Web Search Engines

Popular Search Engines & types

Google	Search by keywords
Alta Vista	
Bing	
Yahoo	Search by categories
Ask jeeves	Interview simulation

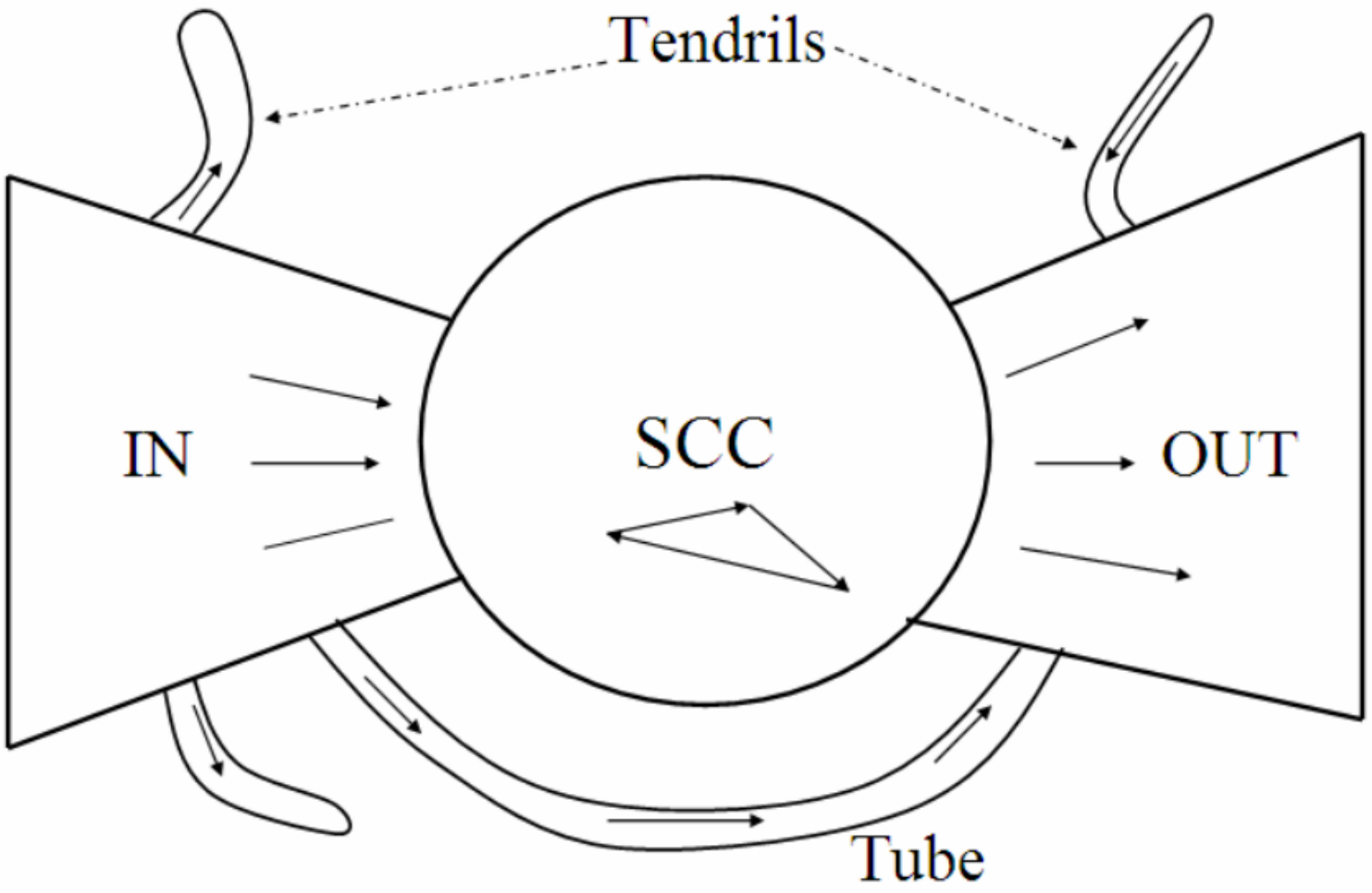
Without search engines, the web wouldn’t work

- Without search, content is hard to find.
- Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don’t get ad revenue from it?
- Interest aggregation
  - Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other
- Somebody needs to pay for the web.
  - Servers, web infrastructure, content creation
  - A large part today is paid by search ads.

Issues with web search engines

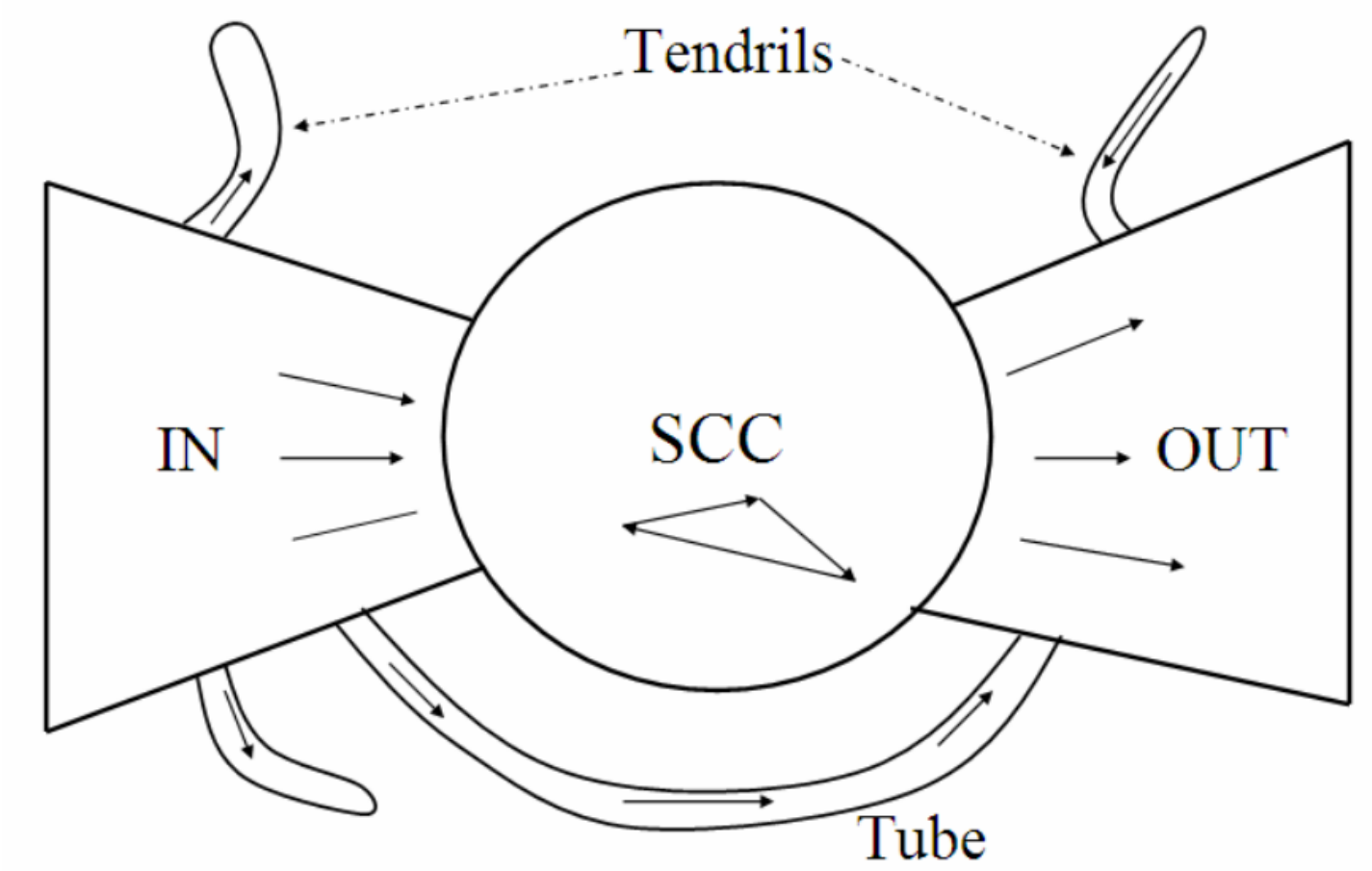
- Dynamic data
- Quality is variable & user has to make judgement
- Factual knowledge is not objective
- Scope of web is not fixed

Structure of the web



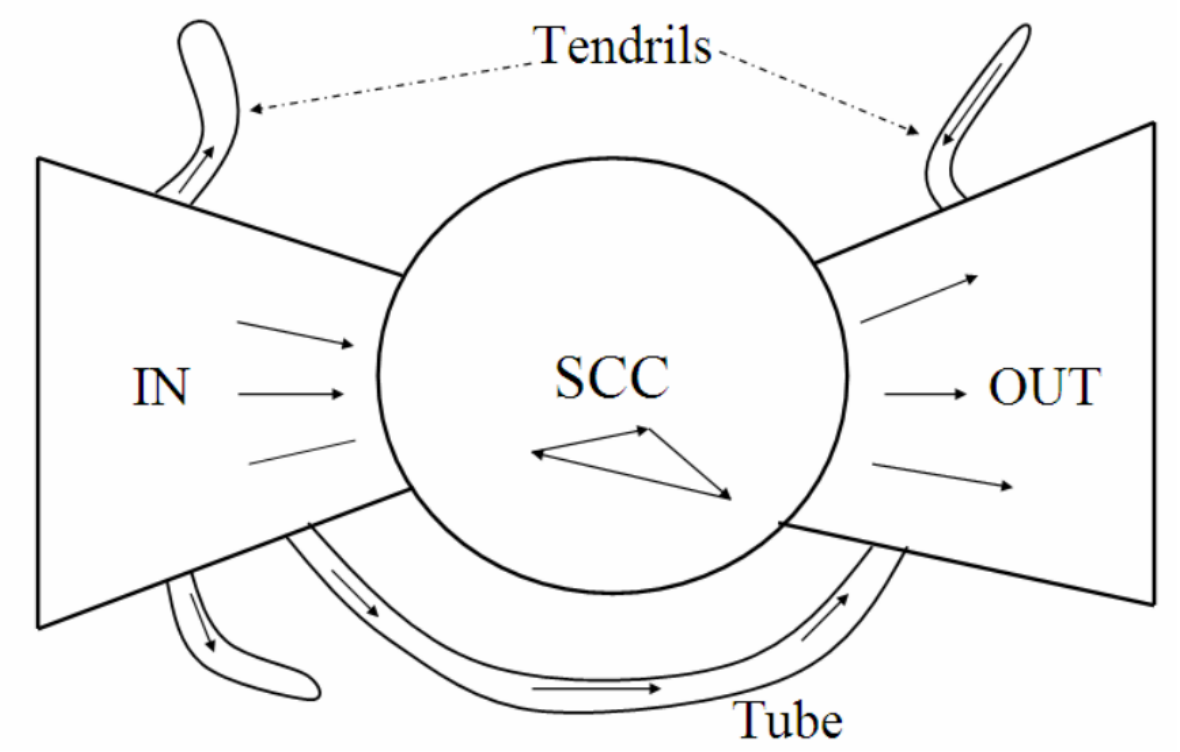
Bow-tie structure of web

Bow-tie structure of web



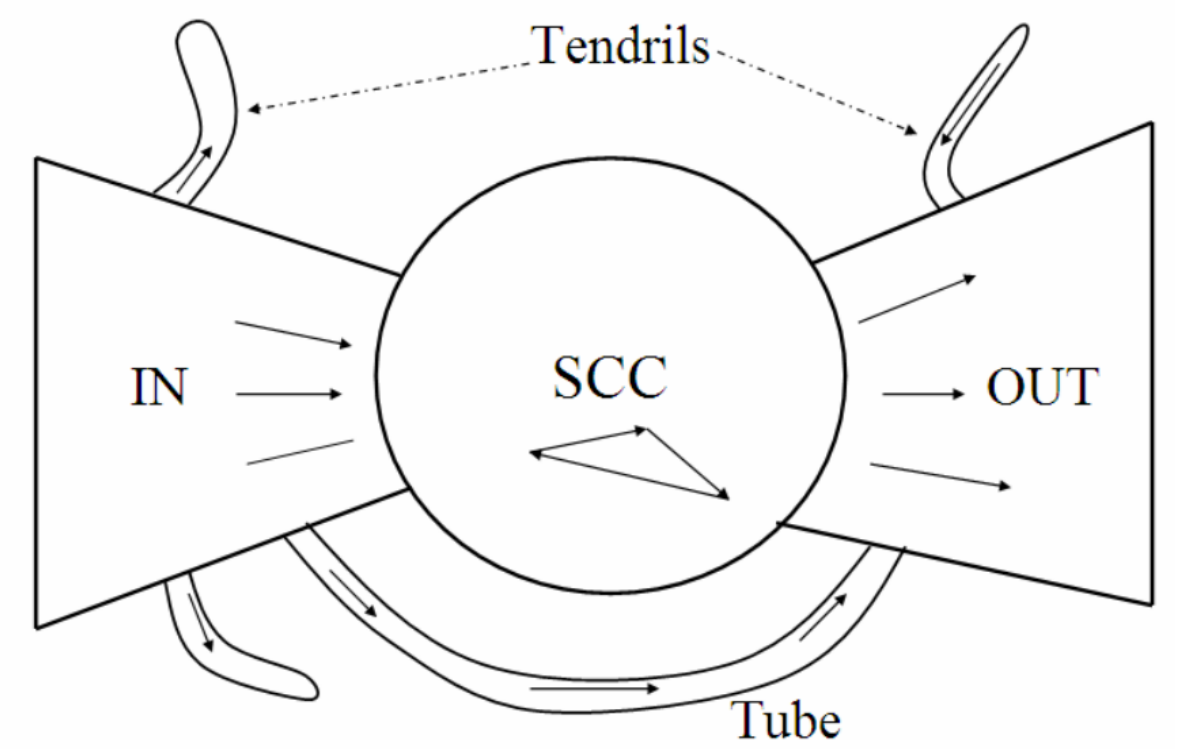
Strongly connected component (SCC) in the center

Bow-tie structure of web



Lots of pages that get linked to, but don't link (OUT)
Lots of pages that link to other pages, but don't get linked to (IN)

Bow-tie structure of web

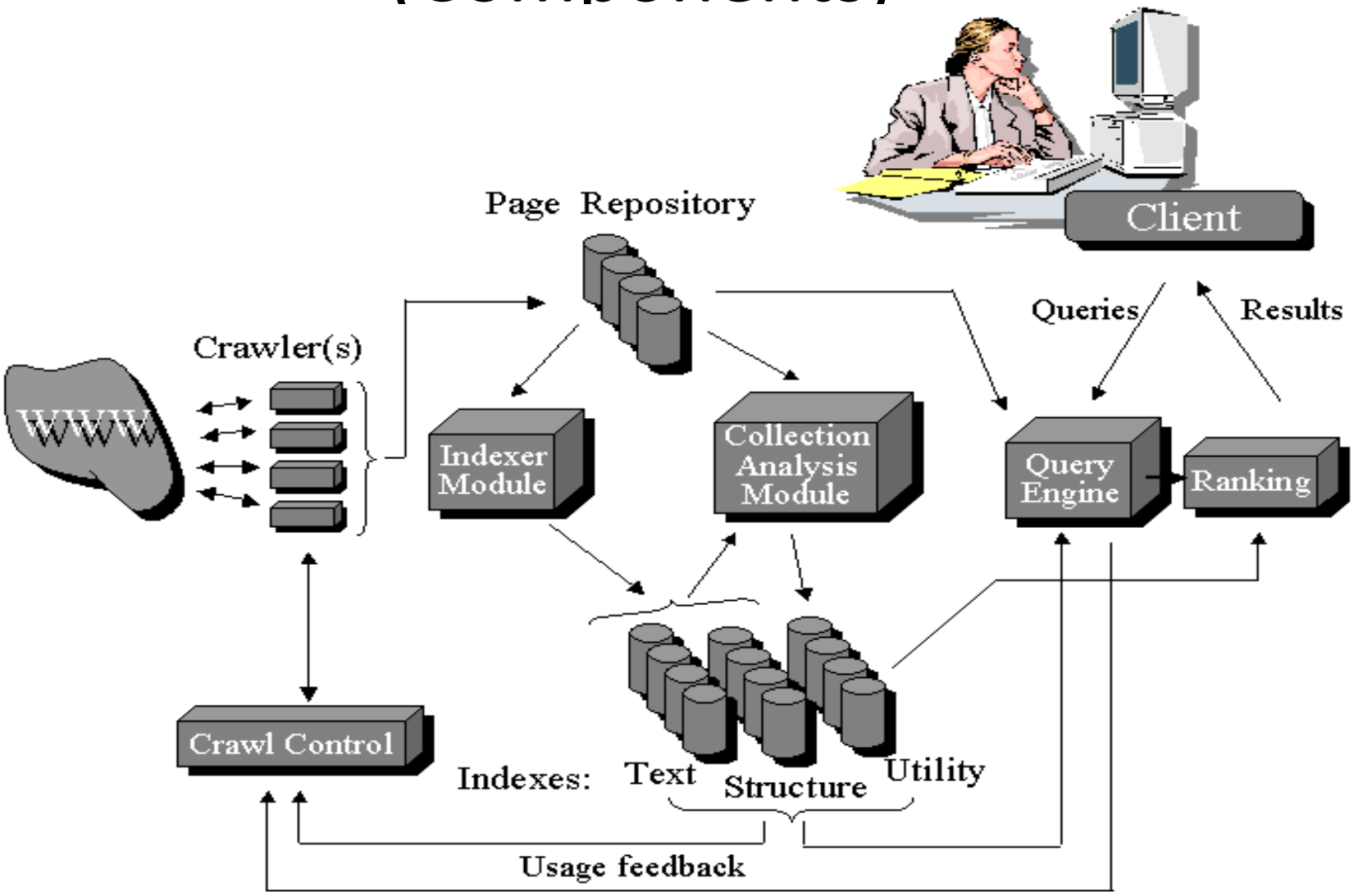


Tendrils: that either lead nowhere from IN, or from nowhere to OUT.
Tubes: small sets of pages outside SCC that lead directly from IN to OUT

Users

- Web queries are short

Web search architecture (Components)



Search Index (indexer)

- Inverted index

  chess → [www.chess.co.uk, www.chessclub.com, www.uschess.org]

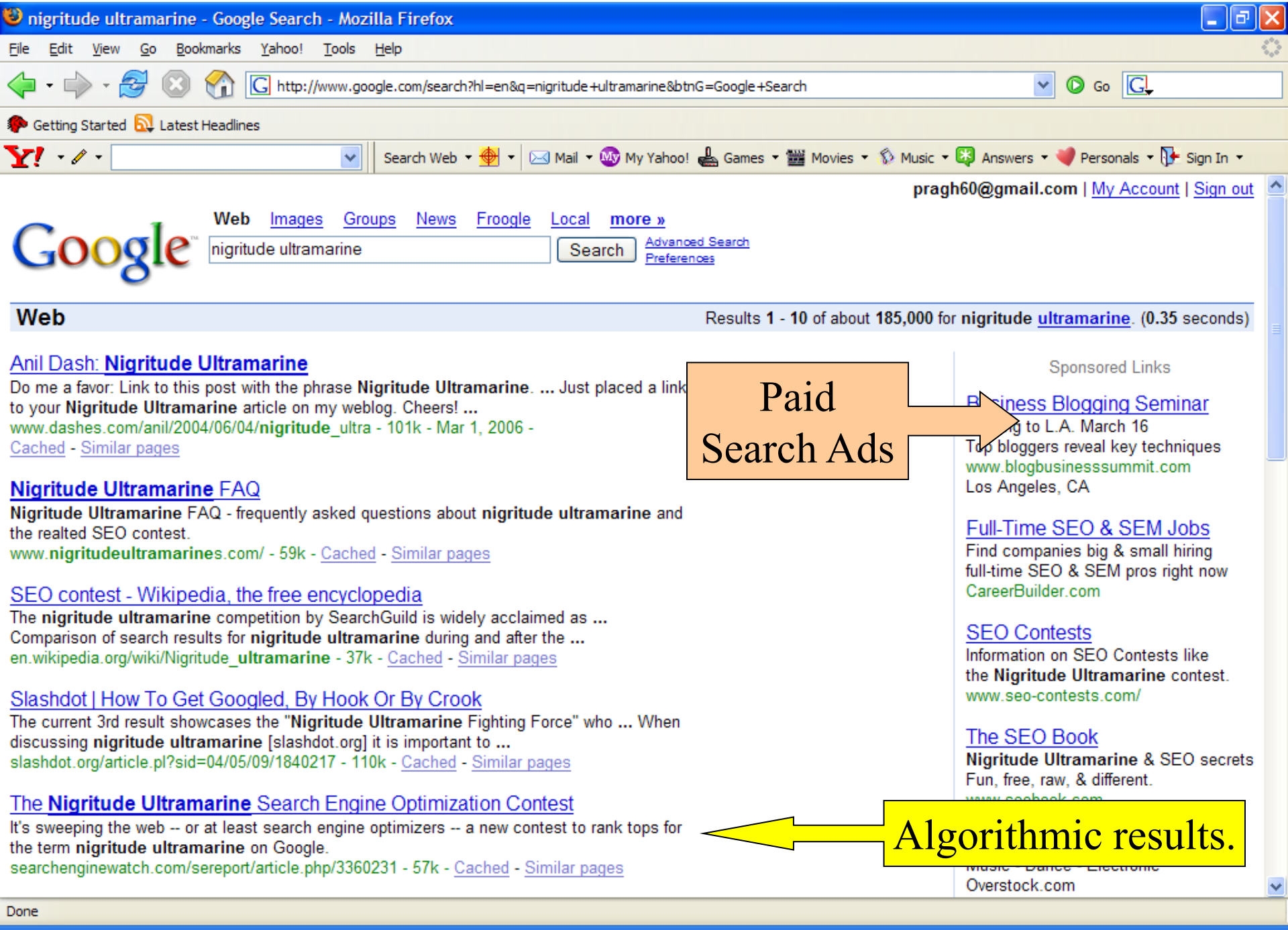
- Information of hyperlinks in link database
  - Organized like inverted index
  - Source URL contains all destination URLs

Query Engine

- Algorithmic heart
- Interface between search index, the user and the web
- Two steps:
  - Retrieves the results as per matching keywords
  - Ranking the web pages

Search Interface

- Provides look and feel of search engine
- Allows user to submit queries
- Browse result list
- Click on chosen web page
- User should be able to differentiate between sponsored links and organic links


nigritude ultramarine - Google Search - Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
http://www.google.com/search?hl=en&q=nigritude+ultramarine&btnG=Google+Search
Go
Getting Started
Latest Headlines
Search Web
Mail
My Yahoo!
Games
Movies
Music
Answers
Personals
Sign In
pragh60@gmail.com | My Account | Sign out
Google
Web Images Groups News Froogle Local more »
nigritude ultramarine
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 185,000 for nigritude ultramarine. (0.35 seconds)
Anil Dash: Nigritude Ultramarine
Do me a favor: Link to this post with the phrase Nigritude Ultramarine. ... Just placed a link to your Nigritude Ultramarine article on my weblog. Cheers! ...
www.dashes.com/anil/2004/06/04/nigritude_ultra - 101k - Mar 1, 2006 -
Cached - Similar pages
Nigritude Ultramarine FAQ
Nigritude Ultramarine FAQ - frequently asked questions about nigritude ultramarine and the realted SEO contest.
www.nigritudeultramarines.com/ - 59k - Cached - Similar pages
SEO contest - Wikipedia, the free encyclopedia
The nigritude ultramarine competition by SearchGuild is widely acclaimed as ...
Comparison of search results for nigritude ultramarine during and after the ...
en.wikipedia.org/wiki/Nigritude_ultramarine - 37k - Cached - Similar pages
Slashdot | How To Get Googled, By Hook Or By Crook
The current 3rd result showcases the "Nigritude Ultramarine Fighting Force" who ... When discussing nigritude ultramarine [slashdot.org] it is important to ...
slashdot.org/article.pl?sid=04/05/09/1840217 - 110k - Cached - Similar pages
The Nigritude Ultramarine Search Engine Optimization Contest
It's sweeping the web -- or at least search engine optimizers -- a new contest to rank tops for the term nigritude ultramarine on Google.
searchenginewatch.com/sereport/article.php/3360231 - 57k - Cached - Similar pages
Paid Search Ads
Sponsored Links
Business Blogging Seminar
g to L.A. March 16
Top bloggers reveal key techniques
www.blogbusinesssummit.com
Los Angeles, CA
Full-Time SEO & SEM Jobs
Find companies big & small hiring full-time SEO & SEM pros right now
CareerBuilder.com
SEO Contests
Information on SEO Contests like the Nigritude Ultramarine contest.
www.seo-contests.com/
The SEO Book
Nigritude Ultramarine & SEO secrets
Fun, free, raw, & different.
Overstock.com
Algorithmic results.
Done

User Needs

- Need [Brod02, RL04]
  - **Informational** – want to learn about something (~40% / 65%)

    Low hemoglobin
  - **Navigational** – want to go to that page (~25% / 15%)

    United Airlines
  - **Transactional** – want to do something (web-mediated) (~35% / 20%)
    - Access a service

      Seattle weather
    - Downloads

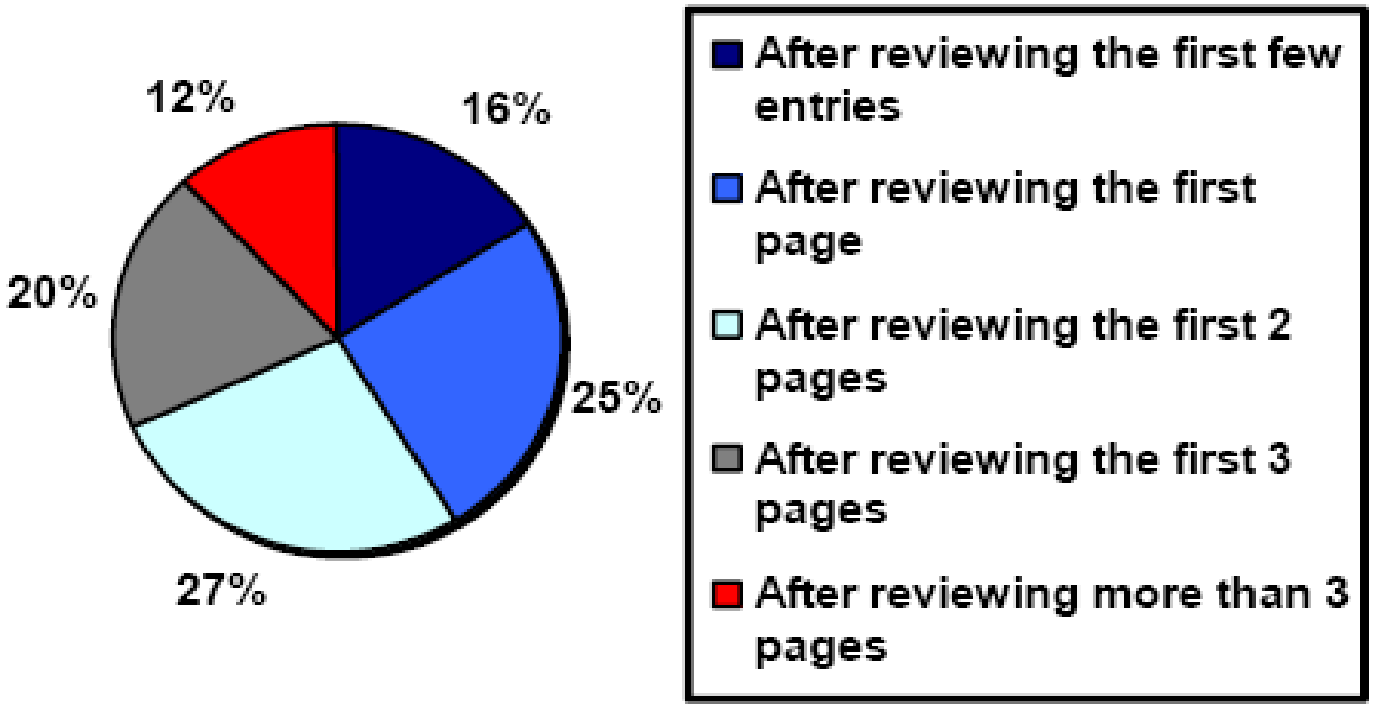
      Mars surface images
    - Shop

      Canon S410
  - **Gray areas**

    Car rental Brasil
    - Find a good hub
    - Exploratory search "see what's there"

How far do people look for results?

"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"



(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

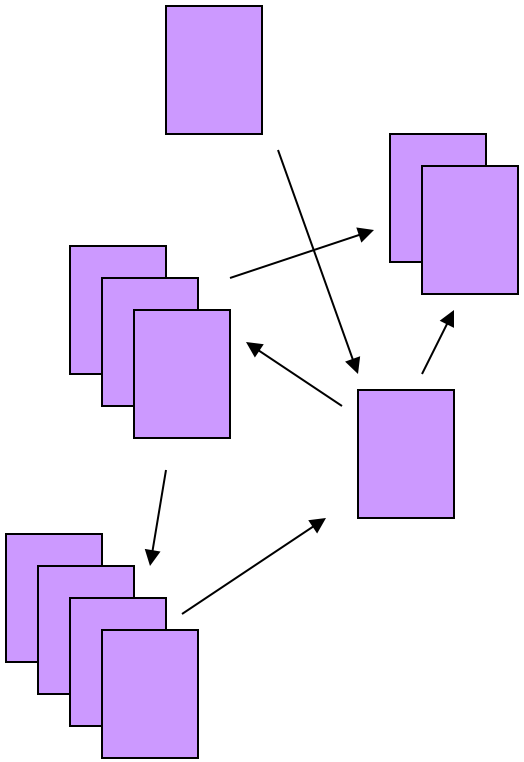
Users' empirical evaluation of results

- Quality of pages varies widely
  - Relevance is not enough
  - Other desirable qualities (non IR!!)
    - Content: Trustworthy, diverse, non-duplicated, well maintained
    - Web readability: display correctly & fast
    - No annoyances: pop-ups, etc
- Precision vs. recall
  - On the web, recall seldom matters
- What matters
  - Precision at 1? Precision within top-K?
  - Comprehensiveness – must be able to deal with obscure queries
    - Recall matters when the number of matches is very small
- User perceptions may be unscientific, but are significant over a large aggregate

Users' empirical evaluation of engines

- Relevance and validity of results
- UI – Simple, no clutter, error tolerant
- Trust – Results are objective
- Coverage of topics for polysemic queries
- Pre/Post process tools provided
  - Mitigate user errors (auto spell check, search assist,…)
  - Explicit: Search within results, more like this, refine ...
  - Anticipative: related searches, instant searches (next slide)
    - Impact on stemming, spell-check, etc
  - Web addresses typed in the search box

The Web document collection



The Web

- No design/co-ordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions …
- Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…
- Scale much larger than previous text collections … but corporate records are catching up
- Growth – slowed down from initial "volume doubling every few months" but still expanding
- Content can be *dynamically generated*

The trouble with paid search ads …

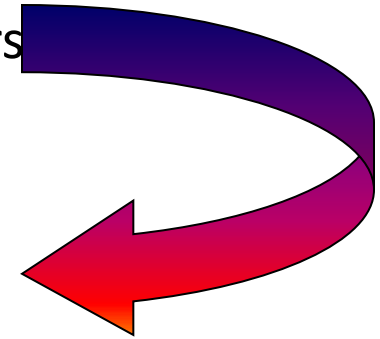
- It costs money. What's the alternative?
- *Search Engine Optimization (SEO):*
  - "Tuning" your web page to rank highly in the algorithmic search results for select keywords
  - Alternative to paying for placement
  - Thus, intrinsically a marketing function
- Performed by companies, webmasters and consultants ("Search engine optimizers") for their clients
- Some perfectly legitimate, some very shady

Search engine optimization (Spam)

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forums
  - E.g., Web master world (www.webmasterworld.com)

Simplest forms: Keyword Stuffing

- First generation engines relied heavily on *tf/idf*
  - The top-ranked pages for the query **`maui resort`** were the ones containing the most **`maui`**'s and **`resort`**'s
- SEOs -- dense repetitions of chosen terms
  - e.g., **`maui resort maui resort maui resort`**
  - Often, the repetitions would be in the same color as the background of the web page
    - Repeated terms got indexed by crawlers
    - But not visible to humans on browsers



Pure word density cannot be trusted as an IR signal

The war against spam

- Quality signals - Prefer authoritative pages based on:
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
- Policing of URL submissions
  - Anti robot test
- Limits on meta-keywords
- Robust link analysis
  - Ignore statistically implausible linkage (or text)
  - Use link analysis to detect spammers (guilt by association)

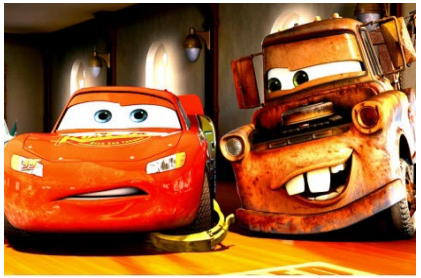
- Spam recognition by machine learning
  - Training set based on known spam
- Family friendly filters
  - Linguistic analysis, general classification techniques, etc.
  - For images: flesh tone detectors, source text analysis, etc.
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect pattern detection

Duplicate documents

- The web is full of duplicated content
- Strict duplicate detection = exact match
  - Not as common
- But many, many cases of near duplicates
  - E.g., Last modified date the only difference between two copies of a page

Eg, Near-duplicate videos

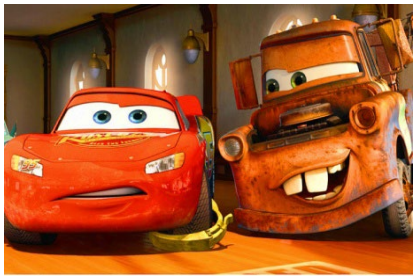
< Original Video>



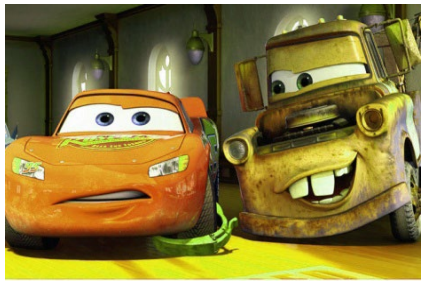
Contrast

Brightne

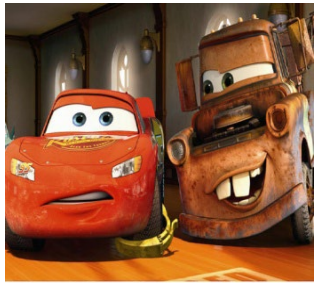
Crop



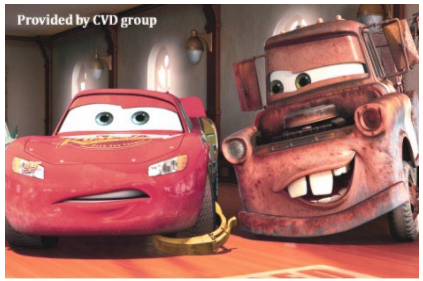
Color Enhancement



Color



TV



Multi-editing

Low resolution

Noise/Blur

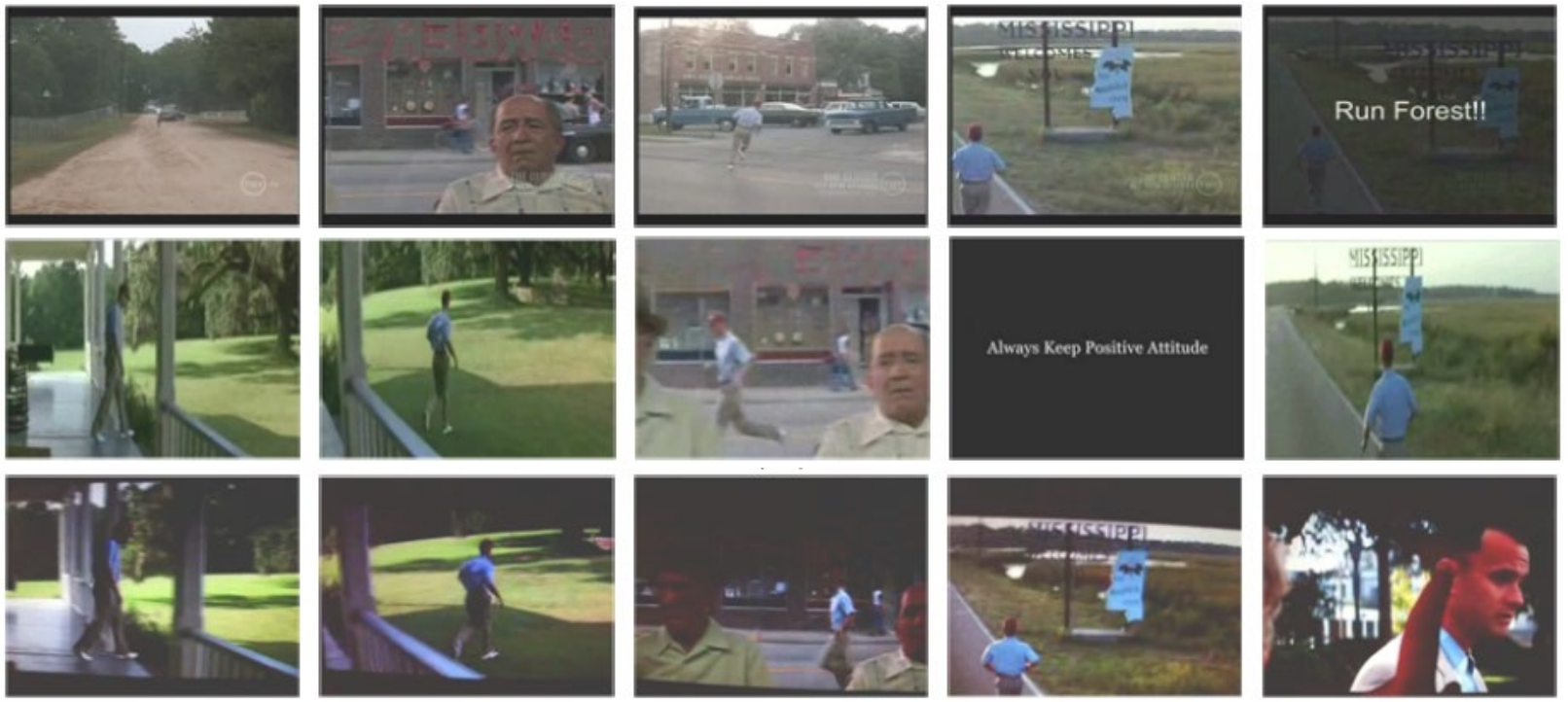
Small Logo

Eg, Near-duplicate videos

Original video

Elongated

Copied video



Duplicate/Near-Duplicate Detection

- *Duplication*: Exact match can be detected with fingerprints
- *Near-Duplication*: Approximate match
  - Compute syntactic similarity with an edit-distance measure
  - Use similarity threshold to detect near-duplicates
    - E.g., Similarity > 80% => Documents are "near duplicates"
    - Not transitive though sometimes used transitively

Computing Similarity

- Features:
  - Segments of a document (natural or artificial breakpoints)
  - Shingles (Word N-Grams)
  - ***a rose is a rose is a rose*** ***my rose is a rose is yours***

    a_rose_is_a

    rose_is_a_rose

    is_a_rose_is

    a_rose_is_a
- Similarity Measure between two docs (= sets of shingles)
  - Set intersection
  - Specifically (Size_of_Intersection / Size_of_Union)

Shingles + Set Intersection

- Issue: Computing exact set intersection of shingles between all pairs of documents is expensive

Shingles + Set Intersection

- Issue: Computing exact set intersection of shingles between all pairs of documents is expensive
  - Solution → Approximate using a cleverly chosen subset of shingles from each (called a *sketch*)
- Estimate (size_of_intersection / size_of_union) based on a short sketch