UNIT - III

TOPIC: Automatic Indexing

Introduction

Indexing

process of transformation of an item that extracts the semantics of the topics discussed in the item.

extracted information used to create

1. processing tokens
2. Searchable data structure

The index can be based on

1. full text of the item
2. automatic or manual generation of a subset of terms/phrases to represent the item .

Classes of Automatic Indexing:

Automatic indexing

process of analyzing an item to extract the information to be permanently kept in an index.

associated with generation of the searchable data structures that are associated with an item.

An index is the data structure created to support the search strategy.

Search strategies can be classified as

Statistical

natural language

concept

Statistical indexing:

Uses frequency of occurrence of events to calculate a number that is used to indicate the potential relevance of an item. Probabilistic systems

calculate a probability value (Probabilistic Weighting) Bayesian Model and Vector (Weighting) approaches

Calculate a relative relevance value (e.g., confidence level).

Natural Language

Goal of natural language processing:

o To enhance the indexing of the item by using the semantic information in addition to the statistical information.

o To improve the precision of searches

o To reduce the number of false hits.

The goal of indexing is to represent the semantic concepts of an item in the information system to support finding relevant information.

Single words have conceptual context, but frequently.

Concept Indexing Goal

To use concepts instead of terms as the basis for the index

To produce a reduced dimension vector space.

Concept indexing

can start with a number of unlabeled concept classes

Let the information in the items define the concepts classes.

Hypertext Linkages:

Hypertext data structure

is a new class of information representation.

generated manually although user interface tools may simplify the process.

hypertext linkages

Create an additional information retrieval dimension.

Traditional items can be viewed as two dimensional constructs.

The text of the items is one dimension which represents the information in the items.

DOCUMENT AND TERM CLUSTERING

Clustering index terms

to create a statistical thesaurus

to increase recall by expanding searches with related terms.

Clustering items

to create document clusters.

Search can retrieve items similar to an item of interest.

Introduction to Clustering

Thesaurus Generation

discusses a variety of specific techniques to create thesaurus clusters.

Thesaurus Generation:

Manual Clustering

Automatic Clustering

Complete Term Relation Method

Clustering Using Existing Clusters

One Pass Assignments

Item Clustering:

Clustering of items is very similar to term clustering for the generation of thesauri.

Manual item clustering is inbuilt in any library or filing system.

In this case someone reads the item and determines the category or categories to which it belongs.

When physical clustering occurs, each item is usually assigned to one category.

With the introduction of indexing, an item is physically stored in a primary category but it can be found in other categories.

Hierarchy of Clusters:

Hierarchical clustering in Information Retrieval focuses on the area of Hierarchical Agglomerative Clustering Methods (HACM).

Agglomerative means the clustering process starts with unclustered items and performs pair wise similarity measures to determine the clusters.

Divisive is the term applied to starting with a cluster and breaking it down into smaller clusters.