Indexing and Abstracting

Lecture 09 -- Automatic Indexing

Kuang-hua Chen
Department of Library and Information Science
National Taiwan University

Outline

- What's the subject indexing?
- Types of subject indexing
- The taxonomy for subject indexing
- Index Structures
- The models for automatic indexing
- 3-tier model for automatic indexing
- Natural language processing techniques
- Future trends

Subject Indexing

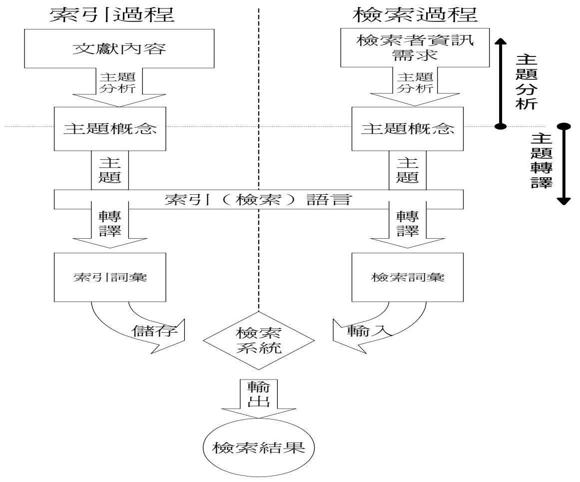
- The function of indexing is to describe the "aboutness" of documents
- Indexing terms are used to present content of document
- Challenge
  - how to select a set of indexing terms to represent the document contain thousands of words faithfully
- Process
  - subject analysis
  - subject translation

Process of Subject Indexing

- Subject analysis
  - Analyze content of texts and distill the subject concepts
  - Basis of subject indexing
- Subject translation
  - Translate the subject concepts to index terms
  - Two approaches
    - natural language indexing (free-text indexing)
    - controlled vocabulary indexing

Tasks of Indexing

- Analysis of the subject content of the document
- Review of indexing policies and authorities to aid in the correct assignment of terms
- Presentation of the index terms in the appropriate order of the indexing system
- Weighting of index terms
- Quality control of the index terms



Indexing Consistency

- The degree of agreement in the representation of the essential information content of the document by certain sets of indexing terms selected individually and independently by each of the indexers in the group.

Indexing Consistency Rating

- all studies indicated that consistency was very low
- a figure of 30% often was used
- Indexing consistency can vary on several factors
  - familiarity with the indexing policies
  - experience with the specific subject
  - the document most recently indexed
  - the time allowed to complete the task

How to Measure Consistency

- inter-indexer consistency
  - the overlap in index term assignment by two or more indexers for the same document
- intra-indexer consistency
  - the same indexer indexes the same document at two different times

Increase Indexing Consistency

- Manual Aids
  - Vocabulary control
  - thesauri
  - scope notes
- indexer may choose not to use manual aids
  - takes additional time
  - relevance of the aid to the problem is not apparent
  - indexer believes there is no problem at all

Machine-Readable Indexing Aids

- The indexer’s tools included authority files, policy manuals, handbooks, textbooks, etc.
- Machine Readable Indexing resources are available.

Pre- versus Post-coordination

- Pre-coordinated indexing term
  - complex/compound concepts are represented in a single term
- Post-coordinated indexing term
  - concepts are joined at the time of retrieval.

Controlled versus Uncontrolled

- Controlled Indexing
  - may be selected from a hierarchical thesaurus
  - may be selected from a list of classification level subject headings
- Uncontrolled Indexing
  - natural language terms (free terms) from texts with or without standardization

Automatic versus Manual

- Automatic indexing
  - Apply computers to proceed the indexing task
- Manual indexing
  - Human indexers proceed the indexing task

Indexing Scheme

- Use 3-tuple to represent possible indexing scheme
  - The first element denotes pre-coordinated (+) or post-coordinated (-)
  - The second element denotes controlled (+) or uncontrolled (-)
  - The third element denoted automatic (+) or manual (-)
- IS(-, +, +) represents post-coordinated, controlled, and automatic indexing

Automatic Indexing

- Most works are devoted to automatic free-text indexing
- Few works concern the automatic controlled-vocabulary indexing

Indexing Aims

- The effectiveness of any content analysis or indexing system is controlled by two parameters
  - indexing exhaustivity
    - the degree to which all aspects of the subject matter of a text item are actually recognized
  - term specificity
    - the degree of breadth or narrowness of the terms

Term Specificity

- Broad terms cannot distinguish relevant from irrelevant items
- Narrow terms retrieve relatively fewer items, but most of the retrieved materials are likely to be helpful to users

Approaches for Automatic Indexing

- Semantic Approach
  - based on understanding texts
  - domain-dependent
- Syntactic Approach
  - based on syntactic analysis of texts
  - language-dependent
- Statistical Approach
  - based on the statistics of terms
  - portable

Term Frequency

- Function words
  - for example, "and", "or", "of", "but", …
  - *the frequencies of these words are high in all texts*
- Content words
  - words that actually relate to document content tend to occur with greatly *varying frequencies in the different texts* of a collection
  - the frequency of content word may be used to indicate term importance for content representation.

A Frequency-Based Indexing Method

- Eliminate common function words from the document texts by consulting a special dictionary, or stop list, containing a list of high frequency function words
- Compute the <u>term frequency</u> tf_{ij} for all remaining terms T_j in each document D_i, specifying the number of occurrences of T_j in D_i
- Choose a threshold frequency T, and assign to each document D_i all term T_j for which $tf_{ij} > T$

Document Frequency (DF)

- The number of documents which contain the designated word for a certain collection
- $df_j = df(T_j)$=NumberOfDocumentContain(T_j)

Compose a Single Frequency-Based Indexing Model

- Best indexing terms are those that occur frequently in individual documents but rarely in the remainder of the collection
- A simple combined term importance indicator is

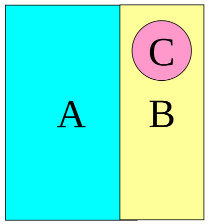
$$w_{ij} = tf_{ij} \times \log \frac{N}{df_j}$$

An Improved Indexing Policy for Free-Term Indexing

- Eliminating common function words
- Computing the value of w_{ij} for each term T_j in each document D_i
- Assigning to the documents a collection of all terms with sufficiently high ($tf \times idf$) factors

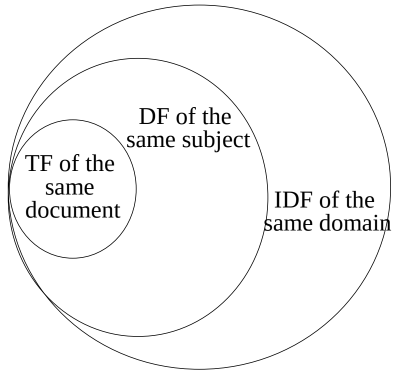
Problems of Traditional Model for Control-Vocabulary Indexing

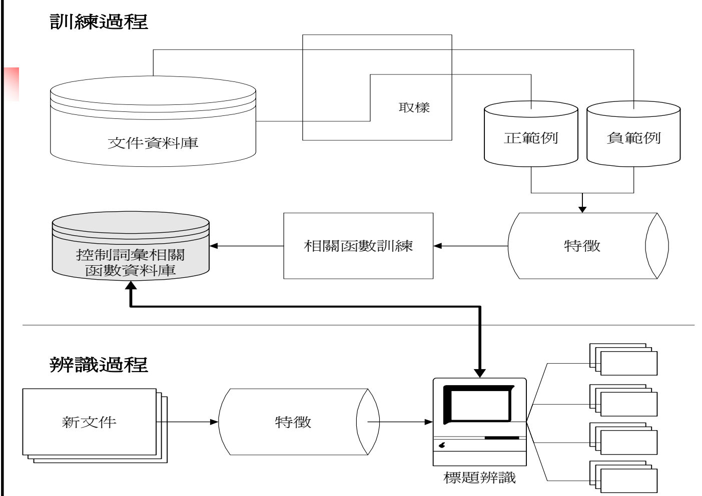
- Term statistics
  - Term frequency (TF)
  - Document frequency (DF)
  - Inverse document frequency (IDF = log(N/DF))
- Traditional model: TF×IDF
- High DF words
  - Common words
  - Domain-specific words
  - Subject-specific words



A+B+C is the words with high DF and low IDF
A = Common words
B = Domain-specific words
C = Subject-specific words

3-Tier Model for Automatic Indexing

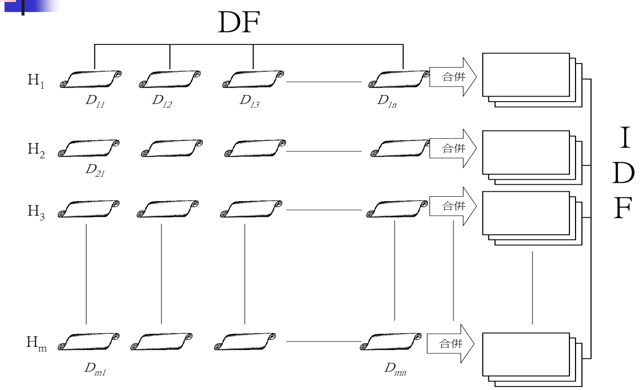




Basic Idea

Subject Headings	Learning Result
H_1	$R_1 = \{w_{11}, w_{12}, w_{13}, w_{14}, \ldots w_{1l_1}\}$
H_2	$R_2 = \{w_{21}, w_{22}, w_{23}, w_{24}, \ldots w_{2l_2}\}$
H_3	$R_3 = \{w_{31}, w_{32}, w_{33}, w_{34}, \ldots w_{3l_3}\}$
$\vdots$	$\vdots$
H_j	$R_j = \{w_{j1}, w_{j2}, w_{j3}, w_{j4}, \ldots w_{jl_j}\}$
$\vdots$	$\vdots$
H_m	$R_m = \{w_{m1}, w_{m2}, w_{m3}, w_{m4}, \ldots w_{ml_m}\}$

The Scheme for Term Weight



DF versus IDF

	DF original set	IDF combined set
Common Words	High	Low
Domain-specific Words	High	Low
Subject-specific Words	High	High

Term Weighting

$$Weight = TF \times DF_{originalset} \times IDF_{combinedset}$$

$$Weight_{ik} = TF_{ik} \times OSDF_{nk} \times CSIDF_{mk}$$

$Weight_{ik}$ = weight of term k in document i

TF_{ik} = frequency of term k in document i

$OSDF_{nk}$ = document frequency of term k in original document collection n

$CSIDF_{mk}$ = inverse document frequency of term k in combined document collection m

Training Stage

- Select experimental texts and controlled vocabulary
- Select testing subjects
- Train parameters for the proposed model

	Training Set	Testing Set	(Total)
Positive	40,000	20,000	60,000
Negative		400	400
(Total)	40,000	20,400	60,400

Testing Stage

- Compute the indexing score for testing texts

$$Indexing\ Score = \frac{\sum (OSDF \times CSIDF) \times (TF)}{\text{number of words in the document}}$$

$$\text{(weight of a word} = 0\text{, when the word isn't included in } R_j\text{)}$$

- Thresholding

$$\begin{aligned}&IF\\&\quad IS > M_j\\&THEN\\&\quad \mathrm{H}_j \text{ is assigned to the document}\end{aligned}$$

Evaluation Criteria

- Indexing precision

$$\text{Indexing Precision} = \frac{\text{正確索引之文件數}}{\text{模型索引之文件數}}$$

- indexing recall

$$\text{Indexing Recall} = \frac{\text{模型正確索引之文件數}}{\text{文件集中應索引之文件數}}$$

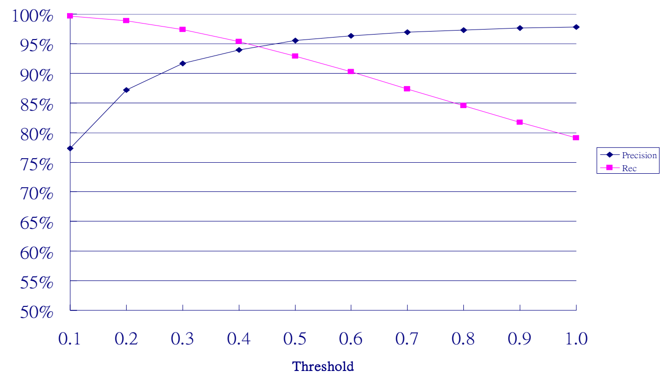
Abstract Part

Threshold	Training Set		Testing Set	
	Precision(%)	Recall(%)	Precision(%)	Recall(%)
0.1	77.31	99.62	76.78	96.63
0.2	87.17	98.83	86.47	92.90
0.3	91.68	97.36	91.01	89.49
0.4	93.92	95.32	93.32	86.19
0.5	95.49	92.88	95.00	83.39
0.6	96.33	90.24	95.91	80.62
0.7	96.92	87.31	96.56	77.87
0.8	97.32	84.51	97.01	75.51
0.9	97.62	81.69	97.35	73.15
1.0	97.83	79.09	97.59	71.09

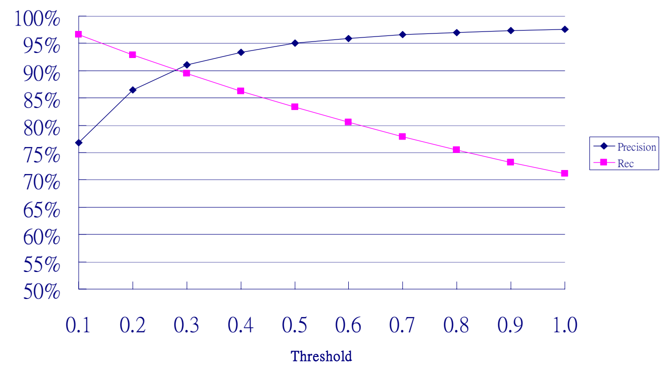
Title Part

Threshold	Training Set		Testing Set	
	Precision(%)	Recall(%)	Precision(%)	Recall(%)
0.1	80.52	76.48	80.87	78.24
0.2	86.85	74.86	86.78	74.38
0.3	89.94	72.86	89.67	70.76
0.4	91.92	70.73	91.54	67.34
0.5	92.98	68.50	92.57	64.48
0.6	93.60	66.09	93.18	61.76
0.7	94.10	63.78	93.71	59.58
0.8	94.51	61.47	94.14	57.67
0.9	94.88	59.30	94.56	55.67
1.0	95.19	57.12	94.92	53.92

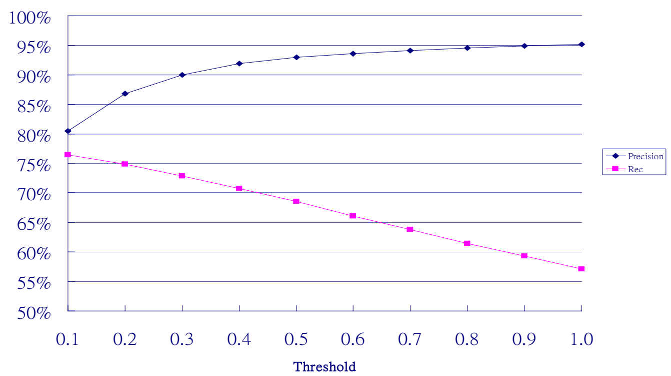
Abstract Part in Training



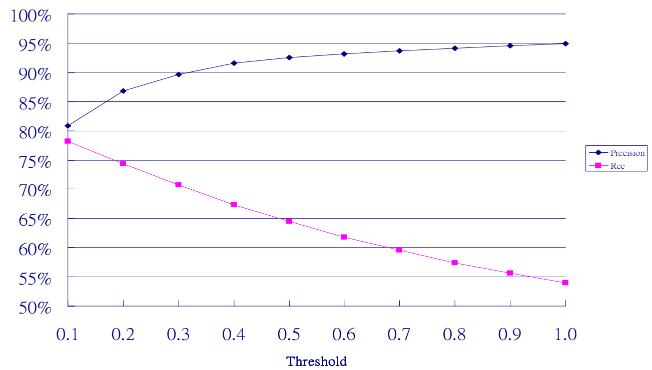
Abstract Part in Testing



Title Part in Training



Title Part in Testing



Comparison to Traditional Model

Threshold	Training Set		Testing Set	
	Precision(%)	Recall(%)	Precision(%)	Recall(%)
0.1	86.67	97.80	84.86	84.30
0.2	93.33	89.01	90.51	60.65
0.3	95.08	73.26	91.18	39.20
0.4	96.10	54.45	91.54	23.92
0.5	97.09	38.04	92.62	14.30
0.6	98.53	24.84	95.55	7.94
0.7	99.81	15.46	99.34	4.52
0.8	99.89	9.47	99.60	2.46
0.9	100.00	5.70	100.00	1.36
1.0	100.00	3.45	100.00	0.71

Related Research

Threshold	Training Set		Testing Set	
	Positive(%)	Negative(%)	Positive(%)	Negative(%)
0.3	97.85	7.67	90.54	7.67
0.4	96.11	4.98	87.39	4.98
0.5	94.00	4.25	84.64	4.25
Leung&Kan	89.70	4.88	87.72	6.01

Comparisons for Abstract

- Training part
  - Precision > 90%, when threshold between 0.27 and 0.61
  - Both precision and recall > 94%, when threshold = 0.43
- Testing part
  - Both precision and recall > 90%, when threshold = 0.27
- Training part and testing part
  - Recall > 96% and keep precision > 77%
  - Precision > 97% and keep recall > 71%

Comparisons for Title

- Training Part
  - Precision > 90% and recall > 70%, when threshold = 0.4
  - Both precision and recall > 76%, when threshold = 0.1
- Testing Part
  - Precision > 90% and recall > 70%, when threshold = 0.3
  - Both precision and recall > 78%, when threshold= 0.1
- Training part and testing part
  - Precision = 90%, we can keep the recall above 70%
- The appropriate threshold for this application is 0.27