Module IV

CHAPTER 4

Text Processing

Syllabus

Text and Multimedia languages and properties: Metadata, Markup Languages, Multimedia; Text Operations: Document Preprocessing, Document Clustering.

Self-learning Topics : Digital Library : Greenstone

4.1 TEXT AND MULTIMEDIA LANGUAGES AND PROPERTIES

GQ. Define document and list all the characteristics of a document with the help of diagram.

GQ. Define and explain a term document with an example.

- Text is the primary form utilized for knowledge exchange.
- Text has been developed everywhere, in a wide variety of forms and languages.
- The document designates a single unit of information.
- A document is a piece of text in digital or other form.
- A document can be any physical item (a file, an email) or a fully formed logical unit (a book or a research article), an entry in a dictionary or a judge's opinion on a case.
- The syntax and structure of a document are determined by the application or the person who generated it.
- The author specifies the semantics of the document.

- The presentation style of a document might dictate how it should be presented or printed.
- Its syntax and structure, which are tied to a particular application, determine the presentation style.

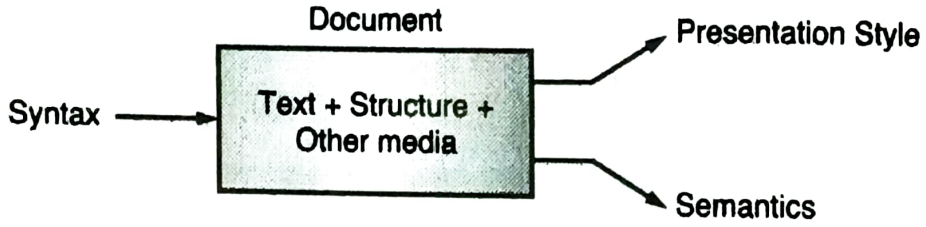

Fig. 4.1.1 : Characteristics of a document

- The document syntax
  - expresses structure, presentation style, semantics or external actions
  - one or more of elements may be given together or implicit in the document's content
  - structural element (such as a section) can have fixed formatting style
  - can be implicit in its content or expressed in a simple declarative language or expressed in a programming language
- Fig. 4.1.1 gives all the characteristics of a document.
- Syntax languages may be proprietary and specific but open and generic languages are more flexible.
- Text can be written in natural language, which is difficult for computers to process.
- The current trend is to use document languages that provide information on structure, format, and semantics that are readable by humans and computers.
- Document style
  - can be embedded in the document : TeX and RTF
  - can be complemented by macros : LaTeX
  - defines how a document is visualized or printed
- Fig. 4.1.2 represent a document with few styles.

- Understanding search engine queries is crucial since they are short chunks of text that are different from normal text and whose semantics are sometimes confusing due to polysemy.
- They are also difficult to infer the user intent behind a query.

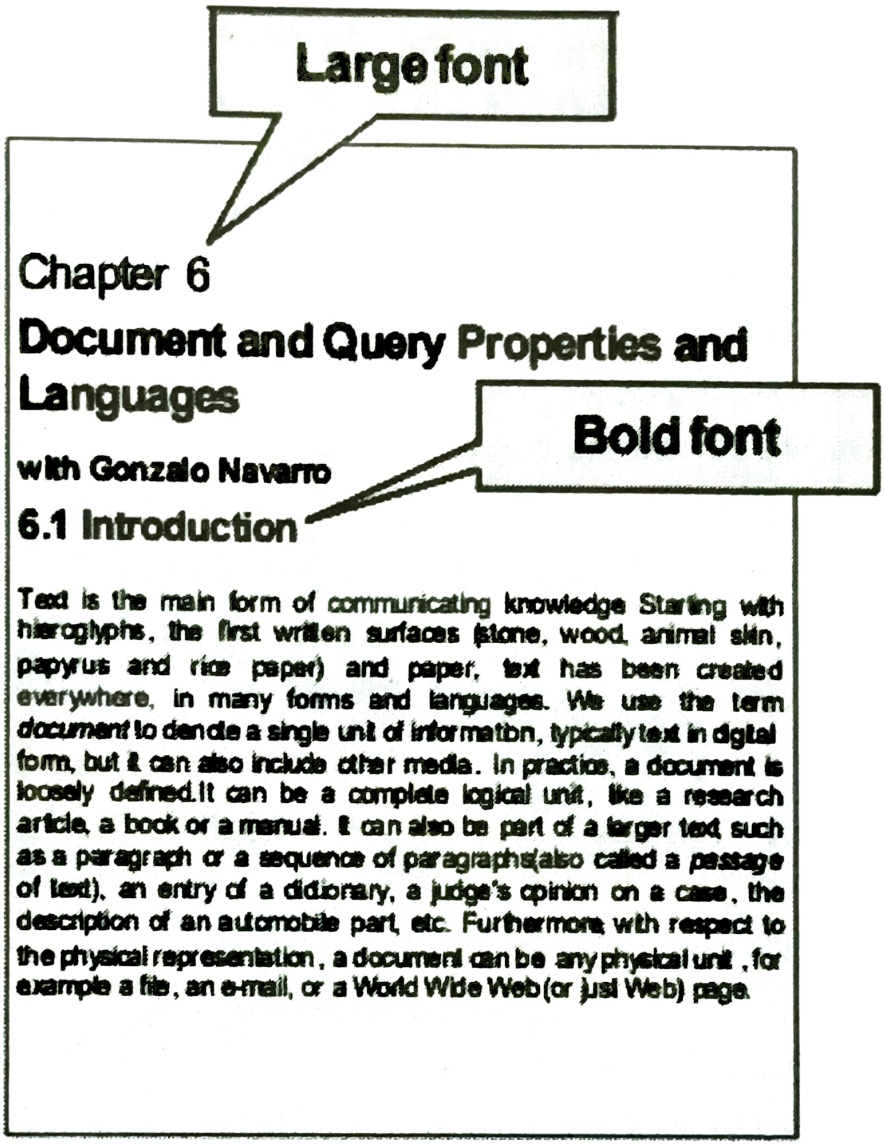

Fig. 4.1.2 : An example of document style

4.1.1 Metadata

GQ. Define briefly the term 'metadata'.
GQ. Explain the use of metadata in Web documents.
GQ. Write a short note on : RDF.

- Metadata is information about how the data is organized, the different data domains, and how they relate to one another.

- Metadata is 'data about the data'.
- The names of relations and attributes in a database that correspond to their domain are known as metadata.
- Most documents and text collections also have metadata.
- **Descriptive Metadata** is a metadata that is external to the meaning of the document and relates more to how it was created.
- Common forms of metadata for documents include the author, the date of publication, the source of the publication, the document length and the document genre.
- The Dublin Core Metadata Element Set suggests 15 fields for document descriptions.
- Marchionini refers to this type of information as Descriptive Metadata.
- **Semantic Metadata**
  - characterizes the subject matter within the document's contents
  - is associated with a wide number of documents
  - its availability is increasing
- Specific ontologies can be used to standardize semantic terms.
- An important metadata format is Machine Readable Cataloging Record (MARC).
  - it is the most used format for library records
  - includes fields for distinct attributes of a bibliographic entry such as title, author, publication venue
  - in the U.S.A., a particular version of MARC is used: USMARC
- Metadata is also used 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
  - privacy levels for access to a document
  - applications to electronic commerce
- Resource Description Framework (RDF) is a new standard for Web metadata.

- RDF enables interoperability across applications and allows characterizing Web resources to facilitate automated processing.
- RDF does not assume any specific application or semantic domain.
- It comprises of a description of nodes and attached attribute/value pairs.
- Nodes are made up of Uniform Resource Identifiers (URIs), which include Uniform Resource Locators (URLs).
- Attributes are properties of nodes with their values in the form of text strings or other nodes (Web resources or metadata instances).
- Metadata can be used to describe non-textual things, such as a collection of keywords for an image.

4.1.2 Markup Languages

GQ. What is Markup? Discuss various markup Languages in detail.
GQ. Define about xml language?
GQ. Differentiate hypertext and xml data structure?
GQ. Express standard languages for multimedia applications with proper example.

- The term "markup" refers to the additional syntax required to express formatting actions, structure details, text semantics, attributes, etc.
- The marks are called tags.
- Marked text is enclosed by an initial and an ending tag to prevent ambiguity.
- Examples of Markup Languages
  - SGML : Standard Generalized Markup Language
  - HTML : HyperText Markup Language
  - XML : eXtensible Markup Language

SGML

- SGML (ISO 8879) stands for Standard Generalized Markup Language, i.e., a meta-language for tagging text.
- SGML includes rules for constructing a markup language based on tags and a description of the document structure called a document type definition (DTD).

- An SGML document is defined by a document type definition and the text itself is marked with tags that specify the structure.
- The document type definition is used to
  - describe and name the pieces that a document is composed of
  - define how those pieces relate to each other
- Part of the DTD can be defined by an SGML document type declaration.
- Other components of the DTD, such as the semantics of elements and attributes or application norms, cannot be described explicitly in SGML.
  - Comments can be used, however, to express them informally
  - More complete information is usually present in separate documentation.
- Tags are denoted by angle brackets
  - Tags are used to identify the beginning and ending of an element such as (<tagname> element </tagname>)
  - Ending tags include a slash before the tag name
  - Attributes are specified inside the beginning tag
- Fig. 4.1.3 shows an example of a SGML DTD for electronic messages.

```
<!--SGML DTD for electronic messages -->

<!ELEMENT email              -- (prolog, contents) >
<!ELEMENT prolog             -- (sender, address+, subject?, Cc*) >
<!ELEMENT (sender | address | subject | Cc) - 0 (#PCDATA) >
<!ELEMENT contents           -- (par | image | audio)+ >
<!ELEMENT par                -0 (ref | #PCDATA)+ >
<!ELEMENT ref                -0 EMPTY >
<!ELEMENT (image | audio) --  (#NDATA) >

<!ATTLIST email
        id          ID                  #REQUIRED
        date_sent   DATE                #REQUIRED
        status      (secret | public)    public >
<!ATTLIST ref
        id          IDREF               #REQUIRED >
<!ATTLIST  (image | audio )
        id          ID                  #REQUIRED >
```

Fig. 4.1.3 : DTD for structuring electronic mails

- Fig. 4.1.4 shows an example of use of previous DTD.

```
<!--Example of use of previous DTD-->
<!DOCTYPE email SYSTEM "email.dtd">
<email id=12345jm date_sent=02102022>
        <prolog>
        <sender> Rugved More </sender>
        <address> Albert Gonsalves </address>
        <address> Mumbai </address>
        <subject> Pictures of my house in city town
        <Cc> Saumil More </Cc>
        </prolog>
        <contents>
                <par>
                Kindly check the attached images of my house and the
                splendid sea view from my bedroom
                   (photo <ref idref= "F2">).
                </par>
                <image id=F1> "photo1.gif" </image>
                <image id=F2> "photo2.jpg" </image>
                <par>
                Regards from the South, Rugved.
        </contents>
</email>
```

Fig. 4.1.4 : An example of use of DTD for structuring electronic mails

- Document description does not specify how a document is printed
  - Output specifications gives the directions on how to format a document which are often added to SGML documents, such as
    (1) DSSSL : Document Style Semantic Specification Language
    (2) FOSI : Formatted Output Specification Instance
  - These standards define mechanisms for associating style information with SGML document instances
  - They allow defining that the data identified by a tag should be typeset in some particular font
- One important use of SGML is in the Text Encoding Initiative (TEI-started in 1987)
  - Includes number of US associations related to the humanities and linguistics.

- SGML DTDs provide several document formats
- primary objective is to create guidelines for the preparation and interchange of electronic texts for scholarly research, as well as the industry
- the most popular format is TEI Lite

HTML

- HTML stands for HyperText Markup Language which is an instance of SGML.
- Since its creation in 1992, HTML has undergone several revisions, with version 4.0 being the most recent.
- HTML5 is under development and continually be updated with new features, called as "living standard".
- Most documents on the Web are stored and transmitted in HTML.
- HTML is simple language well suited for
  - Hypertext
  - Multimedia
  - The display of small and simple document
- Although there is an HTML DTD, most HTML instances do not explicitly refer to the DTD.
- The HTML tags follow all the SGML conventions and also include formatting directives.
- HTML pages can contain other media embedded in them, such as images or audios.
- HTML also provides fields for metadata, which can be utilized for various applications and purposes.
- Dynamic HTML (DHTML): A page that uses HTML and another program (for example, using JavaScript)
- Fig. 4.1.5 gives an example of an HTML document.

```
<html>
<head>
<title>HTML Practice Example</title>
<meta name=JM content="basic example">
</head>
<body>
<h1>HTML Practice Example</h1>
<p>
<hr><hr>
<p>
HTML is very <b>simple</b> language:
<ul>
<li> link to <b><a href=https://www.xavier.ac.in/>XIE</a></b>
   (a from anchor),
<li> paragraphs (p), headings (h1, h2, etc), font types (b, i),
<li> horizontal rules (hr), unordered lists and items (ul, li),
<li> images (img), tables, forms, etc.
</ul>
<p>
<hr><hr>
<p>
<img align=left src="flower.jpg" height=10 width= 10>
Look at beautiful <b>flower</b>.
</body>
</html>
```

Fig. 4.1.5 : Example of an HTML document

- Fig. 4.1.6 gives an output of above HTML document on a browser.

HTML Practice Example

HTML is very **simple** language:

- link to **XIE** (a from anchor),
- paragraphs (p), headings (h1, h2, etc), font types (b, i),
- horizontal rules (hr), unordered lists and items (ul, li),
- images (img), tables, forms, etc.

Look at beautiful **flower**.

Fig. 4.1.6 : Output of an HTML document on a browser

- Cascade Style Sheets (CSS) were established in 1997 since HTML does not fix presentation style.
- CSS offer
  - a powerful and manageable mechanism for authors to enhance the aesthetics of HTML pages
  - a way to distinguish information about presentation from document content
  - a support in current browsers which is still modest
- The evolution of HTML implies support for both backward and forward compatibility.
- HTML 4.0 has been specified in three flavours: strict, transitional, and frameset
  - Strict HTML only worries about non-presentational syntax, leaving all the displaying information to CSS
  - Transitional HTML makes advantage of all presentational features to create pages that can be read by older browsers without understanding CSS
  - Frameset HTML is utilized when you want to divide the browser window into two or more frames
- Style sheets, internationalization, frames, richer tables and forms, and accessibility features for people with impairments are all supported by HTML 4.0.
- Typical HTML applications employ a fixed, limited set of tags
  - which makes the language specification much easier to implement an applications
  - which significantly restricts HTML in a number of crucial areas
  - HTML does not
    (1) allow users to declare their own tags
    (2) support the specification of nested structures needed to represent database schemas
    (3) provide language specifications that enable consuming applications to validate imported data for structural accuracy

XML

- XML, the eXtensible Markup Language is a simplified version of SGML.
- It is not a markup language, like HTML, but a meta-language, like SGML.
- It allows for machine-readable semantic markup that is also readable by humans.
- It makes it easier to create and use new specific markup languages.
- XML does not have many of the restrictions of HTML.
- However, XML imposes a more rigid syntax on the markup
  - In XML, ending tags must present
  - XML is case sensitive
  - All attribute values must be added between quotes
  - Parsing XML without a DTD is easier
    1. The tags can be obtained while the parsing is done
- XML tags are not predefined, user can define their own tags.
- Fig. 4.1.7 shows an XML document without a DTD analogous to the previous electronic mail DTD given for SGML.

```
<?XML VERSION= "1.0" RMD= "NONE" ?>
<email id=12345jm date_sent=02102022>
        <prolog>
        <sender> Rugved More </sender>
        <address> Albert Gonsalves </address>
        <address> Mumbai </address>
        <subject> Pictures of my house in city town
        <Cc> Saumil More </Cc>
        </prolog>
        <contents>
                <par>
                Kindly check the attached images of my house and the
                splendid sea view from my bedroom
                   (photo <ref idref= "F2">).
                </par>
                <image id=F1> "photo1.gif" </image>
                <image id=F2> "photo2.jpg" </image>
                <par>
                Regards from the South, Rugved.
        </contents>
</email>
```

Fig. 4.1.7 : An XML document without a DTD

- Extensible Style sheet Language (XSL)
  - the XML counterpart of Cascading Style Sheets (CSS)
  - syntax defined based on XML
  - created to modify and style highly-structured, data-rich XML documents
  - using XSL, for instance, it would be possible to automatically extract a document's table of contents
- Extensible Linking Language (XLL)
  - Another extension to XML, defined using XML
  - defines different types of links (external and internal)
- Recent uses of XML include
  - Mathematical Markup Language (MathML)
  - Synchronized Multimedia Integration Language (SMIL)
  - Resource Description Format
- Next generation of HTML should be based in a suite of XML tag sets.

4.1.3 Multimedia

GQ. Express various formats for multimedia applications.

- Recent advances in computer technology have precipitated a new era in the way people create and store data.
- Multimedia usually stands for applications that handle different types of digital data.
- Millions of multimedia documents including images, videos, audio, graphics, and texts can now be digitized.
- Different types of formats are necessary for storing each media.
- Most formats for multimedia can only be processed by a computer.

Text

- With the advent of the computer, it became necessary to represent code characters in binary digits through coding schemes
  - EBCDIC (7 bits), ASCII (8 bits) and UNICODE (16 bits)
- All these coding schemes are based on characters.

- An IR system should be able to retrieve information from many text formats (doc, pdf, html, txt).
- IR systems uses filters to handle most popular documents.
- But good filters might not be possible with proprietary formats.
- Other text formats
  - Rich Text Format (RTF) : for document interchange
  - Portable Document Format (PDF) : for printing and displaying
  - Postscript : for printing, displaying and drawing
  - Multipurpose Internet Mail Exchange (MIME): for encoding email
- Most common compression software and associated formats
  - Compress (Unix), ARJ (PCs) : for compressing text
  - ZIP (Unix) (gzip in Unix and Winzip in Windows) : for compressing text.

Image Formats

- Direct representations of a bit-mapped display, such as XBM, BMP, or PCX, are the most basic image formats.
- 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 attempts to remove portions of the image that are less noticeable to the human eye
  - This format is parametric, meaning that the loss may be adjusted
- Another common image format is the Tagged Image File Format (TIFF)
  - exchange of documents between different applications and computers
- TIFF provides for metadata, compression, and varying number of colors.

- Truevision Targa image file (TGA- Truevision Graphics Adapter.) : another format related to video game boards.
- Portable Network Graphics (PNG) : bit-mapped image format for use in the Internet.
- Various other image forms, such as fax and fingerprints, are associated to specific applications.

Audio

- For proper storage, audio must be converted to digital format.
- The most popular audio file formats are AU, MIDI, and WAVE.
- MIDI: a common format for exchanging music between computers and electronic devices.
- Other formats, such RealAudio or CD formats, are employed for audio libraries.

Movies

- Main format for animations is Moving Pictures Expert Group (MPEG)
  - operates by coding the changes in successive frames
  - utilizes the temporal image redundancy that any video contains
  - incorporates the audio signal linked with the video
  - specific cases for audio (MP3), video (MP4), etc.
- Other video formats are AVI, FLI and QuickTime
  - AVI may include compression (CinePac)
  - QuickTime, developed by Apple, also includes compression

Graphics and Virtual Reality

- Three-dimensional graphics can be displayed in a variety of formats.
- Computer Graphics Metafile (CGM) allows for the open exchange of structured graphical objects and the attributes associated with them.
- The Virtual Reality Modeling Language (VRML) is a 3D graphics and multimedia interchange format that can be utilized in a wide range of application fields, including
  - engineering and scientific visualization

- multimedia presentations
- entertainment and educational titles
- web pages and shared virtual worlds

- VRML has emerged as the de facto Web modelling language.

HyTime

- Hypermedia/Time-based Structuring Language
  - Multimedia document markup standard
  - an SGML architecture that specifies a document's generic hypermedia structure
- HyTime hypermedia concepts include
  - complex locating of document objects
  - relationships (hyperlinks) between document objects
  - numeric, measured associations between document objects
- The HyTime architecture has three parts
  - the base linking and addressing architecture
  - the scheduling architecture (derived from the base architecture)
  - the rendition architecture (which is an application of the scheduling architecture)
- HyTime does not directly provide graphical interfaces, user navigation or user interaction.
- These components of document processing are derived from the HyTime structures in a similar way to how style sheets are used in SGML documents.

4.2 TEXT OPERATIONS

4.2.1 Document Preprocessing

GQ. Describe **five** text operations of document preprocessing.
GQ. Describe lexical analysis of the text.
GQ. Explain about stemming process.
GQ. Describe **the process of** thesaurus generation.

GQ. Explain the various phases of text preprocessing with the help of the logical view of the document.

(1) Lexical analysis of the text
(2) Elimination of stopwords
(3) Stemming
(4) Selection of index terms
(5) Construction of term categorization structures

(1) Lexical analysis of the text

- It is the process of turning a stream of characters into a stream of words.
- The basic goal is to identify the words in the text.
- It treats spaces (reduced to one space as word separator), digits (remove all words containing sequences of digits), hyphens (remove), punctuation marks (remove) and the case of the letters (converts all the text to either lower or upper case).

(2) Elimination of stopwords

- The main objective is filtering out words with very low discrimination values for retrieval purposes.
- Stopwords are the words
  - that are too common among the documents
  - which occurs in 80% of the documents
  - For example, articles, prepositions, conjunctions, etc.
- Stopword elimination significantly shrinks the size of the indexing structure but may decrease recall.
- Problem : Search for "to be or not to be"
  - Elimination process might leave only the term 'be' which makes it difficult to recognize the documents of that phrase

(3) Stemming

- The major goal is to get rid of affixes (prefixes and suffixes) and make it possible to retrieve documents with query terms that have syntactic variants.

- Syntactic variations are plurals, gerund forms and past tense suffixes.
- A stem is generated by removing the affixes of a word.
- Examples of generated stems are
  - connected, connecting, connection, connections--> connect
  - effectiveness --> effective --> effect
  - picnicking --> picnic
  - stresses --> stress
  - king --> k
- Stems help to enhance retrieval efficiency.
- Stemming help to reduce the size of the indexing structure.
- Stemming strategies by Frakes
  - affix removal: intuitive, simple and effective implementation
  - table lookup: simple but dependent on data on stems
  - successor variety: more complex than affix removal
  - n-gram: more clustering procedure than stemming

(4) Selection of index terms

- A sentence is usually composed of nouns, pronouns, articles, verbs, adjectives, adverbs, and connectives.
- Since the majority of semantics are carried by noun words, the motivation for selecting index terms is to use nouns in the text.
- Identification of noun groups is a good approach for selecting index terms.
- A noun group is a set of nouns whose syntactic distance in the text does not exceed a predefined threshold (for example, information technology).

(5) Construction of term categorization structures (such as thesaurus)

- Thesaurus is a word that has Greek and Latin roots and refers to the treasury of words.
- Treasury consists of
  - A precompiled list of important words in a given domain of knowledge

- A set of words for each word in the above list
- Normalization of the vocabulary

- An example from Peter Roget thesaurus is given below

  cowardly *adj*

  Ignobly lacking in courage: cowardly turncoats

  Syns: chicken (slang), chicken-hearted, craven, dastardly, faint-hearted, gutless, lily-livered, pusillanimous, unmanly, yellow (slang), yellow-bellied (slang).

- The idea of adopting a regulated vocabulary for indexing and searching is what inspired the creation of a thesaurus.
- The Purpose of a Thesaurus by Foskett
  - to provide a standard vocabulary for indexing and searching
  - to assist users with locating terms for proper query formulation
  - to provide classified hierarchies that allow the broadening and narrowing of the current query request
- Thesaurus index terms, term relationships, and layout designs for these term relationships are its key constituents.
- Thesaurus Term Relationships
  - BT : broader
  - NT : narrower
  - RT : non-hierarchical, but related
- Fig. 4.2.1 represent a logical view of the document after each of the above phases is completed.

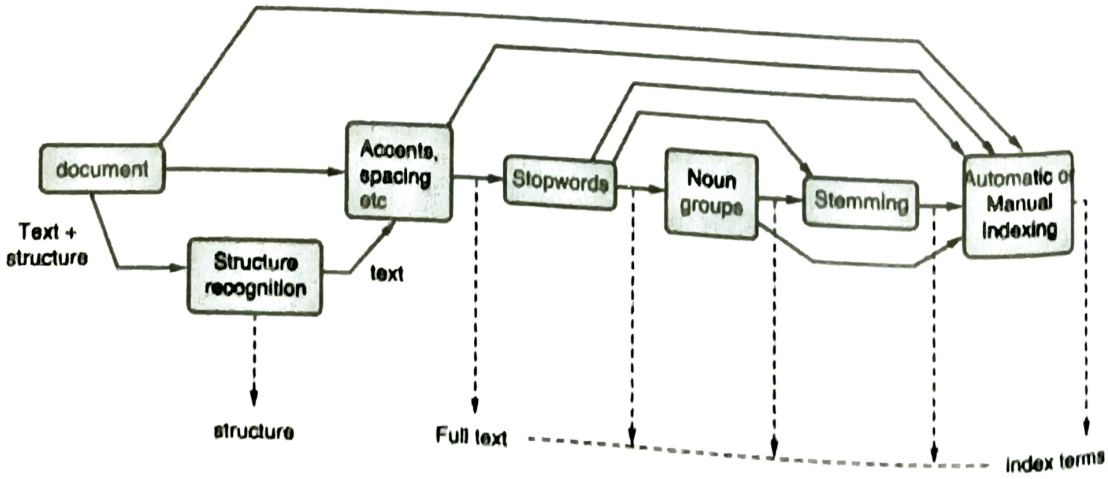

Fig. 4.2.1 : Logical view of the document throughout the various phases of text preprocessing

4.2.2 Document Clustering

GQ. What are the two types of clustering?
GQ. Differentiate between document clustering and term clustering?
GQ. Describe any one technique for term clustering with help of an example?
GQ. Define term clustering? Explain item clustering with suitable example?
GQ. Describe document clustering?
GQ. Explain about document clustering? Explain it with the help of example?
GQ. Describe the technique for term clustering?
GQ. Define clustering? What are the general guidelines for clustering?
GQ. List all steps of query expansion through local context analysis.

- **Clustering :** the grouping of documents which satisfy a set of common properties.
- **Document clustering** is an operation on the collection of documents and not on the text.
- Two types of operation of clustering documents: global and local.
- Global clustering strategy: the documents are organized into groups based on how frequently they appear throughout the whole collection.
- Local clustering strategy: the grouping of document is affected by the context defined by the current query and its local set of retrieved documents.
- Attempting to obtain a description for a larger cluster of relevant documents automatically is based on identification of terms related to the query terms such as synonyms, stemming, variations, terms with a distance of at most k words from a query term.
- In a global strategy, the entire collection of documents is used to create a global thesaurus that chooses terms for query extension.
- The local strategy involves looking through the documents that were returned for a specific query q at query time to identify terms for query expansion.

- Two basic types of local strategy :
  (1) Local clustering (2) Local context analysis
- Local strategies suit for environment of intranets, not for web documents.

Query Expansion Through Local Clustering

- Local feedback strategies are that expands the query with terms correlated to the query terms.
- Such correlated terms are those present in local clusters built from the local document set.

Definition : A $V(s)$ be a non-empty subset of words which are grammatical variants of each other. A canonical form s of $V(s)$ is called a stem.

- **Example**

  If $V(s)$ = {connect, connecting, connected} then s=connect
- For a given query q:
  - D_l : the local document set, the set of documents retrieved for a given query q
  - V_l : local vocabulary, the set of all distinct words in the local document set
  - S_l : the set of all distinct stems derived from the set V_l
- Strategies for building local clusters
  (1) Association clusters
  (2) Metric clusters
  (3) Scalar clusters

(1) Association clusters

- An association cluster is based on the co-occurrence of stems or words inside the documents by using the idea of synonymity association.
- Generation of Association clusters
  - $fs_{i,j}$: the frequency of a stem s_i in a document d_j, $d_j \in D_l$
  - Let $\vec{m} = (m_{ij})$ be an association matrix with $| S_l |$ row and $| D_l |$ columns, where $m_{ij} = fs_{i,j}$

- The matrix $\vec{s} = \vec{m}\,\vec{m}^t$ is a local stem-stem association matrix.
- Each element $s_{u,v}$ in $\vec{s}$ expresses a correlation $c_{u,v}$ between the stems s_u and s_v

$$c_{u,v} = \sum_{d_j \in D_l} fs_{u,j} \times fs_{v,j} \quad \ldots(4.2.1)$$

- The correlation factor $c_{u,v}$ quantifies the absolute frequencies of co-occurrence
- The association matrix $\vec{s}$ is unnormalized.

  if we adopt,

$$S_{u,v} = c_{u,v} \quad \ldots(4.2.2)$$

- Normalize the correlation factor to normalize the association matrix using

$$S_{u,v} = \frac{c_{u,v}}{c_{u,u} + c_{v,v} - c_{u,v}} \quad \ldots(4.2.3)$$

☞ **Build local association clusters**

- Consider the *u-th* row in the association matrix
- Let $S_u(n)$ be a function which takes the *u-th* row and returns the set of n largest values $s_{u,v}$, where v varies over the set of local stems and $v \neq u$
- Then $s_u(n)$ defines a local association cluster around the stem s_u.

(2) Metric Clusters

- Metric cluster is based on the idea that two terms which occur in the same sentence seem more correlated than two terms which occur far apart in a document.
- It might be worthwhile to factor in the distance between two terms in the computation of their correlation factor.
- Let the distance $r(k_i, k_j)$ between two keywords k_i and k_j in a same document
  - If k_i and k_j are in distinct documents we take $r(k_i, k_j) = \infty$
  - $V(s_u)$ is the set of keywords with s_u as their stems
  - $V(s_v)$ is the set of keywords with s_v as their stems

- A local stem-stem metric correlation matrix $\vec{s}$ is defined as each element $s_{u,v}$ of correlation matrix expresses a metric correlation $c_{u,v}$ between the setms s_u, and s_v

$$c_{u,v} = \sum_{k_i \in V(s_u)} \sum_{k_j \in V(s_v)} \frac{1}{r(ki, kj)} \quad \ldots(4.2.4)$$

- The correlation factor $c_{u,v}$ quantifies the absolute inverse distances
  - The association matrix $\vec{s}$ is unnormalized.
  - If we adopt,

$$S_{u,v} = c_{u,v} \quad \ldots(4.2.5)$$

  - Normalize the correlation factor to make normalized association matrix

$$S_{u,v} = \frac{c_{u,v}}{|V(s_u)| \times |V(s_v)|} \quad \ldots(4.2.6)$$

☞ **Build local metric clusters**

- Given a local metric matrix $\vec{s}$
- Consider the *u-th* row in the association matrix
- Let $S_u(n)$ be a function which takes the *u-th* row and returns the set of n largest values $s_{u,v}$, where v varies over the set of local stems and $v \neq u$
- Then $s_u(n)$ defines a local association cluster around the stem s_u

(3) Scalar Clusters

- Two stems with similar neighbourhoods have some synonymity relationship.
- The way to quantify such neighbourhood relationships is to arrange all correlation values $s_{u,i}$ in a vector $\vec{s}_u$, to arrange all correlation values $s_{v,i}$ in another vector $\vec{s}_v$, and to compare these vectors through a scalar measure.
- Let $\vec{s}_u = (s_{u,1}, s_{u,2}, \ldots, s_{u,n})$ and $\vec{s}_v = (s_{v,1}, \ldots s_{v,2}, s_{v,n})$ be two vectors of correlation values for the stems s_u and s_v

- Let $\vec{s} = (s_{u,v})$ be a scalar association matrix.
- Each $s_{u,v}$ can be defined as

$$S_{u,v} = \frac{\vec{s_u} \cdot \vec{s_v}}{|\vec{s_u}| \times |\vec{s_v}|} \quad \ldots(4.2.7)$$

- Let $S_u(n)$ be a function which returns the set of n largest values $s_{u,v}$, $v \neq u$. Then $S_u(n)$ defines a scalar cluster around the stem s_u.

☞ Interactive Search Formulation

- Stems that belong to clusters associated to the query stems(or terms) can be used to expand the original query.
- A stem s_u which belongs to a cluster (of size n) associated to another stem s_v is said to be a neighbour of s_v.
- Fig. 4.2.2 represents a stem s_u as a neighbour of the stem s_v within a neighbourhood $S_v(n)$.

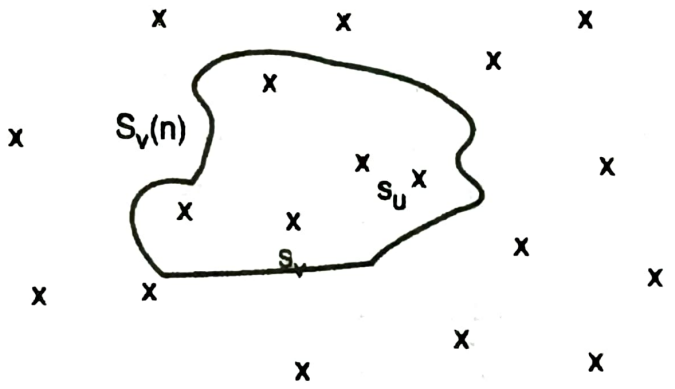

Fig. 4.2.2 : Stem s_u as a neighbour of the stem s_v

- For each stem , select m neighbour stems from the cluster $S_v(n)$ (which might be of type association, metric, or scalar) and add them to the query.
- Hopefully, the additional neighbour stems will retrieve new relevant documents.
- $S_v(n)$ may composed of stems obtained using correlation factors normalized and unnormalized.
  - normalized cluster tends to group stems which are more rare
  - unnormalized cluster tends to group stems due to their large frequencies

- Using information about correlated stems to improve the search
  - Let two stems s_u and s_v be correlated with a correlation factor $c_{u,v}$
  - If $c_{u,v}$ is larger than a predefined threshold then a neighbour stem of s_u can also be interpreted as a neighbour stem of s_v and vice versa
  - This provides greater flexibility, particularly with Boolean queries
  - Consider the expression $(s_u + s_v)$ where the + symbol stands for disjunction
  - Let s_u' be a neighbour stem of s_u
  - Then one can try both $(s_u'+s_v)$ and (s_u+s_u') as synonym search expressions, because of the correlation given by $c_{u,v}$

☞ Query Expansion Through Local Context Analysis

The local context analysis procedure operates in three steps :

(1) Retrieve the top n ranked passages using the original query This is accomplished by breaking up the documents initially retrieved by the query in fixed length passages (for instance, of size 300 words) and ranking these passages as if they were documents.

(2) For each concept c in the top ranked passages, the similarity $sim(q, c)$ between the whole query q (not individual query terms) and the concept c is computed using a variant of tf-idf ranking.

(3) The top m ranked concepts (according to $sim(q, c)$) are added to the original query q. To each added concept is assigned a weight given by $1\text{-}0.9 \times i/m$ where i is the position of the concept in the final concept ranking. The terms in the original query q might be stressed by assigning a weight equal to 2 to each of them.

⏭ 4.3 GENERAL QUESTIONS

Q. 1 Explain XML retrieval?

☑ **Ans. :**

Document-oriented XML retrieval

(1) Document vs. data- centric XML retrieval (recall)

(2) Focused retrieval

(3) Structured documents

(4) Structured document (text) retrieval

(5) XML query languages

(6) XML element retrieval

(7) (A bit about) user aspects Explain the above in details.

Q. 2 Define clustering.

☑ **Ans. :**

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. Help users understand the natural grouping or structure in a data set. Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms.

Chapter Ends...

❑❑❑