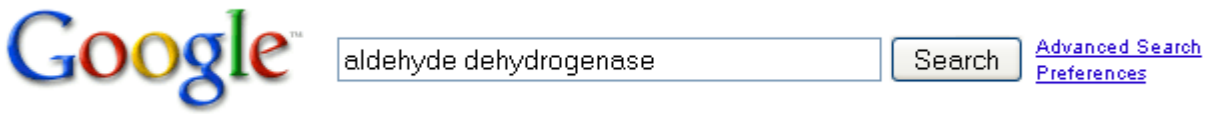
Operator	Description
"<phrase>"	Phrase search. By putting double quotes around a list of words, Google considers the exact words in that exact order without any change.
site:	Search within a specific website. Specify that your search results must come from a specific website.
-	Terms you want to exclude
*	Wildcard. Google treats the star as a placeholder for any unknown term(s) and then find the best matches.
+	Search term exactly as it appears
OR	Disjunction of terms

Most practical query models are hybrid!

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Sample Advance Google's Query #1

- Simple and common Google search query
  - aldehyde dehydrogenase



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Results of Google Query #1

Google aldehyde dehydrogenase Search Advanced Search Preferences

Web Show options... Results 1 - 10 of about 429,000 for aldehyde dehydrogenase. (0.36 secc

Aldehyde dehydrogenase - Wikipedia, the free encyclopedia
Chimera Image of a Monomer of **Aldehyde Dehydrogenase** 2 with a space filling model of NAD+ in the active site. (ALDH2, pdb code: 1o02) ...
en.wikipedia.org/wiki/**Aldehyde_dehydrogenase** - Cached - Similar -

Alcohol **dehydrogenase** - Wikipedia, the free encyclopedia
5 May 2009 ... This may be a correlating evolution with the rise of **aldehyde dehydrogenase**, which has been suggested as one of the more recognizable recent ...
en.wikipedia.org/wiki/Alcohol_**dehydrogenase** - Cached - Similar -
More results from en.wikipedia.org »

aldehyde dehydrogenase (enzyme) -- Britannica Online Encyclopedia
Britannica online encyclopedia article on **aldehyde dehydrogenase** (enzyme), ...is converted by this action to acetaldehyde, itself a highly toxic substance, ...
www.britannica.com/EBchecked/topic/13550/**aldehyde-dehydrogenase** - Cached - Similar -

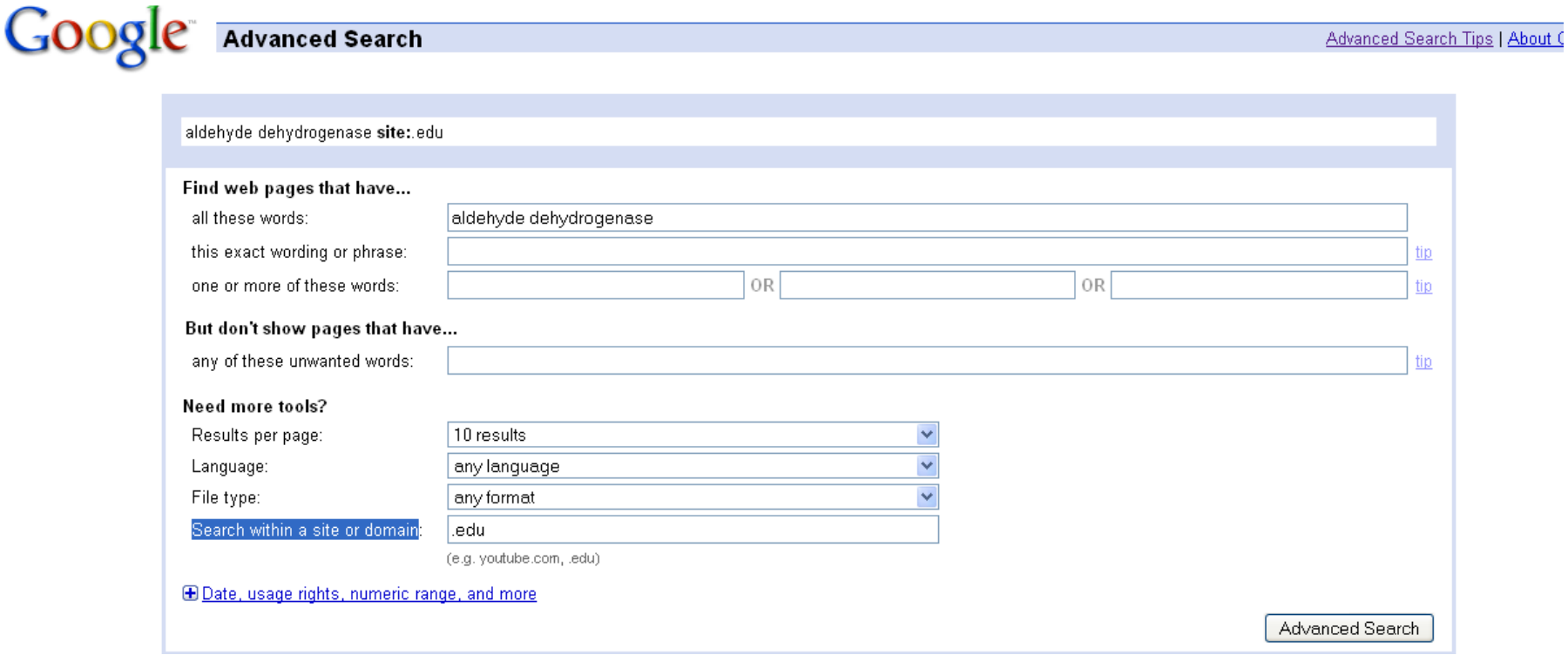
Alcohol Metabolism in Asian-American Men with Genetic ...
Genotypes for **aldehyde dehydrogenase** deficiency and alcohol sensitivity. ... Alcohol and **aldehyde dehydrogenase** genotypes and alcoholism in Chinese men. ...
www.annals.org/cgi/content/full/127/5/376 - Similar -
by TL Wall - 1997 - Cited by 44 - Related articles - All 4 versions

Aldehyde Dehydrogenase Gene Superfamily Resource Database
Our laboratory continues to compile data for the **aldehyde dehydrogenase** (ALDH) gene superfamily to provide a cohesive informational resource regarding this ...
www.aldh.org/ - Cached - Similar -

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Sample Advance Google's Query #2

- Google search query within a site or domain
  - All educational sites
    - site:.edu aldehyde dehydrogenase



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Results of Google Query #2

Google site:.edu aldehyde dehydrogenase Search Advanced Search Preferences

Web Show options... Results **1** - **10** of about **21,800** for **site:.edu aldehyde dehydrogenase**. (**0.31** secc

All in the Family: Relationships Within the **Aldehyde Dehydrogenase** ...
14 Sep 1999 **...** Hempel, a University of Pittsburgh biologist, has for 20 years focused on a family of enzymes called **aldehyde dehydrogenase** (ALDH). **...**
www.psc.edu/science/hempel.html - Cached - Similar -

Aldehyde Dehydrogenase
Aldehyde dehydrogenase: Maintaining critical active site geometry at motif 8 in the class 3 enzyme. European Journal of Biochemistry Vol. 268. pp. 722-726. **...**
www.psc.edu/~deerfiel/ALDH-SciVis.html - Cached - Similar -

10 Entries for <font color="blue ...
8854, ALDH1A2, 1 17 194 464, **aldehyde dehydrogenase** 1 family, member A2.RALDH2; MGC26444; RALDH2-T; RALDH(II).**aldehyde dehydrogenase** 1A2 isoform 3.
dags.stanford.edu/cancer/cgi_bin/cancer_list.cgi?attribute_name=**aldehyde+ dehydrogenase**+activity&attribute... - Cached - Similar -

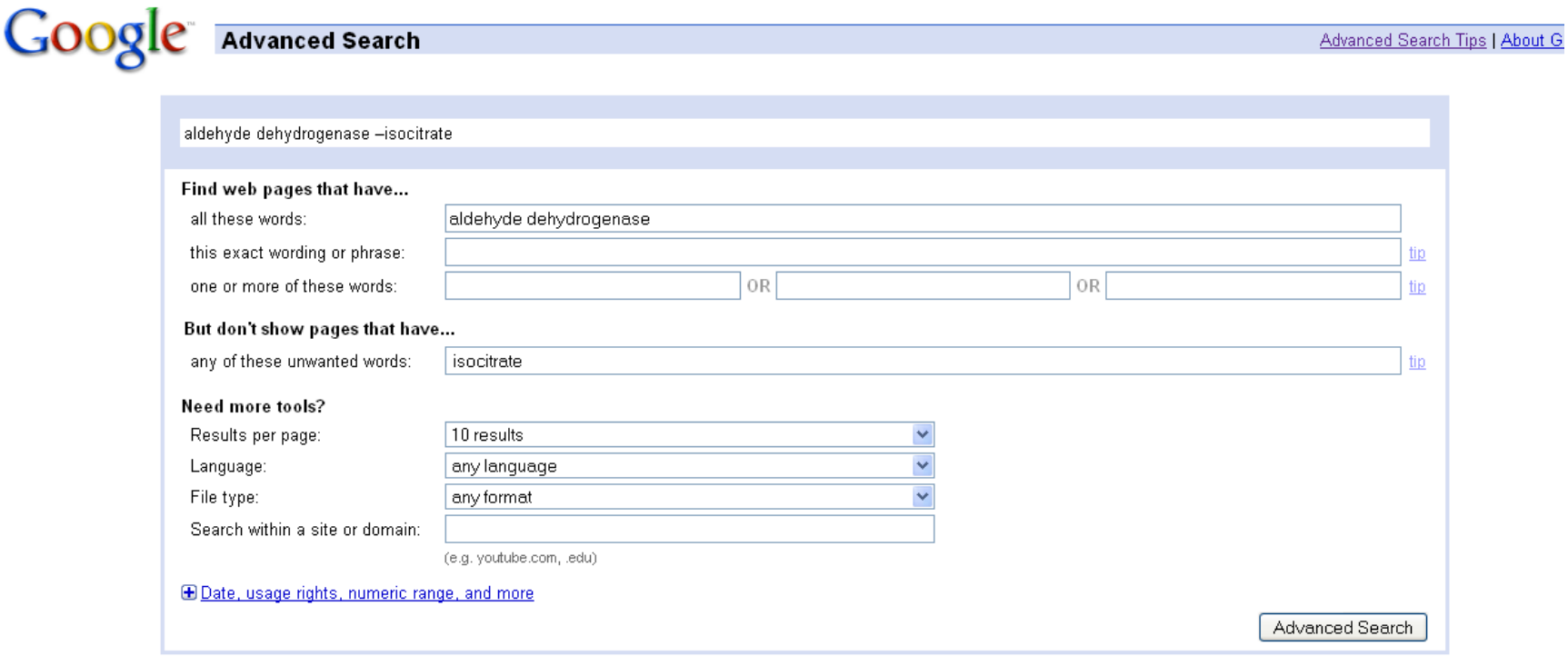
Aldehyde Dehydrogenase-2 Genotype Detection in Fingernails among ...
Document details from CiteSeerX (Isaac Councill, Lee Giles): ABSTRACT A genetic study on **aldehyde dehydrogenase** 2 (ALDH2) genotype was performed in a rural **...**
citeseer.ist.psu.edu/642580.html - Similar -
by A Paiboon Mongconthawornchai - All 2 versions

GO:0004029 ! **aldehyde dehydrogenase** (NAD) activity - GONUTS
26 Feb 2009 **...** name: **aldehyde dehydrogenase** (NAD) activity namespace: molecular_function def: "Catalysis of the reaction: an **aldehyde** + NAD+ + H2O = an **...**
gowiki.tamu.edu/wiki/index.php/Category:GO:0004029_!_**aldehyde_ dehydrogenase**_(NAD)_activity - Cached - Similar -

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Sample Advance Google's Query #3

- Google search query of the phrase aldehyde dehydrogenase, but the results not contains the word alcohol
  - aldehyde dehydrogenase –isocitrate



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Results of Google Query #3

Google

aldehyde dehydrogenase -isocitrate | Search | Advanced Search | Preferences

Web Show options... Results **1** - **10** of about **1,540,000** for **aldehyde dehydrogenase -isocitrate**. (**0.29** sec

Scholarly articles for **aldehyde dehydrogenase -isocitrate**

Alcohol and **aldehyde dehydrogenase** genotypes and ... - Thomasson - Cited by 343
Induction of class 3 **aldehyde dehydrogenase** in the mouse ... - Törrönen - Cited by 227
Molecular abnormality of an inactive **aldehyde** ... - Yoshida - Cited by 178

Aldehyde dehydrogenase - Wikipedia, the free encyclopedia
Chimera Image of a Monomer of **Aldehyde Dehydrogenase** 2 with a space filling model of NAD+ in the active site. (ALDH2, pdb code: 1o02) **...**
en.wikipedia.org/wiki/**Aldehyde_dehydrogenase** - Cached - Similar -

Long-chain-**aldehyde dehydrogenase** - Wikipedia, the free encyclopedia
Long-chain-**aldehyde dehydrogenase** (or fatty **aldehyde dehydrogenase**) is an **aldehyde dehydrogenase** enzyme associated with Sjögren-Larsson syndrome. **...**
en.wikipedia.org/wiki/Long-chain-**aldehyde_dehydrogenase** - Cached - Similar -
More results from en.wikipedia.org »

Alcohol Metabolism in Asian-American Men with Genetic **...**
Genotypes for **aldehyde dehydrogenase** deficiency and alcohol sensitivity. **...** Alcohol and **aldehyde dehydrogenase** genotypes and alcoholism in Chinese men. **...**
www.annals.org/cgi/content/full/127/5/376 - Similar -
by TL Wall - 1997 - Cited by 43 - Related articles

OMIM - **ALDEHYDE DEHYDROGENASE** 2 FAMILY; ALDH2
MIM +100650 · Description · Cloning · Gene Function · Gene Structure · Mapping · Molecular Genetics · Animal Model · Allelic Variants **...**
www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=100650 - Cached - Similar -

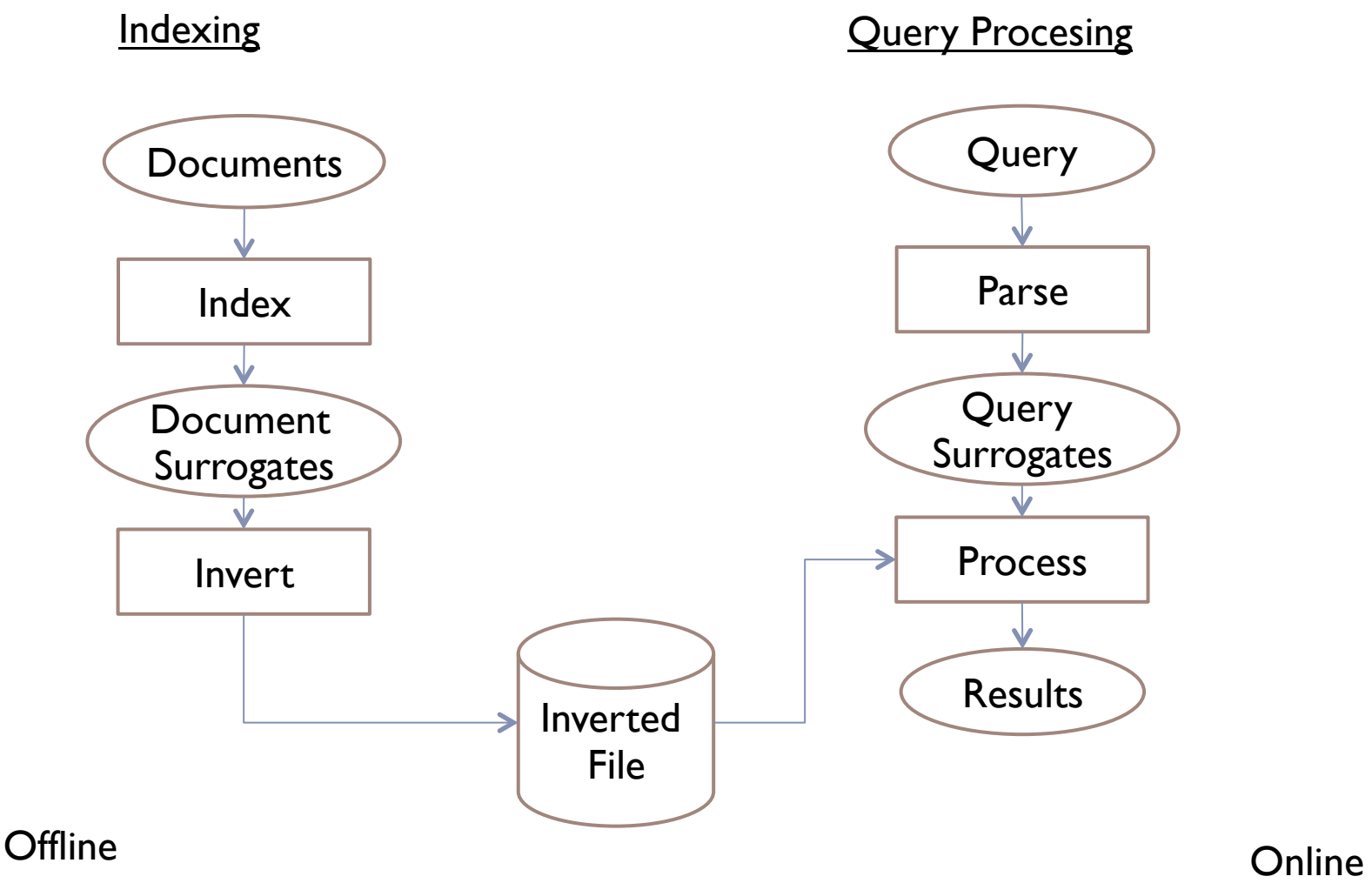
aldehyde dehydrogenase (enzyme) -- Britannica Online Encyclopedia
Britannica online encyclopedia article on **aldehyde dehydrogenase** (enzyme), ...is converted

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Unstructured Data Repositories: Outline

- Introduction and Examples
- Query Models
- Implementation Issues
- References

Architecture of an Information Retrieval System



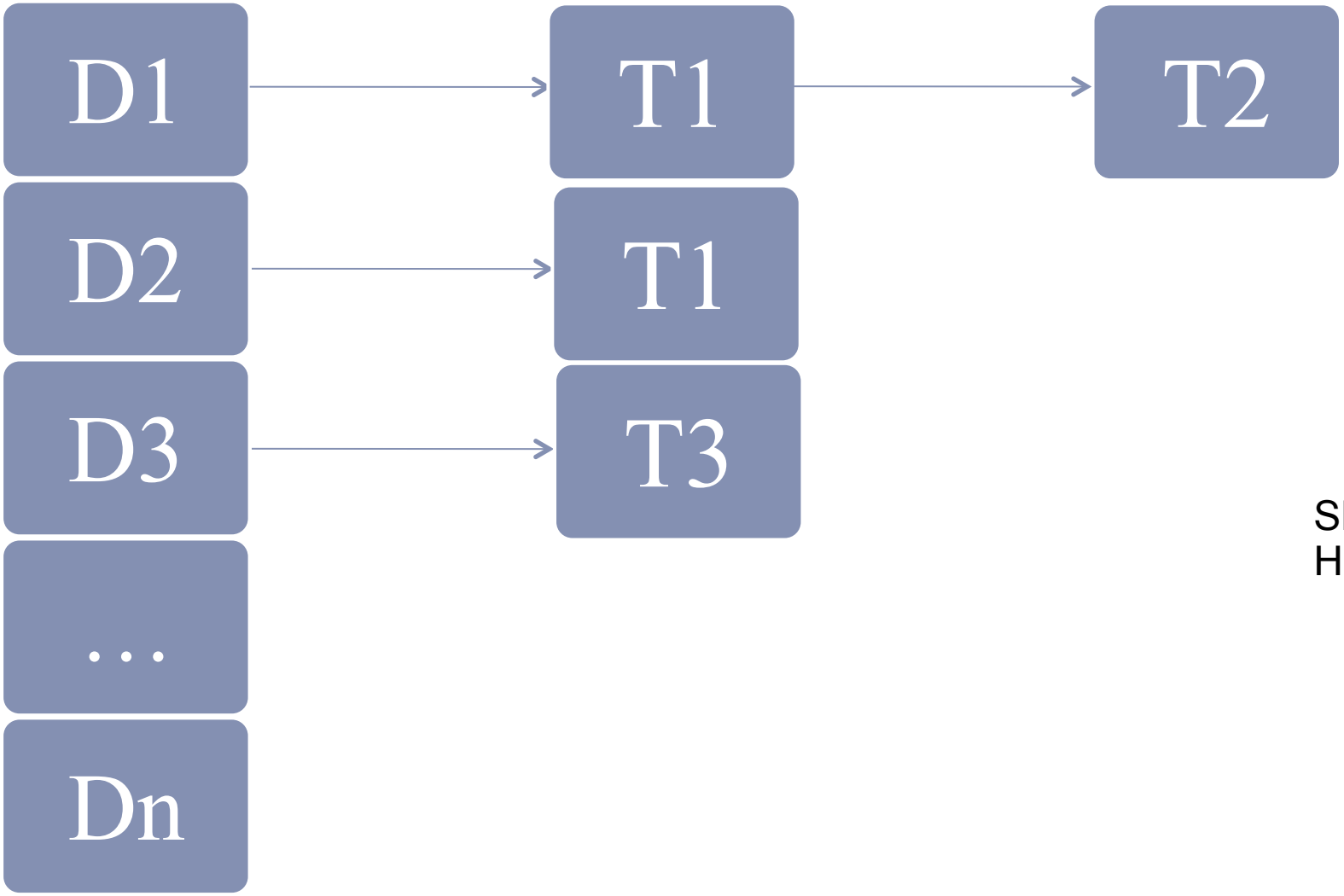
These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Implementation Issues Subtopics

- The Inverted File Data Structure
- Query Processing Using Inverted Files
- Inverted File Generation Algorithm
- Inverted file management for scalability
- Automatic Indexing

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

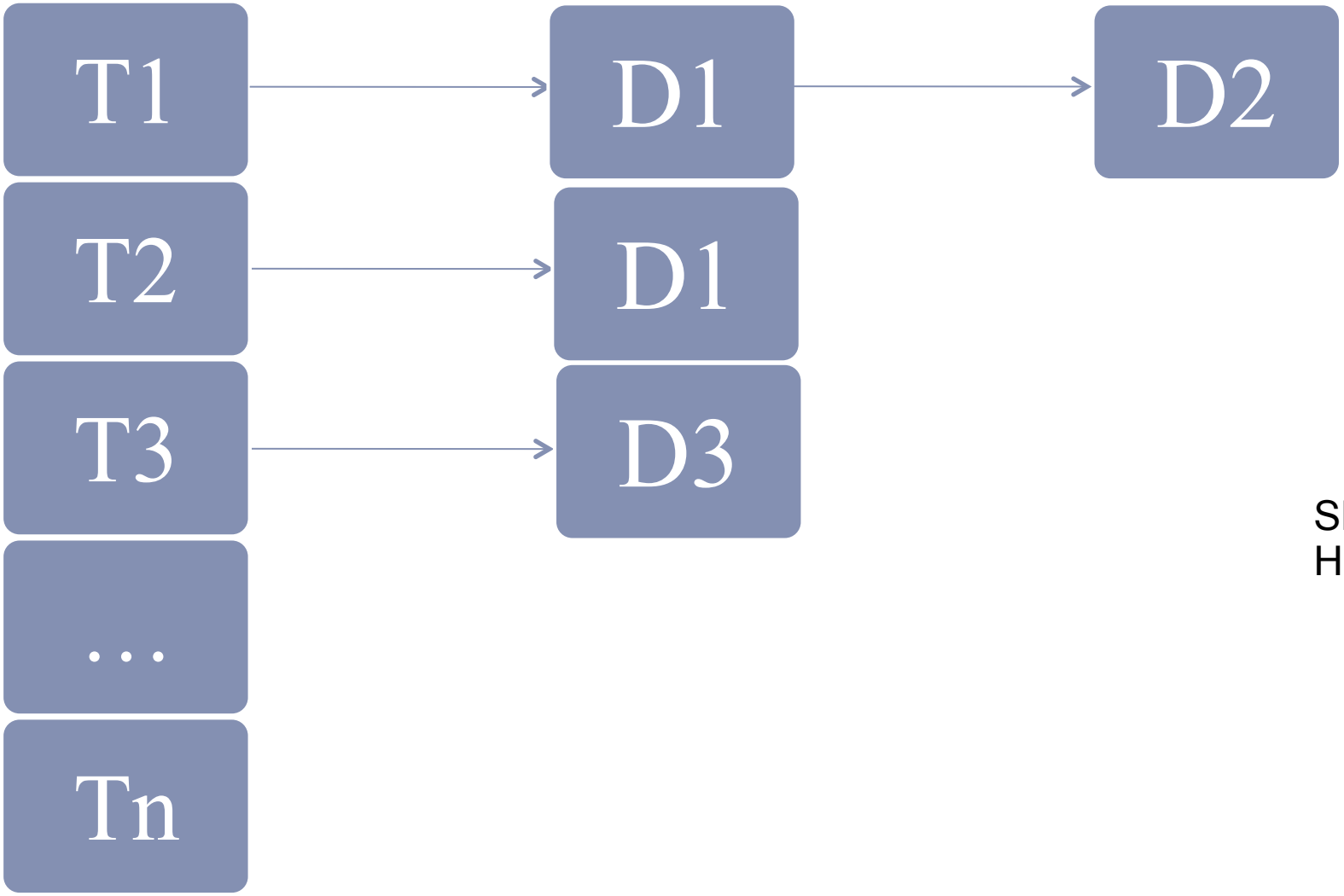
The Inverted File



SLIDE HIDDEN

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

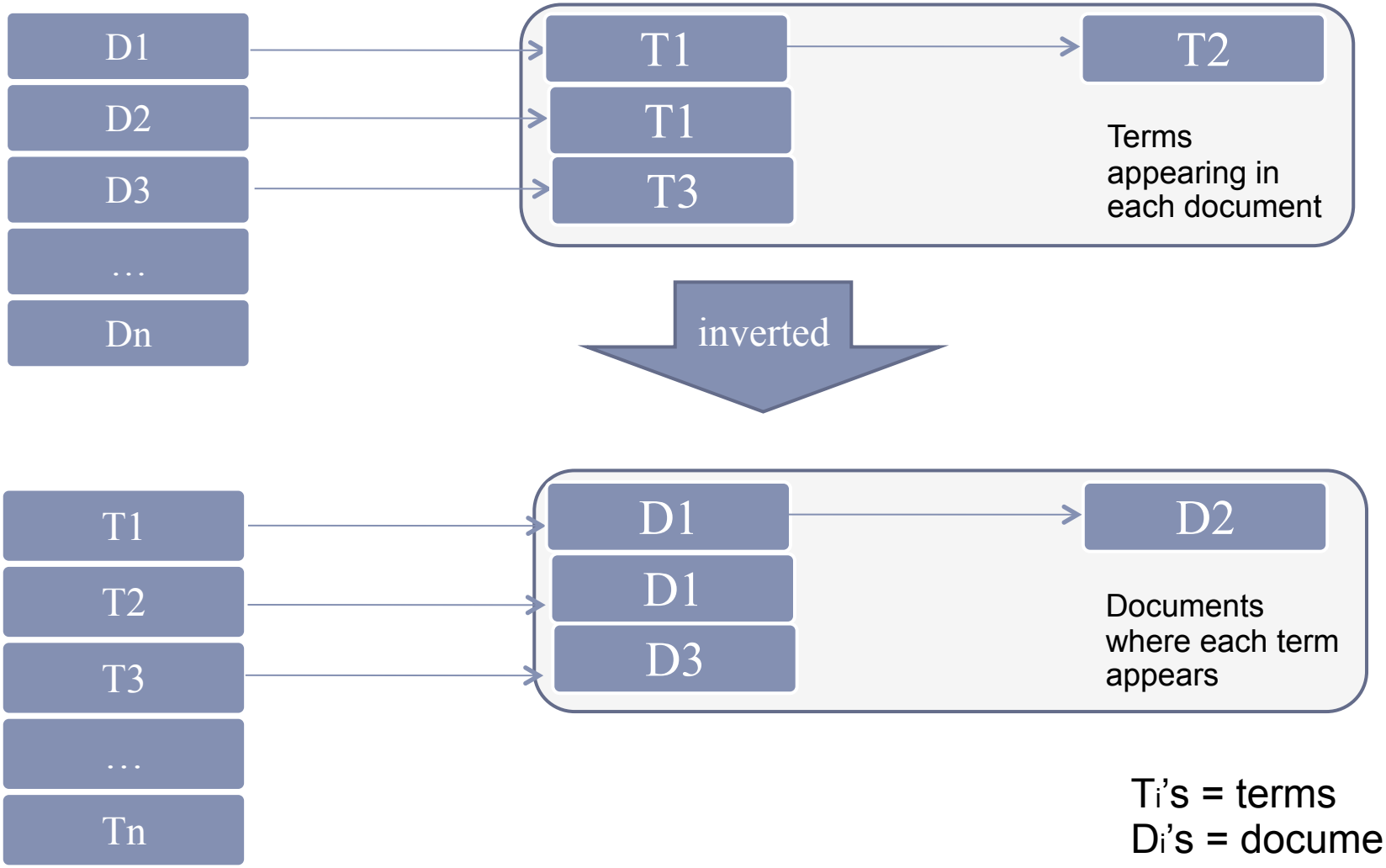
The Inverted File



SLIDE HIDDEN

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File Data Structure



T_i's = terms
D_i's = documents

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Implementation Issues Subtopics

- The Inverted File Data Structure
- Query Processing Using Inverted Files
- Inverted File Generation Algorithm
- Inverted file management for scalability
- Automatic Indexing

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

An Example Document Set

Relevant
Non relevant
assumed

- Example of Document 1

aldehyde dehydrogenase

- Example of Document 2

aldehyde isocitrate dehydrogenase

- Example of Document 3

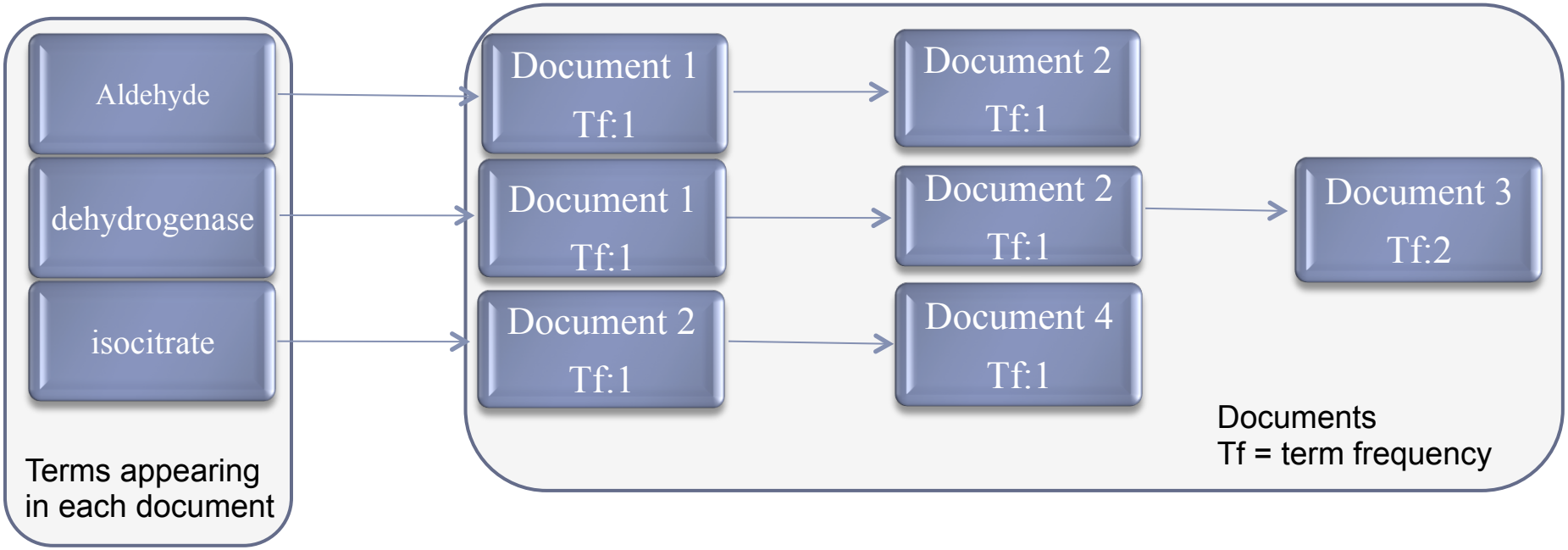
dehydrogenase dehydrogenase

- Example of Document 4

isocitrate

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File Data Structure for Example Document Set



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Inverted File

Example of boolean queries

(or aldehyde dehydrogenase)

Document 1

Document 2

Document 3

Total relevant retrieved = 2
Total retrieved = 3
Total relevant = 10

Precision = 2/3 = 66.7%
Recall = 3/10 = 20%

(and aldehyde dehydrogenase)

Document 1

Document 2

Total relevant retrieved = 1
Total retrieved = 2
Total relevant = 10

Precision = 1/2 = 50%
Recall = 1/10 = 10%

(and
(and aldehyde dehydrogenase)
(not isocitrate))

Document 1

En este slide no se pq cambio los valores de los total relevant

Total relevant retrieved = 1
Total retrieved = 1
Total relevant = 10

Precision = 1/1 = 100%
Recall = 1/10 = 10%

Relevant | Non relevant

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Example of vector space queries

aldehyde

Document 3
Tf:2

Document 1
Tf:1

Document 2
Tf:1

Sim(Q,D)=

(aldehyde:10
dehydrogenase:25)

Document 3
Tfa:2
Tfb: 0

Document 1
Tfa:1
Tfb: 1

Document 2
Tfa:1
Tfb: 1

Sim(Q,D)=

(isocitrate:-3
aldehyde:3
dehydrogenase:10)

Document 3
Tfa:2
Tfb:0
Tfc:0

Document 1
Tfa:0
Tfb:1
Tfc:1

Document 2
Tfa:1
Tfb:1
Tfc:1

Document 4
Tfa:1
Tfb:0
Tfc:0

Sim(Q,D)=

Verif orden
este correcto

Escribir las
similaridades

SLIDE
HIDDEN

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Inner product similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 1
3	aldehyde	1
10	dehydrogenase	1
-3	isocitrate	0

$$\mathrm{sim}_{\mathrm{inner}}(\mathrm{Q},\mathrm{D}_1) = \sum_{i=1}^{t} q_i \bullet d_i = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D$$

$$\mathrm{sim}_{\mathrm{inner}}(\mathrm{Q},\mathrm{D}_1) = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D = 3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (0) = 13$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Dice vector similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 1
3	aldehyde	1
10	dehydrogenase	1
-3	isocitrate	0

$$\mathrm{sim}_{\mathrm{dice}}(\mathrm{Q},\mathrm{D}_1) = \frac{2\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2{}_i + \sum_{i=1}^{t} d^2{}_i} = \frac{2(t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2}$$

$$\mathrm{sim}_{\mathrm{dice}}(\mathrm{Q},\mathrm{D}_1) = \frac{2(3\cdot 1 + 10\cdot 1 + (-3)\cdot(0))}{3^2 + 1^2 + 10^2 + 1^2 + (-3)^2 + (0)^2} = \frac{2(13)}{120} = \frac{26}{120} = 0.2167$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Cosine similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 1
3	aldehyde	1
10	dehydrogenase	1
-3	isocitrate	0

$$\mathrm{sim}_{\cos}(Q, D_1) = \frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sqrt{\sum_{i=1}^{t} q^2{}_i \bullet \sum_{i=1}^{t} d^2{}_i}} = \frac{t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D}{\sqrt{((t_1^Q)^2 + (t_2^Q)^2 + (t_3^Q)^2) \cdot ((t_1^D)^2 + (t_2^D)^2 + (t_3^D)^2)}}$$

$$\mathrm{sim}_{\cos}(Q, D_1) = \frac{3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (0)}{\sqrt{(3^2 + 10^2 + (-3)^2) \cdot (1^2 + 1^2 + 0^2)}} = \frac{13}{\sqrt{118 \cdot 2}} = \frac{13}{\sqrt{236}} = 0.8462$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Jaccard similarity :

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 1
3	aldehyde	1
10	dehydrogenase	1
-3	isocitrate	0

$$\mathrm{sim}_{\mathrm{Jaccard}}(Q, D_1) =$$

$$\frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2{}_i \bullet \sum_{i=1}^{t} d^2{}_i - \sum_{i=1}^{t} q_i \bullet d_i} = \frac{t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2 - (t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}$$

$$\mathrm{sim}_{\mathrm{Jaccard}}(Q, D_1) = \frac{3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (-0)}{3^2 + 1^2 + 10^2 + 1^2 + (-3)^2 + (0)^2 - (3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (0))} = \frac{13}{107} = 0.1215$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Inner product similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 2
3	aldehyde	1
10	dehydrogenase	1
-3	isocitrate	1

$$\mathrm{sim}_{\mathrm{inner}}(\mathrm{Q},\mathrm{D}_1) = \sum_{i=1}^{t} q_i \bullet d_i = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D$$

$$\mathrm{sim}_{\mathrm{inner}}(\mathrm{Q},\mathrm{D}_1) = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D = 3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (1) = 10$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Dice vector similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 2
3	aldehyde	1
10	dehydrogenase	1
-3	isocitrate	1

$$\mathrm{sim}_{\mathrm{dice}}(\mathrm{Q},\mathrm{D}_1) = \frac{2\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2{}_i + \sum_{i=1}^{t} d^2{}_i} = \frac{2(t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2}$$

$$\mathrm{sim}_{\mathrm{dice}}(\mathrm{Q},\mathrm{D}_1) = \frac{2(3\cdot 1 + 10\cdot 1 + (-3)\cdot(1))}{3^2 + 1^2 + 10^2 + 1^2 + (-3)^2 + (1)^2} = \frac{2(10)}{121} = \frac{20}{121} = 0.1653$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Cosine similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 2
3	aldehyde	1
10	dehydrogenase	1
-3	isocitrate	1

$$\mathrm{sim}_{\cos}(Q, D_1) = \frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sqrt{\sum_{i=1}^{t} q^2{}_i \bullet \sum_{i=1}^{t} d^2{}_i}} = \frac{t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D}{\sqrt{((t_1^Q)^2 + (t_2^Q)^2 + (t_3^Q)^2) \cdot ((t_1^D)^2 + (t_2^D)^2 + (t_3^D)^2)}}$$

$$\mathrm{sim}_{\cos}(Q, D_1) = \frac{3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (1)}{\sqrt{(3^2 + 10^2 + (-3)^2) \cdot (1^2 + 1^2 + 1^2)}} = \frac{10}{\sqrt{118 \cdot 3}} = \frac{10}{\sqrt{354}} = 0.5315$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Jaccard similarity :

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 2
3	aldehyde	1
10	dehydrogenase	1
-3	isocitrate	1

$$\mathrm{sim}_{\mathrm{Jaccard}}(Q, D_1) =$$

$$\frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2{}_i \bullet \sum_{i=1}^{t} d^2{}_i - \sum_{i=1}^{t} q_i \bullet d_i} = \frac{t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2 - (t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}$$

$$\mathrm{sim}_{\mathrm{Jaccard}}(Q, D_1) = \frac{3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (1)}{3^2 + 1^2 + 10^2 + 1^2 + (-3)^2 + (1)^2 - (3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (1))} = \frac{10}{111} = 0.0901$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Inner product similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 3
3	aldehyde	0
10	dehydrogenase	2
-3	isocitrate	0

$$\mathrm{sim}_{\mathrm{inner}}(\mathrm{Q},\mathrm{D}_1) = \sum_{i=1}^{t} q_i \bullet d_i = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D$$

$$\mathrm{sim}_{\mathrm{inner}}(\mathrm{Q},\mathrm{D}_1) = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D = 3 \cdot 0 + 10 \cdot 2 + (-3) \cdot (0) = 20$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Dice vector similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 3
3	aldehyde	0
10	dehydrogenase	2
-3	isocitrate	0

$$\mathrm{sim}_{\mathrm{dice}}(\mathrm{Q},\mathrm{D}_1) = \frac{2\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2{}_i + \sum_{i=1}^{t} d^2{}_i} = \frac{2(t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2}$$

$$\mathrm{sim}_{\mathrm{dice}}(\mathrm{Q},\mathrm{D}_1) = \frac{2(3\cdot 0 + 10\cdot 2 + (-3)\cdot(0))}{3^2 + 0^2 + 10^2 + 2^2 + (-3)^2 + (0)^2} = \frac{2(20)}{122} = \frac{40}{122} = 0.3279$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Cosine similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 3
3	aldehyde	0
10	dehydrogenase	2
-3	isocitrate	0

$$\mathrm{sim}_{\cos}(Q, D_1) = \frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sqrt{\sum_{i=1}^{t} q^2{}_i \bullet \sum_{i=1}^{t} d^2{}_i}} = \frac{t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D}{\sqrt{((t_1^Q)^2 + (t_2^Q)^2 + (t_3^Q)^2) \cdot ((t_1^D)^2 + (t_2^D)^2 + (t_3^D)^2)}}$$

$$\mathrm{sim}_{\cos}(Q, D_1) = \frac{3 \cdot 0 + 10 \cdot 2 + (-3) \cdot (0)}{\sqrt{(3^2 + 10^2 + (-3)^2) \cdot (0^2 + 2^2 + 0^2)}} = \frac{20}{\sqrt{118 \cdot 4}} = \frac{16}{\sqrt{472}} = 0.9206$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Jaccard similarity :

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 3
3	aldehyde	0
10	dehydrogenase	2
-3	isocitrate	0

$$\mathrm{sim}_{\mathrm{Jaccard}}(Q, D_1) =$$

$$\frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2{}_i \bullet \sum_{i=1}^{t} d^2{}_i - \sum_{i=1}^{t} q_i \bullet d_i} = \frac{t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2 - (t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}$$

$$\mathrm{sim}_{\mathrm{Jaccard}}(Q, D_1) = \frac{3 \cdot 0 + 10 \cdot 2 + (-3) \cdot (0)}{3^2 + 0^2 + 10^2 + 2^2 + (-3)^2 + (0)^2 - (3 \cdot 0 + 10 \cdot 2 + (-3) \cdot (0))} = \frac{20}{102} = 0.1961$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Inner product similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 4
3	aldehyde	0
10	dehydrogenase	0
-3	isocitrate	1

$$\mathrm{sim}_{\mathrm{inner}}(\mathrm{Q},\mathrm{D}_1) = \sum_{i=1}^{t} q_i \bullet d_i = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D$$

$$\mathrm{sim}_{\mathrm{inner}}(\mathrm{Q},\mathrm{D}_1) = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D = 3 \cdot 0 + 10 \cdot 0 + (-3) \cdot (1) = -3$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Dice vector similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 4
3	aldehyde	0
10	dehydrogenase	0
-3	isocitrate	1

$$\mathrm{sim}_{\mathrm{dice}}(\mathrm{Q},\mathrm{D}_1) = \frac{2\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2{}_i + \sum_{i=1}^{t} d^2{}_i} = \frac{2(t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2}$$

$$\mathrm{sim}_{\mathrm{dice}}(\mathrm{Q},\mathrm{D}_1) = \frac{2(3\cdot 0 + 10\cdot 0 + (-3)\cdot(1))}{3^2 + 0^2 + 10^2 + 0^2 + (-3)^2 + (1)^2} = \frac{2(-3)}{119} = \frac{-6}{119} = -0.0504$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Cosine similarity

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 4
3	aldehyde	0
10	dehydrogenase	0
-3	isocitrate	1

$$\mathrm{sim}_{\cos}(Q, D_1) = \frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sqrt{\sum_{i=1}^{t} q^2{}_i \bullet \sum_{i=1}^{t} d^2{}_i}} = \frac{t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D}{\sqrt{((t_1^Q)^2 + (t_2^Q)^2 + (t_3^Q)^2) \cdot ((t_1^D)^2 + (t_2^D)^2 + (t_3^D)^2)}}$$

$$\mathrm{sim}_{\cos}(Q, D_1) = \frac{3 \cdot 0 + 10 \cdot 0 + (-3) \cdot (1)}{\sqrt{(3^2 + 10^2 + (-3)^2) \cdot (0^2 + 0^2 + 1^2)}} = \frac{-3}{\sqrt{118 \cdot 1}} = \frac{-3}{\sqrt{118}} = -0.2762$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

Jaccard similarity :

(isocitrate:-3 aldehyde:3 dehydrogenase:10)

Query		Document 4
3	aldehyde	0
10	dehydrogenase	0
-3	isocitrate	1

$$\mathrm{sim}_{\mathrm{Jaccard}}(Q, D_1) =$$

$$\frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2{}_i \bullet \sum_{i=1}^{t} d^2{}_i - \sum_{i=1}^{t} q_i \bullet d_i} = \frac{t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2 - (t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}$$

$$\mathrm{sim}_{\mathrm{Jaccard}}(Q, D_1) = \frac{3 \cdot 0 + 10 \cdot 0 + (-3) \cdot (1)}{3^2 + 0^2 + 10^2 + 0^2 + (-3)^2 + (1)^2 - (3 \cdot 0 + 10 \cdot 0 + (-3) \cdot (1))} = \frac{-3}{122} = 0.0246$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Inverted File

Example of vector space query: (isocitrate:-3 aldehyde:3 dehydrogenase:10)

Inner similarity	Dice similarity	Cosine similarity	Jaccard similarity
Document 3 20	Document 3 0.3279	Document 3 0.9206	Document 3 0.1961
Document 1 13	Document 1 0.2167	Document 1 0.8462	Document 1 0.1215
Document 2 10	Document 2 0.1653	Document 2 0.5315	Document 2 0.0901
Document 4 -3	Document 4 -0.0504	Document 4 -0.2762	Document 4 -0.0246

Total relevant retrieved =
Total retrieved =
Total relevant =

Precision = = %
Recall = = %

Relevant	Non relevant

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

The Inverted File

- Example of Extended Boolean Query

SLIDE
HIDDEN

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Implementation Issues Subtopics

- The Inverted File Data Structure
- Query Processing Using Inverted Files
- Inverted File Generation Algorithm
- Inverted file management for scalability
- Automatic Indexing

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Inverted File Generation Algorithm

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Implementation Issues Subtopics

- The Inverted File Data Structure
- Query Processing Using Inverted Files
- Inverted File Generation Algorithm
- Inverted file management for scalability
- Automatic Indexing

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Scaleable Inverted File Management

Inverted file management for scalability
- Offline inverted file
- Lazy query evaluation

Under construction

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Implementation Issues

- The Inverted File Data Structure
- Query Processing Using Inverted Files
- Inverted File Generation Algorithm
- Inverted file management for scalability
- Automatic Indexing

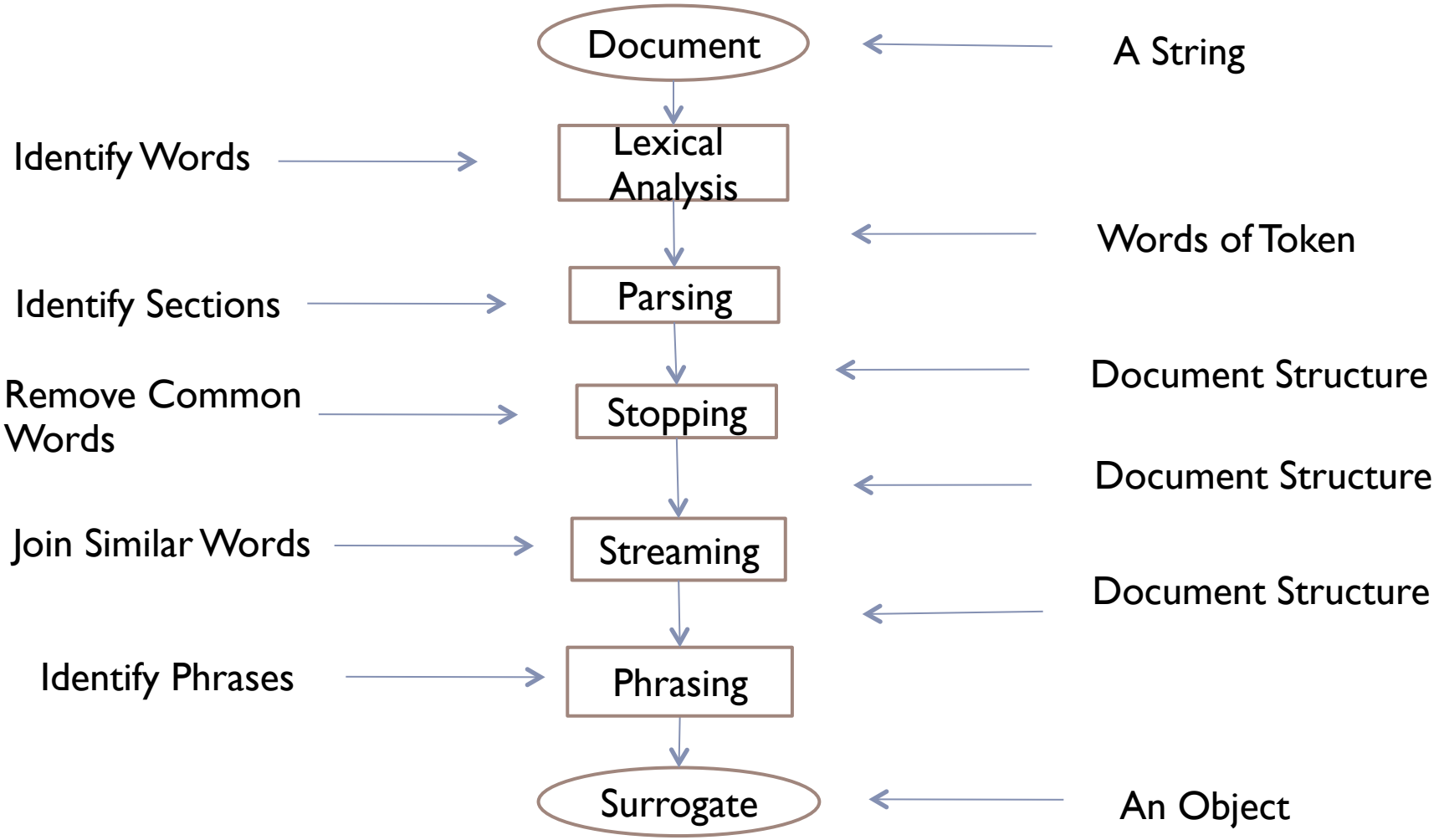
These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Automatic Indexing Subtopics

- Automatic Indexing at a Glance
- Stopping
- Stemming
- Phrase recognition
- Context Identification
- Document Surrogate Representations

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Automatic Indexing at a Glance



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Elements of a Document Surrogate

- Typically organized by terms or phrases
- For each term/phrase store:
  - Frequency
  - Each section where it appears
  - Positions where it appears
  - Others…

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Stopping: Removing Common Terms

- Justification
  - Common terms match too many documents
  - Want to keep inverted file small
- Procedure (Typical)
  - Use a standard list of stop words
  - Drop any term in the list
- Careful: Language specific

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Stemming: Adjoining Common Terms

- Justification
  - Terms have many variations
  - Want to keep inverted file small
  - Want to represent concepts
- Procedure (Typical)
  - Most people use some well known language specific algorithm
- Aqui va una referencia, pero el link del pdf no existe

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Google's English Stopwords

I
a
about
an
are
as
at
be
by
com
de
en
for
from
how
in
is
it
la
of
on
or
that
the
this
to
was
what
when
where
who
will
with
und
the
www

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Google's Spanish Stopwords

un
una
unas
unos
uno
sobre
todo
también
tras
otro
algún
alguno
alguna
algunos
algunas
ser
es
soy
eres
somos
sois
estoy
esta
estamos
estais
estan
como
en

para
atras
porque
por qué
estado
estaba
ante
antes
siendo
ambos
pero
por
poder
puede
puedo
podemos
podeis
pueden
fui
fue
fuimos
fueron
hacer
hago
hace
hacemos
haceis
hacen

cada
fin
incluso
primero
desde
conseguir
consigo
consigue
consigues
conseguimos
consiguen
ir
voy
va
vamosa
vais
van
vaya
gueno
ha
tener
tengo
tiene
tenemos
teneis
tienen
el
la

lo
las
los
su
aqui
mio
tuyo
ellos
ellas
nos
nosotros
vosotros
vosotras
si
dentro
solo
solamente
saber
sabes
sabe
sabemos
sabeis
saben
ultimo
largo
bastante
haces
muchos

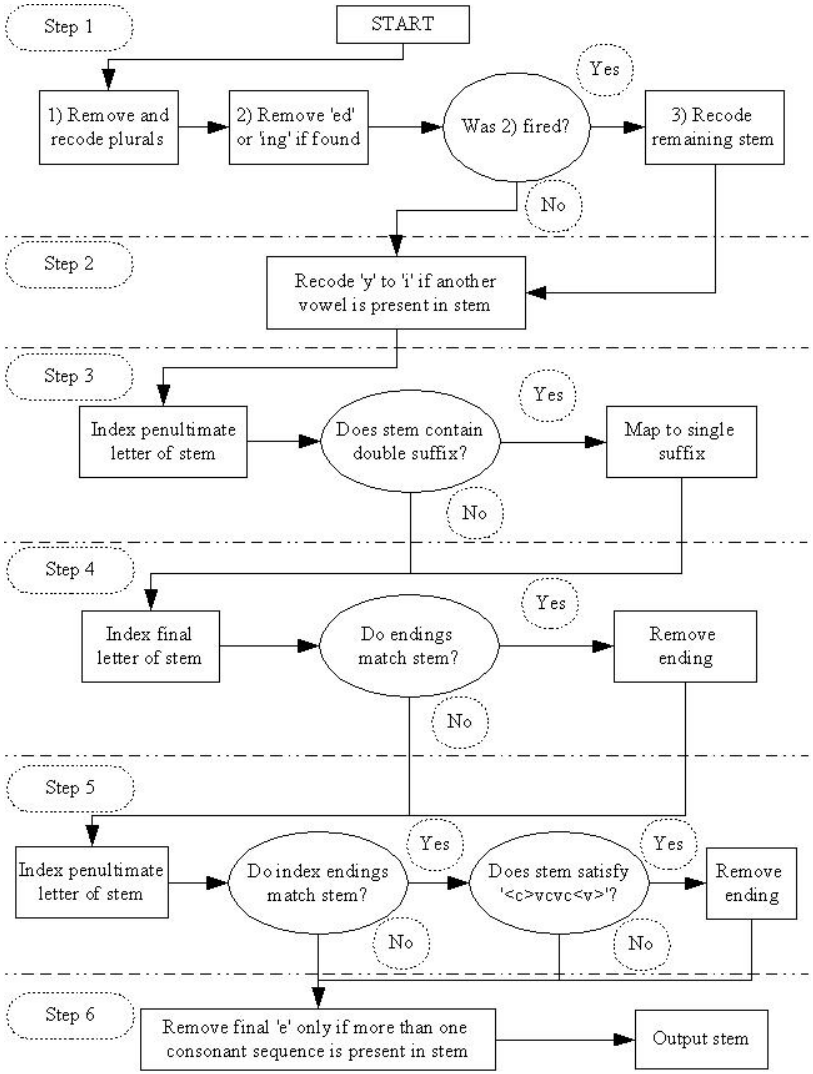
aquellos
aquellas
sus
entonces
tiempo
verdad
verdadero
verdadera
cierto
ciertos
cierta
ciertas
intentar
intento
intenta
intentas
intentamos
intentais
intentan
dos
bajo
arriba
encima
usar
uso
usas
usa
usamos

usais
usan
emplear
empleo
empleas
emplean
ampleamos
empleais
valor
muy
era
eras
eramos
eran
modo
bien
cual
cuando
donde
mientras
quien
con
entre
sin
trabajo
trabajar
trabajas
trabaja

trabajamos
trabajais
trabajan
podria
podrias
podriamos
podrian
podriais
yo
aquel

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Porter's Stemming Algorithm



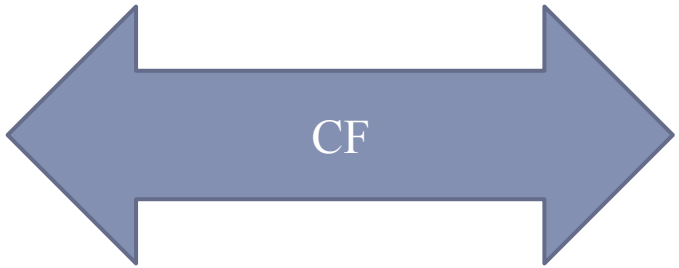
These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

How to deal with other Common Terms?

Collection Frequency = what fraction of all documents contains term

CF

Very uncommon 0%



Very common 100%

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Inverse Document Frequency (IDF)

Let D2 set of all documents

$$idf_i = \log \frac{|D|}{|\{d_i t_i \in D\}|}$$

$$idf_i = \frac{\text{Total number of documents}}{\text{number of document containing term } i}$$

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Revisiting Similarity Measures

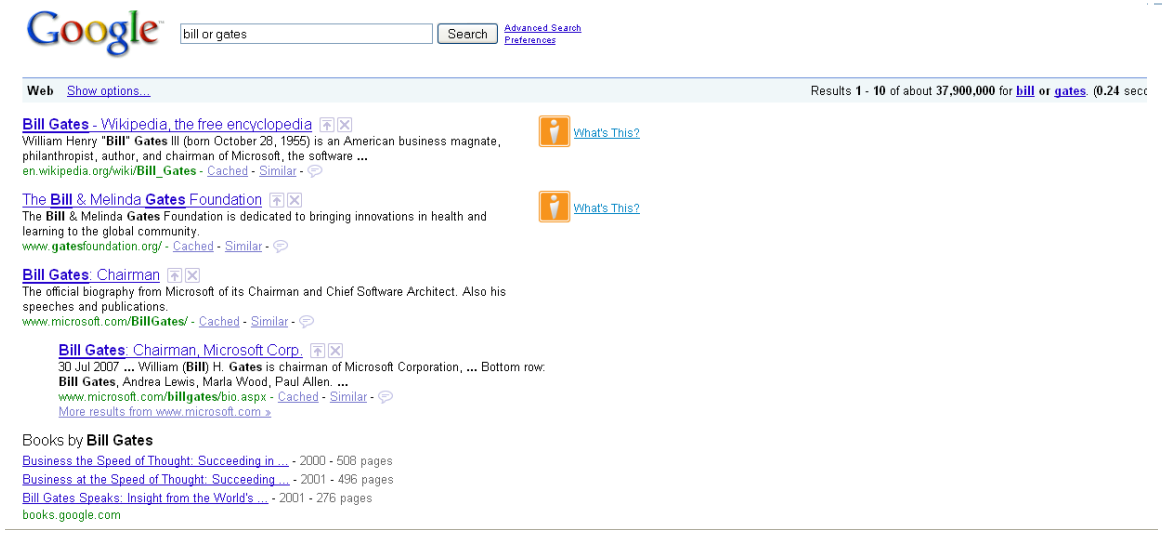
Similarity Measure sim(X,Y)	Evaluation for Binary Term Vectors	Evaluation for Weighted Term Vectors
Inner product	$\lvert Q \cap D \rvert$	$\sum_{i=1}^{t} q_i \bullet d_i$
Dice coefficient	$2\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert + D}$	$\frac{2\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2 i + \sum_{i=1}^{t} d^2 i}$
Cosine coefficient	$\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert^{1/2} \bullet \lvert D \rvert^{1/2}}$	$\frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sqrt{\sum_{i=1}^{t} q^2 i} \bullet \sum_{i=1}^{t} d^2 i}$
Jaccard coefficient	$\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert + \lvert D \rvert - \lvert Q \cap D \rvert}$	$\frac{\sum_{i=1}^{t} q_i \bullet d_i}{\sum_{i=1}^{t} q^2 i + \sum_{i=1}^{t} d^2 i - \sum_{i=1}^{t} q_i \bullet d_i}$

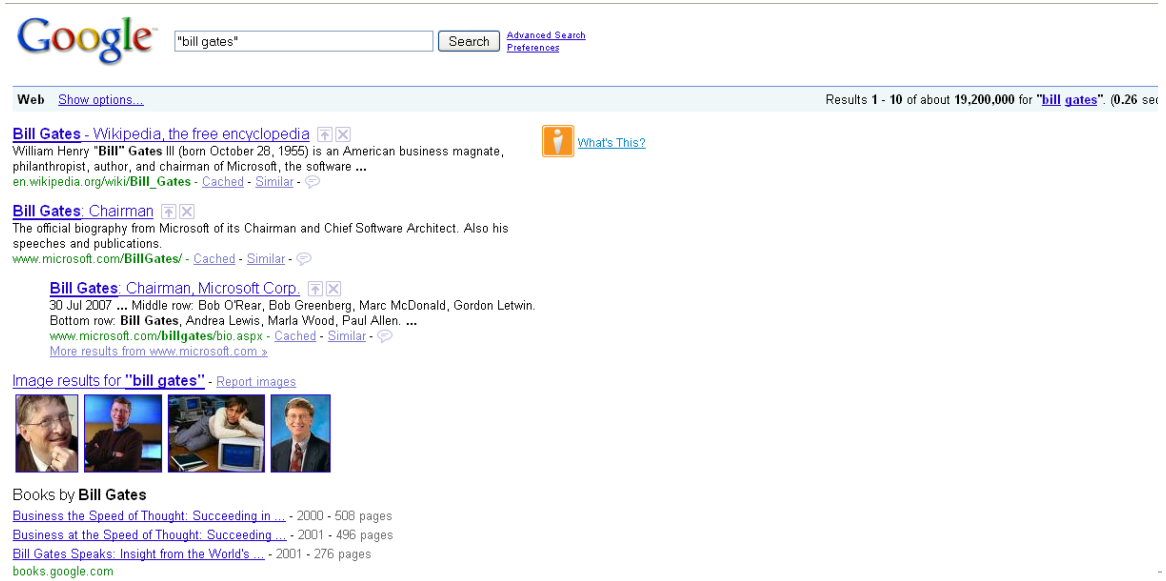
Legend:
$X=(x_1,x_2,\ldots,x_t)$
$\lvert X \rvert$ = number of terms in X
$\lvert X \cap Y \rvert$ =number of terms appearing jointly in X and Y

Table from:
Gerald Salton,
Automatic Text Processing
Page 318

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Phrase Recognition can Improve Decision and Recall





Searching:
bill or gates
vs.
“bill gates”

These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

A Simple Phrase Recognition Algorithm

```
for each document d
    for each k: 1 to m
        for each step p of k consecutive terms
            p = <t1, t2, …, tk>
            countp ++
```

Declare all p with counts over a threshold to be phrases.

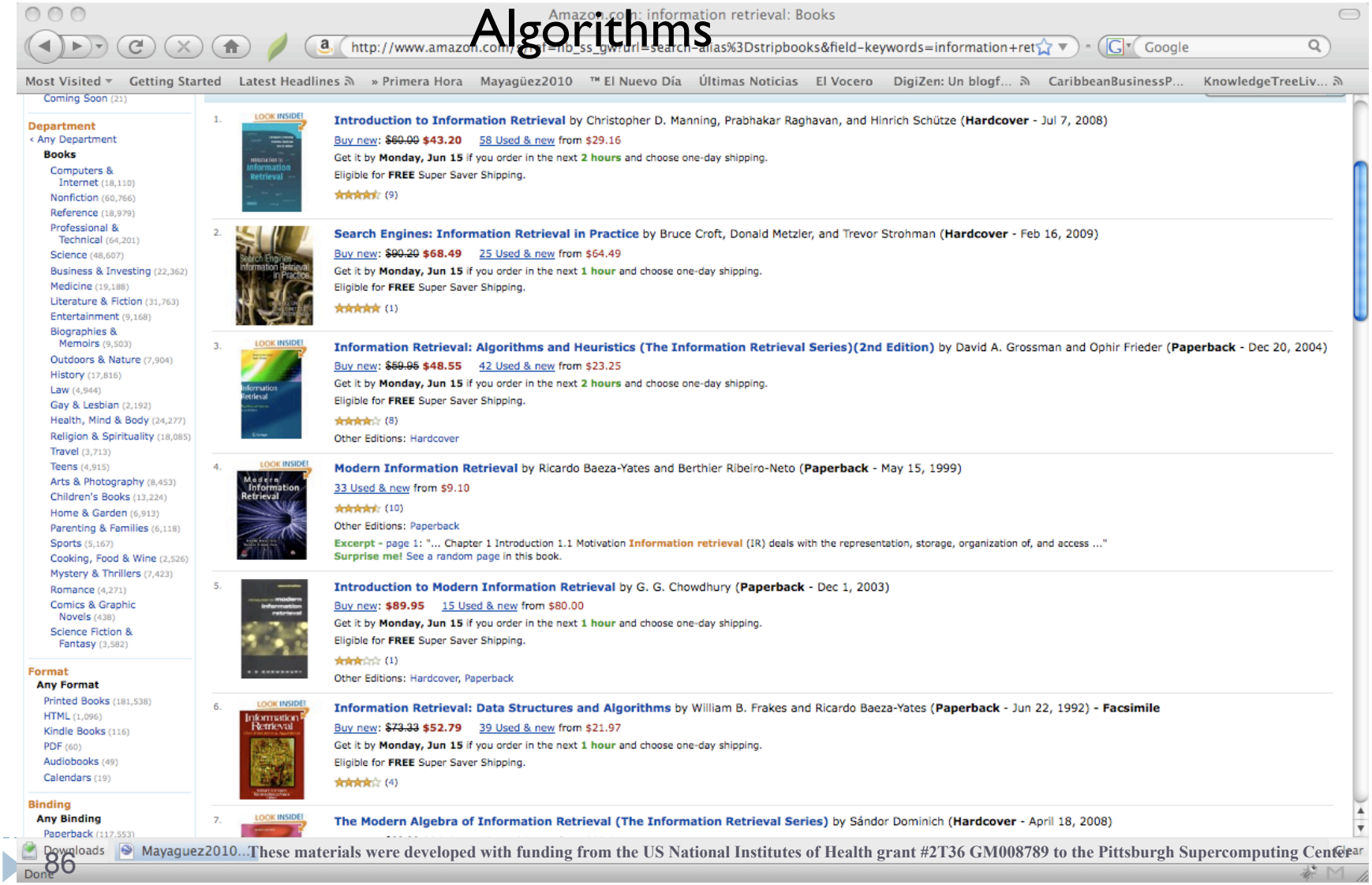
These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center

Outline

- Introduction and Examples
- Query Models
- Implementation Issues
- References

References

- Gerald Salton, Automatic Text Processing
- Baeza Yates, Information Retrieval Algorithms



These materials were developed with funding from the US National Institutes of Health grant #2T36 GM008789 to the Pittsburgh Supercomputing Center