Information Processing Technology of Internet of Things

Chapter 3
Information Retrieval

Wu Liu

Beijing Key Lab of Intelligent Telecomm. Software and Multimedia
Beijing University of Posts and Telecommunications

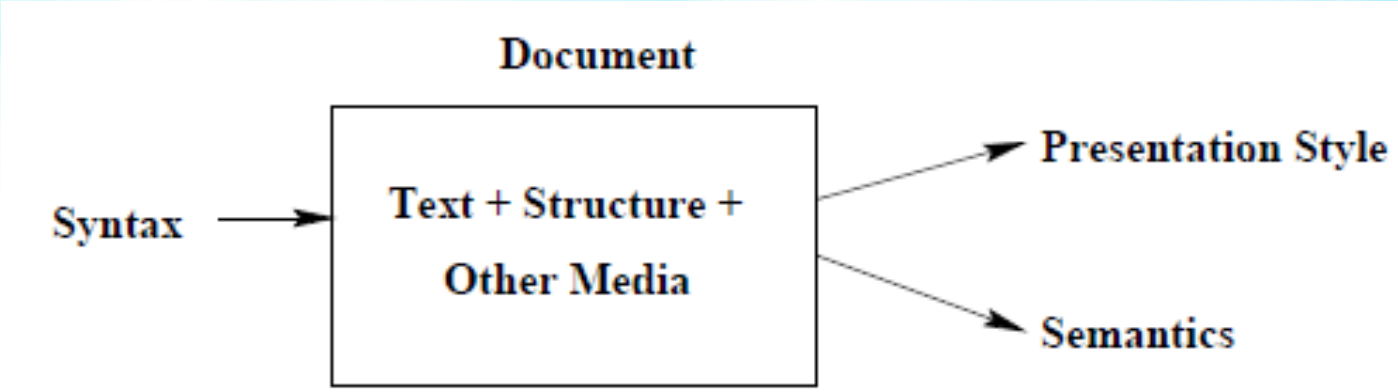
3.3 Documents

3.3.1 Introduction

Introduction

- The **document**
  - denotes a single unit of information
  - has a **syntax and structure**
  - has a **semantics, specified by the author**
  - may have a **presentation style**
    - given by its syntax and structure
    - related to a specific application
    - specifies how to display or print document

Introduction



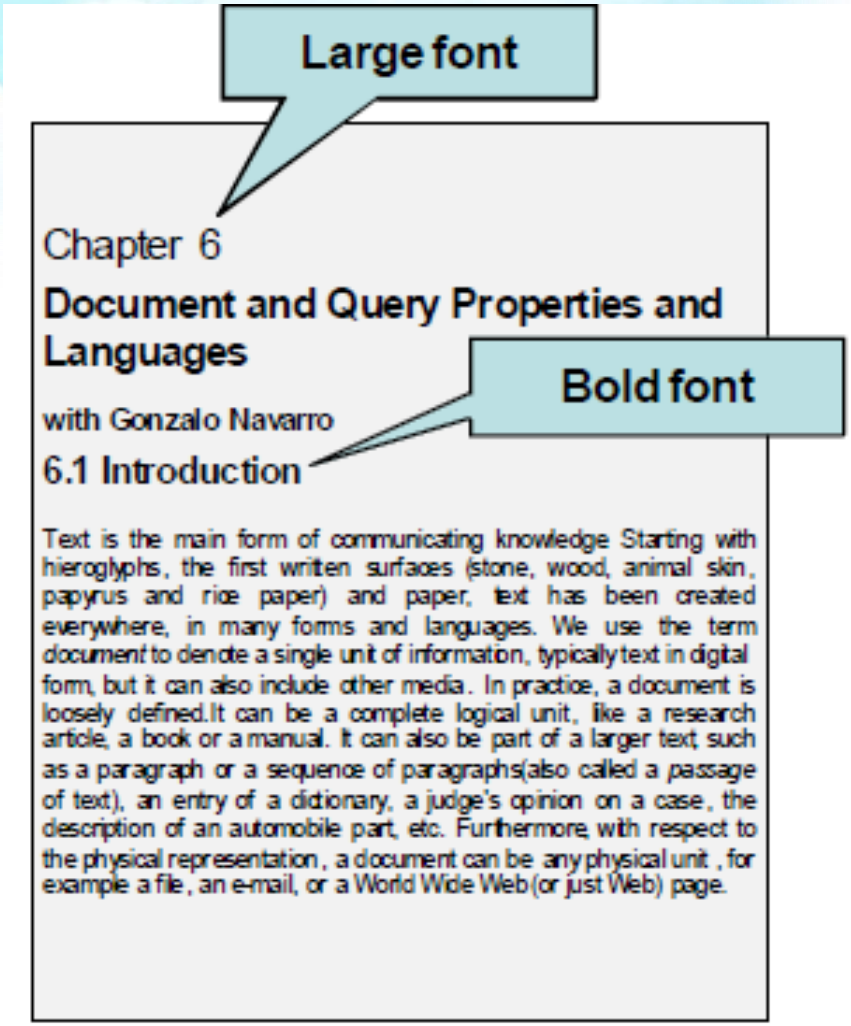
- The **document syntax**
  - expresses structure, presentation style, semantics
  - one or more of elements might be implicit or given together
  - structural element (e.g., a section) can have fixed formatting style

Introduction

- The **document syntax** can be
  - **implicit** in its content
  - expressed in a **simple declarative language**
  - expressed in a **programming language**
    - the language syntax might be proprietary and specific
    - open and generic languages are more flexible
- Text can also be written in **natural language**
  - hard to process using a computer
- Current trend: use document languages that
  - provide information on structure, format, and semantics
  - are readable by humans and computers

Introduction

- **Document style**
- defines how a document is visualized or printed
- can be embedded in the document: TeX and RTF
- can be complemented by macros: LaTeX



Introduction

- **Queries in search engines**
  - can be considered as short pieces of text
  - differ from normal text
  - understanding them is very important
  - semantics often ambiguous due to polysemy
  - not simple to infer user intent behind a query

3.3.2 Metadata

Metadata

- **Metadata** is information on the organization of the data, the various data domains, and their relationship
  - metadata is **data about the data**
  - in a database, names of relations and attributes constitute metadata
  - metadata is associated with most documents and text collections

Descriptive Metadata

- Common forms of metadata for documents
  - author of the text
  - date of publication
  - source of the publication
  - document length
- Dublin Core Metadata Element Set proposes 15 fields to describe a document
- this type of information is **Descriptive Metadata**
  - Descriptive metadata are external to the meaning of the document and pertain more to how it was created

Semantic Metadata

- **Semantic Metadata**
  - characterizes the subject matter within the document contents
  - is associated with a wide number of documents
  - its availability is increasing
- An important metadata format is **MARC (Machine Readable Cataloging Record)**
  - most used format for library records
  - includes fields for distinct attributes of a bibliographic entry such as title, author, publication venue

Metadata in Web Documents

- The increase in Web data has led to many initiatives to add metadata information to Web pages for various purposes such as
  - cataloging and content rating
  - intellectual property rights and digital signatures
  - applications to electronic commerce
- **RDF (Resource Description Framework)**
  - new standard for Web metadata
  - allows describing Web resources to facilitate automated processing

Metadata in Web Documents

- RDF does not assume any particular application or semantic domain
- It consists of a description of nodes and attached **attribute/value pairs**
  - Nodes can be any Web resource, that is, any **Uniform Resource Identifier (URI) including Uniform Resource Locators (URLs)**
  - Attributes are properties of nodes and their values are text strings or other nodes (Web resources or metadata instances)

3.3.3 Document Formats

Text

- An IR system should be able to retrieve information from many text formats (doc, pdf, html, txt)
- Other text formats
  - Rich Text Format (RTF): for document interchange
  - Portable Document Format (PDF): for printing and displaying
  - Postscript: for printing and displaying
- Other interchange formats are used to encode electronic mail
  - Multipurpose Internet Mail Exchange (MIME): for encoding email
  - Compress (Unix), ARJ (PCs): for compressing text
  - ZIP (Unix) (gzip in Unix and Winzip in Windows): for compressing text

Image Formats

- The simplest image formats are direct representations of a bit-mapped display such as XBM, BMP or PCX
- Images of these formats have a lot of redundancy and can be compressed efficiently
  - Example of format that incorporates compression: **Compuserve's Graphic Interchange Format (GIF)**
- To improve compression ratios, lossy compression was developed
  - uncompressing a compressed image does not yield exactly the original image
- This is done by the **Joint Photographic Experts Group** (**JPEG**) format
  - JPEG tries to eliminate parts of the image that have less impact in the human eye
  - This format is parametric, in the sense that the loss can be tuned

Audio

- Audio must be digitalized to be stored properly
- Most common formats for audio: **AU, MIDI** and **WAVE**
  - MIDI: standard format to interchange music between electronic instruments and computers
- For audio libraries other formats are used such as **RealAudio or CD formats**

Movies

- Main format for animations is **Moving Pictures Expert Group (MPEG):**
  - works by coding the changes in consecutive frames
  - profits from the temporal image redundancy that any video has
  - includes the audio signal associated with the video
  - specific cases for audio (MP3), video (MP4), etc.
- Other video formats are **AVI, FLI and QuickTime**
  - AVI may include compression
  - QuickTime, developed by Apple, also includes compression

3.3.4 Markup Languages

Markup Languages

- Markup is defined as extra syntax used to describe **formatting actions, structure information, text semantics, attributes**
- Examples of Markup Languages
  - SGML: Standard Generalized Markup Language
  - XML: eXtensible Markup Language
  - HTML: Hyper Text Markup Language

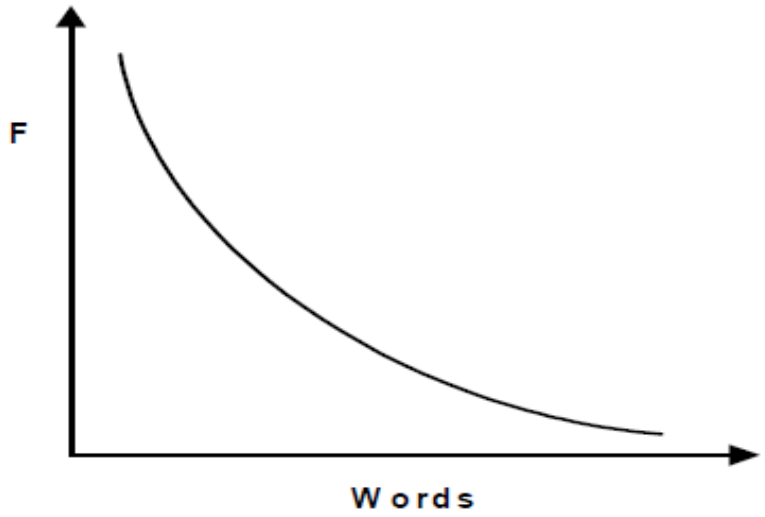
3.3.5 Text Properties

Modeling Natural Language

- We can divide the symbols of a text in two disjoint subsets:
  - symbols that separate words; and
  - symbols that belong to words
- It is well known that symbols are not uniformly distributed in a text
  - For instance, in English, the vowels are usually more frequent than most consonants

Modeling Natural Language

- **How the different words are distributed** inside each document
- An approximate model is the **Zipf's Law**
- Figure below illustrates the distribution of frequencies of the terms in a text
  - words arranged in decreasing order of their frequencies

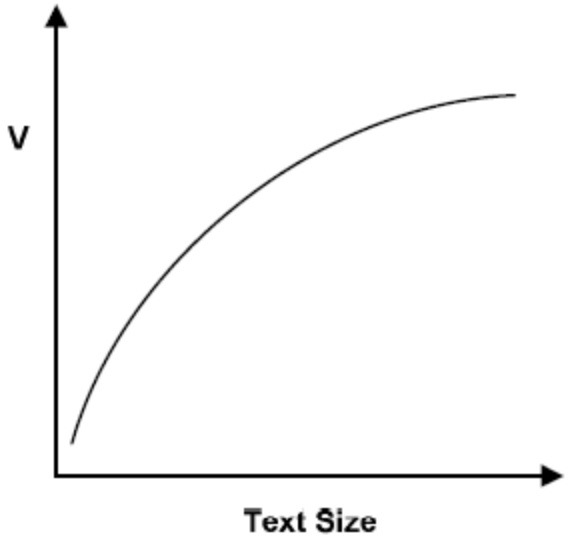


Modeling Natural Language

- Since the distribution of words is very skewed, words that are too frequent, called stopwords, can be disregarded
- A **stopword** is a word which does not carry meaning in natural language
  - Examples of stopwords in english: *a*, *the*, *by*, *and*
  - Fortunately, the most frequent words are stopwords
    - Therefore, half of the words appearing in a text do not need to be considered

Modeling Natural Language

- A issue is the **distribution of words** in the documents of a collection
- the **number of distinct words in a** document (the **document vocabulary)**
  - The figure below illustrates that vocabulary size grows sub-linearly with text size



Text Similarity

- Similarity is measured by a **distance function**
  - for strings of the same length, distance between them is the number of positions with different characters
  - for instance, the distance is 0 if they are equal
  - this is called the **Hamming distance**
- A distance function should also be **symmetric**
  - In this case, the order of the arguments does not matter
- A distance function should also satisfy the **triangle inequality**:
  - distance(a, c) ≤ distance(a, b) + distance(b, c)

Text Similarity

- **Edit (or Levenshtein) distance**
  - important distance function over strings
  - it is the minimal number of char insertions, deletions, and substitutions needed to make two strings equal
  - edit distance between color and colour is 1
  - edit distance between survey and surgery is 2
- **Longest common subsequence (LCS)**
  - all non-common characters of two (or more) strings are deleted
  - remaining sequence of characters is the LCS of both strings
  - LCS of survey and surgery is surey

Text Similarity

- Similarity can be extended to documents
- Consider lines as single symbols and compute the longest common sequence of lines between two files
  - Measure used by the diff command in Unix
  - Problems with this approach
    - very time consuming
    - does not consider lines that are similar

Text Similarity

- **Resemblance measure**
  - If W(dj) is the set of all distinct words in document dj , then the resemblance function between two documents di and dj is defined as

$$R(d_i, d_j) = \frac{|W(d_i) \cap W(d_j)|}{|W(d_i) \cup W(d_j)|}$$

    where $0 \leq R(di, dj) \leq 1$
  - Notice that this is a more efficient document similarity measure
- This resemblance measure can be easily transformed in a distance function D(di, dj)
  - $D(di, dj) = 1 - R(di, dj)$

3.3.5 Document Preprocessing

Document Preprocessing

- Document preprocessing can be divided into five text operations:
  - Lexical analysis of the text
  - Elimination of stopwords
  - Stemming of the remaining words
  - Selection of index terms or keywords
  - Construction of term categorization structures (thesaurus)

Thesauri

- Motivation for building a thesaurus: a **controlled vocabulary** for indexing and searching
- Terms are the **indexing** components of a thesaurus
  - a term can be composed of a word, a group of words, or a phrase
  - it is normally a noun (most concrete part of speech)
  - it usually denotes a **concept**
  - can be expressed as a combination of an adjective with a noun: **polar bear**

On the Use of Thesauri in IR

- Query formation process
  - User forms a query
  - Query terms might be erroneous and improper
  - Solution: reformulate the original query
  - Usually, this implies expanding original query with related terms
  - Thus, it is natural to use a thesaurus for finding related terms

3.4 Queries

Query Languages

- **Different kind of queries normally** posed to text retrieval systems is in part dependent on the retrieval model the system adopts
  - That is, a full-text system will not answer the same kind of queries as those answered by a system based on keyword ranking
- Languages for information retrieval allow the answer to be ranked
- There are a number of techniques to enhance the usefulness of the queries
  - Some examples are the expansion of a word to the set of its synonyms or the use of a thesaurus
  - Some words which are very frequent and do not carry meaning (called *stopwords) may be removed*
  - We refer to words that can be used to match query terms as keywords

Keyword Based Querying

- A **query is the formulation of a user information need**
- Keyword based queries are popular, since they are intuitive, easy to express, and allow for fast ranking
- However, a query can also be a more complex combination of operations involving several words

Word Queries

- The most elementary query that can be formulated in a text retrieval system is the **word**
- Some models are also able to see the internal division of words into letters
  - In this case, the alphabet is split into **letters and separators**
  - A word is a sequence of letters surrounded by separators

Word Queries

- The result of word queries is the set of documents containing at least one of the words of the query
- Further, the resulting documents are ranked according to the degree of similarity with respect to the query
- To support ranking, two common statistics on word occurrences inside texts are commonly used
  - The first is called **term frequency and counts the number of** times a word appears inside a document
  - The second is called **inverse document frequency and counts** the number of documents in which a word appears

Context Queries

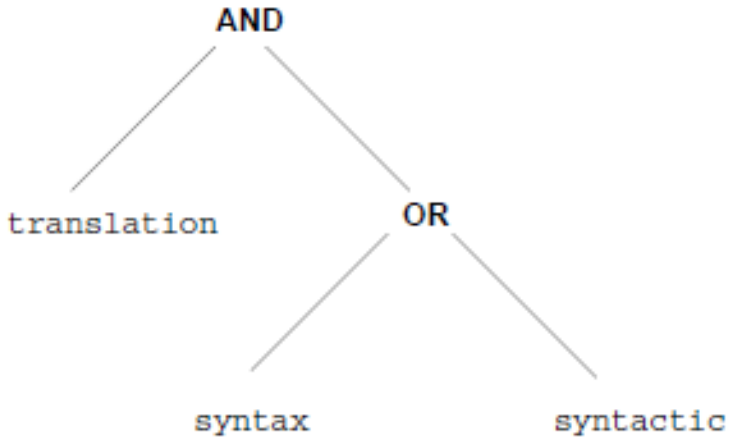
- Many systems complement queries with the ability to search words in a given **context**
- Words which appear near each other may signal higher likelihood of relevance than if they appear apart
- We may want to form phrases of words or find words which are proximal in the text
  - Phrase
    - Is a sequence of single-word queries
    - An occurrence of the phrase is a sequence of words
    - Can be ranked in a fashion somewhat analogous to single words
  - Proximity
    - Is a more relaxed version of the phrase query
    - A maximum allowed distance between single words or phrases is given
    - The ranking technique can be depend on physical proximity

Boolean Queries

- The oldest way to combine keyword queries is to use boolean operators
- A **boolean query has a syntax composed of**
  - **atoms**: basic queries that retrieve documents
  - **boolean operators**: work on their operands (which are sets of documents) and deliver sets of documents
- This scheme is in general **compositional:** operators can be composed over the results of other operators

Boolean Queries

- A **query syntax tree is naturally defined**
- Consider the example of a query syntax tree below



- It will retrieve all the documents which contain the word **translation** as well as either the word **syntax** or the word **syntactic**

Boolean Queries

- The operators most commonly used, given two basic queries or boolean sub-expressions e1 and e2, are:
  - e1 OR e2: the query selects all documents which satisfy e1 or e2
  - e1 AND e2: selects all documents which satisfy both e1 and e2
  - e1 BUT e2: selects all documents which satisfy e1 but not e2
  - NOT e2: the query selects all documents which not contain e2

Boolean Queries

- With classic boolean systems, no ranking of the retrieved documents is provided
  - A document either satisfies the boolean query or it does not
- This is quite a limitation because it does not allow for partial matching between a document and a user query
- To overcome this limitation, the condition for retrieval must be relaxed
  - For instance, a document which partially satisfies an AND condition might be retrieved
- The **NOT** operator is usually not used alone as the complement of a set of documents is the rest of the document collection

Boolean Queries

- A **fuzzy-boolean set of operators has been proposed**
- The idea is that the meaning of AND and OR can be relaxed, so that they retrieve more documents
- The documents are ranked higher when they have a larger number of elements in common with the query