Information Retrieval

– An Introduction –

– The view of an open-minded –

– computer scientist –

www.studentsfocus.com
What is Information Retrieval?

• The process of actively seeking out information relevant to a topic of interest (van Rijsbergen)

  – Typically it refers to the automatic (rather than manual) retrieval of documents
  • Information Retrieval System (IRS)

  – “Document” is the generic term for an information holder (book, chapter, article, webpage, etc)

www.studentsfocus.com
IR in practice

• Information Retrieval is a research-driven theoretical and experimental discipline
  – The focus is on different aspects of the information–seeking process, depending on the researcher’s background or interest:
    • Computer scientist – fast and accurate search engine
    • Librarian – organization and indexing of information
    • Cognitive scientist – the process in the searcher’s mind
    • Philosopher – Is this really relevant?
    • ...
  – Progress influenced by advances in Computational Linguistics, Information Visualization, Cognitive Psychology, HCI, ...

• Experimental vs. operational systems
• Analogy to car manufacturing
Fundamental concepts in IR

• What is **information**?

• **Meaning vs. form**

• **Data vs. Information Retrieval**

• **Relevance**
Disclaimer

• Relevance and other key concepts in IR were discussed in the previous class, so we won’t do it again.
  – We’ll take a simple view: a document is relevant if it is about the searcher’s topic of interest

• We will discuss text documents, not other media
  – Most current tools that search for images, video, or other media rely on text annotations
  – Real content retrieval of other media (based on shape, color, texture, ...) are not mature yet

www.studentsfocus.com
The stages of IR

- Indexed and structured information
- Searching
- Browsing

Creation → Indexing, organizing → Indexed and structured information → Retrieval

www.studentsfocus.com
The formalized IR process

Real world

Collection of documents

Document representations

Matching

Results

Anomalous state of knowledge

Information need

Query

www.studentsfocus.com
What do we want from an IRS?

• Systemic approach
  – Goal (for a known information need):
    Return as many relevant documents as possible and as few non-relevant documents as possible

• Cognitive approach
  – Goal (in an interactive information-seeking environment, with a given IRS):
    Support the user’s exploration of the problem domain and the task completion.
The role of an IR system – a modern view –

• Support the user in
  – exploring a problem domain, understanding its terminology, concepts and structure
  – clarifying, refining and formulating an information need
  – finding documents that match the info need description
    • As many relevant docs as possible
    • As few non-relevant documents as possible

www.studentsfocus.com
How does it do this?

• User interfaces and visualization tools for
  – exploring a collection of documents
  – exploring search results

• Query expansion based on
  – Thesauri
  – Lexical/statistic analysis of text / context and concept formation
  – Relevance feedback

• Indexing and matching model

www.studentsfocus.com
How well does it do this?

• Evaluation
  – Of the components
    • Indexing / matching algorithms
  – Of the exploratory process overall
    • Usability issues
    • Usefulness to task
    • User satisfaction
Role of the user interface in IR

**INPUT**
- Problem definition
- Source selection
- Problem articulation

**OUTPUT**
- Examination of results
- Extraction of information
- Integration with overall task

www.studentsfocus.com
Information Visualization tools for exploration

• Rely on some form of information organization

• Principle:
  – Overview first
  – Zoom
  – Details on demand

• Usability issues
  – Direct manipulation
  – Dynamic, implicit queries
Information Visualization tools

- Repositories
  - University of Maryland HCIL
  - InfoViz repository
    - http://fabdo.fh-potsdam.de/infoviz/repository.html

- Hyperbolic trees
- Themescapes
- Workscapes
- Fisheye view

www.studentsfocus.com
Faceted organization

• Each document is described by a set of attribute (or facet) values

• Example:
  – FilmFinder, HomeFinder
  – Film
    • Attributes (facets): Title, Year, Popularity, Director, Actors

• In design terms, it refers to composition.

www.studentsfocus.com
Hierarchic organization

Role of structure:
• support for exploration (browsing / searching)
• support for term disambiguation
• potential for efficient retrieval

In design terms it refers to inheritance.

www.studentsfocus.com
Structuring a document collection

• Manual, by experts - slow, expensive, infeasible for large corpora

• Supervised categorization
  • Classes or hierarchic structure established by human experts
  • Documents automatically allocated to classes

• Unsupervised classification = clustering
  • Similar documents grouped together, and a structure is expected to emerge
  • Result influenced by the homogeneity/heterogeneity of the documents, by the indexing and clustering methods and parameters

www.studentsfocus.com
Document Clustering

- Finds overall similarities among groups of documents
- Finds overall similarities among groups of documents
- Picks out some themes, ignores others
The Cluster Hypothesis

• “Similar documents tend to be relevant to the same requests”

• Issues:
  – Variants: “Documents that are relevant to the same topics are similar”
  – Simple vs. complex topics
  – Evaluation, prediction

• The cluster hypothesis is the main motivation behind document clustering
Document-document similarity

• Document representative
  – Select features to characterize document: terms, phrases, citations
  – Select weighting scheme for these features:
    • Binary, raw/relative frequency, divergence measure
    • Title / body / abstract, controlled vocabulary, selected topics, taxonomy

• Similarity / association coefficient or dissimilarity / distance metric
Similarity coefficients

- Simple matching
  \[ |X \cap Y| \]

- Dice’s coefficient
  \[ \frac{2 \cdot |X \cap Y|}{|X| + |Y|} \]

- Cosine coefficient
  \[ \frac{\sum x_i y_i}{\sqrt{\sum_i x_i^2 \cdot \sum_i y_i^2}} \]

www.studentsfocus.com
Clustering methods

- **Non-hierarchic methods**
  -> partitions
  – High efficiency, low effectiveness

- **Hierarchic methods**
  -> hierarchic structures - small clusters of highly similar documents nested within larger clusters of less similar documents
  – Divisive => monothetic classifications
  – Agglomerative => polythetic classifications !!
Partitioning method

• Generic procedure:
  – The first object becomes the first cluster
  – Each subsequent object is matched against existing clusters
    • It is assigned to the most similar cluster if the similarity measure is above a set threshold
    • Otherwise it forms a new cluster
  – Re-shuffling of documents into clusters can be done iteratively to increase cluster similarity
HACM’s

• Generic procedure:
  – Each doc to be clustered is a singleton cluster
  – While there is more than one cluster, the clusters with maximum similarity are merged and the similarities recomputed

• A method is defined by the similarity measure between non-singleton clusters

• Algorithms for each method differ in:
  – Space (store similarity matrix ? all of it ?)
  – Time (use all similarities ? use inverted files ?)
Representation of clustered hierarchies

Dendrogram

Tree

Graph

www.studentsfocus.com
Scatter/Gather

• How it works
  – Cluster sets of documents into general “themes”, like a table of contents
  – Display the contents of the clusters by showing topical terms and typical titles
  – User chooses subsets of the clusters and re-clusters the documents within
  – Resulting new groups have different “themes”

• Originally used to give collection overview

• Evidence suggests more appropriate for displaying retrieval results in context
Multi-Dimensional Metaphor for the Document Space
Kohonen Feature Maps on Text
Search strategies

• Analytical strategy (mostly querying)
  – Analyze the attributes of the information need and of the problem domain (mental model)

• Browsing
  – Follow leads by association (not much planning)

• Known site strategy
  – Based on previous searches
  – Indexes or starting points for browsing

• Similarity strategy
  – “more like this”
Non-search activities

• Reading and interpreting

• Annotating or summarizing

• Analysis
  – Finding trends
  – Making comparisons
  – Aggregating information
  – Identifying a critical subset
IRS design trade-offs (high-level)

• General
  – Easy to learn (“walk up and use”)
    • Intuitive
    • Standardized look-and-feel and functionality
  – Simple and easy to use
  – Deterministic and restrictive

• Specialized
  – Complex, require training (course, tutorial)
  – Increased functionality
  – Customizable, non-deterministic
Query specification

• Boolean vs. free text

• Structure analysis vs. bag of words

• Phrases / proximity

• Faceted / weighted queries (TileBars, FilmFinder)

• Graphical support (Venn diagrams, filters)

• Support for query formulation (aid-word list, thesauri, spell-checking)
Query Specification

• Interaction Styles
  – Command Language
  – Form Fillin
  – Menu Selection
  – Direct Manipulation
  – Natural Language

• Example:
  – How do each apply to Boolean Queries

www.studentsfocus.com
Form-Based Query Specification (Altavista)
Form-based Query Specification (Infoseek)
Menu-based Query Specification
(Young & Shneiderman 93)
Putting Results in Context

• Interfaces should
  – give hints about the roles terms play in the collection
  – give hints about what will happen if various terms are combined
  – show explicitly why documents are retrieved in response to the query
  – summarize compactly the subset of interest
KWIC (Keyword in Context)

• An old standard, ignored by internet search engines
  – used in some intranet engines, e.g., Cha-Cha
The formalized IR process

1. Real world
   - Collection of documents
     - Document representations

2. Anomalous state of knowledge
   - Information need
     - Query

3. Matching
   - Results
Indexing

• Association of descriptors (keywords, concepts, metadata) to documents in view of future retrieval

• The knowledge / expectation / behavior of the searcher needs to be anticipated
Manual and automatic indexing

• Manual
  – Human indexers assign index terms to documents
  – A computer system may be used to record the descriptors generated by the human

• Automatic
  – The system extracts “typical”/ “significant” terms
  – The human may contribute by setting the parameters or thresholds, or by choosing components or algorithms

• Semi-automatic
  – The system’s contribution may be support in terms of word lists, thesauri, reference system, etc, following or not the automatic processing of the text
Manual vs. automatic indexing

• Manual
  – Slow and expensive
  – Is based on intellectual judgment and semantic interpretation (concepts, themes)
  – Low consistency

• Automatic
  – Fast and inexpensive
  – Mechanical execution of algorithms, with no intelligent interpretation (aboutness / relevance)
  – Consistent
Vocabulary

• **Vocabulary (indexing language)**
  – The set of concepts (terms or phrases) that can be used to index documents in a collection

• **Controlled**
  – Specific for specialized domains
  – Potential for increased consistency of indexing and precision of retrieval

• **Un-controlled (free)**
  – Potentially all the terms in the documents
  – Potential for increased recall
Thesauri

• Capture relationships between indexing terms
  – Hierarchical
  – Synonymous
  – Related

• Creation of thesauri
  – Manual vs. automatic

• Use of thesauri
  – In manual / semi-automatic / automatic fashion
  – Syntagmatic co-ordination / thesaurus-based query expansion during indexing / searching

www.studentsfocus.com
Query indexing

• Search systems
  – Automatic indexing
  – Synchronization with indexing of documents (vocabulary, algorithms, etc)

• Interactive / browsing systems
  – Support tools (word list, thesauri)
  – Query not necessarily explicit

www.studentsfocus.com
Automatic indexing

• Rationalist approach
  – Natural Language processing / Artificial Intelligence
  – Attempts to define and use grammatical, knowledge and reasoning rules (Chomsky)
  – More computationally intensive

• Empiricist approach
  – Statistical Language Processing
  – Estimate probabilities of linguistic events: words, phrases, sentences (Shannon)
  – Inexpensive, but just as good
Automatic indexing

• There is no “best solution”

• An “engineering” approach is taken: creatively combine theoretical models and techniques, test, make adjustments until the results are satisfying

• Balance between effort/sophistication of method and quality of results needed

• Results depend on the specific document collection and on the type of application
Steps of automatic indexing

1. Collection/document structure
2. Lexical analysis
3. Stopword removal
4. Stemming
5. Data structure

representation
Term significance

Word occurrence frequency is a measure for the significance of terms and their discriminatory power (see Brown corpus).

- too frequent: useless discriminators
- significant terms
- too rare: no significant contribution to the content of the document

www.studentsfocus.com
Weighting

• Heuristics
  – Based on common sense, but adjusted/engineered following experiments. Ex:

  – Terms that occur in only a few documents are often more valuable than ones that occur in many - IDF
  – The more often a term occurs in a document, the more likely it is to be important for that document - TF
  – A term that occurs the same number of times in short document and in a long document is likely to be more valuable for the former - DL

www.studentsfocus.com
Document weighting

• Theoretical models
  – Provide theoretical justification of the formulae
  – Take advantage of mathematical theory
  – Are typically adjusted by heuristics

• Probabilistic
  – Rank documents based on the estimated probability that they are relevant to the query (derived from term counts)

• Language models
  – Rank documents based on the estimated probability that the query is a random sample of document words

www.studentsfocus.com
Ranked retrieval

• The documents are ranked based on their score

• Advantages
  – Query easy to specify
  – The output is ranked based on the estimated relevance of the documents to the query
  – A wide variety of theoretical models exist

• Disadvantages
  – Query less precise (although weighting can be used)
Boolean retrieval

• Documents are retrieved based on their containing or not query terms

• Advantages
  – Very precise queries can be specified
  – Very easy to implement (in the simple form)

• Disadvantages
  – Specifying the query may be difficult for casual users
  – Lack of control over the size of the retrieved set
IR Evaluation

• Why evaluate?
  – “Quality”

• What to evaluate?
  – Qualitative vs. quantitative measures

• How to evaluate?
  – Experimental design; result analysis

• Complex and controversial topic

www.studentsfocus.com
Actors involved

• Funders
  – Cost to implement, estimated savings
  – User satisfaction, public recognition

• Librarian, library scientist
  – Functionality
  – Support for search strategies
  – User satisfaction

www.studentsfocus.com
Actors involved

• Information scientist, mathematician
  – Underlying mathematical model for representing information
  – Weighting scheme, document-query matching
  – System effectiveness

• Computer scientist, software developer
  – System efficiency (speed, resources needed)
  – Flexibility, extensibility
Need to evaluate

• Technology hype or real need?
  – Landauer, T. – “The trouble with computers”

• Does it justify its cost?
  – Pros: review and improvement of procedures and workflow; increased efficiency; increased control and safety
  – Cons: actual cost of system; work interruption; need for re-training

– Quality – can it be improved?
History

• Systemic approach
  – User outside the system
  – Static/fixed information need
  – Retrieval effectiveness measured
  – Batch retrieval simulations

• User-centered approach
  – User part of the system, interacting with other components, trying to resolve an anomalous state of knowledge
  – Task-oriented evaluation

www.studentsfocus.com
Aspects to evaluate

INPUT

Problem definition
Source selection
Problem articulation

OUTPUT

Examination of results
Extraction of information
Integration with overall task

www.studentsfocus.com
Experimental design decisions

- Whole vs. parts
- Black box vs. diagnostic systems
- Operational vs. experimental system
One possible approach

• *IR*-specific evaluation
  – Systemic
    • Quality of search engine
    • Influence of various modelling decisions (stopword removal, stemming, indexing, weighting scheme, …)
  – Interaction
    • Support for query formulation
    • Support for exploration of search output

• Non-specific evaluation
  – Task-oriented evaluation
    • Usefulness, usability
    • Task completion, user satisfaction
Laboratory vs. operational settings

• Laboratory
  – Typically only one or several components of the system are evaluated
  – Assumptions are made about the other components
  – User behavior is typically simulated (software)
  – Control over experimental variables, repeatability, observability

• Operational
  – More or less “real” users
  – Real of inferred information needs
  – Realism
The traditional (lab) IR experiment

• To start with you need:
  – An IR system (or two)
  – A collection of documents
  – A collection of requests
  – Relevance judgements

• Then you run your experiment:
  – Input the documents
  – Put each request to the system
  – Collect the output
Retrieval effectiveness

All docs

Relevant

Retrieved

www.studentsfocus.com
Precision vs. Recall

\[
\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|}
\]

\[
\text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Rel in Collection}|}
\]

All docs

www.studentsfocus.com
Interactive system’s evaluation

• Definition:
Evaluation = the process of **systematically collecting data** that informs us about what it is like for a **particular user or group of users to use a product/system for a particular task in a certain type of environment.**
Problems

• Attitudes:
  – Designers assume that if they and their colleagues can use the system and find it attractive, others will too
    • Features vs. usability or security
  – Executives want the product on the market yesterday
    • Problems “can” be addressed in versions 1.x
  – Consumers accept low levels of usability
    • “I’m so silly”
Two main types of evaluation

- **Formative evaluation** is done at different stages of development to check that the product meets users’ needs.
  - Part of the user-centered design approach
  - Supports design decisions at various stages
  - May test parts of the system or alternative designs

- **Summative evaluation** assesses the quality of a finished product.
  - May test the usability or the output quality
  - May compare competing systems
What to evaluate

Iterative design & evaluation is a continuous process that examines:

• Early ideas for conceptual model
• Early prototypes of the new system
• Later, more complete prototypes

Designers need to check that they understand users’ requirements and that the design assumptions hold.
Four evaluation paradigms

• ‘quick and dirty’

• usability testing

• field studies

• predictive evaluation
Quick and dirty

• ‘quick & dirty’ evaluation describes the common practice in which designers informally get feedback from users or consultants to confirm that their ideas are in-line with users’ needs and are liked.

• Quick & dirty evaluations are done any time.

• The emphasis is on fast input to the design process rather than carefully documented findings.

www.studentsfocus.com
Usability testing

- Usability testing involves recording typical users’ performance on typical tasks in controlled settings. Field observations may also be used.
- As the users perform these tasks they are watched & recorded on video & their key presses are logged.
- This data is used to calculate performance times, identify errors & help explain why the users did what they did.
- User satisfaction questionnaires & interviews are used to elicit users’ opinions.
Usability testing

• It is very time consuming to conduct and analyze
  – Explain the system, do some training
  – Explain the task, do a mock task
  – Questionnaires before and after the test & after each task
  – Pilot test is usually needed
• Insufficient number of subjects for ‘proper’ statistical analysis
• In laboratory conditions, subjects do not behave exactly like in a normal environment

www.studentsfocus.com
Field studies

- Field studies are done in natural settings
- The aim is to understand what users do naturally and how technology impacts them.
- In product design field studies can be used to:
  - identify opportunities for new technology
  - determine design requirements
  - decide how best to introduce new technology
  - evaluate technology in use

www.studentsfocus.com
Predictive evaluation

• Experts apply their knowledge of typical users, often guided by heuristics, to predict usability problems.
• Another approach involves theoretically based models.
• A key feature of predictive evaluation is that users need not be present.
• Relatively quick & inexpensive.

www.studentsfocus.com
Overview of techniques

• Observing users
  • Don’t interfere with the subjects!

• Asking users’ opinions
  • Interviews, questionnaires

• Asking experts’ opinions
  • Heuristics, role-playing; suggestions for solutions
Overview of techniques

• Testing users’ performance
  • Time taken to complete a task, errors made, navigation path
  • Satisfaction

• Modeling users’ task performance
  • Appropriate for systems with limited functionality
  • Make assumptions about the user’s typical, optimal, or poor behaviour
  • Simulate the user and measure performance

www.studentsfocus.com
Web Information Retrieval

Challenges
Approaches

www.studentsfocus.com
Challenges

• Scale, distribution of documents

• Controversy over the unit of indexing
  – What is a document? (hypertext)
  – What does the use expect to be retrieved?

• High heterogeneity
  – Document structure, size, quality, level of abstraction / specialization
  – User search or domain expertise, expectations

• Retrieval strategies
  – What do people want?

• Evaluation
Web documents / data

• No traditional collection
  – Huge
    • Time and space to crawl index
    • IRSs cannot store copies of documents
  – Dynamic, volatile, anarchic, un-controlled
  – Homogeneous sub-collections

• Structure
  – In documents (un-/semi-/fully-structured)
  – Between docs: network of inter-connected nodes
  – Hyper-links - conceptual vs. physical documents

www.studentsfocus.com
Web documents / data

• Mark-up
  – HTML – look & feel
  – XML – structure, semantics
  – Dublin Core Metadata
  – Can webpage authors be trusted to correctly mark-up / index their pages?

• Multi-lingual documents

• Multi-media

www.studentsfocus.com
Theoretical models for indexing / searching

• Content-based weighting
  – As in traditional IRS, but trying to incorporate
    • hyperlinks
    • the dynamic nature of the Web (page validity, page caching)

• Link-based weighting
  – Quality of webpages
    • Hubs & authorities
    • Bookmarked pages
    • Iterative estimation of quality

www.studentsfocus.com
Architecture

• Centralized
  – Main server contains the index, built by an indexer, searched by a query engine
    • Advantage: control, easy update
    • Disadvantage: system requirements (memory, disk, safety/recovery)

• Distributed
  – Brokers & gatherers
    • Advantage: flexibility, load balancing, redundancy
    • Disadvantage: software complexity, update

www.studentsfocus.com
User variability

• Power and flexibility for expert users vs. intuitiveness and ease of use for novice users

• Multi-modal user interface
  – Distinguish between experts and beginners, offer distinct interfaces (functionality)
  – Advantage: can make assumptions on users
  – Disadvantage: habit formation, cognitive shift

• Uni-modal interface
  – Make essential functionality obvious
  – Make advanced functionality accessible

www.studentsfocus.com
Search strategies

• Web directories
• Query-based searching
• Link-based browsing (provided by the browser, not the IRS)
• “More like this”
• Known site (bookmarking)

• A combination of the above

www.studentsfocus.com
Web IRS evaluation

• Effectiveness - problems
  – Search for documents vs. information
  – What is the target collection (the crawled and indexed Web) today?
  – Recall, relative recall, aspectual recall
  – Levels of relevance, quality, hubs & authorities
  – User-centered, task-oriented evaluation
    • Task completion, user satisfaction

• Usability
  – Is there anything specific for Web IRSs?

www.studentsfocus.com
More advanced topics of IR research
Support for Relevance Feedback

- RF can improve search effectiveness ... but is rarely used
- Voluntary vs. forced feedback
- At document vs. word level
- “Magic” vs. control
Term clustering

• Based on `similarity’ between terms
  – Collocation in documents, paragraphs, sentences

• Based on document clustering
  – Terms specific for bottom-level document clusters are assumed to represent a topic

• Use
  – Thesauri
  – Query expansion

[www.studentsfocus.com]
User modelling

• Build a model / profile of the user by recording
  – the `context’
  – topics of interest
  – preferences

based on interpreting (his/her actions):
  – Implicit or explicit relevance feedback
  – Recommendations from `peers’
  – Customization of the environment
Personalised systems

• Information filtering
  – Ex: in a TV guide only show programs of interest

• Use user model to disambiguate queries
  – Query expansion
  – Update the model continuously

• Customize the functionality and the look-and-feel of the system
  – Ex: skins; remember the levels of the user interface
Autonomous agents

• Purpose: find relevant information on behalf of the user
• Input: the user profile
• Output: pull vs. push
• Positive aspects:
  – Can work in the background, implicitly
  – Can update the master with new, relevant info
• Negative aspects: control

• Integration with collaborative systems
Questions ?

www.studentsfocus.com
Information Hierarchy

• Data
  – The raw material of information

• Information
  – Data organized or presented in some context

• Knowledge
  – Information read, heard or seen and understood

• Wisdom
  – Distilled and integrated knowledge and understanding

www.studentsfocus.com
Meaning vs. Form

• Meaning
  – Indicates what the document is about, or the topic of the document
  – Requires intelligent interpretation by a human or artificial intelligence techniques

• Form
  – Refers to the the content *per se*, i.e. the words that make up the document

www.studentsfocus.com
# Data vs. Information Retrieval

<table>
<thead>
<tr>
<th>Matching</th>
<th>Exact match</th>
<th>Partial match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Deterministic</td>
<td>Probabilistic</td>
</tr>
<tr>
<td>Classification</td>
<td>Monothetic</td>
<td>Polythetic</td>
</tr>
<tr>
<td>Query specification</td>
<td>Complete</td>
<td>Incomplete</td>
</tr>
<tr>
<td>Error response</td>
<td>Sensitive</td>
<td>Insensitive</td>
</tr>
</tbody>
</table>

(van Rijsbergen, C.J. (1979) http://www.dcs.gla.ac.uk/Keith/Preface.html)
Relevance

- Depends on the individual and on the context
- Relevance vs. aboutness (for a topic)
- Relevance vs. usefulness (for a task)
- Relevance judgements in test collections
  - Allow for system evaluation

www.studentsfocus.com
Collection/document structure

• Examples
  – Cacm
  – Reuters
  – Email
  – Web

• Issues
  – Identify indexing units / documents
    • Mark-up, parsers
  – Structured documents - what to index ?
  – Weighting scheme

www.studentsfocus.com
Lexical analysis

• Break up the text in words or “tokens”
• Question: “what is a word” ?

• Problem cases
  – Numbers: “M16”, “2001”
  – Hyphenation: “MS-DOS”, “OS/2”
  – Punctuation: “John’s”, “command.com”
  – Case: “us”, “US”
  – Phrases: “venetian blind”
Stopwords

- Very frequent words, with no power of discrimination
- Typically function words, not indicative of content
- The stopwords set depends on the document collection and on the application

www.studentsfocus.com
Stemming

• Identify morphological variants, creating “classes”
  – system, systems
  – forget, forgetting, forgetful
  – analyse, analysis, analytical, analysing

• Use in an IR system
  – Replace each term by the class representative (root or most common variant)
  – Replace each word by all the variants in its class
Stemming errors

• **Too aggressive**
  – organization / organ
  – police / policy
  – arm / army
  – execute / executive

• **Too timid**
  – european / europe
  – create / creation
  – search / searcher
  – cylinder / cylindrical

www.studentsfocus.com
Inverted files

B-tree

search-index

dictionary

postings lists

www.studentsfocus.com
Inverted files

KW_INV

<table>
<thead>
<tr>
<th>token</th>
<th>totdoc</th>
<th>head</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;aardvarck&quot;</td>
<td>20</td>
<td>65</td>
</tr>
</tbody>
</table>

POSTING

<table>
<thead>
<tr>
<th>docno</th>
<th>freq</th>
<th>next</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>