
ADVANCED TOPICS IN INFORMATION RETRIEVAL AND WEB SEARCH

Lecture 2: Information Retrieval vs. Search Engines

S. M. Vahidipour

Vahidipour@kashanu.ac.ir

Outline

- ❑ **Main issues in IR and SE**
- ❑ Search engine architecture
 - Indexing
 - Querying

IR and Search Engines

- A search engine is the practical application of information retrieval techniques to large scale text collections
- Big issues include main IR issues but also some others...

Information Retrieval

- Relevance: Effective ranking
- Evaluation: Testing and measuring
- Information needs: User interaction

Search Engines

- Performance: Efficient search and indexing
- Incorporating new data: Coverage and freshness
- Scalability: Growing with data and users
- Adaptability: Tuning for applications
- Specific problems: e.g., Spam

Additional



Main Issues in IR

Relevance

- A relevant document contains the information a user was looking for when he/she submitted the query.

Evaluation

- How well does the ranking meet the expectation of the user.

Users and information needs

- Users of a search engine are the ultimate judges of quality.

Relevance

- ❑ Simple (and simplistic) definition:
 - A relevant document contains the information that a person was looking for when they submitted a query to the search engine.
- ❑ Many factors influence a person's decision about what is relevant
 - Task at hand, context, novelty, style, serendipity
- ❑ Topical relevance vs. user relevance
 - “Storm in Tehran last Sunday” is topically relevant to query “آب و هوا”...
 - ... but might not be relevant to user because
 - ❖ Read it before
 - ❖ Is five years old
 - ❖ Is in a foreign language, etc.



Relevance

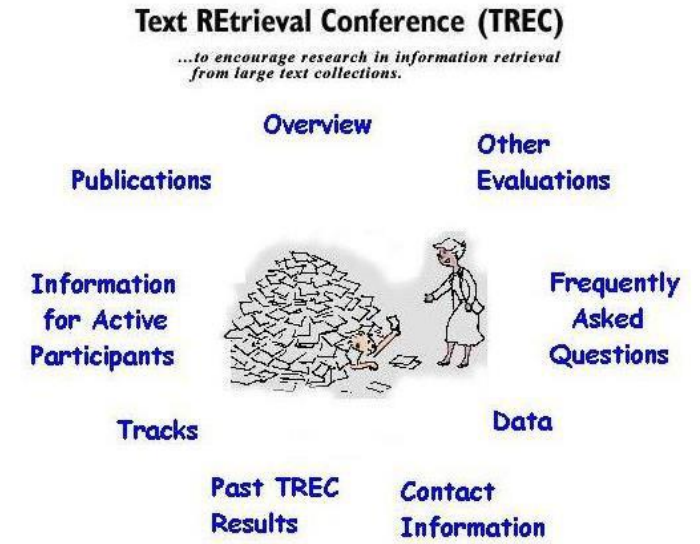
- Retrieval models define a view of relevance
 - Formal representation of the process of matching a query and a document
 - Simple text matching is not sufficient: Vocabulary mismatch problem (synonyms and homonyms)
- Ranking algorithms used in search engines are based on retrieval models
 - Real-world search engines consider topical and user relevance
- Most models describe statistical properties of text rather than linguistic
 - i.e. counting simple text features, such as words, instead of parsing and analyzing the sentences
 - Linguistic features can be part of a statistical model

Evaluation

- Experimental procedures and measures for comparing system output with user expectations
 - Originated in Cranfield experiments in the 60s
 - First large scale “benchmark”
- IR evaluation methods now used in many fields
- Recall and precision are two examples of effectiveness measures

Evaluation

- Typically use test collection (corpus) of documents, queries, and relevance judgments
 - Most commonly used are TREC collections (Text REtrieval Conf.)



- Clickthrough data to evaluate retrieval models and search engines.

[Free SEO Tools & Search Engine Optimization Software](http://tools.seobook.com/)
tools.seobook.com/

Free link analysis and keyword research **tools** brought to you by **SEO** expert Aaron Wall.
Keyword Tool - SEO for Firefox - Keyword Density Analyzer - Rank Checker
Todd Malicoat and aaron wall +1'd this

[SEO Tools - Search Engine Optimization Software | SEOMoz](http://www.seomoz.org/tools)
www.seomoz.org/tools

With **SEOMoz tools** you can better handle everyday **SEO** tasks. You can analyze keywords, research backlinks, do on-page analysis, find accessibility issues ...

[Free SEO Tools](http://www.seotools.net/)
www.seotools.net/

SEO Tools - 77 free online tools to help you perform search engine optimization. ... Use our free **SEO tools** for your comprehensive search engine optimization, ...

[SEO Tools - Search Engine Optimization, Google Optimization](http://www.seochat.com/seo-tools/)
www.seochat.com/seo-tools/

The **SEO Tools** found in this section were designed to assist you in configuring your ... Please select one of the **SEO Tools** from below to begin optimizing your ...

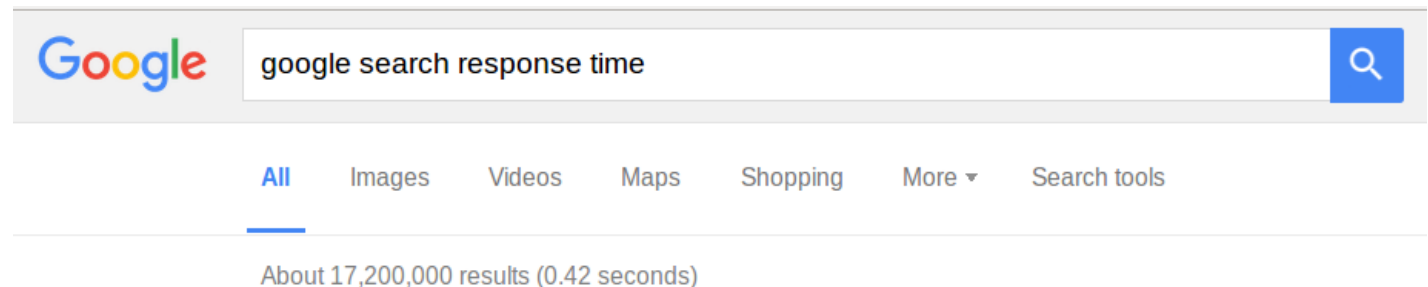
Users and Information Needs

- Search evaluation is user-centered
- Keyword queries are often poor descriptions of actual information needs
 - Query for “cats” could mean places to buy cats or the musical.
 - Search queries (in particular one-word queries) are under-specified.
- Interaction and context are important for understanding user intent
- Query refinement techniques such as
 - Query suggestion
 - Query expansion and relevance feedback
- Improve ranking



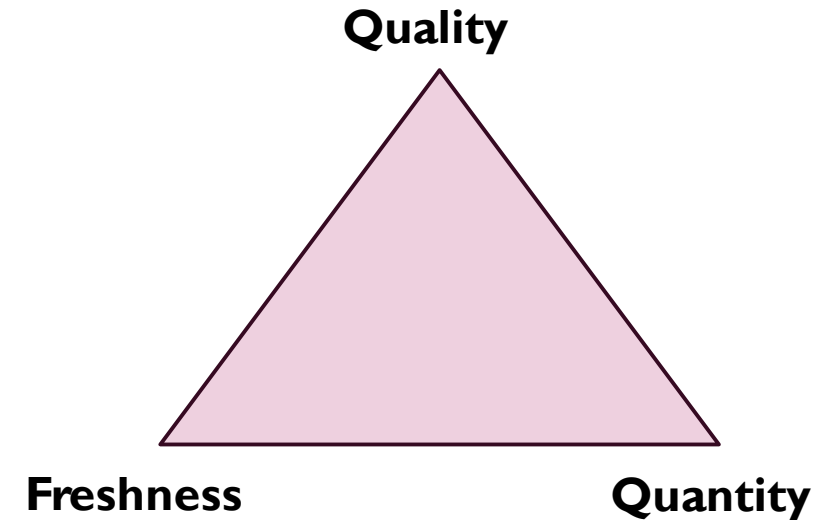
Performance

- Measuring and improving the efficiency of search
 - Reduce response time
 - Increase query throughput
 - Increase indexing speed
- Indexes are data structures designed to improve search efficiency.
 - Designing and implementing them are major issues for search engines.
- Same tradeoffs as in DBMS



Dynamic Data

- The “collection” for most real applications is constantly changing in terms of updates, additions, deletions
- Acquiring or “crawling” the documents is a major task
- Typical measures are
 - Coverage (how much has been indexed)
 - Freshness (how recently was it indexed)



Scalability

- Making everything work with billions of users every day, and many terabytes of documents
- Distributed processing is essential
- But: Large \neq scalable

Google in 2006	Google in 2008	Google in 2018
> 25 billion pages	• 1 trillion pages (1,000,000,000,000)	• 130 trillion pages

Adaptability

- Changing and tuning search engine components
 - ranking algorithm
 - indexing strategy
 - interface for different applications
- Adapt to different requirements for different applications / users
 - New APIs
 - New uses for search

Spam

- For Web search, spam in all its forms is one of the major issues
- Affects the efficiency of search engines and, more seriously, the effectiveness of the results
- Many types of spam
 - e.g., spamdexing or term spam, link spam, “optimization”
 - <http://en.wikipedia.org/wiki/Spamdexing>
- New subfield called adversarial IR, since spammers are “adversaries” with different goals

Outline

- Main issues in IR and SE
- **Search engine architecture**
 - Indexing
 - Querying

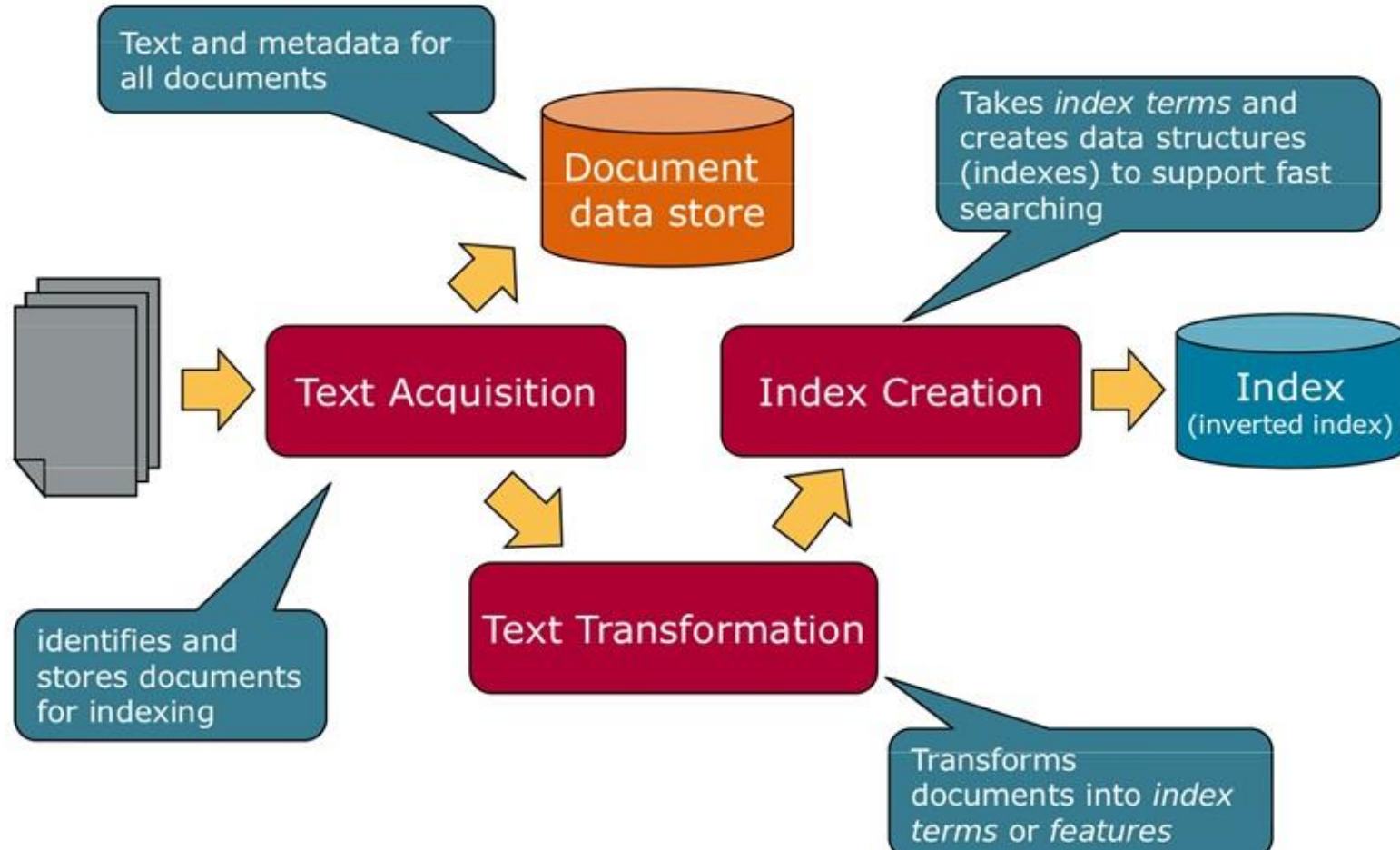
Search Engine Architecture

- Basic Building Blocks
- Indexing
 - Text Acquisition
 - Text Transformation
 - Index Creation
- Querying
 - User Interaction
 - Ranking
 - Evaluation

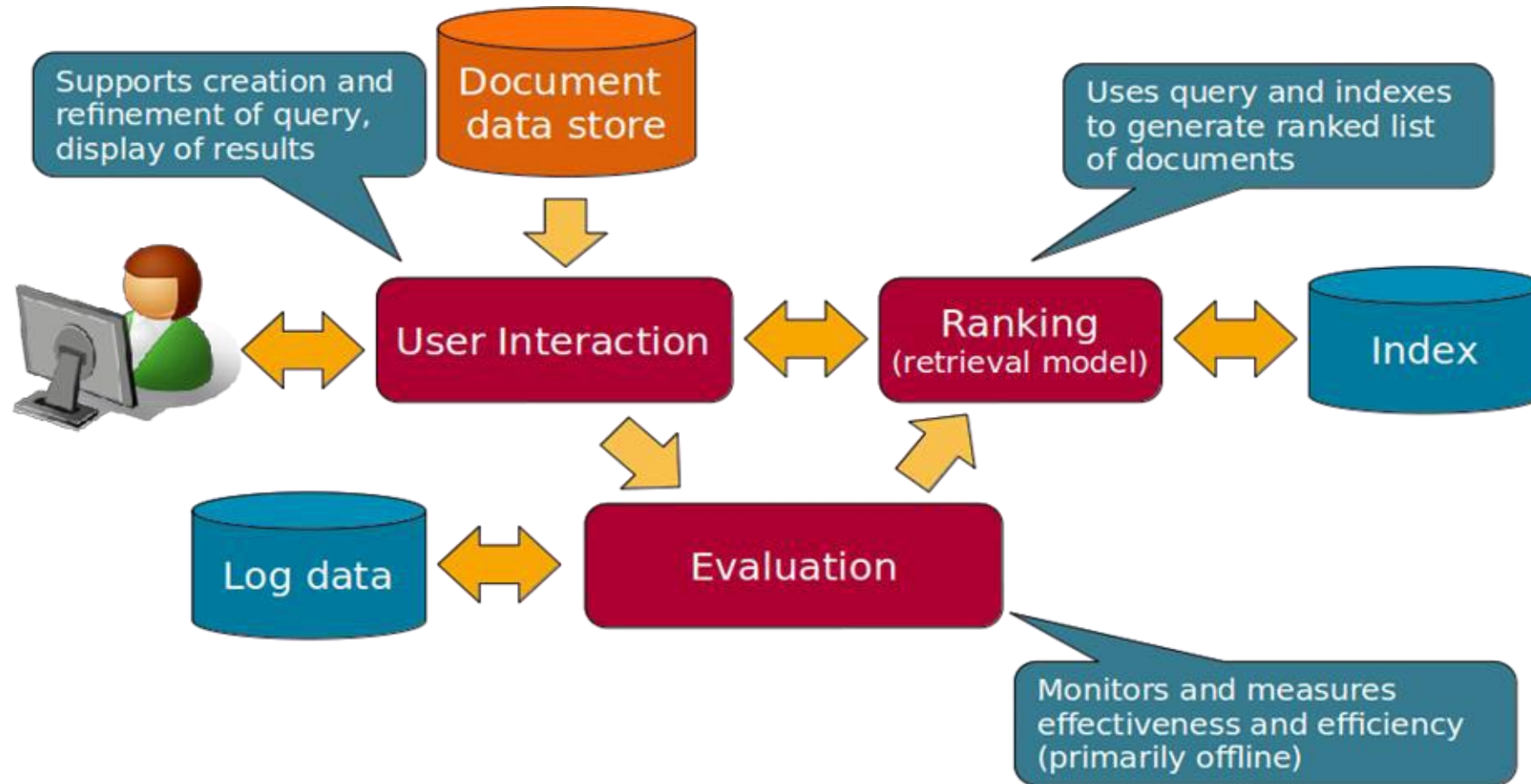
Search Engine Architecture

- Determined by two main requirements
 - Effectiveness (quality of results)
 - As good as possible
 - Efficiency (response time and throughput)
 - As quickly as possible
- Other requirements fall into these categories
 - Changing documents -> Effectiveness and efficiency
 - Personalization: Effectiveness
 - Spam: Effectiveness and efficiency
 - ...

The Indexing Process



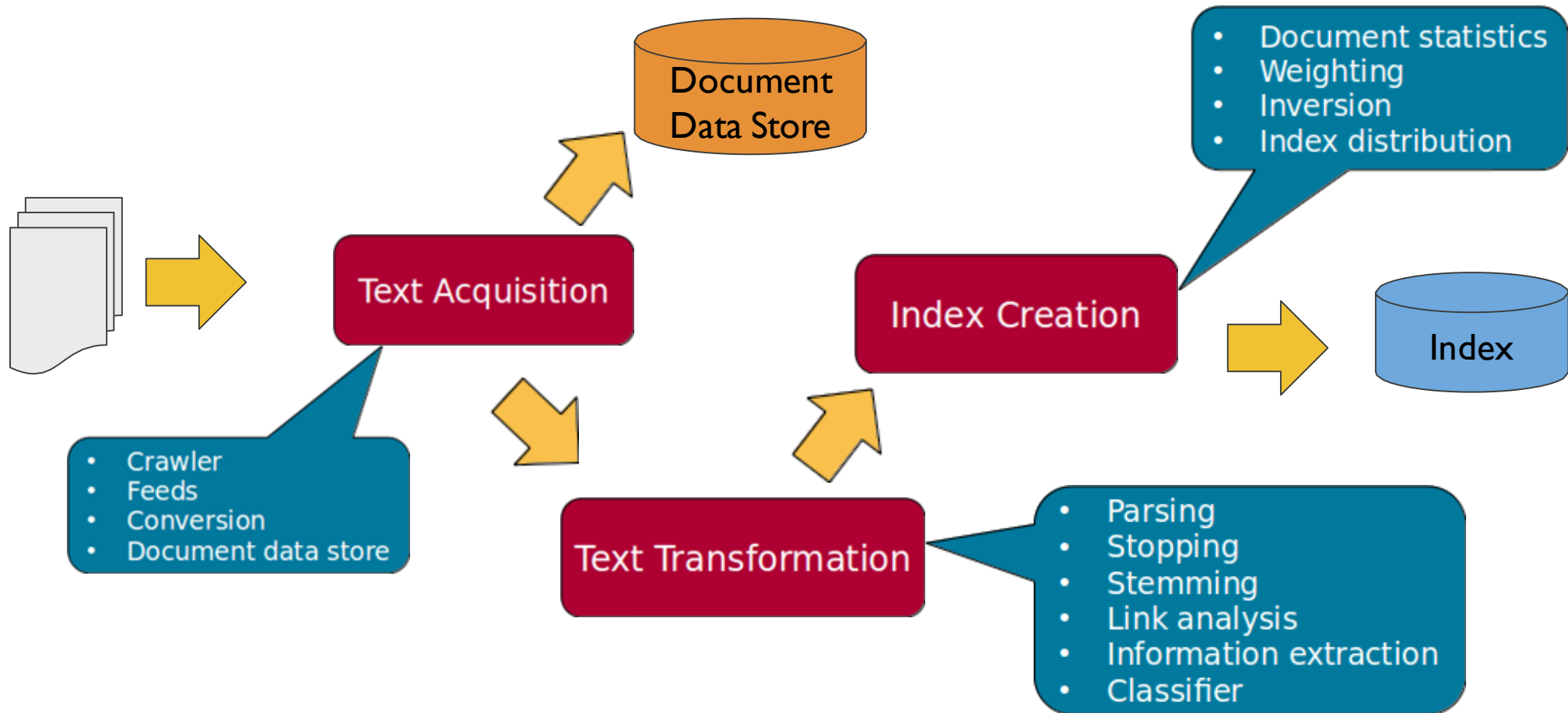
The Query Process



Outline

- Main issues in IR and SE
- **Search engine architecture**
 - **Indexing**
 - Querying

The Indexing Process



Text Acquisition – Crawler

- Identifies and acquires documents for search engine
- Many types
 - Web, enterprise, desktop
- Web crawlers follow links to find documents
 - Must efficiently find huge numbers of web pages (coverage) and keep them up-to-date (freshness)
 - Single site crawlers for site search
 - Topical or focused crawlers for vertical search
- Document crawlers for enterprise and desktop search
 - Follow links and scan directories



TextAcquisition – Feeds

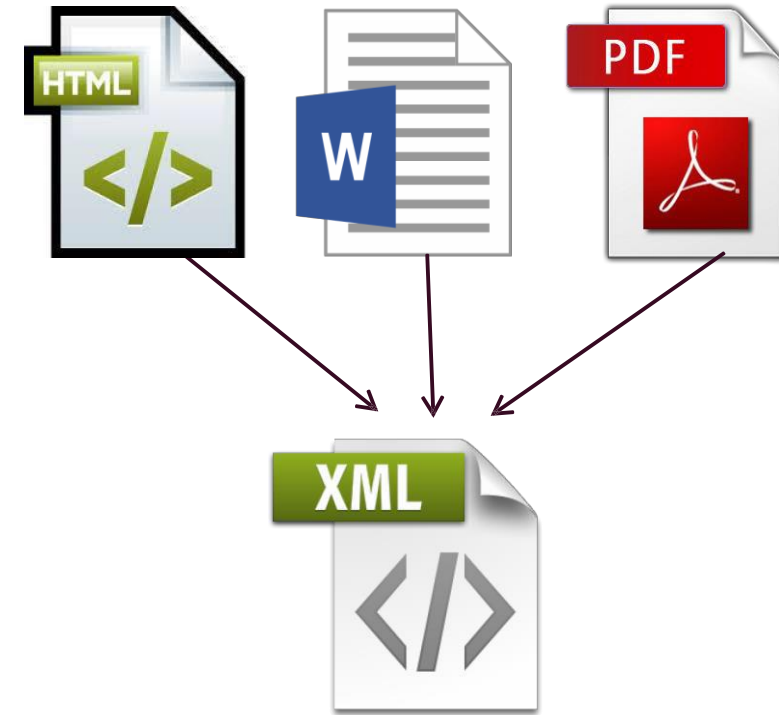
- Real-time streams of documents
 - Web feeds for news, blogs, video, radio, TV
- RSS is common standard
 - RSS “reader” can provide new XML documents to search engine



```
Quelltext von: http://www.hpi.uni-potsdam.de/fileadmin/hpi_settings/rss.php - Mozilla Firefox
Datei Bearbeiten Ansicht Hilfe
<?xml version="1.0" encoding="utf-8"?>
<rss version="2.0" xmlns:content="http://purl.org/rss/1.0/modules/content/">
  <channel>
    <title>HPI RSS Newsfeed</title>
    <link>http://www.hpi.uni-potsdam.de/</link>
    <description>News der HPI-Webseite</description>
    <language>de</language>
    <image>
      <title>HPI RSS Newsfeed</title>
      <url>http://www.hpi.uni-potsdam.de/uploads/RTEmagicC_rss.gif</url>
      <link>http://www.hpi.uni-potsdam.de/</link>
      <width>16</width>
      <height>16</height>
      <description>News der HPI-Webseite</description>
    </image>
    <generator>TYPO3 - get.content.right</generator>
    <doc>http://blogs.law.harvard.edu/tech/rss</doc>
    <lastBuildDate>Wed, 01 Apr 2009 10:32:00 +0200</lastBuildDate>
    <item>
      <title>HPI baut Angebot bei iTunes U weiter aus</title>
      <link>http://www.hpi.uni-potsdam.de/news/beitrag/hpi-baut-angebo
      <description>Mit neuen Beiträgen weitet das Hasso-Plattner-Insti
      <content:encoded><![CDATA[<p class="bodytext">Mit neuen Beiträge
    </p>
    <p class="bodytext"><a href="http://itunes.hpi.uni-potsdam.de" title="Opens external lin
      <pubDate>Wed, 01 Apr 2009 10:32:00 +0200</pubDate>
    </item>
    <item>
      <title>Internet-Experten treffen sich zum IPv6-Gipfel</title>
      <link>http://www.hpi.uni-potsdam.de/news/beitrag/internet-expert
      <description>Zum zweiten Mal findet am Hasso-Plattner-Institut a
      <content:encoded><![CDATA[<p class="bodytext">Zum zweiten Mal fi
    </p>
    <p class="bodytext">Der erste Tag der Veranstaltung soll eine Übersicht über internationa
    </p>
    <p class="bodytext">Der Schwerpunkt des zweiten Tages liegt auf der Sicht internationale
    </p>
    <p class="bodytext">Zusätzlich wird parallel ein halbtägiges IPv6-Tutorium angeboten.
    </p>
    <p class="bodytext">&nbsp;</p>
    <p class="bodytext">Weitere Informationen in der aktuellen <a href="http://www.hpi.uni-p
      <pubDate>Fri, 27 Mar 2009 09:00:00 +0100</pubDate>
    </item>
    <item>
      <title>HPI lädt zum MINT-Camp 2009</title>
```

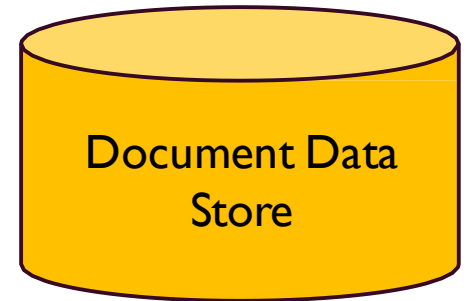
TextAcquisition – Conversion

- Convert variety of document formats into a consistent text-plus-metadata format
 - e.g., HTML, XML, Word, PDF, etc. → XML
- Convert text encoding for different languages
 - Using a standard like UTF-8
 - Be consistent throughout application
- Non-content data (tags, metadata) is either removed or stored as metadata.
- First step towards text transformation

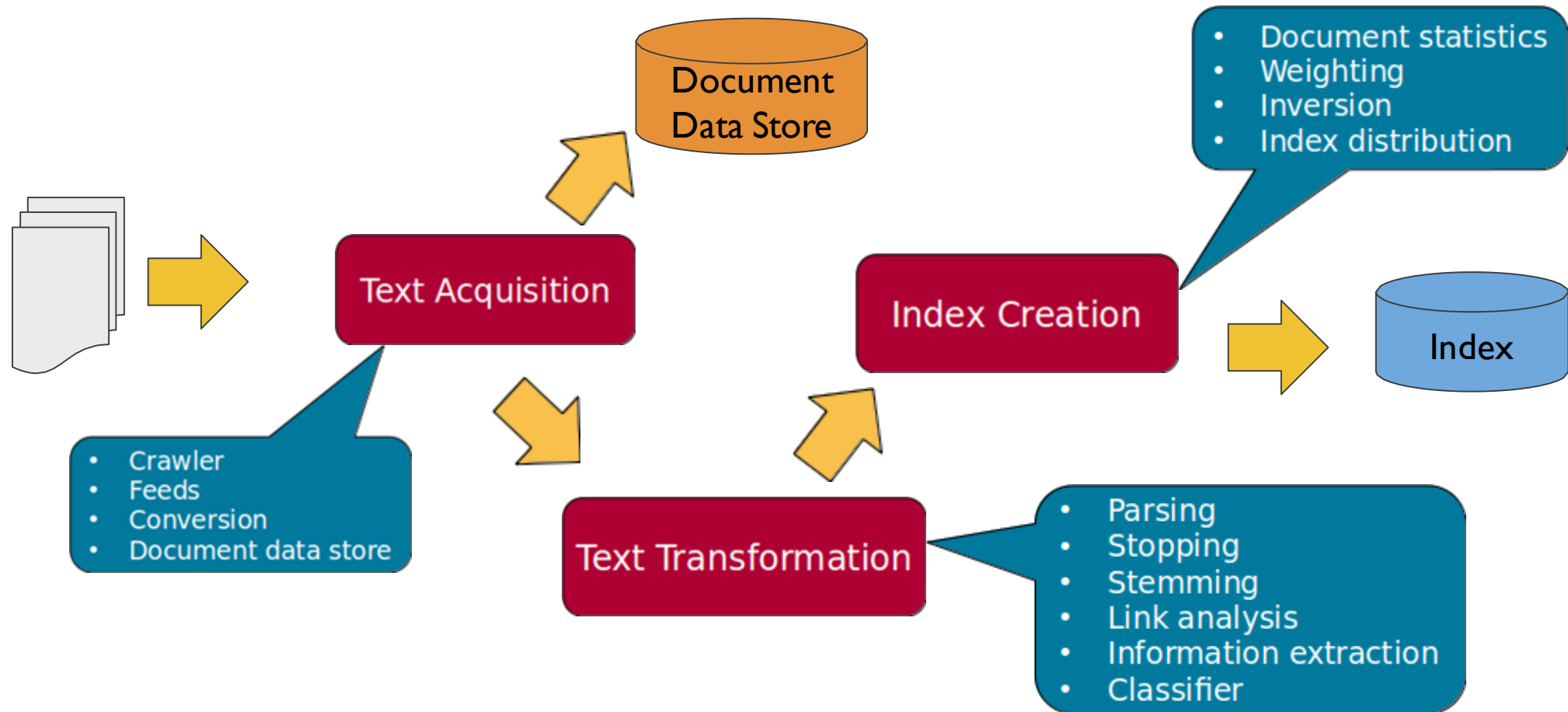


TextAcquisition – Document Data Store

- Two parts
 - Unstructured text
 - Structured metadata
- Stores text, metadata, and other related content for documents
 - Metadata is information about document
 - Type, creation date, ...
 - Other content includes links, anchor text
- Why store documents? They are available on the Web anyway...
 - Provide fast access to document contents for search engine components
 - Result list generation, document summary, snippet



The Indexing Process



Text Transformation – Parsing

- Processing the sequence of text tokens in the document to recognize structural elements
 - Titles, links, headings, etc.
- Markup languages such as HTML, XML often used to specify structure
 - Tags used to specify document elements
 - E.g., `<h2> Overview </h2>`
 - Document parser uses syntax of markup language (or other formatting) to identify structure
 - E.g. email format, MS Word metadata etc.



Text Transformation – Parsing

Input: Friends, Romans, Countrymen, lend me your ears;

Output: Friends Romans Countrymen lend me your ears

- Tokenizer recognizes “words” in the text.
 - Must consider issues like capitalization, hyphens, apostrophes, non-alpha characters, separators



neill	aren't
oneill	arent
o'neill	are n't
o' neill	aren t?
o neill?	

朝鲜外务省发言人11月1日在平壤宣布，朝鲜将重返六方会谈，但前提条件是朝鲜与美国在六方会谈框架内讨论解除美国对朝鲜核问题。

针对朝鲜方面“*Where are the words?*”表示欢迎。

美联社11月1日报道说：“长期以来一直拒绝与平壤进行直接对话的美国总统布什认为，各方达成一致、同意恢复六方会谈应归功于中国的斡旋。”

Text Transformation – Stopping

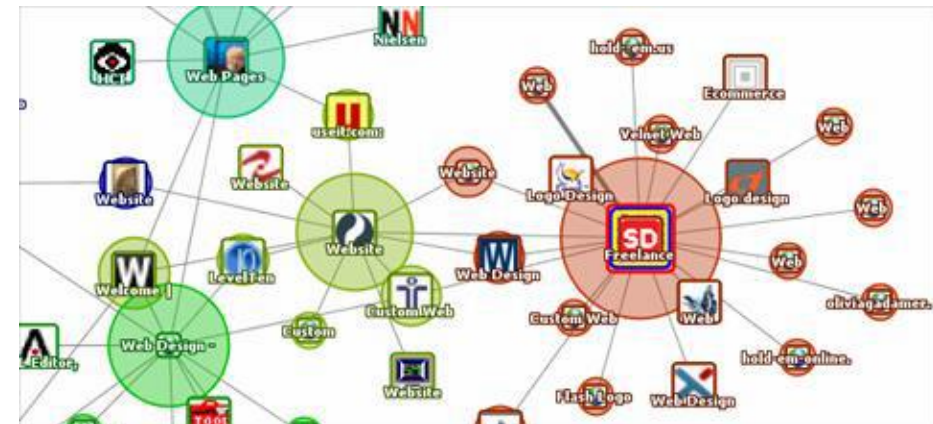
- Remove common words
 - “and”, “or”, “the”, “in”, ...
- Some impact on efficiency and effectiveness
- Can be a problem for some queries
 - “To be or not to be”

Text Transformation – Stemming

- Group words derived from a common stem
 - “computer”, “computers”, “computing”, “compute”
 - “fish”, “fishing”, “fisherman”
- Usually effective, but not for all queries
- Benefits vary for different languages
 - Arabic: Very complicated morphology
 - Chinese: Few word variations anyway

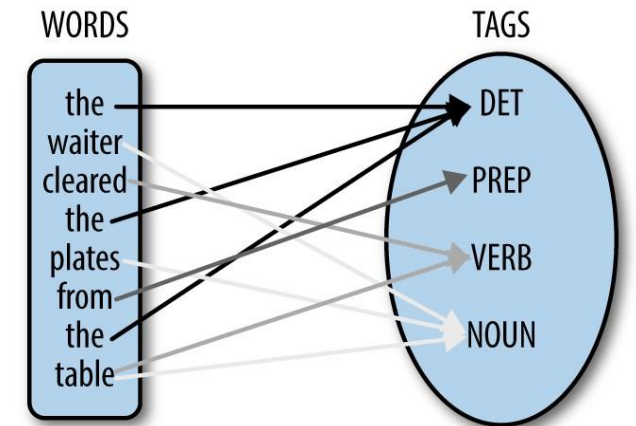
Text Transformation – Link Analysis

- Makes use of links and anchor text in web pages.
 - Stored and indexed separately
 - ` University of Kashan`
- Link analysis identifies popularity and community information
 - e.g., PageRank
- Anchor text can significantly enhance the representation of pages pointed to by links
- Significant impact on web search
 - Less importance in other applications



Text Transformation – Information Extraction

- Identify classes of index terms that are important for some applications
- Simple: Bold-face, heading, title
- Part of speech tagging
- Named entity recognizers (NER) identify classes such as
 - People
 - Locations
 - Companies
 - Dates
 - etc.

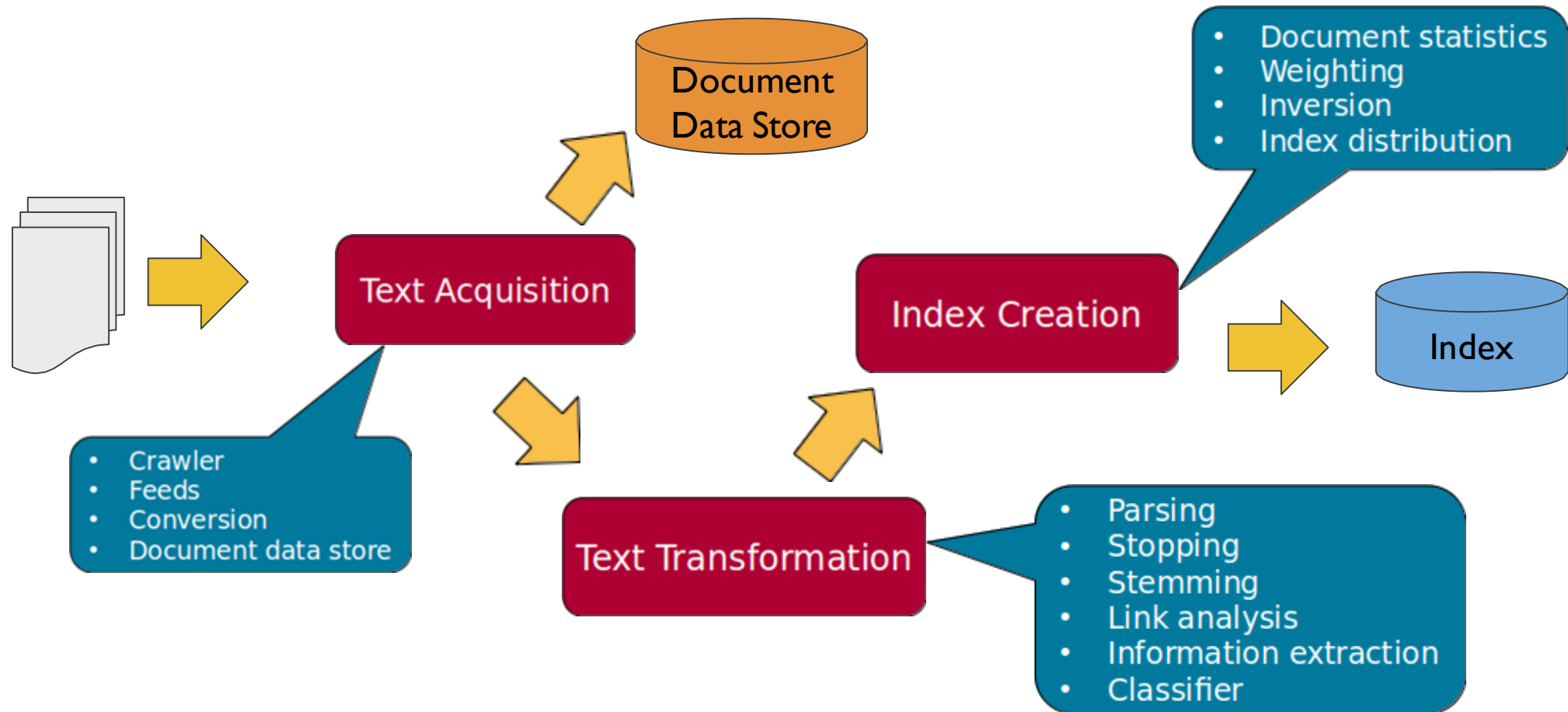


Text Transformation – Classifier

- Identifies class-related metadata for documents
 - i.e., assigns labels to documents
 - e.g., topics, reading levels, sentiment, genre
 - Spam!
 - Advertisements in documents
- Use depends on application



The Indexing Process



Index Creation - Document Statistics

- Statistical information about words, features and documents
- Gathers counts and positions of words and other features
 - Within a document
 - Across groups of documents
 - Across all documents
- Used in ranking algorithm

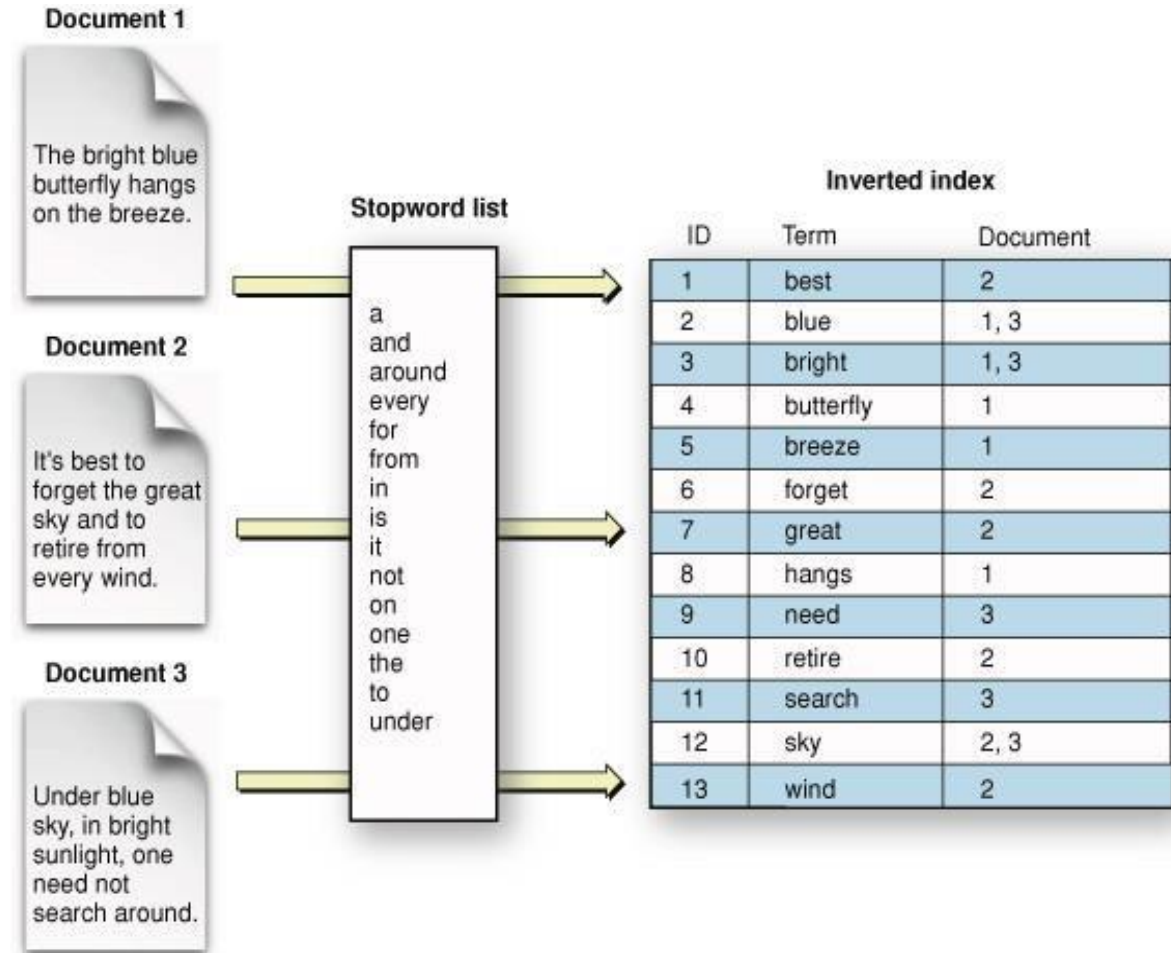


Index Creation –Weighting

- Computes weights for index terms
 - Relative importance of words in documents
 - Used in ranking algorithm
- Global weight
 - Query-dependent weight
- TF.IDF weight
 - Combination of term frequency in document
 - and inverse document frequency in the collection

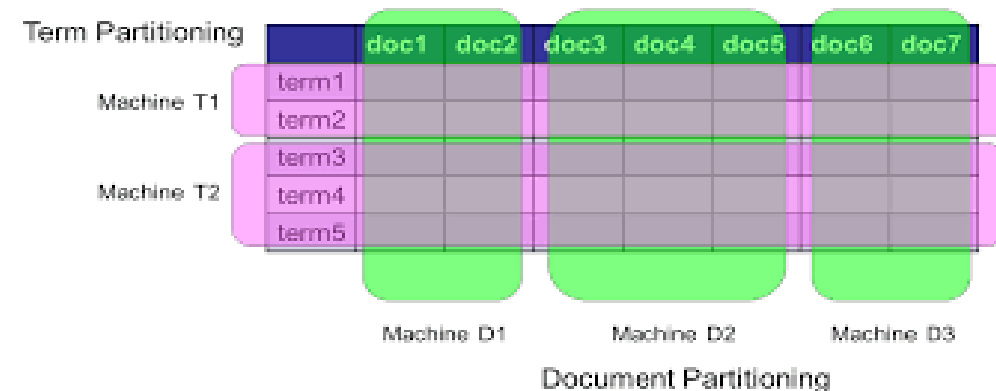
Index Creation – Inversion

- Core of indexing process
- Converts document-term information to term-document for indexing
 - Difficult for very large numbers of documents
 - Classical Map/Reduce use case
- Format of inverted file is designed for fast query processing
 - Must also handle updates
 - Compression used for efficiency



Index Creation – Index Distribution

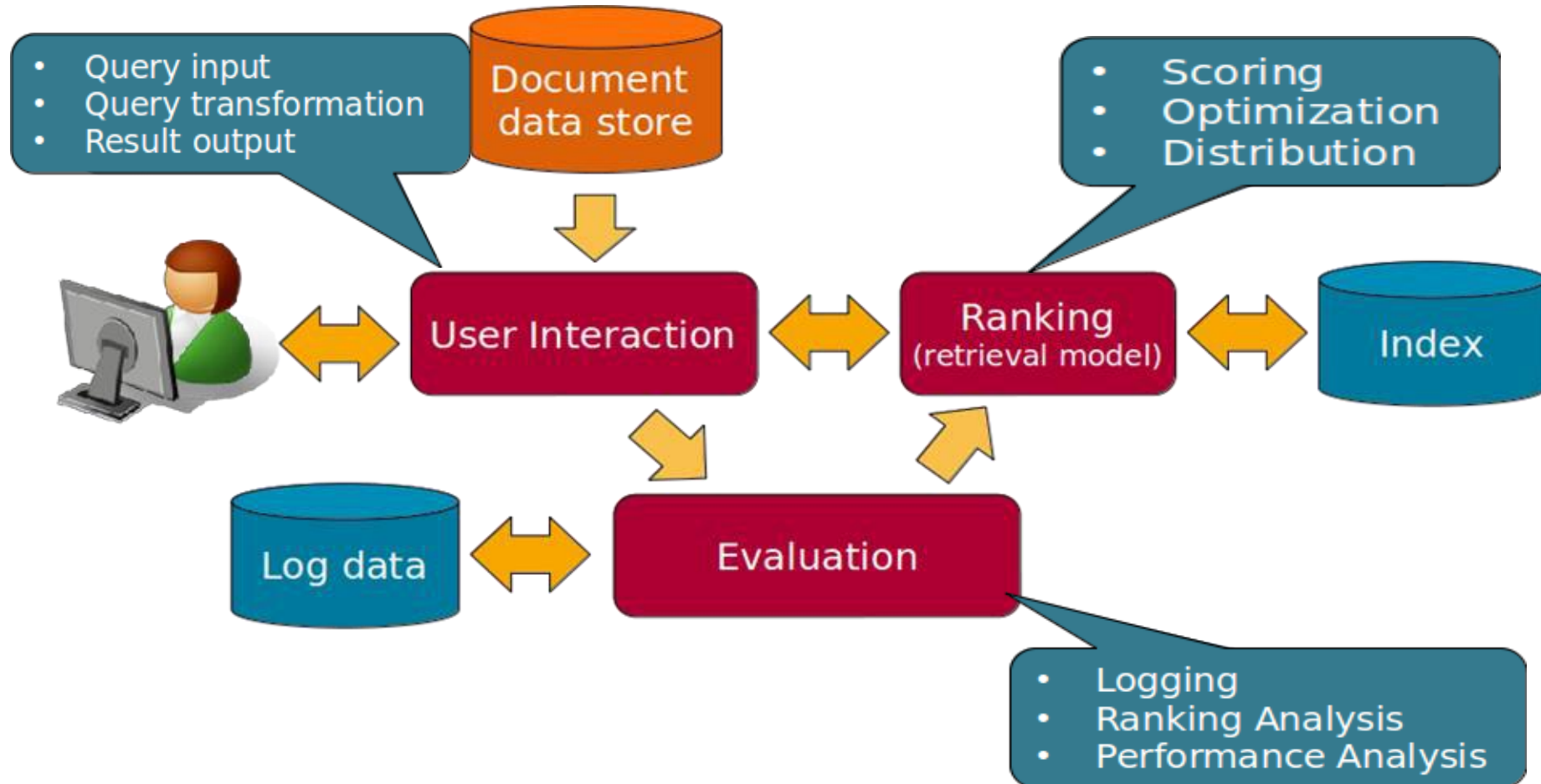
- Distribute indexes
 - across multiple computers
 - and/or multiple sites
- Essential for fast query processing with large numbers of documents
- Many variations
 - Document distribution: Distribute index for subsets of documents
 - Term distribution: Distribute index for subset of terms
 - Replication



Outline

- Main issues in IR and SE
- **Search engine architecture**
 - Indexing
 - **Querying**

The Query Process



User Interaction - Query input

- Provides interface and parser for internal query language
- Most web queries are very simple
 - Other applications may use forms
- Query language used to describe more complex queries and results of query transformation
 - +, -, “ “, ~, site:, AND, OR, ...
 - Similar to SQL language used in database applications
 - Not for “end users”
 - IR query languages also allow structure specifications, but focus on content

User Interaction - Query input



Google Search

I'm Feeling Lucky



Advanced Search

Find pages with...

all these words:

this exact word or phrase:

any of these words:

none of these words:

numbers ranging from:

to

To do this in the search box.

Type the important words: tri-colour rat terrier

Put exact words in quotes: "rat terrier"

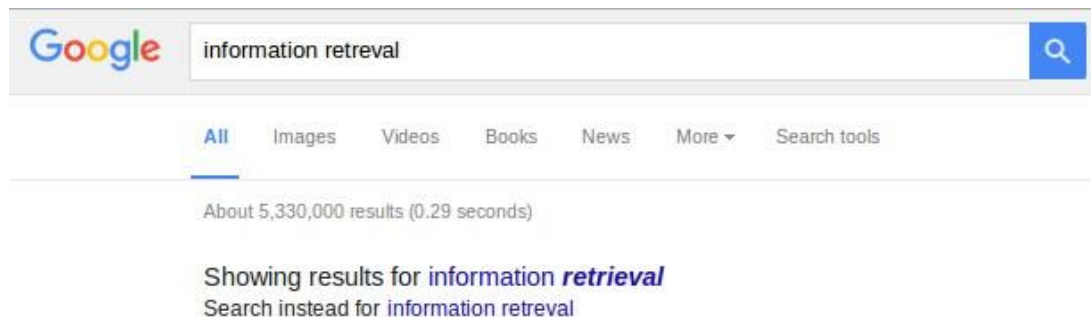
Type OR between all the words you want: miniature OR standard

Put a minus sign just before words that you don't want:
-rodent, -"Jack Russell"

Put two full stops between the numbers and add a unit of measurement:
10..35 kg, £300..£500, 2010..2011

User Interaction – Query transformation

- Improves initial query
 - both before and after initial search
- Includes same text transformation techniques used for documents
 - Tokenization, stemming, stopping
- Spell checking and query suggestion provide alternatives to original query
 - Based on query logs
- Query expansion and relevance feedback modify the original query with additional terms



Searches related to information retrieval

- [information retrieval pdf](#)
- [information retrieval system](#)
- [information retrieval techniques](#)
- [information retrieval papers](#)
- [information retrieval ppt](#)
- [information retrieval journal](#)
- [information retrieval definition](#)
- [information retrieval algorithms](#)

User Interaction – Results output

- Constructs the display of ranked documents for a query
- Generates snippets to show how queries match documents
- Highlights important words and passages
- Retrieves appropriate advertising in many applications
- May provide clustering and other visualization tools
- May translate results from foreign languages

User Interaction – Results output

Google camel

About 151,000,000 results (0.59 seconds)

Camel - Wikipedia, the free encyclopedia
<https://en.wikipedia.org/wiki/Camel> - Wikipedia -
A camel is an even-toed ungulate within the genus *Camelus*, bearing distinctive fatty deposits known as "humps" on its back. The two surviving species of camel ...
Bactrian camel · Dromedary · Camel (disambiguation) · Australian feral camel

Camel (band) - Wikipedia, the free encyclopedia
[https://en.wikipedia.org/wiki/Camel_\(band\)](https://en.wikipedia.org/wiki/Camel_(band)) - Wikipedia -
Camel are an English progressive rock band formed in 1971. Led by founding member Andrew Latimer, they have produced 14 original studio albums, 14 ...
Members: Andrew Latimer, Colin Bass, Denis ... Origin: Guildford, Surrey, England
Genres: Progressive rock, symphonic rock Years active: 1971 –present

In the news

HC dismisses plea to permit camel slaughtering in Tamil Nadu
The Hindu - 2 days ago
The Bench made it clear that the court has not completely banned camel slaughtering, but has temporarily denied permission since there is no ...

Camel slaughter ban to stay: HC
Times of India - 1 day ago

Chicago zoo names baby camel 'Alexander Camelton'
6abc.com - 1 day ago

More news for camel

Apache Camel: Index
camel.apache.org/ - Apache Camel -
Apache Camel uses URIs to work directly with any kind of Transport or messaging model such as HTTP, ActiveMQ, JMS, JBI, SCA, MINA or CXF, as well as ...

Welcome to the Official Camel Website
www.camelproductions.com/ -
Official site with news, tour information, timeline, merchandise and jukebox. Home site of founder Andy Latimer.

camel active
www.camelactive.de/ -
camel active - ist eine internationale Lifestyle-Marke für Männer und Frauen.

Amazing facts about camels | OneKind
www.onekind.com/education/animals/camel/ - OneKind -

Camel
Animal

A camel is an even-toed ungulate within the genus *Camelus*, bearing distinctive fatty deposits known as "humps" on its back.

Scientific name: *Camelus*
Lifespan: Dromedary: 40 years
Speed: 65 km/h (Maximum, in Short Bursts, Running)
Height: Dromedary: 1.8 – 2 m
Gestation period: Dromedary: 15 months, Bactrian: 14 months
Mass: Dromedary: 400 – 600 kg, Bactrian camel: 460 – 680 kg

Lower classifications

Dromedary Bactrian camel Cama Hybride camel

See results about

Camel (Rock band)
Albums: The Snow Goose, Mirage, Moonmadness
Members: Andrew Latimer, Peter Bardens, Andy Ward

Google Camel (band)

About 8,570,000 results (0.45 seconds)

Camel (band) - Wikipedia, the free encyclopedia
[https://en.wikipedia.org/wiki/Camel_\(band\)](https://en.wikipedia.org/wiki/Camel_(band)) - Wikipedia -
Camel are an English progressive rock band formed in 1971. Led by founding member Andrew Latimer, they have produced 14 original studio albums, 14 ...

Camel (album)
Camel is the first studio album by English progressive rock band ...

Andrew Latimer
Andrew Latimer (17 May 1949, Guildford, Surrey, England) is ...

Mirage (Camel album)
Mirage is the second studio album by the English progressive rock ...
More results from wikipedia.org »

I Can See Your House from Here
I Can See Your House From Here is the seventh studio album by ...

The Snow Goose
The Snow Goose is the third album by the band Camel, released in ...

Moonmadness
Moonmadness is an album released in March 1976 by ...

Camel - Snow Goose Excerpts - YouTube
<https://www.youtube.com/watch?v=ZTVnCyDoQIQ> -
Dec 9, 2006 - Uploaded by Artur Rico
... and "Rayader Goes To Town" tapped May 9, 1975, at BBC studios. The Camel band with the original ...

Camel LIVE 2013 - Lady Fantasy - De Melkweg, Amsterdam (Andy ...)
<https://www.youtube.com/watch?v=Lw-dMAuI8M> -
Oct 27, 2013 - Uploaded by sixstringsavoring
Full HD Stereo video of Camel performing "Lady Fantasy" during their 2013 The ...
Camel is by far my ...

Camel Discography at Discogs
<https://www.discogs.com/artist/140646-Camel> - Discogs -
It also marked Camel's return to the UK charts at number 45. The band had three further studio albums, a live album and numerous personnel changes before ...

CAMEL discography and reviews - Progarchives
www.progarchives.com/artist.asp?id=50 -
The roots of CAMEL go as far as 1964, when the Latimer brothers Andrew and Bryan form part of a band called THE PHANTOM FOUR, after gaining some fame, ...

Camel Albums: songs, discography, biography, and listening guide ...
rateyourmusic.com/artist/camel -
Camel discography and songs: Music profile for Camel, formed November 1971. ... about the musical

Camel
Rock band

camelproductions.com

Camel are an English progressive rock band formed in 1971. Led by founding member Andrew Latimer, they have produced 14 original studio albums, 14 singles plus numerous other compilation and live albums.
Wikipedia

Origin: Guildford, United Kingdom (1971)
Members: Andrew Latimer, Peter Bardens, Andy Ward, Colin Bass, more
Record labels: Camel Productions, Decca Records, Deram Records, MCA Records, Arista Records, MCA Inc.

Songs

Lady Fantasy
Mirage - 1974

Rhayader Goes to Town
The Snow Goose - 1975

Long Goodbyes
Stationary Traveller - 1984

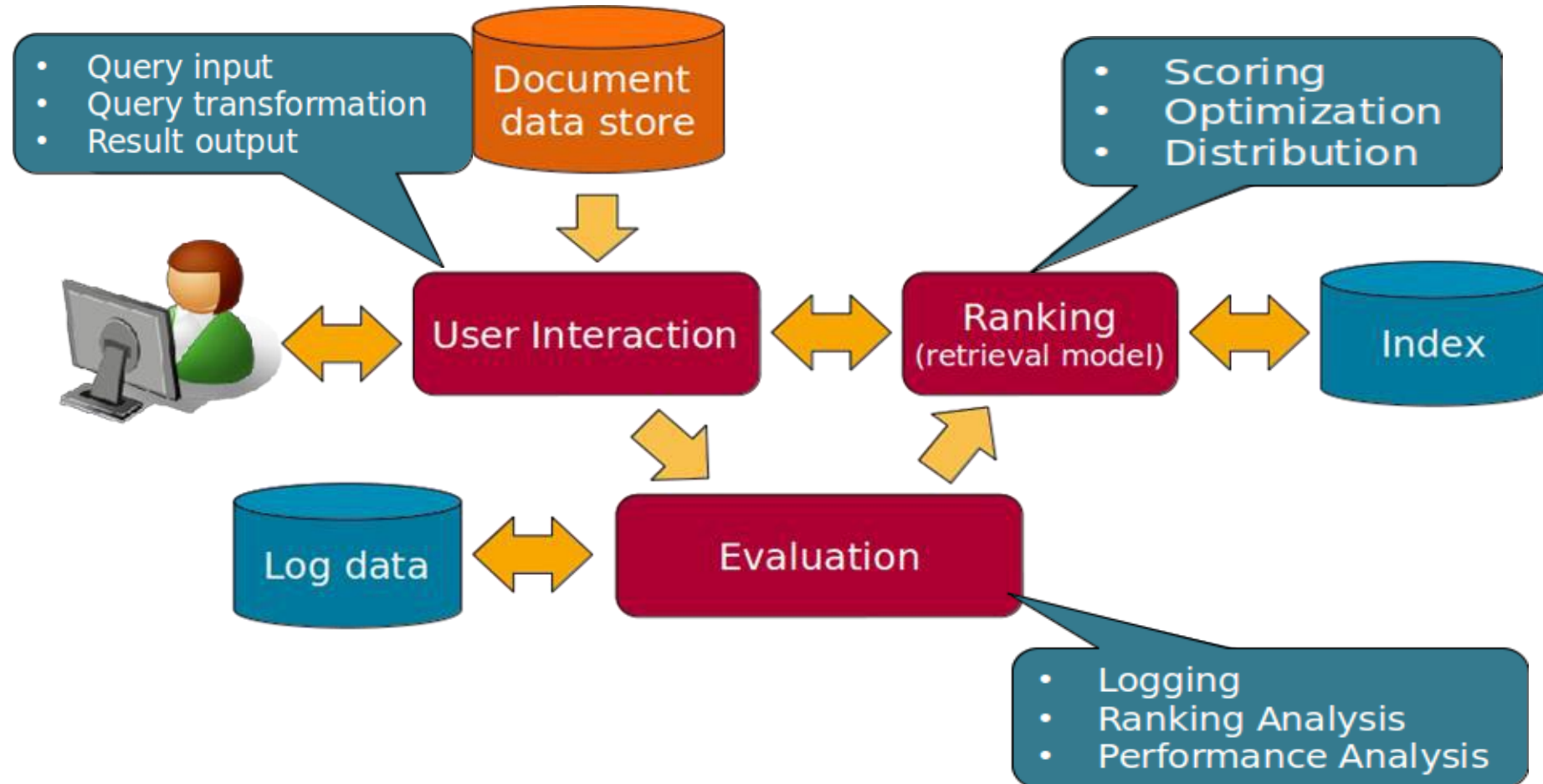
View 25+ more

Albums

The Snow Goose
Mirage
Moonmad...
Stationary Traveller
Rajaz

View 25+ more

The Query Process



Ranking – Scoring

- \approx database query processing
- Calculates scores for documents using a ranking algorithm
 - Based on retrieval model
- Core component of search engine
- Many variations of ranking algorithms and retrieval models
- Key requirement: Fast execution!

Ranking – Performance optimization

- Designing ranking algorithms for efficient processing
 - Term-at-a time vs. document-at-a-time processing
 - Safe vs. unsafe optimizations
 - Trade-off between speed and quality

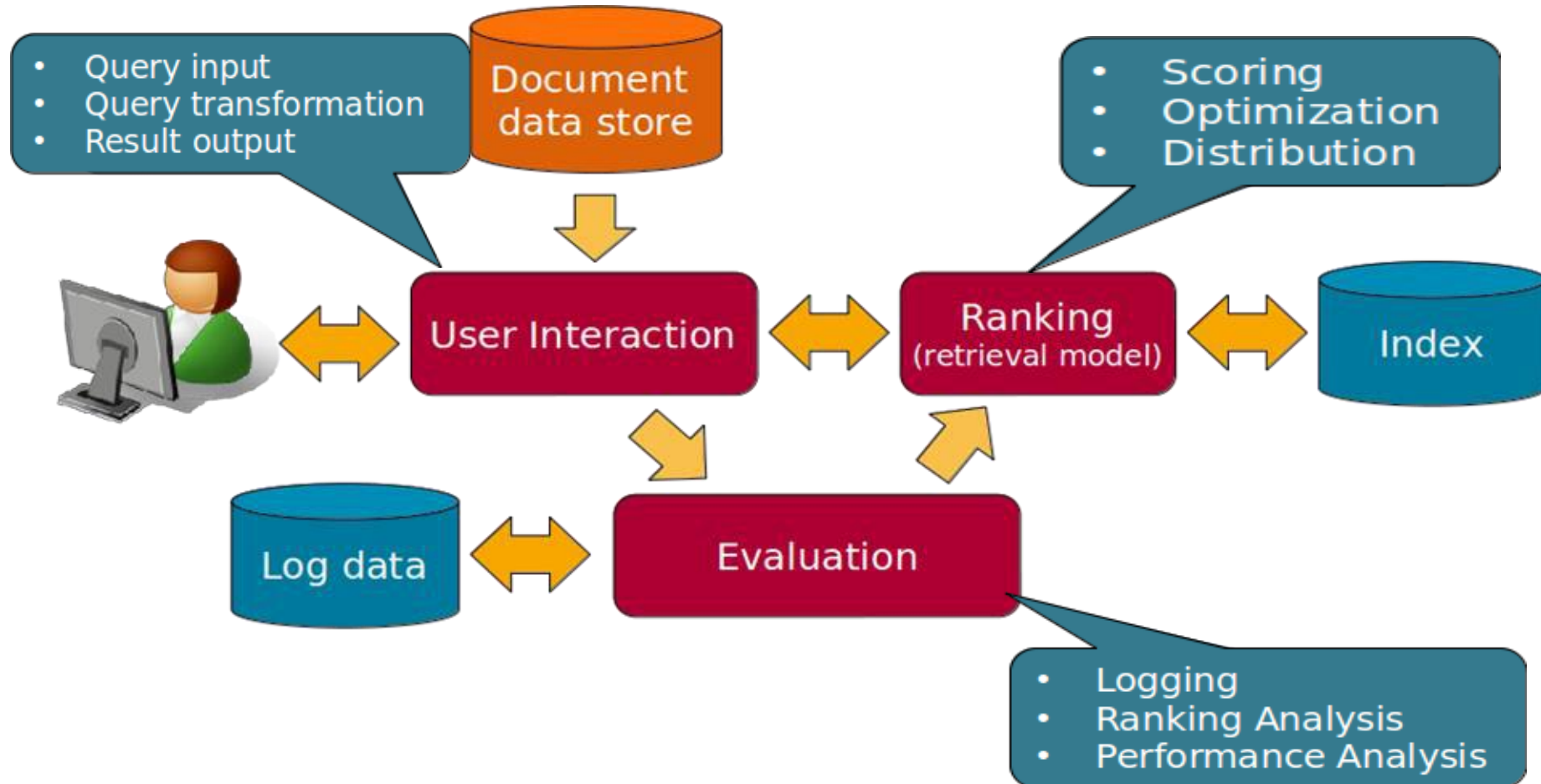


Ranking – Distribution

- Processing queries in a distributed environment
- Query broker distributes queries and assembles results



The Query Process

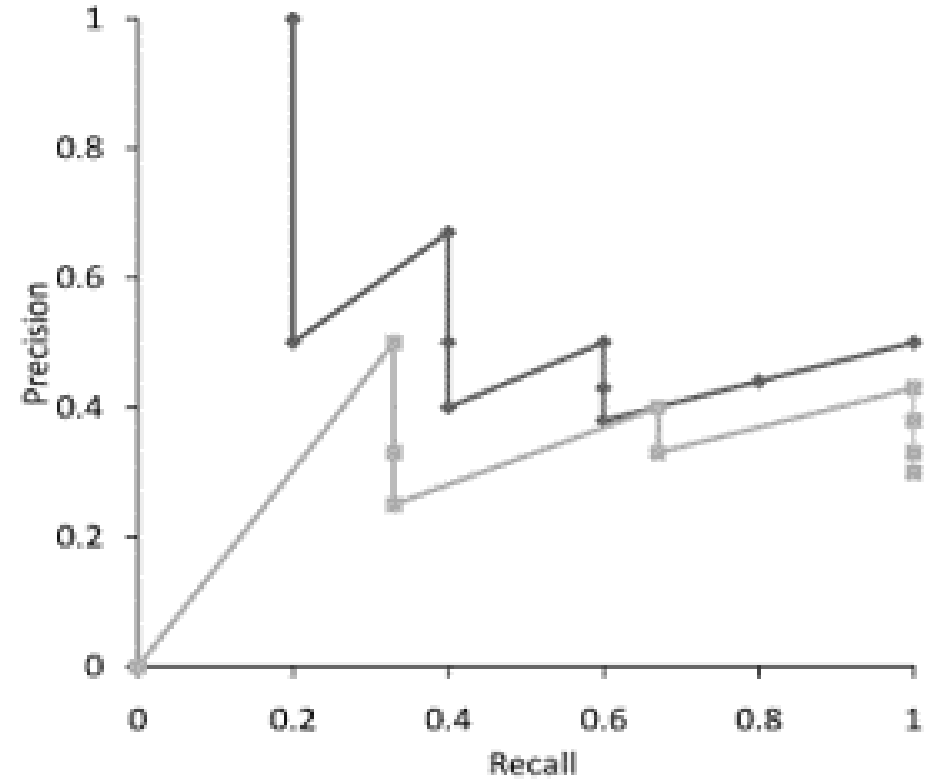


Evaluation – Logging

- Logging user queries and interaction is crucial for improving search effectiveness and efficiency.
- Query logs and clickthrough data (& dwell time) used for
 - Query suggestion
 - Spell checking
 - Query caching
 - Ranking
 - Advertising search
 - ...
- Assumption: Pages clicked on are relevant to query.

Evaluation – Ranking Analysis

- Measuring and tuning ranking effectiveness
- Variety of measures



Evaluation – Performance Analysis

- Measuring and tuning system efficiency
- Response time, throughput
- Simulation



How Does It Really Work?

- This course explains the components of a search engine in more detail.
- Often many possible approaches and techniques for a given component
 - Focus is on the most important alternatives
 - Explain a small number of approaches in detail rather than many approaches
 - “Importance” based on research results and use in actual search engines
 - Alternatives described in references (see book)



Questions?