INFS 427: AUTOMATED INFORMATION RETRIEVAL

(1st Semester, 2018/2019)

Session 05 – SUBJECT ANALYSIS & REPRESENTATION

Lecturer: Mrs. Florence O. Entsua-Mensah, DIS
Contact Information: fentsua-mensah@ug.edu.gh

UNIVERSITY OF GHANA
College of Education
School of Continuing and Distance Education
2014/2015 – 2016/2017

Session Overview

- One of the major functions of an information retrieval system is to match the contents of documents with users queries.
- The system personnel have to prepare a surrogate for every document, and all such surrogates must be maintained in an organized manner.
- This activity is achieved through '**subject analysis'**.
- The session therefore examines the concept of subject analysis and representation.

Session Outline

The key topics to be covered in the session are:

- **Topic 1:** Understanding Subject Analysis
- **Topic 2:** Determining the 'subject matter' of a document
- **Topic 3:** Subject Indexing Systems

Recommended Reading

Chowdhury, G. G. (2010). *Introduction to modern information retrieval*. London: Facet publishing. **(Chapter 5)**.

Korfhage, R. R. (2006). *Information Storage and Retrieval*. Wiley India Pvt. Limited.

Taylor, A. G. (2009). *The organization of information*. (3rd ed.). Westport, CT: Libraries Unlimited. **– Chapter 9**

Introduction

- The major function of an IR system is to match users' queries to contents of a document
- This matching will be possible after preparation of a surrogate for each documents
- Document surrogates are "limited representations of full documents" (Korfhage, 2006).
- The construction of document surrogates by assigning specific identifiers or key words to text items is referred to as indexing.
- Indexing based on the conceptual analysis of the subject of documents is known as **subject indexing**.
- **Subject indexing comprises** 2 main intellectual steps:
    1. **Conceptual analysis** – what is the subject of this document? What is it about?
    2. **Representation-** what keyword can best represent this document?

Introduction - 2

- This session focuses on the what, why, when, how, and who of subject analysis, or determining what an information object is about.
- All four subject description processes (classification, subject cataloging, indexing, and abstracting) as well as searching depend on subject analysis.
- This is a vital skill area for information professionals.
- Expert subject analysis requires high levels of verbal aptitude and abstract thinking skills.

Topic One

DEFINITION OF TERMS/CONCEPTS

Definition of Subject Analysis (SA)

- S. A. refers to examination of a bibliographic item by a trained subject specialist to determine the most specific **subject heading(s)** or descriptor(s) that fully describe its content, to serve in the bibliographic record as access points in a subject search of a library catalog, index, abstracting service, or bibliographic database.

(Bastida, 2016)

Definition of Subject Analysis (SA)

- Subject analysis is the task of determining the intellectual content or aboutness of an information object. But this isn't just the documents in the collection.
- If you recall the basic information retrieval (IR) model, two kinds of information go into an IR system:
  - representations of information objects (documents) and
  - representations of information need (queries).

(Taylor, 2009)

Definition of Subject Analysis (SA)

- **Document analysis** is studying a document to determine how to represent it in a record, or what indexing terms or codes to enter.
- **Query analysis** is studying an information request to determine how to formulate a search query, or how to choose appropriate search terms.

(Taylor, 2009)

Definition of Terms/Concepts

- **Subject analysis** is the part of indexing or cataloging that deals with the conceptual analysis of an item (document):
  - Begins with determining what a document is about? what is its form/genre/format of the document?
  - translates that analysis into a particular subject heading system.
  - It is usually the first step in classification

(Robare, 2004)

Definition of Terms/Concepts

Subject heading:

- is a term or phrase used in a subject heading list to represent a concept, event, or name

Classification

- Process of organizing resources by assigning an alphanumerical string that sorts physical objects by subject

(Robare, 2004)

Meaning of subject analysis

Subject analysis is used in two ways in the library and information science (LIS) literature

1. Relates to construction of indexing language and classification systems.
2. Relates to the analysis of the topical content of a document (which is our focus)

- Subject analysis is determining the essence or the subject matter in document texts, databases, controlled and natural languages, information requests, and search strategies.
- The major problem in subject indexing is summarizing the contents in a few words.

Analysis vs. indexing

Analysis:

- Look at the work as a whole to determine its overall contents
- Think of terms that summarize the primary subject focus of the work

Indexing:

- Provide in-depth access to parts of items (chapters, articles, detailed listing of topics)

(Robare, 2004)

Subjects vs. forms/genres

- Subject: what the item is about
- Form: what the item is, rather than what it is about
  - Physical character (video, map, miniature book)
  - Type of data it contains (statistics)
  - Arrangement of information (diaries, indexes)
  - Style, technique (drama, romances)
- Genre: works with common theme, setting, etc.
  - Mystery fiction; Comedy films

(Robare, 2004)

Topic Two

DETERMINING THE 'SUBJECT MATTER' OF A DOCUMENT

- Subject analysis involves four steps:
- 1. **Familiarization**: Becoming acquainted with general content of document and query
- 2. **Extraction**: Identifying and extracting significant concepts and natural-language terms
- 3. **Translation**: Converting extracted terms into controlled vocabulary of system
- 4. **Formalization**: Applying rules for exact format, spelling, punctuation, codes, etc. for input to system

- The steps do not necessarily occur in this order: subject analysis requires evaluation and verification at every stage in a continuous, iterative cycle. Taylor (2004, chapter 9) subsumes steps 1 and 2 under conceptual analysis, or determining aboutness, and steps 3 and 4 under subject analysis, or translating concepts into system terms. Although controlled-vocabulary systems are assumed above, steps 1 and 2 are also applied in natural-language systems. Regardless, the same general considerations come into play:

Determining the subject content

Examine the subject-rich portions of the item being cataloged to identify key words and concepts:

- Title
- Table of contents
- Introduction or preface
- Author's purpose or forward
- Abstract or summary
- Index
- Illustrations, diagrams
- Containers

(Robare, 2004)

Types of concepts to identify

- Topics
- Names of:
  - Persons
  - Corporate bodies
  - Geographic areas
- Time periods
- Titles of works
- Form of the item

(Robare, 2004)

Translating key words & concepts into subject headings

- Controlled vocabulary
  - Thesauri (examples)
    - Art & Architecture Thesaurus (AAT)
    - Thesaurus of ERIC Descriptors
  - Subject heading lists (examples)
    - Library of Congress Subject Headings
    - Sears List of Subject Headings
    - Medical Subject Headings (MeSH)

(Robare, 2004)

Why use controlled vocabulary?

- Controlled vocabularies:
  - identify a preferred way of expressing a concept
  - allow for multiple entry points (i.e., cross-references) leading to the preferred term
  - identify a term's relationship to broader, narrower, and related terms
  - "syndetic structure"

(Robare, 2004)

Function of keywords

- Advantages:
  - provide access to the words used in bibliographic records
- Disadvantages:
  - cannot compensate for complexities of language and expression
  - cannot compensate for context
- Keyword searching is enhanced by assignment of controlled vocabulary!

(Robare, 2004)

Guidelines for determining the subject matter of a document for the purposes of indexing

Dewey Decimal Classification guidelines

- The indexer must:
- examine the title and table of contents,
- chapter headings and subheadings,
- read the forward, preface, and introduction,
- and lastly scan through the main text.

Guidelines for determining the subject matter of a document for the purposes of indexing

- **International Standard Organization guidelines**
- The indexer must examine:
  - the title, abstract (if available)
  - List of contents and introduction
  - The opening chapters and the conclusion
  - Illustrations, diagrams, tables and their captions
  - Words or group of words which are underlined or printed in an unusual typeface
  - Finally indexer identifies main concepts in the document by consulting a checklist of questions.

Checklist for examining documents, determining their subjects and selecting index terms –British Standards Institution (BS 6529:1984)

- ❑ Does the document deal with a specific product, condition and phenomenon?
- ❑ Does the document contain an action concept, an operation or a process?
- ❑ Does the document deal with the agent of this action?
- ❑ Does it refer to particular means of accomplishing the action e.g., special instrument, techniques or methods?

- Where these factors considered in the context of a particular location or environment?
- Are any independent or dependent variables identified?
- Was the subject considered from a special viewpoint not normally associated with that field of study, e.g. a s sociological study of religion? (Chowdhury, 2010, p. 97)
- *Such a checklist demands some level of intellectual capacity on the part of the indexer leading to problems in manual indexing*

Wilson's proposition for determining the subject of a document

Wilson proposes 4 ways:

1. **The purposive method author oriented, i.e. The indexer determines** the purpose of the document to ascertain what the author is narrating, proving, describing, questioning, or explaining by looking for specific clues in the document such as, *I will show that, it shall be proved that etc.*
2. **The figure-ground method –** Indexer determines the aspects of the document that are emphasized or stand out. More or less the indexers impression of the document. May differ from person to person
3. **Constantly-referred-to method** – Subject is determined by counting frequencies of words in the document. The assumption is the word with the highest frequency is the subject of the document. However this might not be true.
4. **The appeal to unity method** – The indexer tries to determine what unifies or makes the document whole or cohesive. Might differ from indexer to indexer.

Modelling the subject analysis process

- Available models for subject analysis can be generalised to a 3 step model:

1. **Document analysis process** - Analysis of the document to determine the subject
2. **Subject description process** – formulation of an indexing phrase or subject description.
3. **Subject analysis process** – translation of the subject description into an indexing language or classification scheme

IMPORTANT FACTORS TO NOTE IN SUBJECT ANALYSIS

Objectivity

- Catalogers must give an accurate, unbiased indication of the contents of an item
- Assess the topic objectively, remain openminded
- Consider the author's intent and the audience
- Avoid personal value judgments
- Give equal attention to works, including:
  - Topics you might consider frivolous
  - Works with which you don't agree

(Robare, 2004)

Informed Subjectivity

Cataloger's judgment

- Individual perspective
- Informed by the cataloger's background knowledge of the subject
- Informed by the cataloger's cultural background
- Consistency in *determining* "What is it about?" leads to greater consistency in *assignment* of subject headings

(Robare, 2004)

Topic Three

SUBJECT INDEXING SYSTEMS

Subject indexing systems

- They are indexing systems based on the analysis of contents of documents.
- They facilitate retrieval of documents by assigning index terms after the subject matter of a document has been analysed.
- Assignment of index terms can be manual or automatic.
- There are 2 types of subject indexing systems:
  - Pre-coordinate systems
  - Post-coordinate systems

Pre-coordinate indexing system

- Coordination of terms are performed at the indexing stage.
- Each index entry represents the full content of a specific document
- Indexer may select terms from an authoritative source such as Library of Congress Subject Headings (LCSH)
- There is no room for manipulation of terms during searching. Searchers can only use the terms or compound terms pre-determined by the indexer (more or less like a controlled vocabulary)
- Examples are Chain indexing, relational indexing, PREserved Context Index System (PRECIS), Postulate-based Permuted Subject Indexing (POPSI)

Post-coordinate indexing system

- A single entry is prepared for each keyword selected to represent the subject of a document. All entries are organized in a file
- During searching users queries are matched against the file of index terms and the relevant documents are retrieved.
- Examples- Uniterm, Peek-a-boo

The Peek-a-boo cards

- Also known as Optical coincidence was used in the mid 1950s
- Each card is divided into small units of numbered squares or a grid from 500 to up to 10,000.
- Each grid/square represents the document number where the index term/phrase appears.
- There is a space at the top of each card for posting the index term.
- During searching, the user has to pick up all the cards marching a query.
- The cards are placed in a box within a source of light.
- Light will pass through the cards that have punches at the same location.
- The numbers represented in those locations will contain all the terms present in the queries.

Basic steps in indexing

1. Analysis of subject
2. Identification of keywords
3. Standardization of keywords
4. Choice of indexing system
    1. Post-coordinate will require
        1. preparation of entries under each keyword with reference to the document identification number, and
        2. preparation of reference entries (i.e. See also entries).
    2. Pre-coordinate will require
        1. preparation of main entry using all keywords organized in a way prescribed by the system,
        2. preparation of index entries using each significant term as an entry element and the main entry as the context
        3. Preparation of reference entries
        4. Filling of entries

Exhaustivity and Specificity

- Exhaustivity and specificity are 2 parameters used to control or ensure the effectiveness of an indexing system.
- **Exhaustivity –** the extent to which all index terms and concepts in a document are covered.
  - This requires the selection of as many keywords as possible to represent all the ideas and concepts in the document
- **Specificity –** the extent to which all topics/concepts are indexed in detail. For eg. indexing language on the topic 'leaves' which excludes types of leaves is less specific. If it refers to some but not all types of leaves, it is less exhaustive

Problems of manual indexing

- Lack of consistency – indexers may not assign the same index term to a given document.
- Varying levels of specificity and exhaustivity are attained based on the different perspectives of indexers
- Use of controlled vocabulary may hinder accuracy; i.e., indexers may not represent the document accurately, especially where new words are introduced to the documents.
- Indexer-user mismatch – same concept may be represented differently by indexer and user e.g., meats and poultry
- Pre-coordination- subsets of terms in manual indexing are often represented by a single term; e.g., gas, oil, coal, are represented by fuel. This may hinder recall.

Summary

- In this session we examined the work of the subject analyst in (automated) information retrieval.

- We also studied some of the ways to conduct subject analysis.

Activity 5.1

- Discuss the role of subject indexing systems in modern information retrieval.

Activity 5.2

- Access this link http://www.ugapress.org/upload/indexing.pdf and make notes on how to systematically index a book

References

Bastida, G. (2016). Subject analysis and representation. Accessed on 23rd August, 2018. Available at: https://slideplayer.com/slide/7802631/

Chowdhury, G. G. (2010). *Introduction to modern information retrieval*. London: Facet publishing.

Korfhage, R. R. (2006). *Information Storage and Retrieval*. Wiley India Pvt. Limited.

Robare, L. (2004). *Basic Subject Cataloguing Using LCSH.: Instructor's manual*. United States. Library of Congress. Retrieved from https://books.google.com.gh/books?id=Um--PAAACAAJ

Taylor, A. G. (2009). *The organization of information*. (3rd ed.). Westport, CT: Libraries Unlimited.