

- 
- ▶ 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.
  - ▶ Unless otherwise specified, all the information contained within is Copyrighted © by Carnegie Mellon University. Permission is granted for use, modify, and reproduce these materials for teaching purposes.

- 
- ▶ 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

# Bioinformatics Data Management

---

Lecture 2

Unstructured Data Repositories

Bienvenido Vélez

UPR Mayaguez

*Reference: Bioinformatics for Dummies*

---



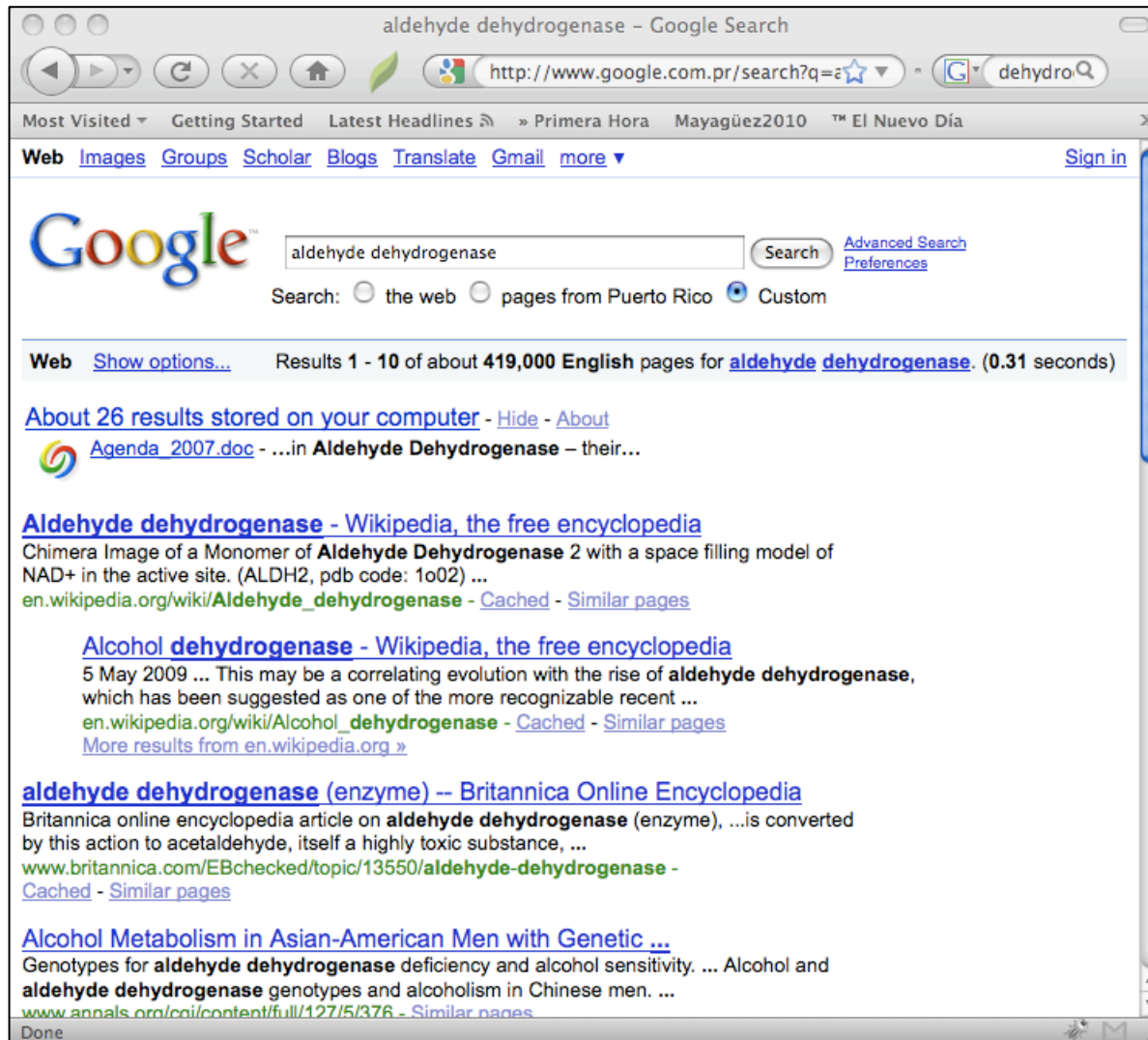
# Unstructured Data Repositories: Outline

---

- ▶ 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

# 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

# PubMed a Domain Specific Unstructured Database

The screenshot shows a web browser window displaying the PubMed search results for 'aldehyde dehydrogenase'. The browser's address bar shows the URL 'http://www.ncbi.nlm.nih.gov/sites/entrez'. The PubMed logo and navigation menu are visible at the top. The search results are displayed in a list format, with the first four items shown. Each item includes a checkbox, a title link, authors, journal information, and PMID. On the right side, there are sections for 'Also try:' and 'Titles with your search terms'. At the bottom right, there is a 'Recent Activity' box showing the current search query.

aldehyde dehydrogenase - PubMed Results

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

dehydrogenase

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

NCBI PubMed A service of the U.S. National Library of Medicine and the National Institutes of Health www.pubmed.gov 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.](#)  
Vondráček 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

# 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

# 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

# The Boolean Query Model: Definition

---

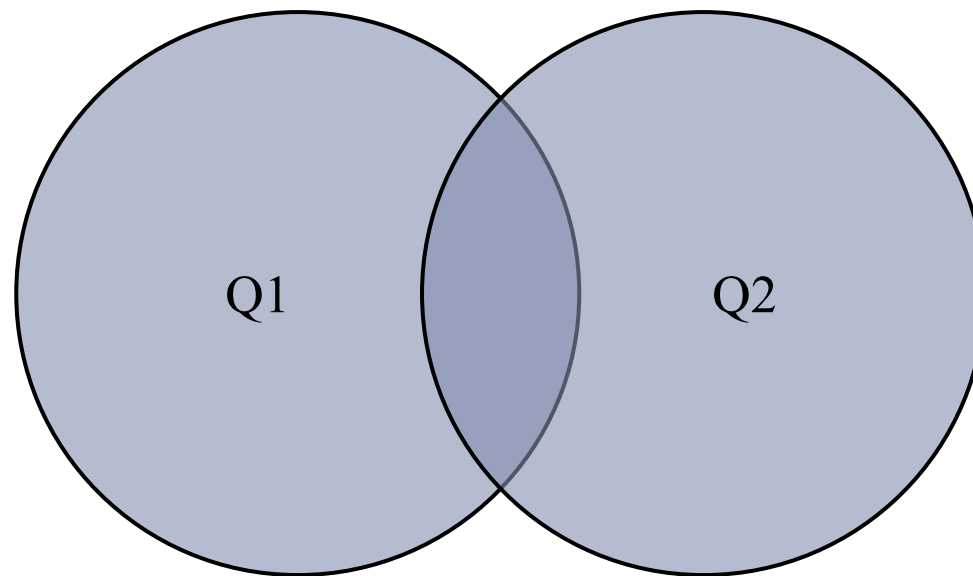
Query consists of:	Document Matches When:
t	Document contains at least one occurrence of term t
(not q)	Document does not match query q
(and q1 q2)	Document matches both queries q1 and q2
(or q1 q2)	Document matches either query q1 or q2

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

# Computing Boolean Queries

---

All documents



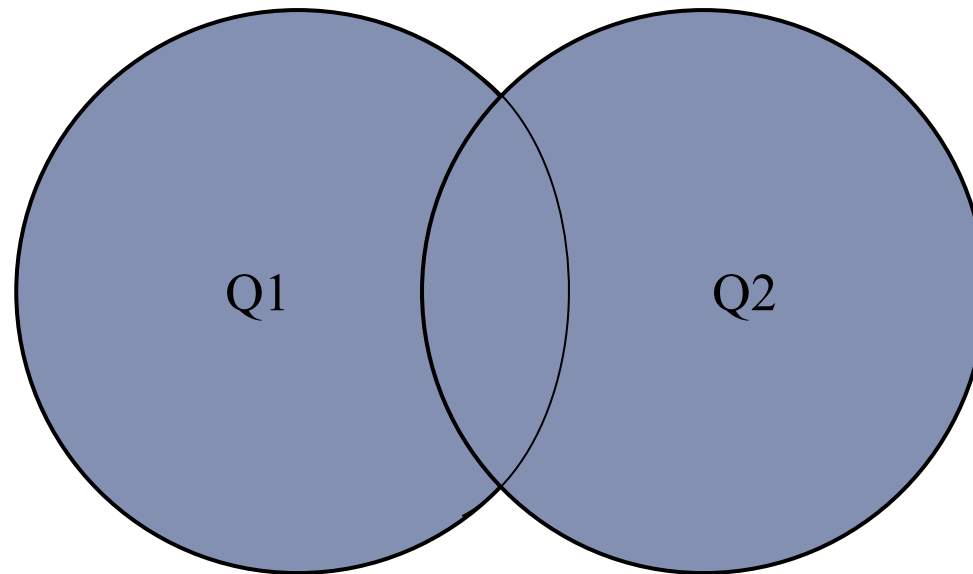
selected

Not  
selected

# Computing Boolean Queries

---

All documents



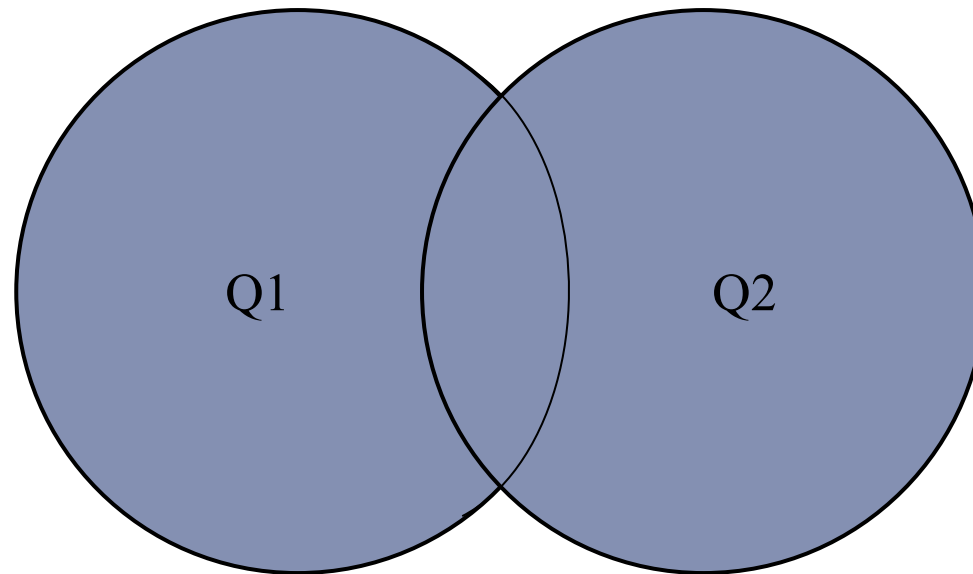
or Q1, Q2



# Computing Boolean Queries

---

All documents



or Q1, Q2

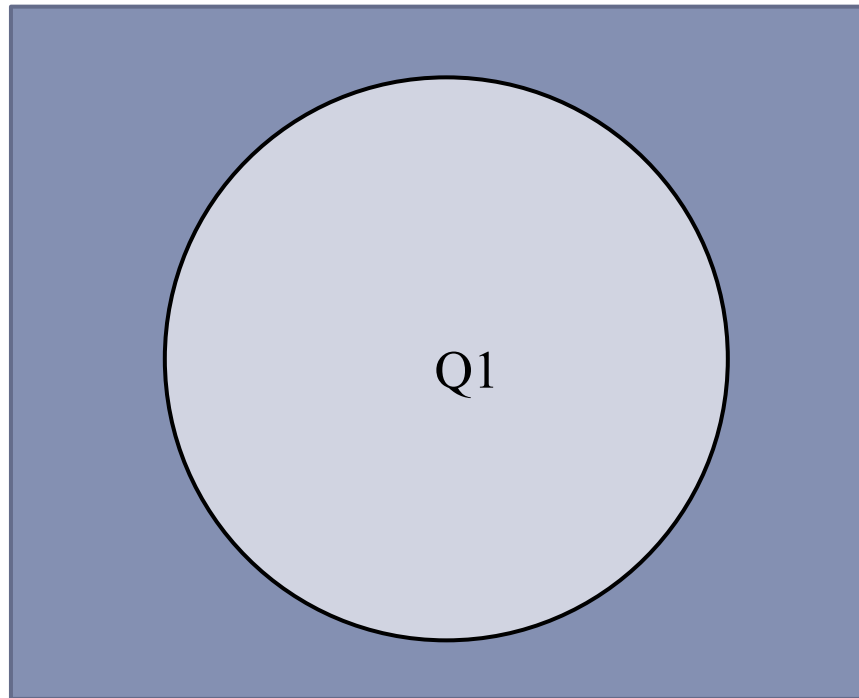


SLIDE  
HIDDEN

# Computing Boolean Queries

---

All documents



not Q1

selected

Not  
selected

# What Makes a Good Boolean Query?

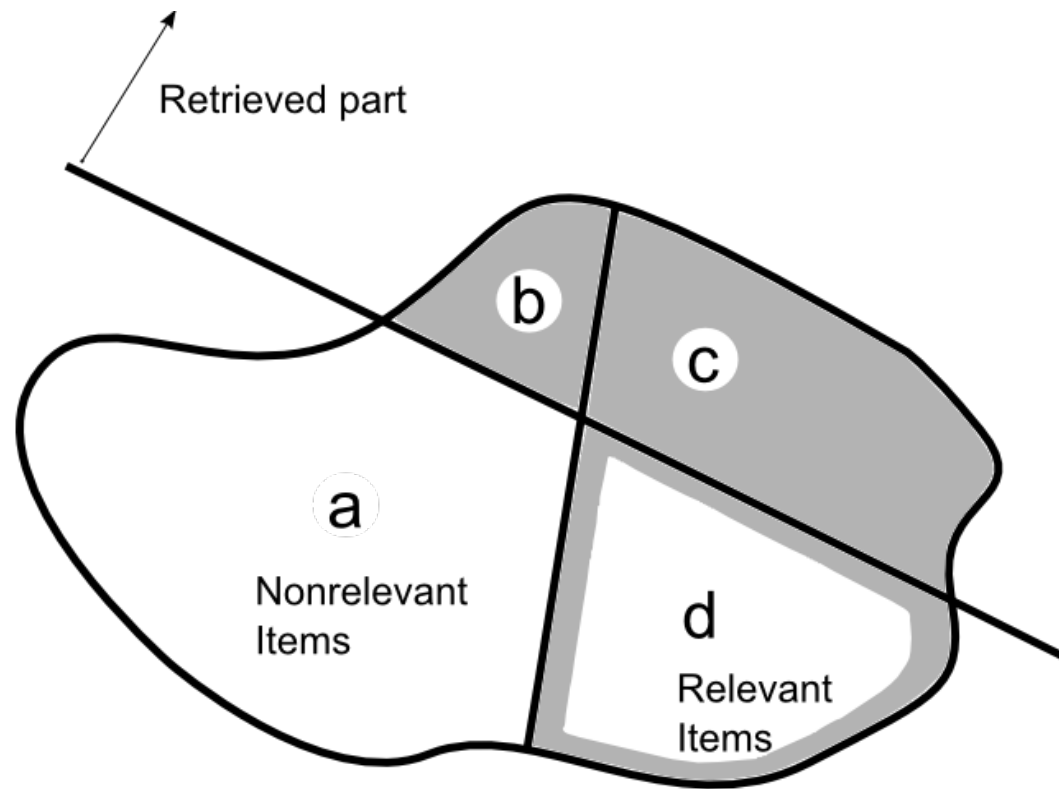
---

- ▶ 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.



# Measuring Boolean Query effectiveness

---



# Measuring Boolean Query Effectiveness

---

## ▶ Precision

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

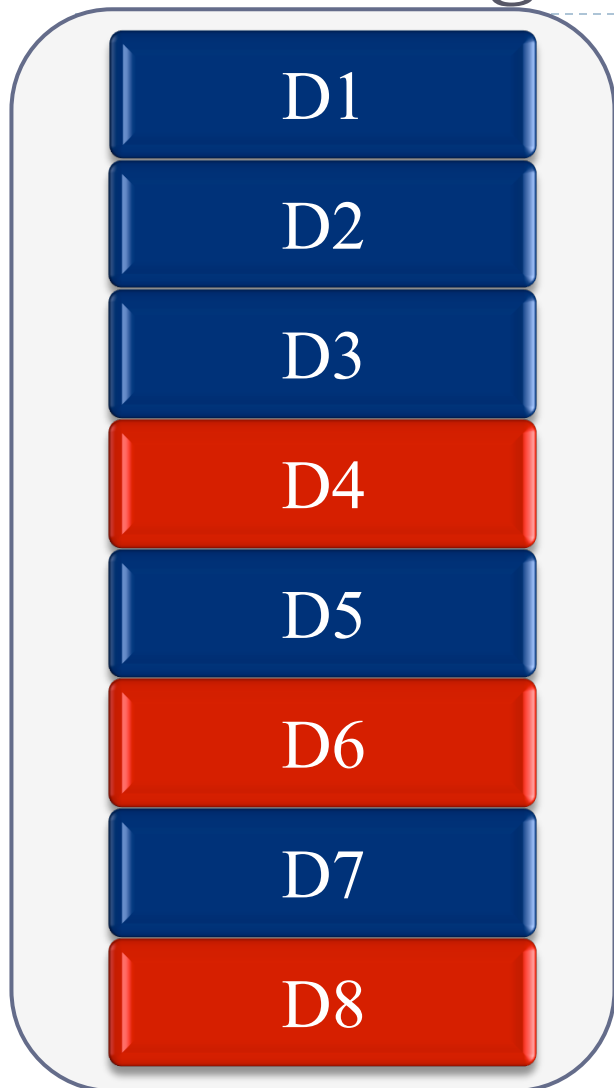
## ▶ Recall

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

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

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

# 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 =  $5/20 = 25\%$



# Query Models Subtopics

---

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

# 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

# The Vector Space Query Model: Definition

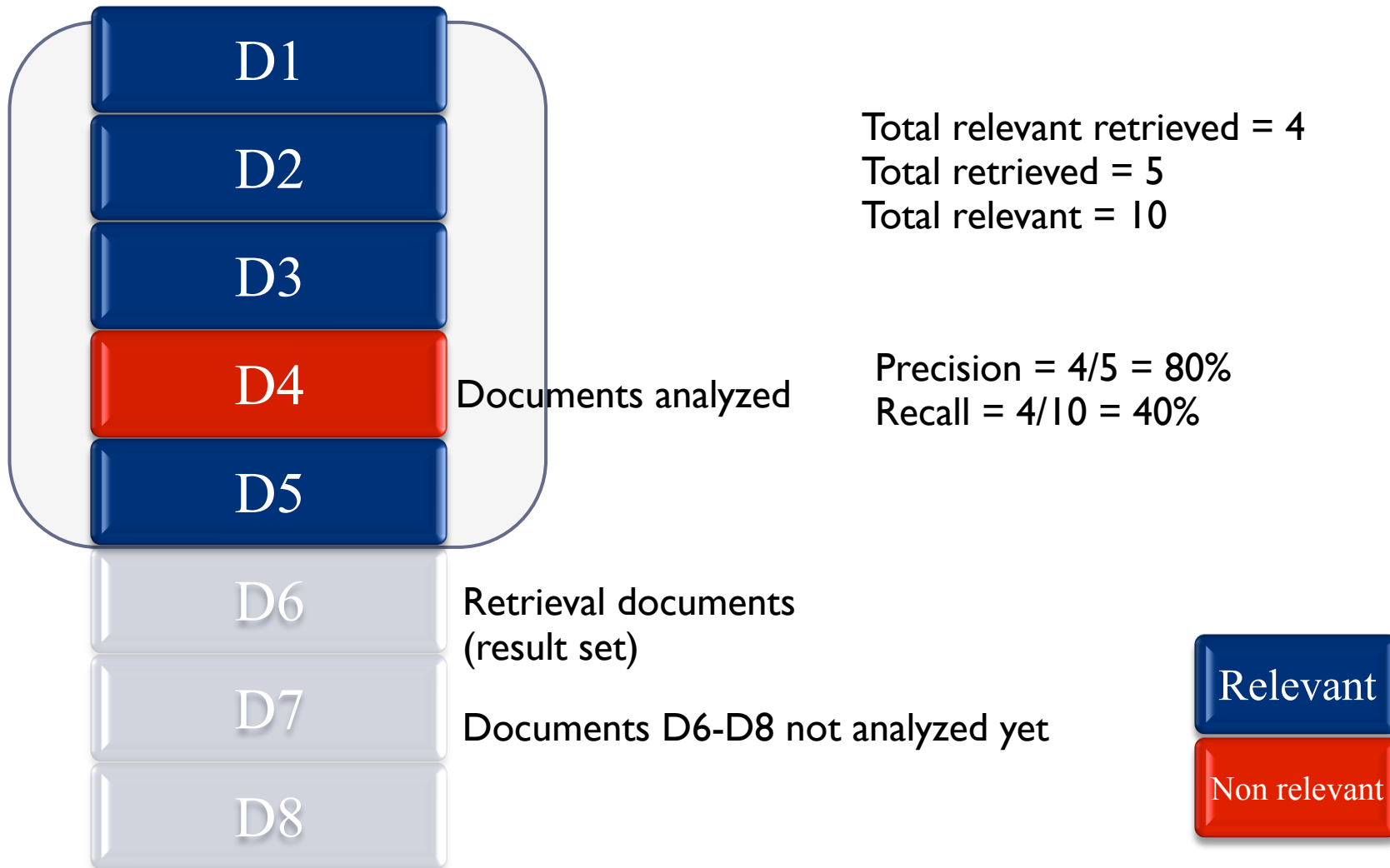
---

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

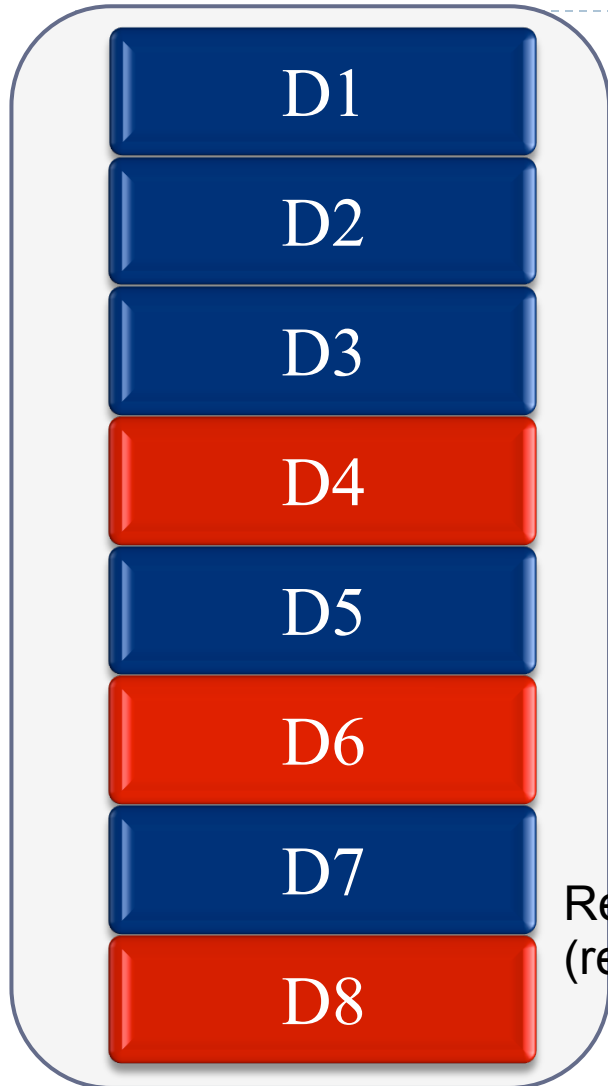
Query consists of:	Document Matches When:
aldehyde	Document contains term aldehyde
(aldehyde:10 dehydrogenase:25)	Document that contains terms aldehyde or dehydrogenase. Documents with dehydrogenase ranked higher.
(isocitrate:-3 aldehyde:3 dehydrogenase:10)	Document may contains terms aldehyde, dehydrogenase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

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

# Measuring Vector Space Query Effectiveness



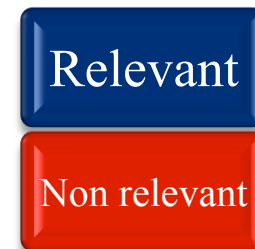
# Measuring Vector Space Query Effectiveness



Retrieval documents  
(result set)

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

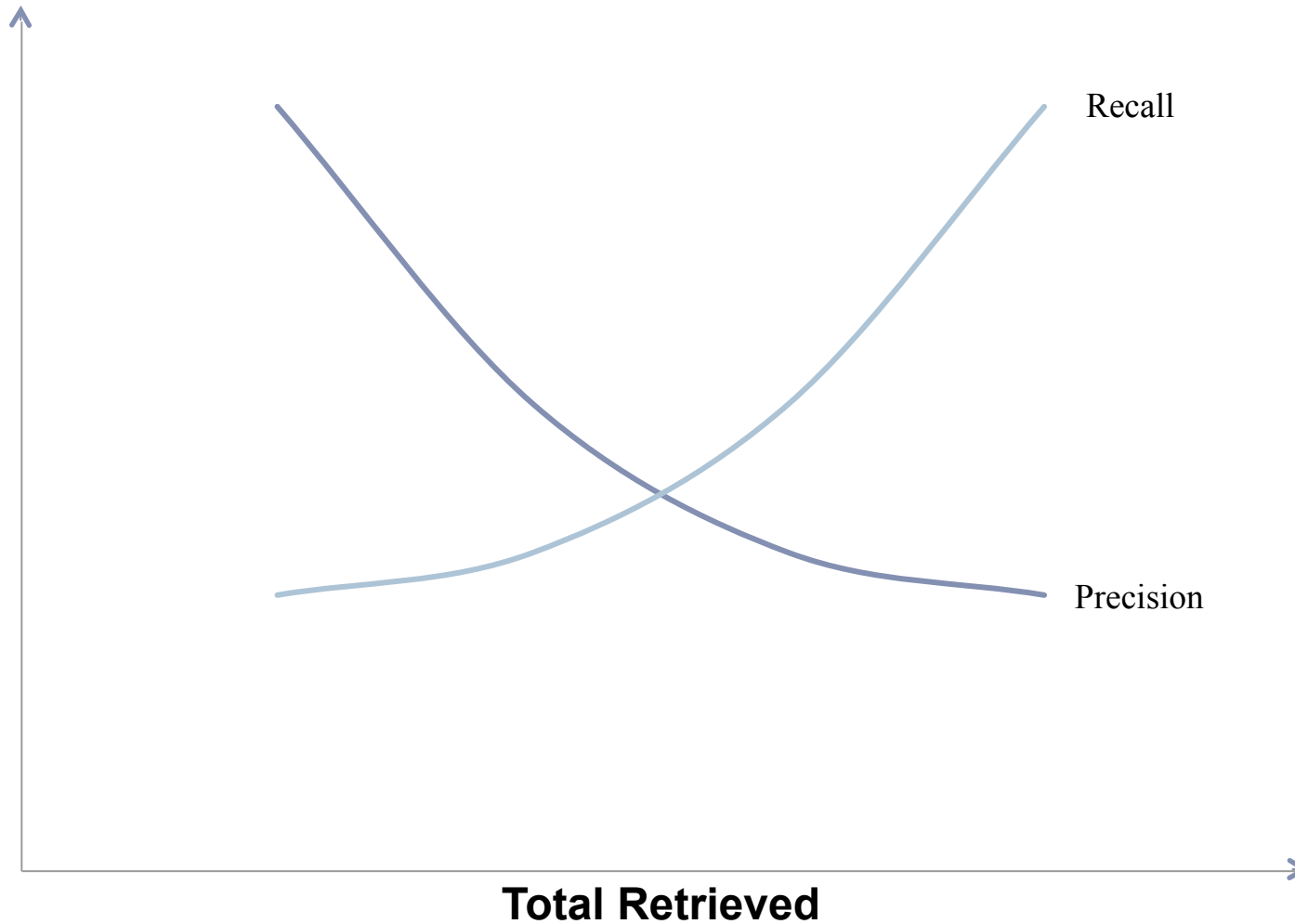
Precision =  $5/8 = 62.5\%$   
Recall =  $4/10 = 50\%$





# In Vector Space Query Model Precision and Recall constitute a tradeoff

---



# Examples of Similarity Measures

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

Legend:

$X = (x_1, x_2, \dots, x_t)$

$|X|$  = number of terms in X

$|X \cap Y|$  = number of terms appearing jointly in X and Y

SLIDE  
HIDDEN

Table from:

Gerald Salton,  
Automatic Text Processing

Page 318

# Examples of Similarity Measures

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

Legend:

$X = (x_1, x_2, \dots, x_t)$

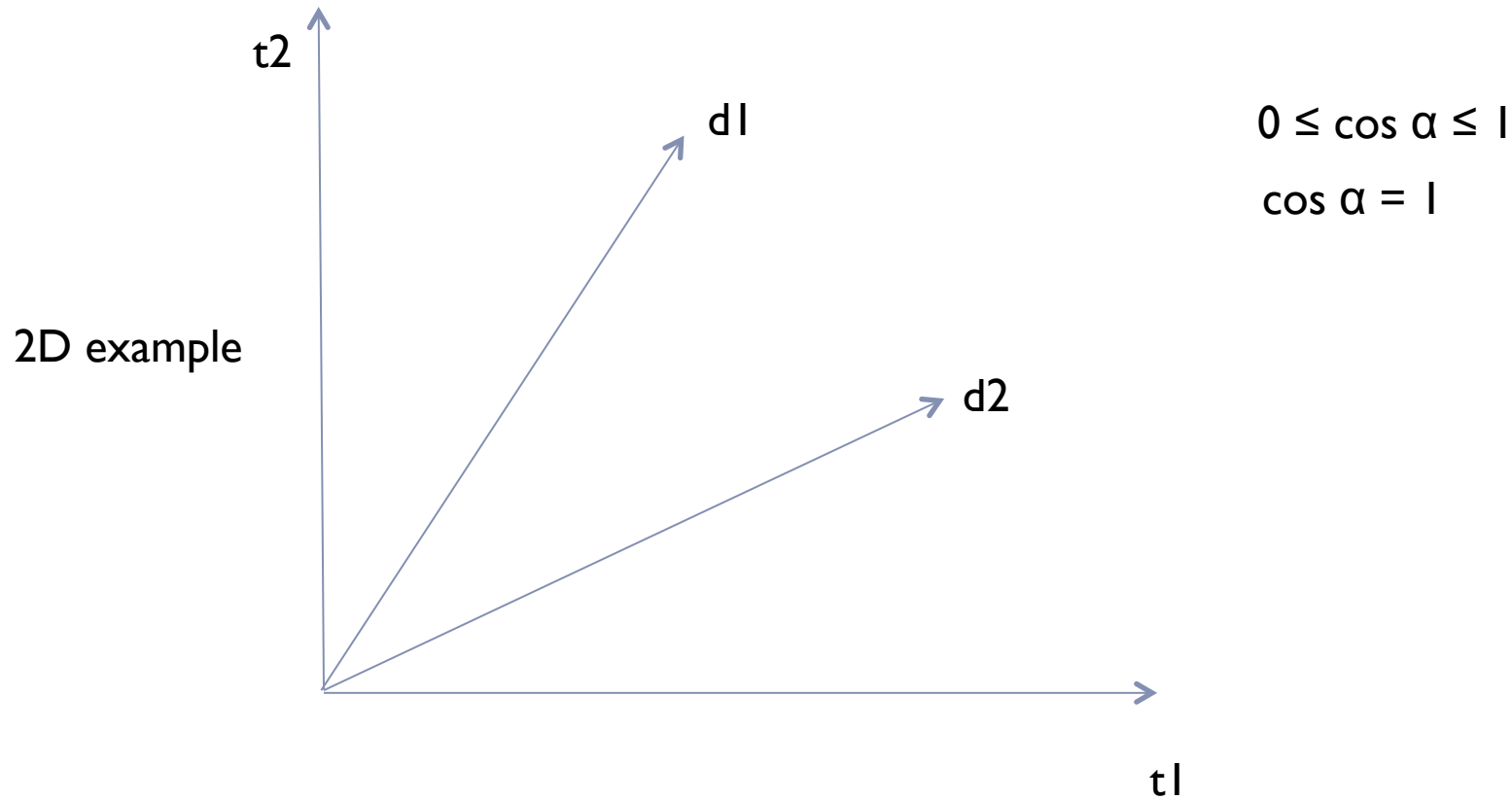
$|X|$  = number of terms in X

$|X \cap Y|$  = number of terms appearing jointly in X and Y

Table from:  
Gerald Salton,  
Automatic Text Processing  
Page 318

# Diced coefficient: A Normalized Similarity Measure

---



# Query Models Subtopics

---

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

# Google's Query Model



Advanced Search

[Advanced Search Tips](#) | [About G](#)

Use the form below and your advanced search will appear here

**Find web pages that have...**

all these words:

this exact wording or phrase:  [tip](#)

one or more of these words:  OR  OR  [tip](#)

**But don't show pages that have...**

any of these unwanted words:  [tip](#)

**Need more tools?**

Results per page:  [v](#)

Language:  [v](#)

File type:  [v](#)

Search within a site or domain:

(e.g. youtube.com, .edu)

[+ Date, usage rights, numeric range, and more](#)

**Annotations:**

- conjunction: points to the "OR" operators in the "one or more of these words" field.
- phrases: points to the "this exact wording or phrase" field.
- disjunction: points to the "any of these unwanted words" field.

# Google's Query Model

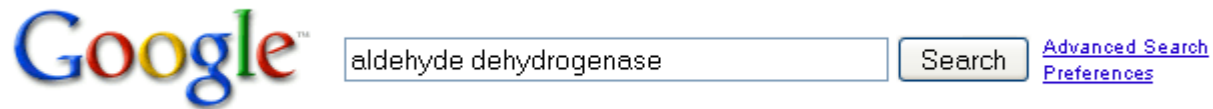
Operator	Description
“<phrase>”	Phrase search. By putting double quotes around a list of words, Google considers the exact words in that exact order without any change.
site:	Search within a specific website. Specify that your search results must come from a specific website.
-	Terms you want to exclude
*	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!

# Sample Advance Google's Query #1

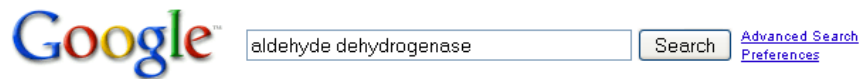
---

- ▶ Simple and common Google search query
  - ▶ aldehyde dehydrogenase





# Results of Google Query #1



Web [Show options...](#)

Results 1 - 10 of about 429,000 for [aldehyde dehydrogenase](#). (0.36 sec)

[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](http://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](http://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](http://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](http://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/](http://www.aldh.org/) - [Cached](#) - [Similar](#) - 

# Sample Advance Google's Query #2

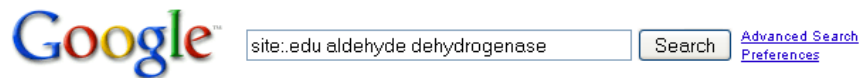
- ▶ Google search query within a site or domain
  - ▶ All educational sites
    - ▶ site:edu aldehyde dehydrogenase

The screenshot shows the Google Advanced Search interface. At the top left is the Google logo, followed by the text "Advanced Search". On the top right, there are links for "Advanced Search Tips" and "About C". The main search area contains the query "aldehyde dehydrogenase site:edu". Below this, there are several sections for refining the search:

- Find web pages that have...**
  - all these words:
  - this exact wording or phrase:
  - one or more of these words:  OR  OR
- But don't show pages that have...**
  - any of these unwanted words:
- Need more tools?**
  - Results per page:
  - Language:
  - File type:
  - Search within a site or domain:   
(e.g. youtube.com, .edu)

At the bottom right, there is a button labeled "Advanced Search".

# Results of Google Query #2



Web [Show options...](#)

Results 1 - 10 of about 21,800 for [site:.edu aldehyde dehydrogenase](#). (0.31 sec)

## [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](http://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/~deerfel/ALDH-SciVis.html](http://www.psc.edu/~deerfel/ALDH-SciVis.html) - [Cached](#) - [Similar](#) - 

## [10 Entries](#)

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...](http://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](http://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](http://gowiki.tamu.edu/wiki/index.php/Category:GO:0004029_|aldehyde_dehydrogenase_(NAD)_activity) - [Cached](#) - [Similar](#) - 

# 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

The screenshot shows the Google Advanced Search interface. At the top left is the Google logo, followed by the text "Advanced Search". On the top right, there are links for "Advanced Search Tips" and "About G". The main search area contains the query "aldehyde dehydrogenase -isocitrate". Below the search bar, there are several sections for refining the search:

- Find web pages that have...**
  - all these words:  [tip](#)
  - this exact wording or phrase:  [tip](#)
  - one or more of these words:  OR  OR  [tip](#)
- But don't show pages that have...**
  - any of these unwanted words:  [tip](#)
- Need more tools?**
  - Results per page:
  - Language:
  - File type:
  - Search within a site or domain:

At the bottom left, there is a link: [+ Date, usage rights, numeric range, and more](#). At the bottom right, there is a button labeled "Advanced Search".

# Results of Google Query #3



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: 1a02) ...

[en.wikipedia.org/wiki/Aldehyde\\_dehydrogenase](http://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](http://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](http://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/dispmim.cgi?id=100650](http://www.ncbi.nlm.nih.gov/entrez/dispmim.cgi?id=100650) - [Cached](#) - [Similar](#) -

## [aldehyde dehydrogenase \(enzyme\)](#) -- Britannica Online Encyclopedia

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

# Unstructured Data Repositories: Outline

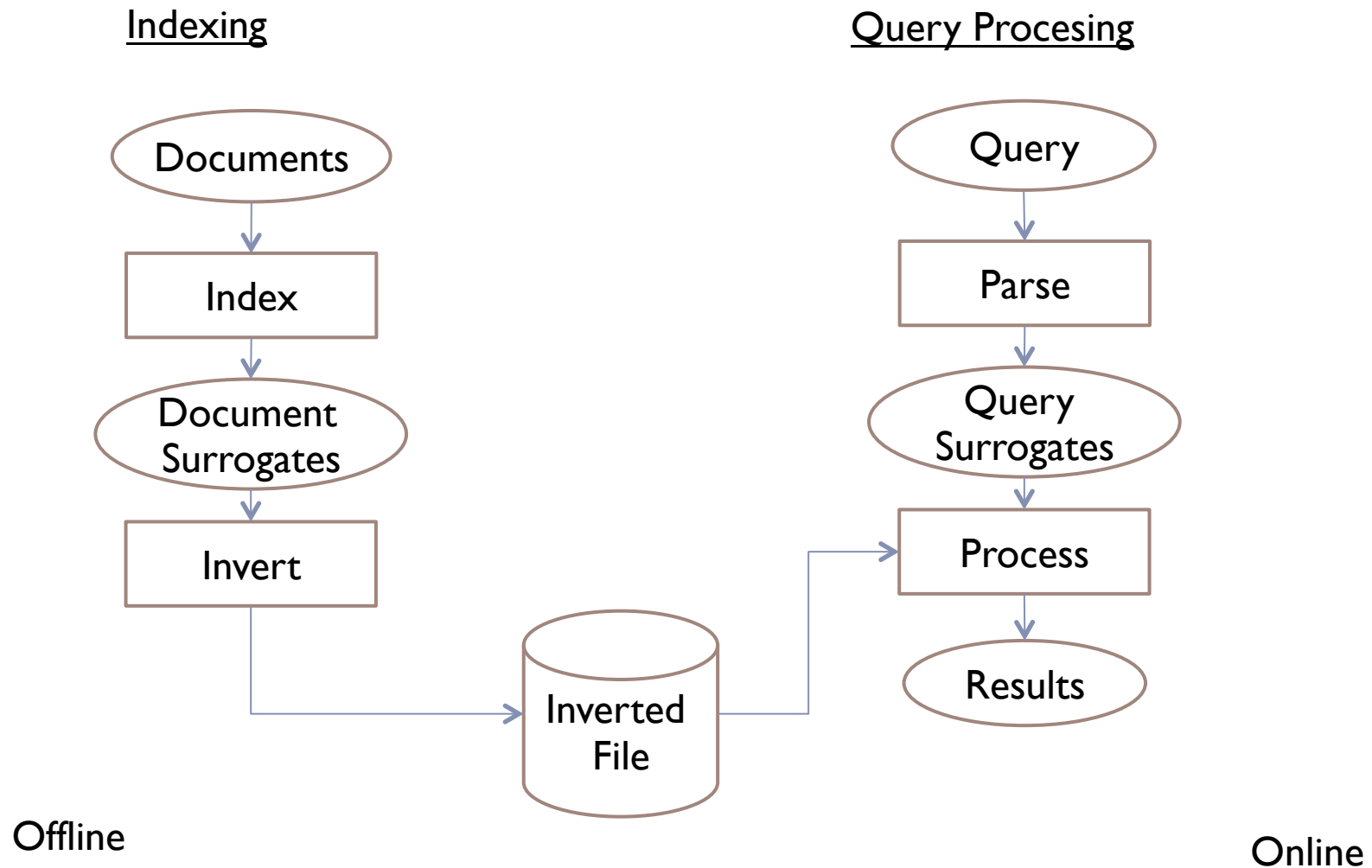
---

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



# Architecture of an Information Retrieval System

---



# Implementation Issues Subtopics

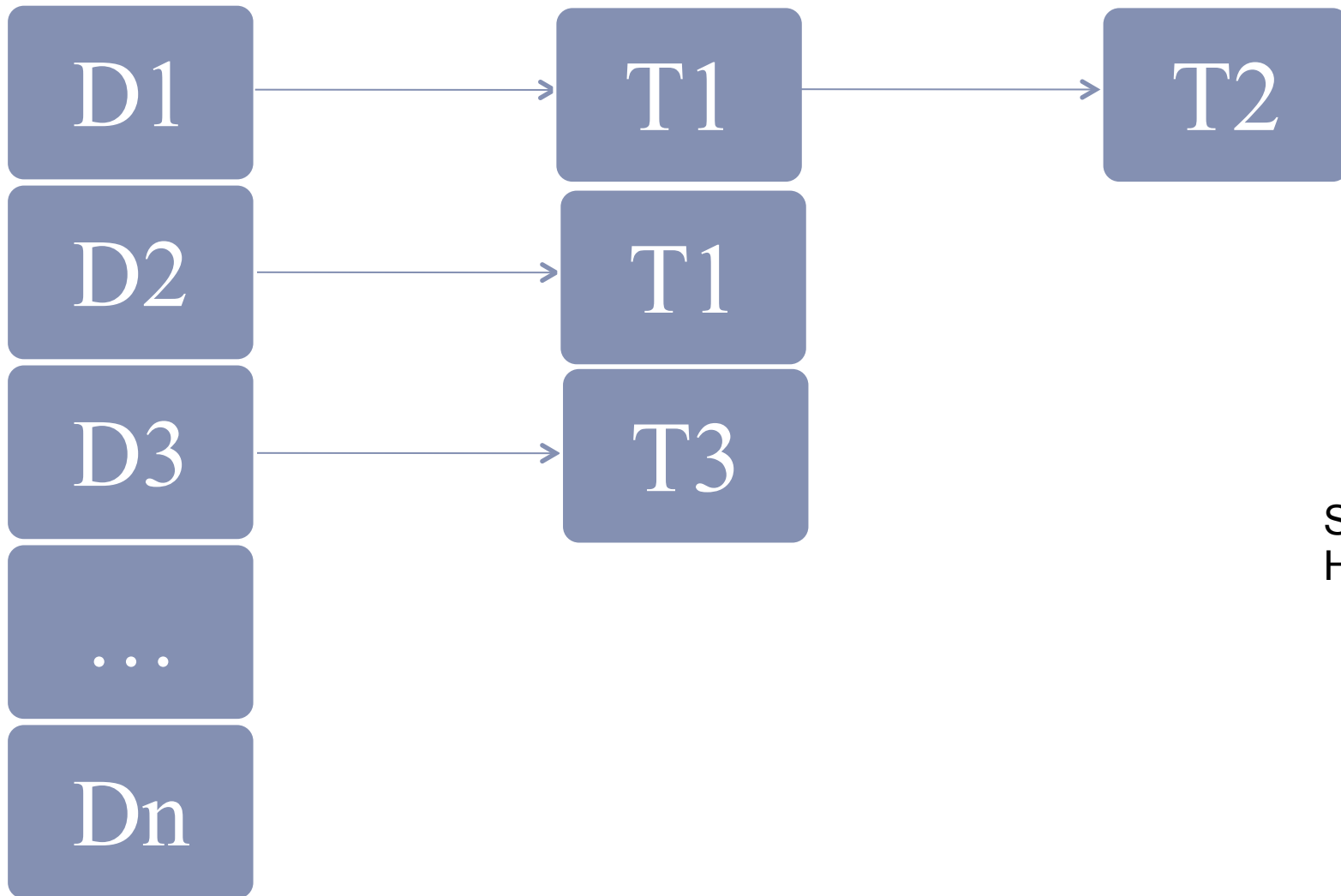
---

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



# The Inverted File

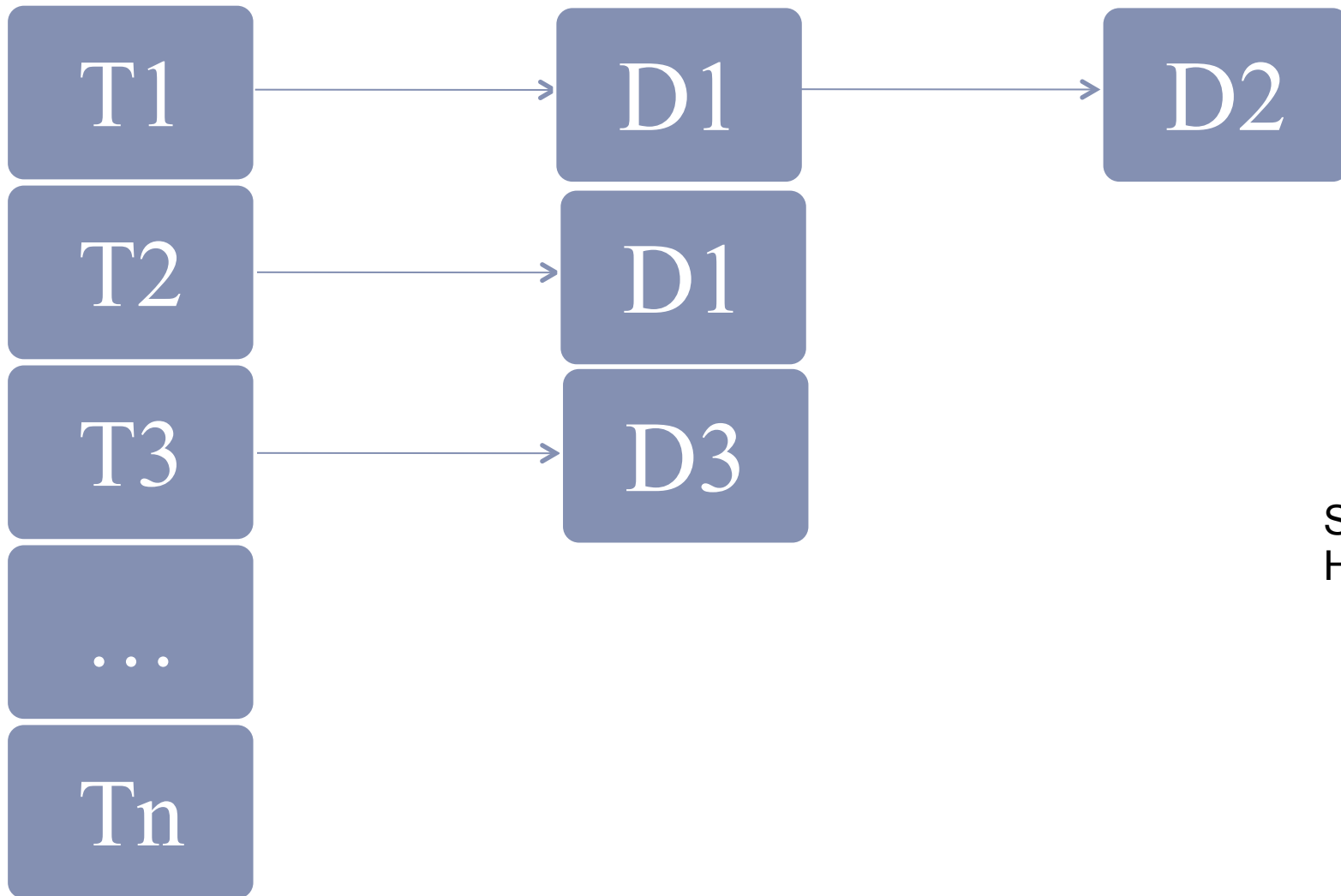
---



SLIDE  
HIDDEN

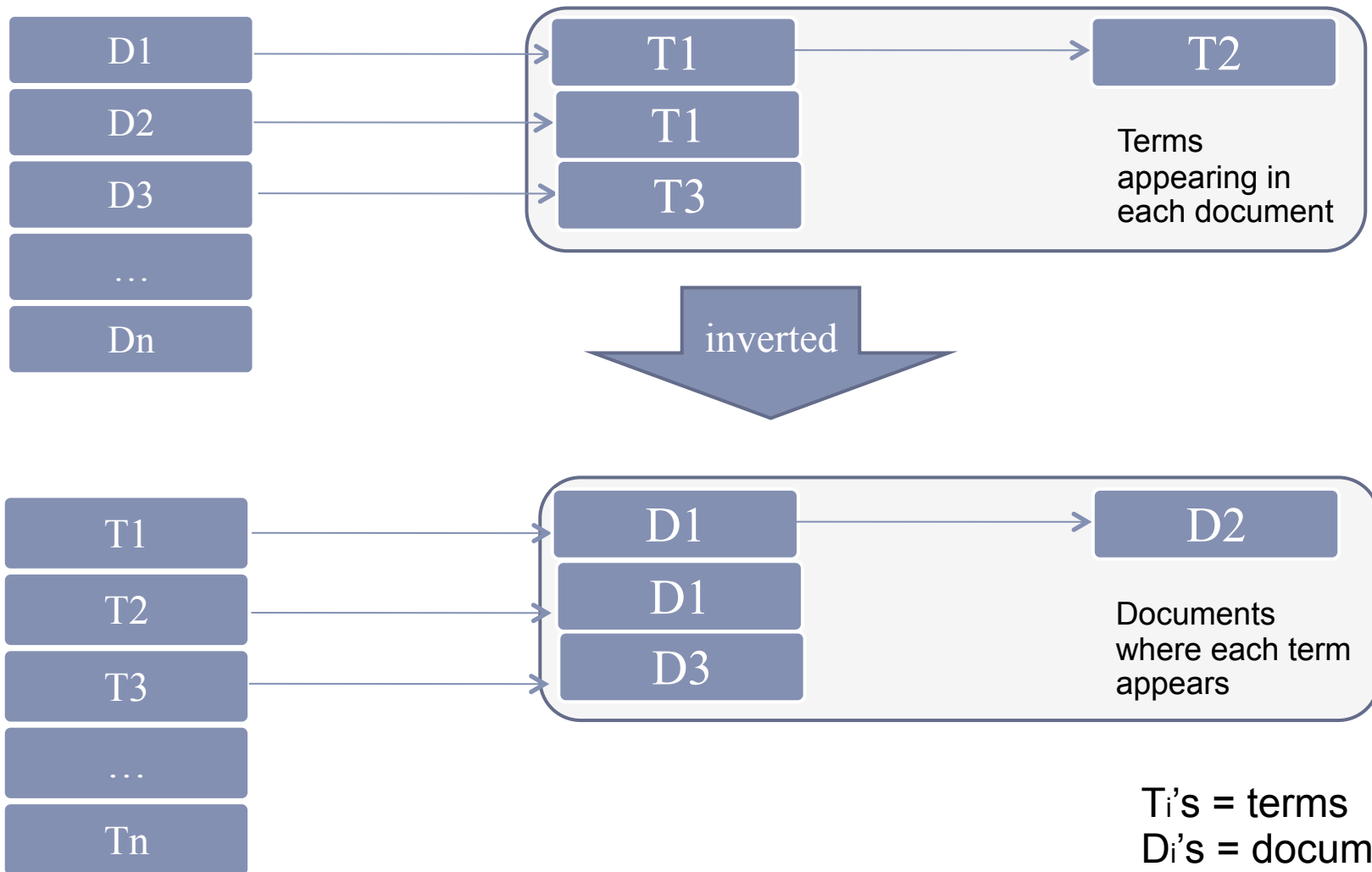
# The Inverted File

---



SLIDE  
HIDDEN

# The Inverted File Data Structure



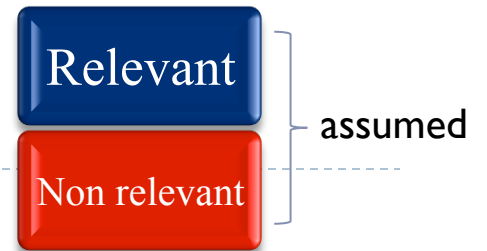
# Implementation Issues Subtopics

---

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

# An Example Document Set

---



## ▶ Example of Document 1

aldehyde dehydrogenase

## ▶ Example of Document 2

aldehyde isocitrate dehydrogenase

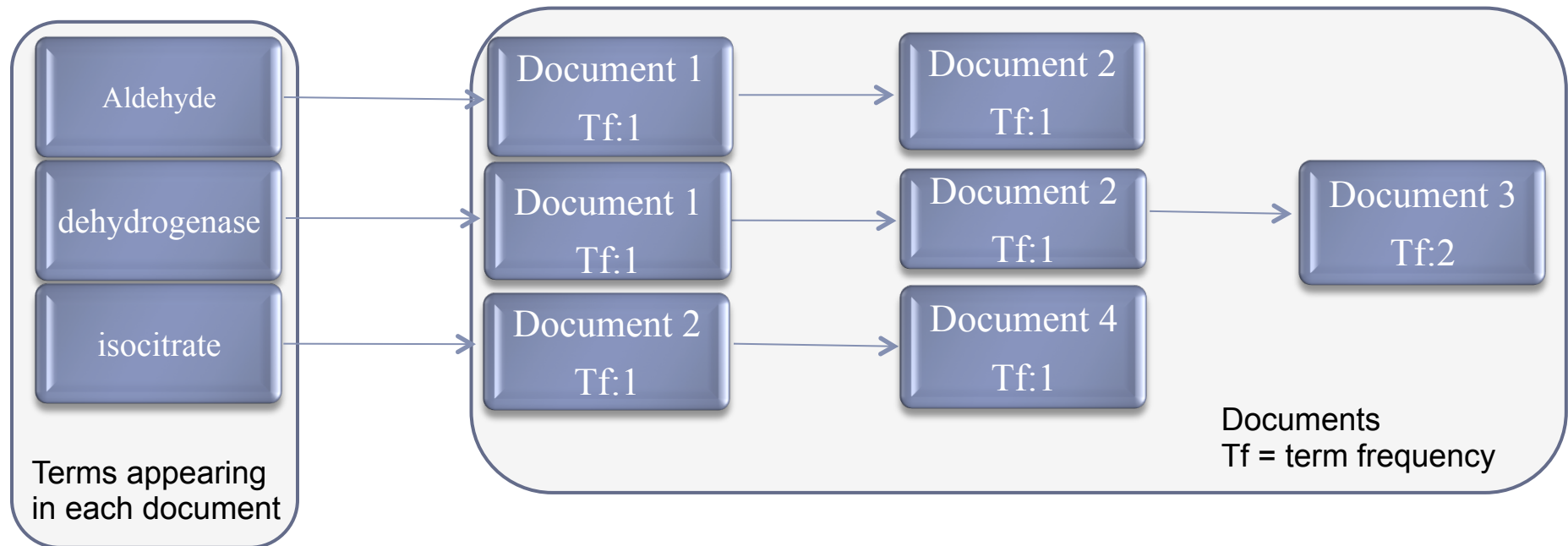
## ▶ Example of Document 3

dehydrogenase dehydrogenase

## ▶ Example of Document 4

isocitrate

# The Inverted File Data Structure for Example Document Set

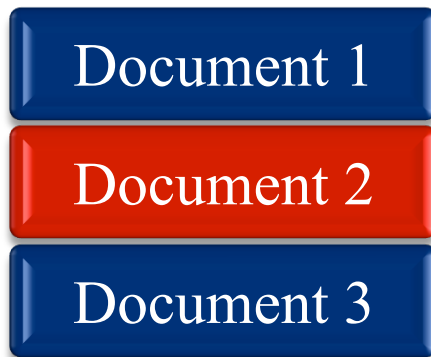


# Inverted File

Example of boolean queries

(or aldehyde dehydrogenase) (and aldehyde dehydrogenase)

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



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

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

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

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

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

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

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



# The Inverted File

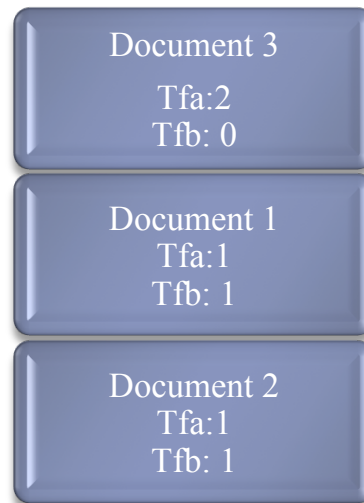
## Example of vector space queries

aldehyde



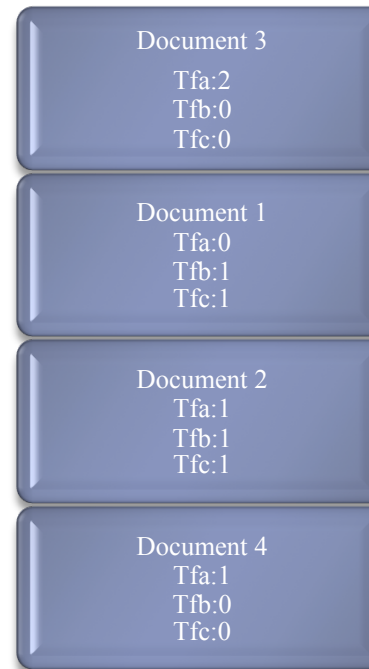
Sim(Q,D)=

(aldehyde:10  
dehydrogenase:25)



Sim(Q,D)=

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



Sim(Q,D)=

Verif orden  
este correcto

Escribir las  
similaridades

SLIDE  
HIDDEN



# The Inverted File

Inner product similarity

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

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

$$\text{sim}_{\text{inner}}(Q, D_1) = \sum_{i=1}^t q_i \cdot d_i = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D$$

$$\text{sim}_{\text{inner}}(Q, D_1) = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D = 3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (0) = 13$$

# The Inverted File

Dice vector similarity

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

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

$$\text{sim}_{\text{dice}}(Q, D_1) = \frac{2 \sum_{i=1}^t q_i \cdot d_i}{\sum_{i=1}^t q_i^2 + \sum_{i=1}^t d_i^2} = \frac{2(t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2}$$

$$\text{sim}_{\text{dice}}(Q, D_1) = \frac{2(3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (0))}{3^2 + 1^2 + 10^2 + 1^2 + (-3)^2 + (0)^2} = \frac{2(13)}{120} = \frac{26}{120} = 0.2167$$

# The Inverted File

## Cosine similarity

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

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

$$\text{sim}_{\cos}(Q, D_1) = \frac{\sum_{i=1}^t q_i \cdot d_i}{\sqrt{\sum_{i=1}^t q_i^2 \cdot \sum_{i=1}^t d_i^2}} = \frac{t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D}{\sqrt{((t_1^Q)^2 + (t_2^Q)^2 + (t_3^Q)^2) \cdot ((t_1^D)^2 + (t_2^D)^2 + (t_3^D)^2)}}$$
$$\text{sim}_{\cos}(Q, D_1) = \frac{3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (0)}{\sqrt{(3^2 + 10^2 + (-3)^2) \cdot (1^2 + 1^2 + 0^2)}} = \frac{13}{\sqrt{118 \cdot 2}} = \frac{13}{\sqrt{236}} = 0.8462$$

# The Inverted File

Jaccard similarity :

Query

3
10
-3

aldehyde  
dehydrogenase  
isocitrate

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

Document 1

1
1
0

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

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

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

# The Inverted File

Inner product similarity

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

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

$$\text{sim}_{\text{inner}}(Q, D_1) = \sum_{i=1}^t q_i \cdot d_i = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D$$

$$\text{sim}_{\text{inner}}(Q, D_1) = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D = 3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (1) = 10$$

# The Inverted File

Dice vector similarity

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

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

$$\text{sim}_{\text{dice}}(Q, D_1) = \frac{2 \sum_{i=1}^t q_i \cdot d_i}{\sum_{i=1}^t q_i^2 + \sum_{i=1}^t d_i^2} = \frac{2(t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2}$$

$$\text{sim}_{\text{dice}}(Q, D_1) = \frac{2(3 \cdot 1 + 10 \cdot 1 + (-3) \cdot (1))}{3^2 + 1^2 + 10^2 + 1^2 + (-3)^2 + (1)^2} = \frac{2(10)}{121} = \frac{20}{121} = 0.1653$$

# The Inverted File

Cosine similarity

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

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

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

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

# The Inverted File

Jaccard similarity :

Query

3
10
-3

aldehyde  
dehydrogenase  
isocitrate

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

Document 2

1
1
1

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

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

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



# The Inverted File

Inner product similarity

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

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

$$\text{sim}_{\text{inner}}(Q, D_1) = \sum_{i=1}^t q_i \cdot d_i = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D$$

$$\text{sim}_{\text{inner}}(Q, D_1) = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D = 3 \cdot 0 + 10 \cdot 2 + (-3) \cdot (0) = 20$$

# The Inverted File

Dice vector similarity

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

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

$$\text{sim}_{\text{dice}}(Q, D_1) = \frac{2 \sum_{i=1}^t q_i \cdot d_i}{\sum_{i=1}^t q_i^2 + \sum_{i=1}^t d_i^2} = \frac{2(t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2}$$

$$\text{sim}_{\text{dice}}(Q, D_1) = \frac{2(3 \cdot 0 + 10 \cdot 2 + (-3) \cdot (0))}{3^2 + 0^2 + 10^2 + 2^2 + (-3)^2 + (0)^2} = \frac{2(20)}{122} = \frac{40}{122} = 0.3279$$

# The Inverted File

Cosine similarity

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

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

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

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

# The Inverted File

Jaccard similarity :

Query

3
10
-3

aldehyde  
dehydrogenase  
isocitrate

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

Document 3

0
2
0

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

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

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

# The Inverted File

Inner product similarity

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

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

$$\text{sim}_{\text{inner}}(Q, D_1) = \sum_{i=1}^t q_i \cdot d_i = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D$$

$$\text{sim}_{\text{inner}}(Q, D_1) = t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D = 3 \cdot 0 + 10 \cdot 0 + (-3) \cdot (1) = -3$$

# The Inverted File

Dice vector similarity

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

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

$$\text{sim}_{\text{dice}}(Q, D_1) = \frac{2 \sum_{i=1}^t q_i \cdot d_i}{\sum_{i=1}^t q_i^2 + \sum_{i=1}^t d_i^2} = \frac{2(t_1^Q \cdot t_1^D + t_2^Q \cdot t_2^D + t_3^Q \cdot t_3^D)}{(t_1^Q)^2 + (t_1^D)^2 + (t_2^Q)^2 + (t_2^D)^2 + (t_3^Q)^2 + (t_3^D)^2}$$

$$\text{sim}_{\text{dice}}(Q, D_1) = \frac{2(3 \cdot 0 + 10 \cdot 0 + (-3) \cdot (1))}{3^2 + 0^2 + 10^2 + 0^2 + (-3)^2 + (1)^2} = \frac{2(-3)}{119} = \frac{-6}{119} = -0.0504$$

# The Inverted File

## Cosine similarity

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

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

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

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

# The Inverted File

Jaccard similarity :

Query

3
10
-3

aldehyde  
dehydrogenase  
isocitrate

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

Document 4

0
0
1

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

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

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



# 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 = = %



# The Inverted File

---

- ▶ Example of Extended Boolean Query

SLIDE  
HIDDEN

# Implementation Issues Subtopics

---

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

# Inverted File Generation Algorithm

---

# Implementation Issues Subtopics

---

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

# Scaleable Inverted File Management

---

Inverted file management for scalability  
Offline inverted file  
Lazy query evaluation

Under construction

# Implementation Issues

---

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

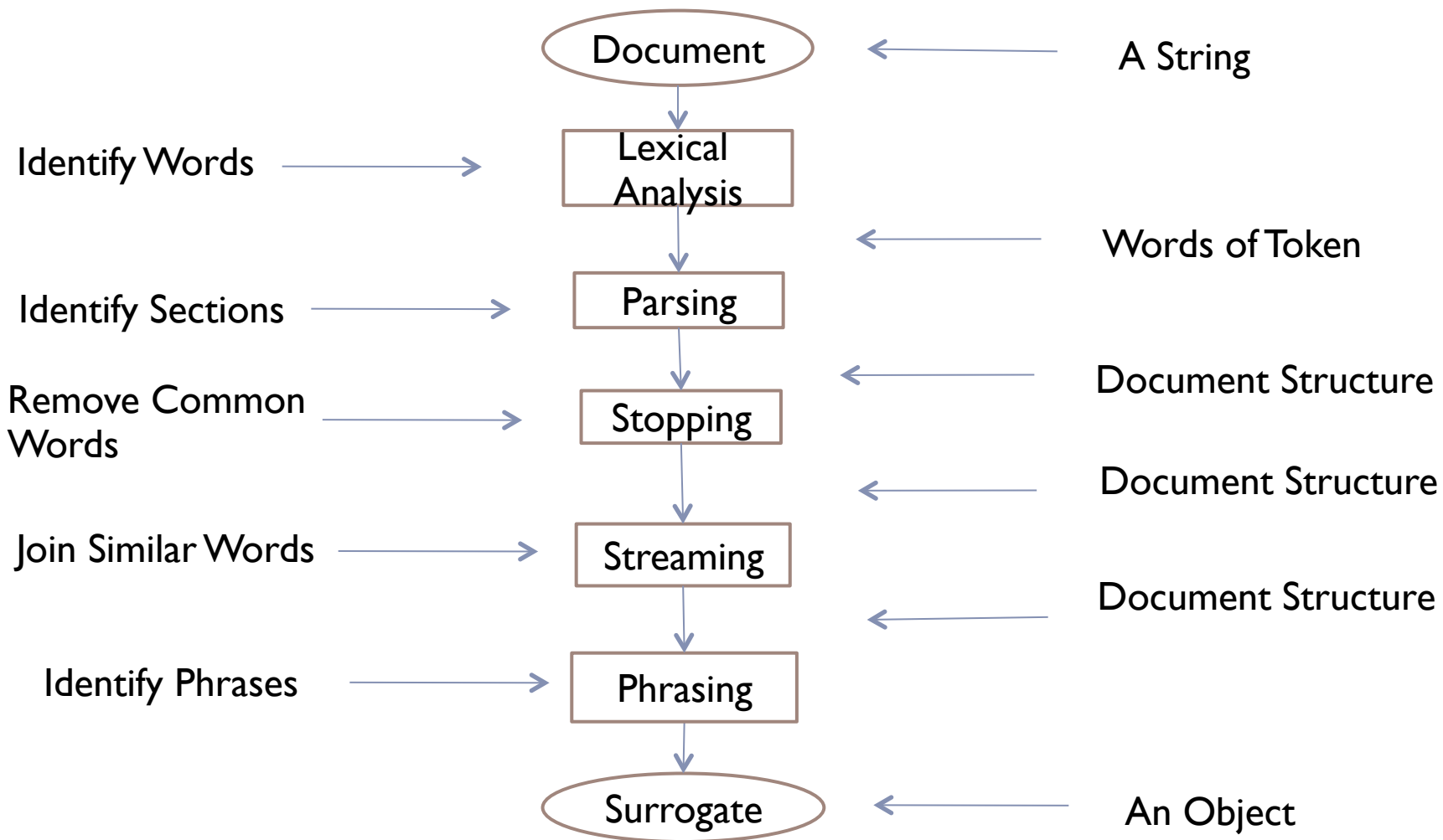
# Automatic Indexing Subtopics

---

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



# Automatic Indexing at a Glance



# 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...

# 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**

# 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**

# Google's English Stopwords

---

**I  
a  
about  
an  
are  
as  
at  
be  
by  
com  
de  
en  
for  
from  
how**

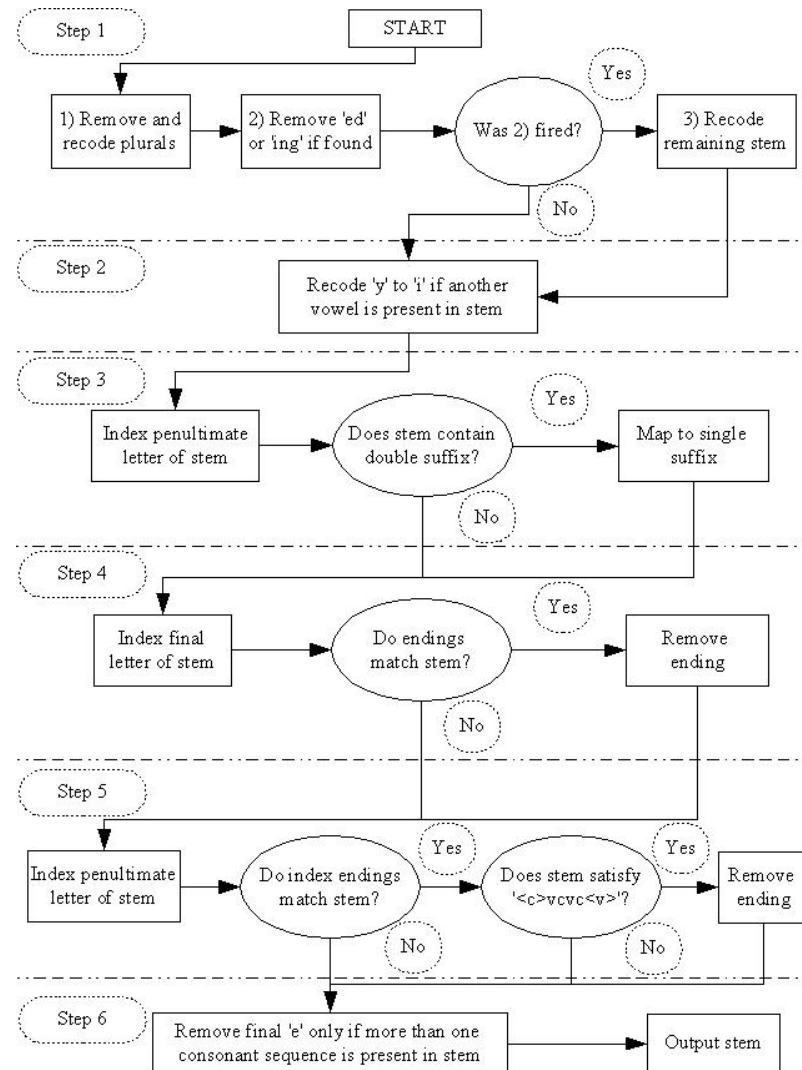
**in  
is  
it  
la  
of  
on  
or  
that  
the  
this  
to  
was  
what  
when  
where**

**who  
will  
with  
und  
the  
www**

# Google's Spanish Stopwords

un	para	cada	lo	aquellos	usais	trabajamos
una	atras	fin	las	aquellas	usan	trabajais
unas	porque	incluso	los	sus	emplear	trabajan
unos	por qué	primero	su	entonces	empleo	podria
uno	estado	desde	aqui	tiempo	empleas	podrias
sobre	estaba	conseguir	mio	verdad	emplean	podriamos
todo	ante	consigo	tuyo	verdadero	empleamos	podrian
también	antes	consigue	ellos	verdadera	empleais	podriais
tras	siendo	consigues	ellas	cierto	valor	yo
otro	ambos	conseguimos	nos	ciertos	muy	aquel
algún	pero	consiguen	nosotros	cierta	era	
alguno	por	ir	vosotros	ciertas	eras	
alguna	poder	voy	vosotras	intentar	eramos	
algunos	puede	va	si	intento	eran	
algunas	puedo	vamosa	dentro	intenta	modo	
ser	podemos	vais	solo	intentas	bien	
es	podeis	van	solamente	intentamos	cual	
soy	pueden	vaya	saber	intentais	cuando	
eres	fui	gueno	sabes	intentan	donde	
somos	fue	ha	sabe	dos	mientras	
sois	fuimos	tener	sabemos	bajo	quien	
estoy	fueron	tengo	sabeis	arriba	con	
esta	hacer	tiene	saben	encima	entre	
estamos	hago	tenemos	ultimo	usar	sin	
estais	hace	teneis	largo	uso	trabajo	
están	hacemos	tienen	bastante	usas	trabajar	
como	haceis	el	haces	usa	trabajas	
en	hacen	la	muchos	usamos	trabaja	

# Porter's Stemming Algorithm



# How to deal with other Common Terms?

---

Collection Frequency = what fraction of all documents contains term

CF

Very  
uncommon  
0%



Very  
common  
100%



# Inverse Document Frequency (IDF)

---

Let  $D$  be set of all documents

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

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

# Revisiting Similarity Measures

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

Legend:

$X = (x_1, x_2, \dots, x_t)$

$|X|$  = number of terms in X

$|X \cap Y|$  = number of terms appearing jointly in X and Y

Table from:  
Gerald Salton,  
Automatic Text Processing  
Page 318

# Phrase Recognition can Improve Decision and Recall

Google  Search [Advanced Search](#) [Preferences](#)

---

Web [Show options...](#) Results 1 - 10 of about 37,900,000 for **bill or gates**. (0.24 sec)

**Bill Gates** - [Wikipedia, the free encyclopedia](#) [What's This?](#)  
William Henry "Bill" Gates III (born October 28, 1955) is an American business magnate, philanthropist, author, and chairman of Microsoft, the software ...  
[en.wikipedia.org/wiki/Bill\\_Gates](http://en.wikipedia.org/wiki/Bill_Gates) - [Cached](#) - [Similar](#) -

**The Bill & Melinda Gates Foundation** [What's This?](#)  
The Bill & Melinda Gates Foundation is dedicated to bringing innovations in health and learning to the global community.  
[www.gatesfoundation.org/](http://www.gatesfoundation.org/) - [Cached](#) - [Similar](#) -

**Bill Gates: Chairman**   
The official biography from Microsoft of its Chairman and Chief Software Architect. Also his speeches and publications.  
[www.microsoft.com/BillGates/](http://www.microsoft.com/BillGates/) - [Cached](#) - [Similar](#) -

**Bill Gates: Chairman, Microsoft Corp.**   
30 Jul 2007 ... William (Bill) H. Gates is chairman of Microsoft Corporation, ... Bottom row: Bill Gates, Andrea Lewis, Marla Wood, Paul Allen. ...  
[www.microsoft.com/billgates/bio.aspx](http://www.microsoft.com/billgates/bio.aspx) - [Cached](#) - [Similar](#) -   
[More results from www.microsoft.com >](#)

Books by **Bill Gates**  
[Business the Speed of Thought: Succeeding in...](#) - 2000 - 508 pages  
[Business at the Speed of Thought: Succeeding...](#) - 2001 - 496 pages  
[Bill Gates Speaks: Insight from the World's ...](#) - 2001 - 276 pages  
[books.google.com](http://books.google.com)

Searching:  
bill or gates  
vs.  
"bill gates"

Google  Search [Advanced Search](#) [Preferences](#)

---


Web [Show options...](#) Results 1 - 10 of about 19,200,000 for **"bill gates"**. (0.26 sec)

**Bill Gates** - [Wikipedia, the free encyclopedia](#) [What's This?](#)  
William Henry "Bill" Gates III (born October 28, 1955) is an American business magnate, philanthropist, author, and chairman of Microsoft, the software ...  
[en.wikipedia.org/wiki/Bill\\_Gates](http://en.wikipedia.org/wiki/Bill_Gates) - [Cached](#) - [Similar](#) -

**Bill Gates: Chairman**   
The official biography from Microsoft of its Chairman and Chief Software Architect. Also his speeches and publications.  
[www.microsoft.com/BillGates/](http://www.microsoft.com/BillGates/) - [Cached](#) - [Similar](#) -

**Bill Gates: Chairman, Microsoft Corp.**   
30 Jul 2007 ... Middle row: Bob O'Rear, Bob Greenberg, Marc McDonald, Gordon Letwin. Bottom row: Bill Gates, Andrea Lewis, Marla Wood, Paul Allen. ...  
[www.microsoft.com/billgates/bio.aspx](http://www.microsoft.com/billgates/bio.aspx) - [Cached](#) - [Similar](#) -   
[More results from www.microsoft.com >](#)

Image results for **"bill gates"** - [Report images](#)



Books by **Bill Gates**  
[Business the Speed of Thought: Succeeding in...](#) - 2000 - 508 pages  
[Business at the Speed of Thought: Succeeding...](#) - 2001 - 496 pages  
[Bill Gates Speaks: Insight from the World's ...](#) - 2001 - 276 pages  
[books.google.com](http://books.google.com)

# A Simple Phrase Recognition Algorithm

---

for each document  $d$

for each  $k: 1$  to  $m$

for each step  $p$  of  $k$  consecutive terms

$p = \langle t_1, t_2, \dots, t_k \rangle$

$\text{count}_p ++$

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

# Outline

---

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



# References

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

## Algorithms

Amazon.com: information retrieval: Books

http://www.amazon.com/s?ref=hb\_ss\_gw?url=search-alias%3Dstripbooks&field-keywords=information+ret

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

Coming Soon (21)

**Department**  
Any Department

**Books**

- Computers & Internet (18,110)
- Nonfiction (60,766)
- Reference (18,979)
- Professional & Technical (64,201)
- Science (48,607)
- Business & Investing (22,362)
- Medicine (19,188)
- Literature & Fiction (31,763)
- Entertainment (9,168)
- Biographies & Memoirs (9,503)
- Outdoors & Nature (7,904)
- History (17,816)
- Law (4,944)
- Gay & Lesbian (2,192)
- Health, Mind & Body (24,277)
- Religion & Spirituality (18,085)
- Travel (3,713)
- Teens (4,915)
- Arts & Photography (8,453)
- Children's Books (13,224)
- Home & Garden (6,913)
- Parenting & Families (6,118)
- Sports (5,167)
- Cooking, Food & Wine (2,526)
- Mystery & Thrillers (7,423)
- Romance (4,271)
- Comics & Graphic Novels (438)
- Science Fiction & Fantasy (3,582)

**Format**  
Any Format

- Printed Books (181,538)
- HTML (1,096)
- Kindle Books (116)
- PDF (60)
- Audiobooks (49)
- Calendars (19)

**Binding**  
Any Binding

- Paperback (117,553)

- Introduction to Information Retrieval** by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (**Hardcover** - Jul 7, 2008)  
**Buy new:** ~~\$60.00~~ **\$43.20** **58 Used & new** from **\$29.16**  
Get it by **Monday, Jun 15** if you order in the next **2 hours** and choose one-day shipping.  
Eligible for **FREE** Super Saver Shipping.  
★★★★☆ (9)
- Search Engines: Information Retrieval in Practice** by Bruce Croft, Donald Metzler, and Trevor Strohman (**Hardcover** - Feb 16, 2009)  
**Buy new:** ~~\$90.20~~ **\$68.49** **25 Used & new** from **\$64.49**  
Get it by **Monday, Jun 15** if you order in the next **1 hour** and choose one-day shipping.  
Eligible for **FREE** Super Saver Shipping.  
★★★★☆ (1)
- Information Retrieval: Algorithms and Heuristics (The Information Retrieval Series) (2nd Edition)** by David A. Grossman and Ophir Frieder (**Paperback** - Dec 20, 2004)  
**Buy new:** ~~\$69.95~~ **\$48.55** **42 Used & new** from **\$23.25**  
Get it by **Monday, Jun 15** if you order in the next **2 hours** and choose one-day shipping.  
Eligible for **FREE** Super Saver Shipping.  
★★★★☆ (8)  
Other Editions: Hardcover
- Modern Information Retrieval** by Ricardo Baeza-Yates and Berthier Ribeiro-Neto (**Paperback** - May 15, 1999)  
**33 Used & new** from **\$9.10**  
★★★★☆ (10)  
Other Editions: Paperback  
**Excerpt** - page 1: "... Chapter 1 Introduction 1.1 Motivation **Information retrieval** (IR) deals with the representation, storage, organization of, and access ..."  
**Surprise me!** See a random page in this book.
- Introduction to Modern Information Retrieval** by G. G. Chowdhury (**Paperback** - Dec 1, 2003)  
**Buy new:** ~~\$89.95~~ **\$15.95** **15 Used & new** from **\$80.00**  
Get it by **Monday, Jun 15** if you order in the next **1 hour** and choose one-day shipping.  
Eligible for **FREE** Super Saver Shipping.  
★★★★☆ (1)  
Other Editions: Hardcover, Paperback
- Information Retrieval: Data Structures and Algorithms** by William B. Frakes and Ricardo Baeza-Yates (**Paperback** - Jun 22, 1992) - **Facsimile**  
**Buy new:** ~~\$73.33~~ **\$52.79** **39 Used & new** from **\$21.97**  
Get it by **Monday, Jun 15** if you order in the next **1 hour** and choose one-day shipping.  
Eligible for **FREE** Super Saver Shipping.  
★★★★☆ (4)
- The Modern Algebra of Information Retrieval (The Information Retrieval Series)** by Sándor Dominich (**Hardcover** - April 18, 2008)