

Search Engines

Information Retrieval in Practice

Storing the Documents

- Many reasons to store converted document text
 - saves crawling time when page is not updated
 - provides efficient access to text for snippet generation, information extraction, etc.
- Database systems can provide document storage for some applications
 - web search engines use customized document storage systems

Storing the Documents

- Requirements for document storage system:
 - Random access
 - request the content of a document based on its URL
 - hash function based on URL is typical
 - Compression and large files
 - reducing storage requirements and efficient access
 - Update
 - handling large volumes of new and modified documents
 - adding new anchor text

Large Files

- Store many documents in large files, rather than each document in a file
 - avoids overhead in opening and closing files
 - reduces seek time relative to read time
- Compound documents formats
 - used to store multiple documents in a file
 - e.g., TREC Web

TREC Web Format

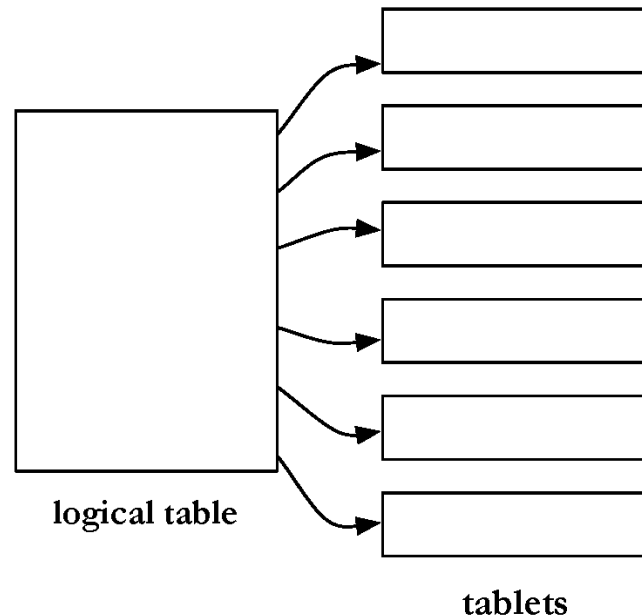
```
<DOC>
<DOCNO>WTX001-B01-10</DOCNO>
<DOCHDR>
http://www.example.com/test.html 204.244.59.33 19970101013145 text/html 440
HTTP/1.0 200 OK
Date: Wed, 01 Jan 1997 01:21:13 GMT
Server: Apache/1.0.3
Content-type: text/html
Content-length: 270
Last-modified: Mon, 25 Nov 1996 05:31:24 GMT
</DOCHDR>
<HTML>
<TITLE>Tropical Fish Store</TITLE>
Coming soon!
</HTML>
</DOC>
<DOC>
<DOCNO>WTX001-B01-109</DOCNO>
<DOCHDR>
http://www.example.com/fish.html 204.244.59.33 19970101013149 text/html 440
HTTP/1.0 200 OK
Date: Wed, 01 Jan 1997 01:21:19 GMT
Server: Apache/1.0.3
Content-type: text/html
Content-length: 270
Last-modified: Mon, 25 Nov 1996 05:31:24 GMT
</DOCHDR>
<HTML>
<TITLE>Fish Information</TITLE>
This page will soon contain interesting
information about tropical fish.
</HTML>
</DOC>
```

Compression

- Text is highly redundant (or predictable)
- Compression techniques exploit this redundancy to make files smaller without losing any of the content
- Compression of indexes covered later
- Popular algorithms can compress HTML and XML text by 80%
 - e.g., DEFLATE (zip, gzip) and LZW (UNIX compress, PDF)
 - may compress large files in blocks to make access faster

BigTable

- Google's document storage system
 - Customized for storing, finding, and updating web pages
 - Handles large collection sizes using inexpensive computers

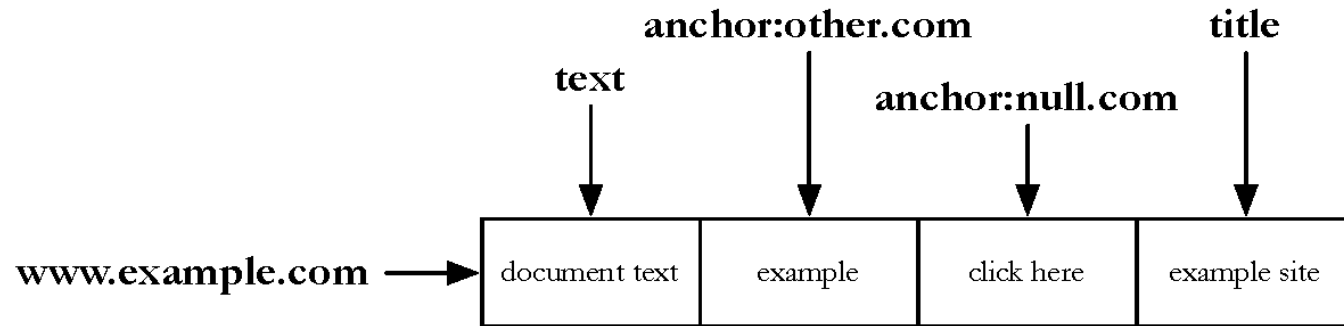


BigTable

- No query language, no complex queries to optimize
- Only row-level transactions
- Tablets are stored in a replicated file system that is accessible by all BigTable servers
- Any changes to a BigTable tablet are recorded to a transaction log, which is also stored in a shared file system
- If any tablet server crashes, another server can immediately read the tablet data and transaction log from the file system and take over

BigTable

- Logically organized into rows
- A row stores data for a single web page



- Combination of a row key, a column key, and a timestamp point to a single *cell* in the row

BigTable

- BigTable can have a huge number of columns per row
 - all rows have the same column groups
 - not all rows have the same columns
 - important for reducing disk reads to access document data
- Rows are partitioned into tablets based on their row keys
 - simplifies determining which server is appropriate

Detecting Duplicates

- Duplicate and near-duplicate documents occur in many situations
 - Copies, versions, plagiarism, spam, mirror sites
 - 30% of the web pages in a large crawl are exact or near duplicates of pages in the other 70%
- Duplicates consume significant resources during crawling, indexing, and search
 - Little value to most users

Duplicate Detection

- *Exact* duplicate detection is relatively easy
- *Checksum* techniques
 - A checksum is a value that is computed based on the content of the document
 - e.g., sum of the bytes in the document file

T	r	o	p	i	c	a	l		f	i	s	h	<i>Sum</i>
54	72	6F	70	69	63	61	6C	20	66	69	73	68	508

- Possible for files with different text to have same checksum
- Functions such as a *cyclic redundancy check* (CRC), have been developed that consider the positions of the bytes

Near-Duplicate Detection

- More challenging task
 - Are web pages with same text context but different advertising or format near-duplicates?
- A near-duplicate document is defined using a threshold value for some similarity measure between pairs of documents
 - e.g., document $D1$ is a near-duplicate of document $D2$ if more than 90% of the words in the documents are the same

Near-Duplicate Detection

- *Search*:
 - find near-duplicates of a document D
 - $O(N)$ comparisons required
- *Discovery*:
 - find all pairs of near-duplicate documents in the collection
 - $O(N^2)$ comparisons
- IR techniques are effective for search scenario
- For discovery, other techniques used to generate compact representations

Fingerprints

1. The document is parsed into words. Non-word content, such as punctuation, HTML tags, and additional whitespace, is removed.
2. The words are grouped into contiguous *n-grams* for some *n*. These are usually overlapping sequences of words, although some techniques use non-overlapping sequences.
3. Some of the n-grams are selected to represent the document.
4. The selected n-grams are hashed to improve retrieval efficiency and further reduce the size of the representation.
5. The hash values are stored, typically in an inverted index.
6. Documents are compared using overlap of fingerprints

Fingerprint Example

Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species.

(a) Original text

tropical fish include, fish include fish, include fish found, fish found in, found in tropical, in tropical environments, tropical environments around, environments around the, around the world, the world including, world including both, including both freshwater, both freshwater and, freshwater and salt, and salt water, salt water species

(b) 3-grams

938 664 463 822 492 798 78 969 143 236 913 908 694 553 870 779

(c) Hash values

664 492 236 908

(d) Selected hash values using $0 \bmod 4$

Simhash

- Similarity comparisons using word-based representations more effective at finding near-duplicates
 - Problem is efficiency
- Simhash combines the advantages of the word-based similarity measures with the efficiency of fingerprints based on hashing
- Similarity of two pages as measured by the cosine correlation measure is proportional to the number of bits that are the same in the simhash fingerprints

Simhash

1. Process the document into a set of features with associated weights. We will assume the simple case where the features are words weighted by their frequency.
2. Generate a hash value with b bits (the desired size of the fingerprint) for each word. The hash value should be unique for each word.
3. In b -dimensional vector V , update the components of the vector by adding the weight for a word to every component for which the corresponding bit in the word's hash value is 1, and subtracting the weight if the value is 0.
4. After all words have been processed, generate a b -bit fingerprint by setting the i th bit to 1 if the i th component of V is positive, or 0 otherwise.

Simhash Example

Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species.

(a) Original text

tropical 2 fish 2 include 1 found 1 environments 1 around 1 world 1
including 1 both 1 freshwater 1 salt 1 water 1 species 1

(b) Words with weights

tropical	01100001	fish	10101011	include	11100110
found	00011110	environments	00101101	around	10001011
world	00101010	including	11000000	both	10101110
freshwater	00111111	salt	10110101	water	00100101
species	11101110				

(c) 8 bit hash values

1 -5 9 -9 3 1 3 3

(d) Vector V formed by summing weights

1 0 1 0 1 1 1 1

(e) 8-bit fingerprint formed from V

Removing Noise

- Many web pages contain text, links, and pictures that are not directly related to the main content of the page
- This additional material is mostly *noise* that could negatively affect the ranking of the page
- Techniques have been developed to detect the content blocks in a web page
 - Non-content material is either ignored or reduced in importance in the indexing process

Noise Example

CNN.com Member Center Sign In Register International Edition

SEARCH THE WEB RSS FEED Search

Home Page World U.S. Weather Business Sports Analysis Politics Law Technology Science & Space Health Entertainment Offbeat Travel Education Special Reports Video Autos I-Reports

IMPACT & WORLD TAKE ACTION SERVICES E-mails RSS Podcasts Mobile CNN Pipeline SEARCH WEB CHANNEL Search

SCIENCE & SPACE

Aquarium plays whale shark matchmaker

Two females flown 8,000 miles for double date in Atlanta

Monday, June 5, 2006, Posted: 5:28 p.m. EDT (21:28 GMT)

ATLANTA, Georgia (CNN) -- Ralph and Norton, meet Alice and Trixie.

The Georgia Aquarium's two male whale sharks got some female companionship on Saturday, when they were joined by two females transported to Atlanta from Taipei, Taiwan.

Researchers are hoping the sharks will mate.

The females -- 11 feet and 14 feet long -- were flown more than 8,000 miles by UPS, which reconfigured a company B-747 freighter with advanced marine life support systems to carry them. (Watch what it took to get the sharks together -- 1:55)

The pilot said they treated the massive fish like first-class passengers.

"As we were doing the descent, we asked to start down a little sooner to make a nice shallow descent, to not make things too uncomfortable back there for the whale sharks," UPS pilot Capt. Bob Crum said.

The plane's center of balance was carefully planned, according to a statement from the aquarium, and veterinarians accompanied the sharks.


The delivery company also brought the two males to Atlanta, where researchers can study the whale sharks' behavior, breeding and development.

The whale sharks -- named after the main characters in the 1950s sitcom "The Honeymooners" -- were delivered to the aquarium in special transportation containers.

The Georgia Aquarium, which opened in November, is the world's largest aquarium. It was a \$250 million gift to Georgia from Bernie Marcus, co-founder of The Home Depot and his wife, Bill, through the Marcus Foundation.

It is the only aquarium outside of Asia to showcase whale sharks, which are the largest fish on Earth.

The aquarium's 6.2-million gallon "Ocean Voyager" tank can hold up to six whale sharks, but it's room for the whale sharks to start a family.



Alice the whale shark swims into the Ocean Voyager tank at the Georgia Aquarium for the first time.

Image: [REDACTED]

YOUR E-MAIL ALERTS

Atlanta (Georgia)

Taiwan

ACTIVATE or Create Your Own

Manage Alerts | What is This?

Subscribe to Time for \$1.00

SPACE

Section Page | Video

Astronauts prepare for third spacewalk

- Astronomers vie to make biggest telescope
- NASA to beam Beatles song to North Star
- U.S. plans for falling satellite

TOP STORIES

Home Page | Video | Most Popular

- Russians choose Putin's successor
- Iran's president makes landmark visit to Iraq
- Israel PM: Attacks on militants go on
- Cable arrested in abandoned baby case

International Edition Languages CNN TV CNN International Headline News Transcripts Advertise with Us About Us

SEARCH THE WEB RSS FEED Search

© 2007 Cable News Network. A Time Warner Company. All Rights Reserved. Terms under which this service is provided to you. Read our privacy guidelines Contact Us Site Map

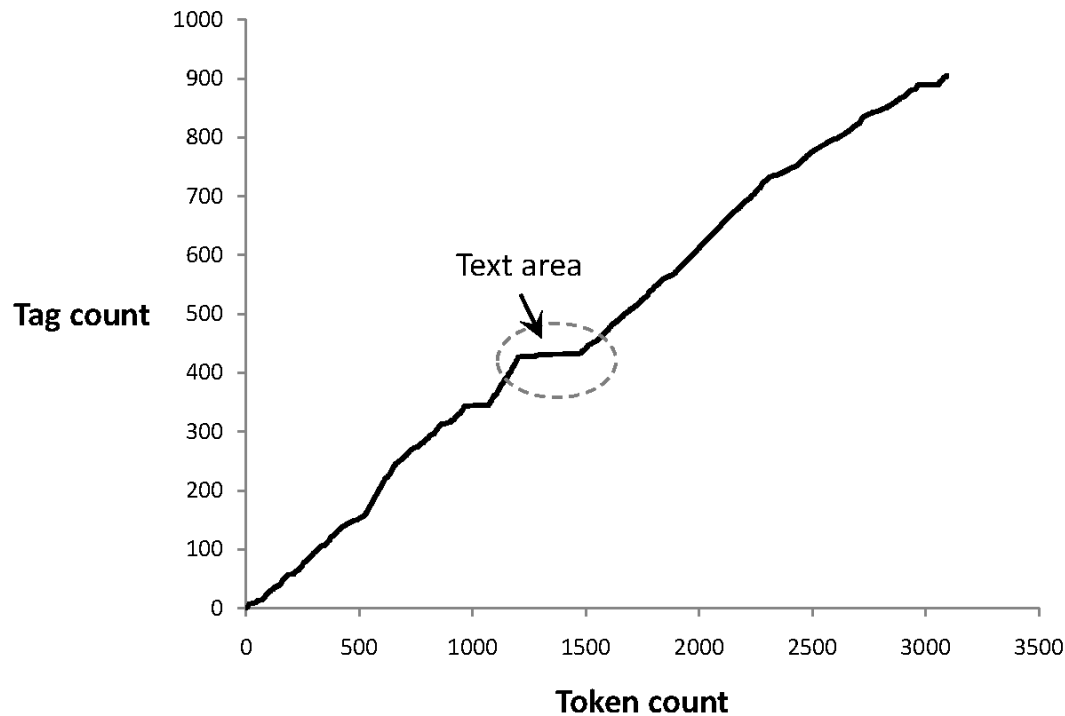
External sites open in new window; not endorsed by CNN.com. See [REDACTED] Pay service with live and archived video. Learn more

Download audio news Add RSS headlines

Content block

Finding Content Blocks

- Cumulative distribution of tags in the example web page



- Main text content of the page corresponds to the “plateau” in the middle of the distribution

Finding Content Blocks

- Represent a web page as a sequence of bits, where $b_n = 1$ indicates that the n th token is a tag
- Optimization problem where we find values of i and j to maximize both the number of tags below i and above j and the number of non-tag tokens between i and j
- i.e., maximize

$$\sum_{n=0}^{i-1} b_n + \sum_{n=i}^j (1 - b_n) + \sum_{n=j+1}^{N-1} b_n$$

