Web Search and Information Retrieval

Cornelia Caragea

Department of Computer Science and Engineering
University of North Texas

Credits for slides: Mooney.

June 14, 2016

Large Digital Information Repositories

- World Wide Web ($> 10^{12}$ links)
- Digital Libraries
- Company intranets and digital assets
- Scientific literature libraries (e.g., CiteSeer, ArnetMiner, Microsoft Academic Search, Google Scholar)
- Medical information portals (e.g., Medline)
- Patent databases (e.g., US Patent Office)
- Online encyclopedias (e.g., Wikipedia)

Various Needs for Information

- Search for documents that fall in a given topic
- Search for specific information
- Search an answer to a question
- Search for information in a different language
- . . .
- Search for images
- Search for music
- Search for a (candidate) friend

Definition of Information Retrieval

- Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

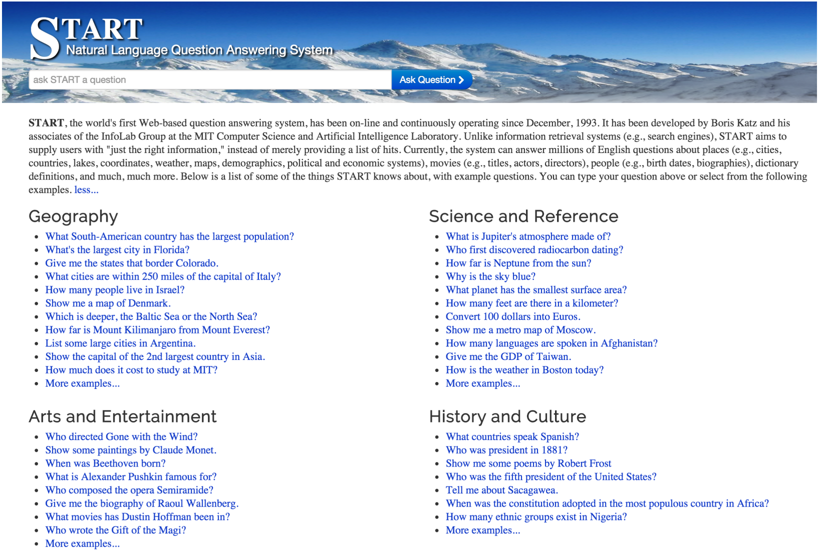
Search

- We live in a search society - belief that (almost) everything is known, we just have to find the information.
- We search for everything - the right book, movie, car, house, vacation trip, bargain, search engine, etc.

Examples of Information Retrieval Systems

- Conventional (library catalog)
  - Search by keyword, title, author, etc.
- Text-based (Google, Bing, DuckDuckGo)
  - Search by keywords. Limited search using queries in natural language.
- Question answering systems (START, Ask)
  - Search in (restricted) natural language
- Other:
  - Cross language information retrieval, music retrieval

START



START, the world's first Web-based question answering system, has been on-line and continuously operating since December, 1993. It has been developed by Boris Katz and his associates of the InfoLab Group at the MIT Computer Science and Artificial Intelligence Laboratory. Unlike information retrieval systems (e.g., search engines), START aims to supply users with "just the right information," instead of merely providing a list of hits. Currently, the system can answer millions of English questions about places (e.g., cities, countries, lakes, coordinates, weather, maps, demographics, political and economic systems), movies (e.g., titles, actors, directors), people (e.g., birth dates, biographies), dictionary definitions, and much, much more. Below is a list of some of the things START knows about, with example questions. You can type your question above or select from the following examples. less...

Geography

- What South-American country has the largest population?
- What's the largest city in Florida?
- Give me the states that border Colorado.
- What cities are within 250 miles of the capital of Italy?
- How many people live in Israel?
- Show me a map of Denmark.
- Which is deeper, the Baltic Sea or the North Sea?
- How far is Mount Kilimanjaro from Mount Everest?
- List some large cities in Argentina.
- Show the capital of the 2nd largest country in Asia.
- How much does it cost to study at MIT?
- More examples...

Arts and Entertainment

- Who directed Gone with the Wind?
- Show some paintings by Claude Monet.
- When was Beethoven born?
- What is Alexander Pushkin famous for?
- Who composed the opera Semiramide?
- Give me the biography of Raoul Wallenberg.
- What movies has Dustin Hoffman been in?
- Who wrote the Gift of the Magi?
- More examples...

Science and Reference

- What is Jupiter's atmosphere made of?
- Who first discovered radiocarbon dating?
- How far is Neptune from the sun?
- Why is the sky blue?
- What planet has the smallest surface area?
- How many feet are there in a kilometer?
- Convert 100 dollars into Euros.
- Show me a metro map of Moscow.
- How many languages are spoken in Afghanistan?
- Give me the GDP of Taiwan.
- How is the weather in Boston today?
- More examples...

History and Culture

- What countries speak Spanish?
- Who was president in 1881?
- Show me some poems by Robert Frost
- Who was the fifth president of the United States?
- Tell me about Sagagawea.
- When was the constitution adopted in the most populous country in Africa?
- How many ethnic groups exist in Nigeria?
- More examples...

IR systems links

- Search for Web pages
  http://www.google.com
- Search for images
  http://www.picsearch.com
- Search for image content
  http://wang.ist.psu.edu/IMAGE/
- Search for answers to questions
  http://www.ask.com
  http://start.csail.mit.edu/
- Music retrieval
  http://www.rotorbrain.com/foote/musicr/

Information Retrieval

- The processing, indexing and retrieval of textual documents.
- Searching for pages on the World Wide Web is perhaps the most widely used IR application
- Concerned firstly with retrieving relevant documents to a query.
- Concerned secondly with retrieving from large sets of documents efficiently.

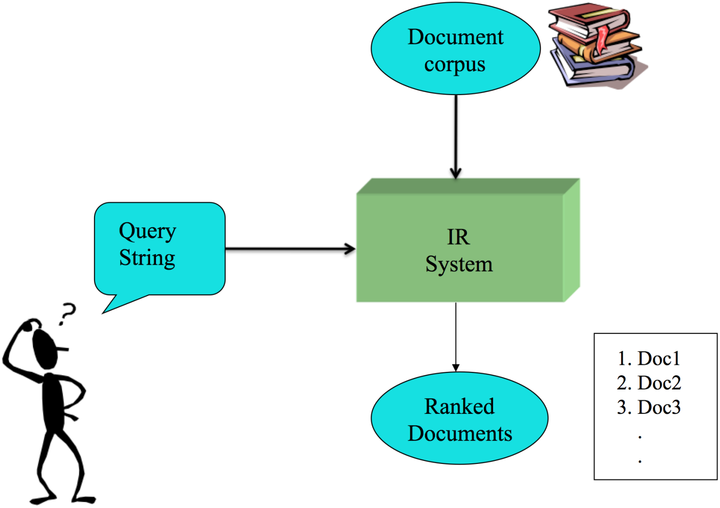
Typical IR Task

- Given:
  - A corpus of textual natural-language documents
  - A user query in the form of a textual string
- Find:
  - A ranked set of documents that are relevant to the query

Key Terms Used in IR

- Query: a representation of what the user is looking for - can be a list of words or a phrase.
- Document: an information entity that the user wants to retrieve
- Collection or corpus: a set of documents
- Index: a representation of information that makes querying easier
- Term: word or concept that appears in a document or a query

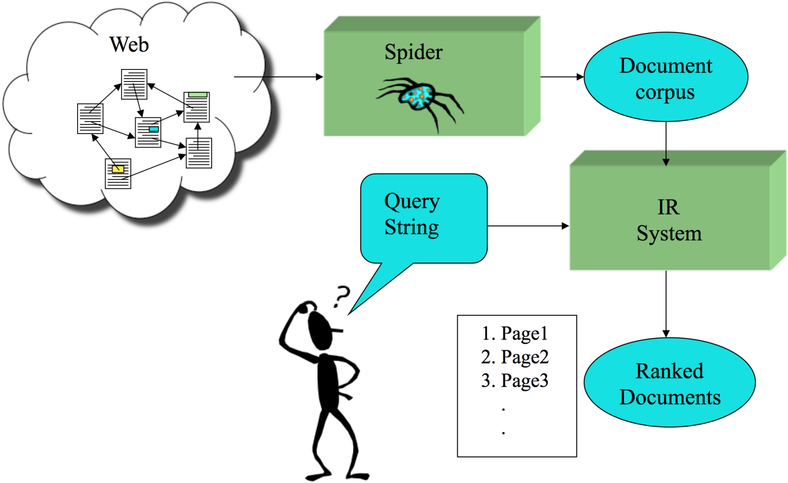
Typical IR System Architecture



Web Search

- Application of IR to HTML documents on the World Wide Web.
- Differences:
  - Must assemble a document corpus by spidering the Web.
  - Documents change uncontrollably.
  - Can exploit the structural layout information in HTML (or XML).
  - Can exploit the link structure of the Web.

Web Search System



Relevance

- Relevance is a subjective judgment and may include:
  - Being on the proper subject.
  - Being timely (recent information).
  - Being authoritative (from a trusted source).
  - Satisfying the goals of the user and his/her intended use of the information (information need)
- Main relevance criterion: an IR system should fulfill a user's information need

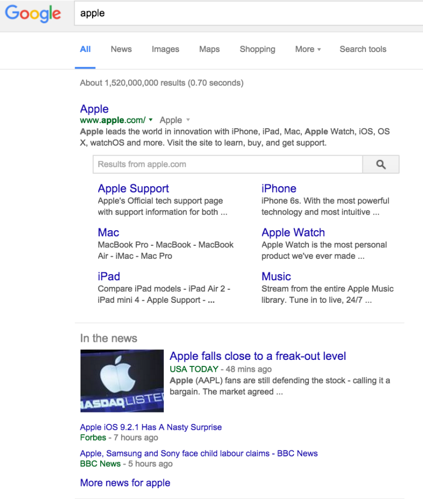
Basic IR Approach: Keyword Search

- Simplest notion of relevance is that the query string appears verbatim in the document.
- Slightly less strict notion is that the words in the query appear frequently in the document, in any order - bag of words representation
    - Example: “unlabeled data homepage classification”
- How does this approach work?
    - Find words/concepts in documents
    - Compare them to words in a query
    - Very effective!

Problems with Keywords

- May not retrieve relevant documents that include synonymous terms.
  - “restaurant” vs. “café”
  - “PRC” vs. “China” (PRC = People’s Republic of China)
- May retrieve irrelevant documents that include ambiguous terms.
  - “bat” (baseball vs. mammal)
  - “bit” (unit of data vs. act of eating)
  - “Apple” (company vs. fruit)

"Apple" (company vs. fruit)



apple

All News Images Maps Shopping More ▾ Search tools

About 1,520,000,000 results (0.70 seconds)

Apple
www.**apple**.com/ ▾ Apple ▾
Apple leads the world in innovation with iPhone, iPad, Mac, **Apple** Watch, iOS, OS X, watchOS and more. Visit the site to learn, buy, and get support.

Results from apple.com

Apple Support
Apple's Official tech support page with support information for both ...

iPhone
iPhone 6s. With the most powerful technology and most intuitive ...

Mac
MacBook Pro - MacBook - MacBook Air - iMac - Mac Pro

Apple Watch
Apple Watch is the most personal product we've ever made ...

iPad
Compare iPad models - iPad Air 2 - iPad mini 4 - Apple Support - ...

Music
Stream from the entire Apple Music library. Tune in to live, 24/7 ...

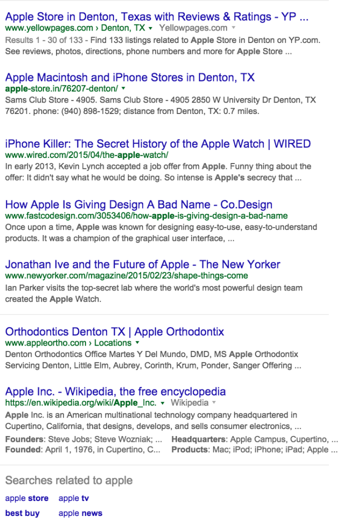
In the news

Apple falls close to a freak-out level
USA TODAY - 48 mins ago
Apple (AAPL) fans are still defending the stock - calling it a bargain. The market agreed ...

Apple iOS 9.2.1 Has A Nasty Surprise
Forbes - 7 hours ago

Apple, Samsung and Sony face child labour claims - BBC News
BBC News - 5 hours ago

More news for apple



Apple Store in Denton, Texas with Reviews & Ratings - YP ...
www.yellowpages.com › Denton, TX ▾ Yellowpages.com ▾
Results 1 - 30 of 133 - Find 133 listings related to **Apple** Store in Denton on YP.com. See reviews, photos, directions, phone numbers and more for **Apple** Store ...

Apple Macintosh and iPhone Stores in Denton, TX
apple-store.in/76207-denton/ ▾
Sams Club Store - 4905. Sams Club Store - 4905 2850 W University Dr Denton, TX 76201. phone: (940) 898-1529; distance from Denton, TX: 0.7 miles.

iPhone Killer: The Secret History of the Apple Watch | WIRED
www.wired.com/2015/04/the-**apple**-watch/
In early 2013, Kevin Lynch accepted a job offer from **Apple**. Funny thing about the offer: It didn't say what he would be doing. So intense is **Apple's** secrecy that ...

How Apple Is Giving Design A Bad Name - Co.Design
www.fastcodesign.com/3053406/how-**apple**-is-giving-design-a-bad-name
Once upon a time, **Apple** was known for designing easy-to-use, easy-to-understand products. It was a champion of the graphical user interface, ...

Jonathan Ive and the Future of Apple - The New Yorker
www.newyorker.com/magazine/2015/02/23/shape-things-come
Ian Parker visits the top-secret lab where the world's most powerful design team created the **Apple** Watch.

Orthodontics Denton TX | Apple Orthodontix
www.appleortho.com › Locations ▾
Denton Orthodontics Office Martes Y Del Mundo, DMD, MS **Apple** Orthodontix Servicing Denton, Little Elm, Aubrey, Corinth, Krum, Ponder, Sanger Offering ...

Apple Inc. - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Apple**_Inc. ▾ Wikipedia ▾
Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, ...
Founders: Steve Jobs; Steve Wozniak; ... **Headquarters**: Apple Campus, Cupertino, ...
Founded: April 1, 1976, in Cupertino, C... **Products**: Mac; iPod; iPhone; iPad; Apple ...

Searches related to apple

apple **store** apple **tv**
best buy apple **news**

Techniques for Intelligent IR

- Take into account the meaning of the words used
- Take into account the order of words in the query
- Adapt to the user based on implicit or explicit feedback
- Extend search with related terms
- Perform automatic spell checking / diacritics restoration
- Take into account the authority of the source.
- Use the link structure of the data.

IR System Components

- Text Operations form index words (tokens)
  - Tokenization
  - Stop-word removal
  - Stemming
- Indexing constructs an inverted index of word to document pointers.
  - Mapping from keywords to document ids

Indexing - Inverted Index

Doc 1
I did enact Julius Caesar: I was killed i’ the Capitol; Brutus killed me.

Doc 2
So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious:

term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i’	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

⟹

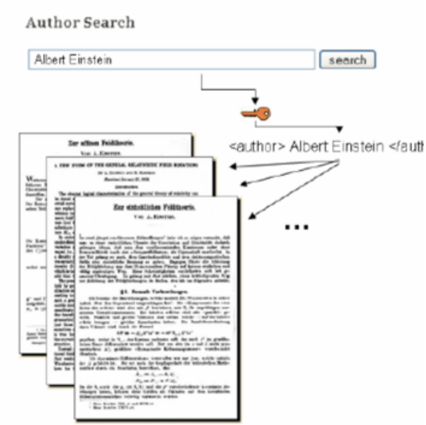
term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i’	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

⟹

term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2
I	1	→	1
i’	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 → 2
told	1	→	2
you	1	→	2
was	2	→	1 → 2
with	1	→	2

Document Indexing

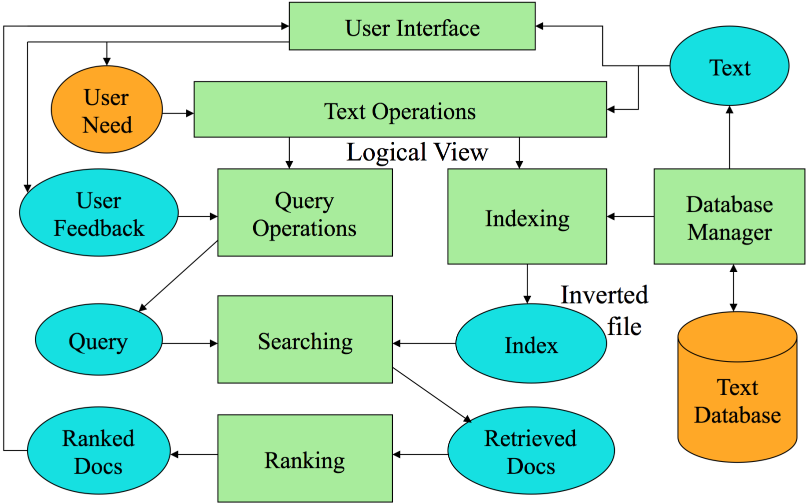
- Index: associates a document with one or more keys (keywords)
- Present key → identify documents that match key
- Efficiency is crucial, fast access



IR System Components

- Searching retrieves documents that contain a given query token from the inverted index.
- Ranking scores all retrieved documents according to a relevance metric.
- User Interface manages interaction with the user:
  - Query input and document output
  - Relevance feedback
  - Visualization of results
- Query Operations transform the query to improve retrieval:
  - Query expansion using a thesaurus
  - Query transformation using relevance feedback

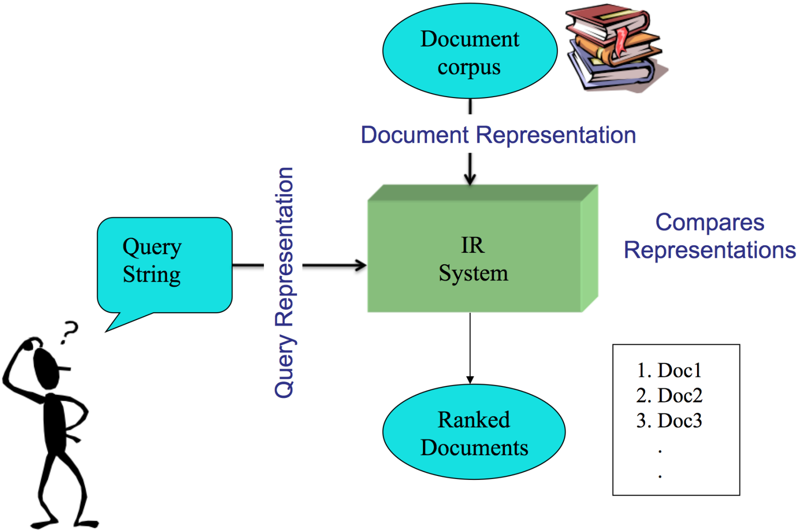
IR System Architecture



Retrieval Models

- A retrieval model specifies the details of:
  - Document representation
  - Query representation
  - How do we compare representations - retrieval function?
- Determines a notion of relevance.
- Notion of relevance can be binary or continuous (i.e. ranked retrieval).

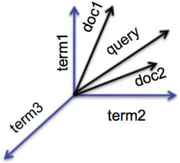
Typical IR System Architecture



A Class of Retrieval Models: Vector Space Models

Vector Space Models are among the most widely used models.

- Key idea: Everything (documents, queries, terms) is a vector in a high-dimensional space.



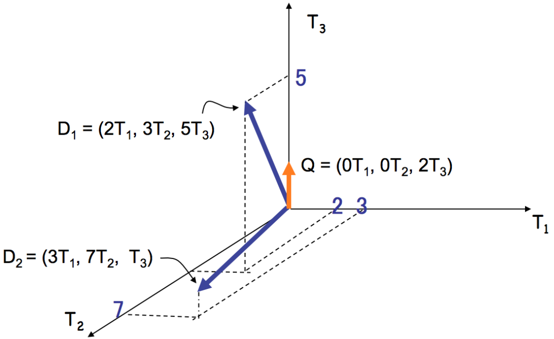
- The geometry of space induces a similarity measure between documents
- The documents are ranked based on their similarity with the query

Issues for Vector Space Models

- How to determine important words in a document?
  - How to select basis vectors (dimensions)
- How to convert objects into vectors?
  - Documents, queries, terms
- Assumption - not all terms are equally useful for representing the document contents, less frequent terms allow identifying a narrower set of documents
- How to compare objects in the vector space?
  - How to determine the degree of similarity between a document and the query?
- In the case of the web, what is a collection and what are the effects of links, formatting information, etc.?

Example Graphical Representation

- $D_1 = (2T_1, 3T_2, 5T_3)$
- $D_2 = (3T_1, 7T_2, 1T_3)$
- $Q = (0T_1, 0T_2, 2T_3)$



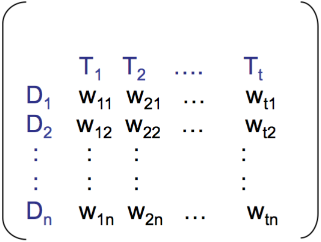
- Is D_1 or D_2 more similar to Q?
- How to measure the degree of similarity? Distance? Angle?

The Vector-Space Model

- Assume t distinct terms remain after preprocessing; call them index terms or the vocabulary.
- These "orthogonal" terms form a basis of a vector space. Dimension $= t = |\text{vocabulary}|$
- Each term, i, in a document or query, j, is given a real-valued weight, w_{ij}.
- Both documents and queries are expressed as t-dimensional vectors:
  $d_j = (w_{1j}, w_{2j}, \cdots, w_{tj})$

Document Collection

- A collection of *n* documents can be represented in the vector space model by a term-document matrix.
- An entry in the matrix corresponds to the "weight" of a term in the document; zero means the term has no significance in the document or it simply does not exist in the document.



$$\left(\begin{array}{ccccc} & T_1 & T_2 & \cdots & T_t \\ D_1 & w_{11} & w_{21} & \cdots & w_{t1} \\ D_2 & w_{12} & w_{22} & \cdots & w_{t2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & w_{1n} & w_{2n} & \cdots & w_{tn} \end{array}\right)$$

Term Weights: Term Frequency

- More frequent terms in a document are more important, i.e. more indicative of the topic.

  f_{ij} = frequency of term i in document j

- May want to normalize *term frequency (tf)*
  - e.g. by dividing by the frequency of the most common term in the document:

$$tf_{ij} = \frac{f_{ij}}{max_i\{f_{ij}\}}$$

Term Weights: Inverse Document Frequency

- Terms that appear in many *different* documents are less indicative of the overall topic.
  - df_i = document frequency of term i = number of documents containing term i
  - idf_i = inverse document frequency of term $i = \log_2(N/df_i)$ (N: total number of documents)
- An indication of a term's *discrimination* power.
- Log used to dampen the effect relative to *tf*.

TF-IDF Weighting

- A typical combined term importance indicator is *tf-idf* weighting:

$$w_{ij} = tf_{ij}\, idf_i = tf_{ij} \log_2(N/df_i)$$

- A term occurring frequently in the document but rarely in the rest of the collection is given high weight.

Computing *tf-idf* - An Example

- Given a document containing terms with given frequencies:

$$A(3), B(2), C(1)$$

- Assume collection contains 10,000 documents and document frequencies of these terms are:

$$A(50), B(1300), C(250)$$

- Compute *tf*, *idf*, *tf-idf*?

$$w_{ij} = tf_{ij} idf_i = (f_{ij}/max_i\{f_{ij}\}) \cdot \log_2(N/df_i)$$

Computing *tf-idf* - An Example

- Given a document containing terms with given frequencies:

$$A(3), B(2), C(1)$$

- Assume collection contains 10,000 documents and document frequencies of these terms are:

$$A(50), B(1300), C(250)$$

- Then:

$$A : tf = 3/3; idf = \log_2(10000/50) = 7.6; tf\text{-}idf = 7.6$$

$$B : tf = 2/3; idf = \log_2(10000/1300) = 2.9; tf\text{-}idf = 2.0$$

$$C : tf = 1/3; idf = \log_2(10000/250) = 5.3; tf\text{-}idf = 1.8$$

Query Vector

- Query vector is typically treated as a document and is also *tf-idf* weighted.
- The alternative is for the user to supply weights for the given query terms.
  - Weighted query terms:
    $Q = <$ database 0.5; text 0.8; information 0.2 $>$
  - Unweighted query terms:
    $Q = <$ database; text; information $>$

Similarity Measures

- A similarity measure is a function that computes the degree of similarity between two vectors.
- Using a similarity measure between the query and each document:
  - It is possible to rank the retrieved documents in the order of presumed relevance.
- Common similarity measures:
  - Inner Product
  - Cosine Similarity

Inner Product

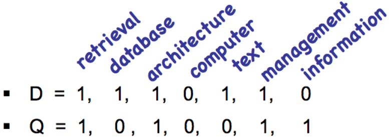
- Similarity between vectors for the document d_j and query q can be computed as the vector inner product (or the dot product):

  $$sim(d_j, q) = d_j \cdot q = \sum_{i=1}^{t} w_{ij} w_{iq}$$

  where w_{ij} is the weight of term i in document j and w_{iq} is the weight of term i in the query
- For binary vectors, the inner product is the number of matched query terms in the document (size of intersection).
- For weighted term vectors, it is the sum of the products of the weights of the matched terms.

Inner Product - Examples

- Binary:



	retrieval	database	architecture	computer	text	management	information
▪ D =	1,	1,	1,	0,	1,	1,	0
▪ Q =	1,	0 ,	1,	0,	0,	1,	1

  Size of vector = size of vocabulary = 7; 0 means corresponding term not found in document or query
  $sim(D, Q) = ?$

- Weighted:

$$D_1 = (2T_1, 3T_2, 5T_3), D_2 = (3T_1, 7T_2, 1T_3),$$

$$Q = (0T_1, 0T_2, 2T_3)$$

$sim(D_1, Q) = ?$
$sim(D_2, Q) = ?$

Inner Product - Examples

- Binary:
  Size of vector = size of vocabulary = 7; 0 means corresponding term not found in document or query
  $sim(D, Q) = 3$

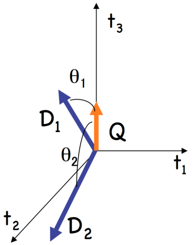
	retrieval	database	architecture	computer	text	management	information
▪ D =	1,	1,	1,	0,	1,	1,	0
▪ Q =	1,	0 ,	1,	0,	0,	1,	1

- Weighted:

$$D_1 = (2T_1, 3T_2, 5T_3), D_2 = (3T_1, 7T_2, 1T_3),$$

$$Q = (0T_1, 0T_2, 2T_3)$$

$sim(D_1, Q) = 2 \cdot 0 + 3 \cdot 0 + 5 \cdot 2 = 10$
$sim(D_2, Q) = 3 \cdot 0 + 7 \cdot 0 + 1 \cdot 2 = 2$



Cosine Similarity Measure

- Cosine similarity measures the cosine of the angle between two vectors.
- Inner product normalized by the vector lengths.

$$CosSim(d_j, q) = \frac{\langle d_j, q \rangle}{\|d_j\| \cdot \|q\|} = \frac{\sum_{i=1}^{t} w_{ij} w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$$

$$D_1 = (2T_1, 3T_2, 5T_3), D_2 = (3T_1, 7T_2, 1T_3),$$

$$Q = (0T_1, 0T_2, 2T_3)$$

$CosSim(D_1, Q) = ?$
$CosSim(D_2, Q) = ?$

Cosine Similarity Measure

- Cosine similarity measures the cosine of the angle between two vectors.
- Inner product normalized by the vector lengths.

$$CosSim(d_j, q) = \frac{\langle d_j, q \rangle}{\|d_j\| \cdot \|q\|} = \frac{\sum_{i=1}^{t} w_{ij} w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$$

$$D_1 = (2T_1, 3T_2, 5T_3), D_2 = (3T_1, 7T_2, 1T_3),$$

$$Q = (0T_1, 0T_2, 2T_3)$$

$$CosSim(D_1, Q) = 10/\sqrt{(4+9+25)(0+0+4)} = 0.81$$
$$CosSim(D_2, Q) = 2/\sqrt{(9+49+1)(0+0+4)} = 0.13$$

D_1 is 6 times better than D_2 using cosine similarity but only 5 times better using inner product.

Vector Space Summary

- Very simple
  - Map everything to a vector
  - Compare using angle between vectors
- Challenge is mostly finding good weighting scheme
  - Variants on *tf-idf* are most common
- Considers both local (tf) and global (idf) word occurrence frequencies.
- Tends to work quite well in practice despite obvious weaknesses.

Problems with Vector Space Model

- Missing semantic information (e.g. word sense).
- Missing syntactic information (e.g. phrase structure, word order, proximity information).
- Does not consider the link structure of documents