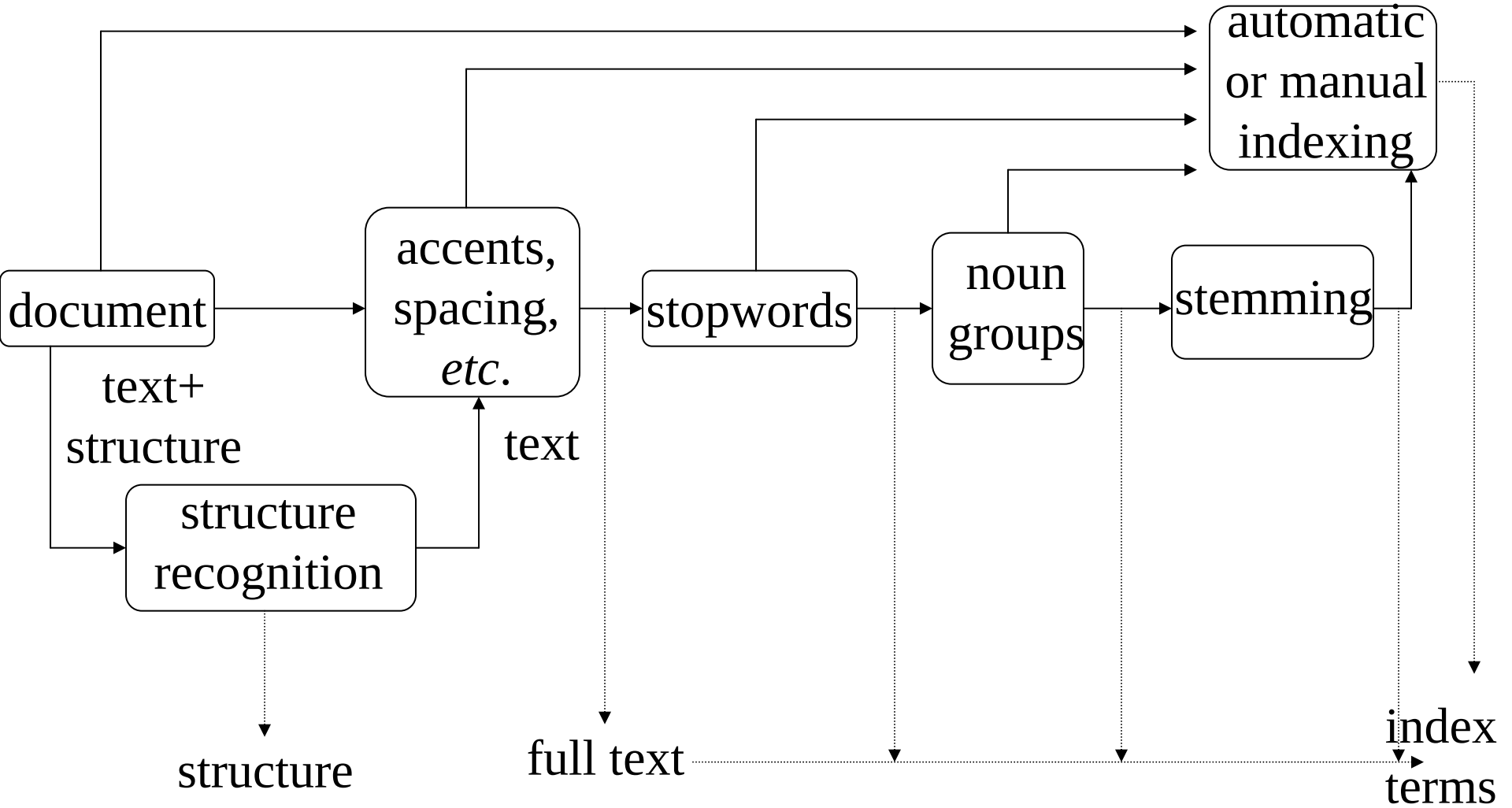
Lecture 2 An Overall View on IR

From full text to a set of index terms



Indexing

- indexing: assign identifiers to text items.
- assign: manual vs. automatic indexing
- identifiers:
  - objective vs. nonobjective text identifiers
    cataloging rules define, e.g., author names, publisher names, dates of publications, …
  - controlled vs. uncontrolled vocabularies
    instruction manuals, terminological schedules, …
  - single-term vs. term phrase

Two Issues

- Issue 1: indexing exhaustivity
  - exhaustive: assign a large number of terms
  - nonexhaustive
- Issue 2: term specificity
  - broad terms (generic)
    cannot distinguish relevant from nonrelevant items
  - narrow terms (specific)
    retrieve relatively fewer items, but most of them are relevant

Parameters of retrieval effectiveness

- Recall

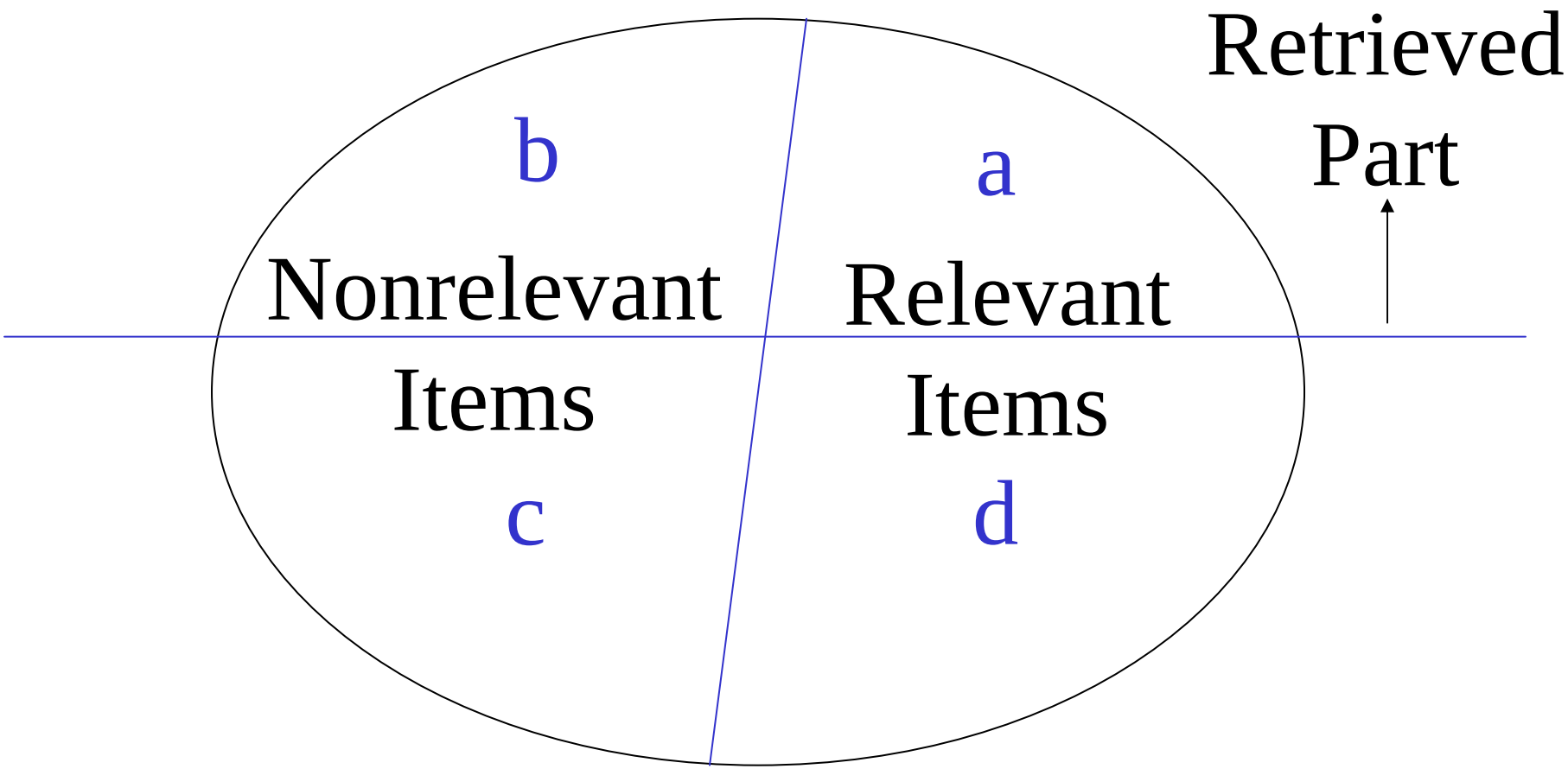
$$R = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items in collection}}$$

- Precision

$$P = \frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}}$$

- Goal

  high recall and high precision



$$\text{Recall} = \frac{a}{a + d}$$

$$\text{Precision} = \frac{a}{a + b}$$

A Joint Measure

- F-score

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

  - β is a parameter that encode the importance of recall and precision.
  - β = 1 or α = ½ : equal weight
  - β>1: precision is more important
  - β<1: recall is more important

Choices of Recall and Precision

- Both recall and precision vary from 0 to 1.
- In principle, the average user wants to achieve both high recall and high precision.
- In practice, a compromise must be reached because simultaneously optimizing recall and precision is not normally achievable.

Choices of Recall and Precision (*Continued*)

- Particular choices of indexing and search policies have produced variations in performance ranging from 0.8 precision and 0.2 recall to 0.1 precision and 0.8 recall.
- In many circumstance, both the recall and the precision varying between 0.5 and 0.6 are more satisfactory for the average users.

Term-Frequency Consideration

- 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
  - varying frequencies in the different texts of a collection
  - indicate term importance for content

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 term frequency 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 $tf_{ij} > T$.

Discussions

- high-frequency terms
  favor recall
- high precision
  the ability to distinguish individual documents from each other
- high-frequency terms
  good for precision when its term frequency is not equally high in all documents.

Inverse Document Frequency

- Inverse Document Frequency (IDF) for term Tj

$$idf_j = \log \frac{N}{df_j}$$

where df*j* (document frequency of term T*j*) is number of documents in which T*j* occurs.

  - fulfil both the recall and the precision
  - occur frequently in individual documents but rarely in the remainder of the collection

New Term Importance Indicator

- weight w_{ij} of a term T_j in a document D_i

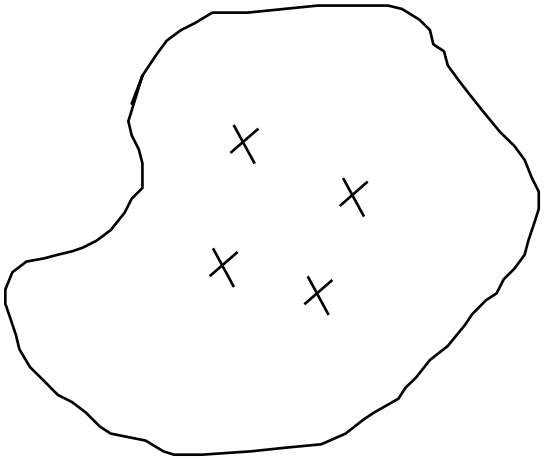
$$w_{ij} = tf_{ij} \times \log \frac{N}{df_j}$$

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

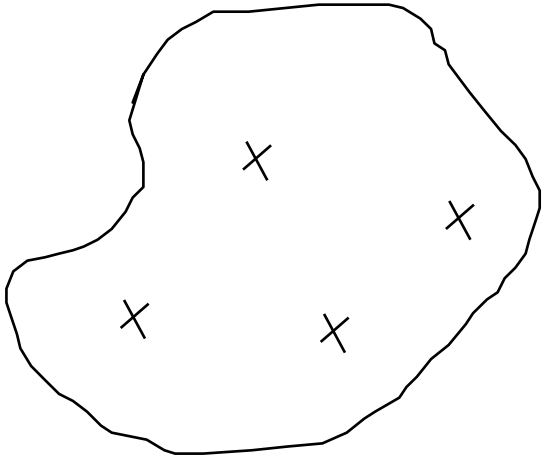
Term-discrimination Value

- Useful index terms
  distinguish the documents of a collection from each other
- Document Space
  - two documents are assigned very similar term sets, when the corresponding points in document configuration appear close together
  - when a high-frequency term without discrimination is assigned, it will increase the document space density

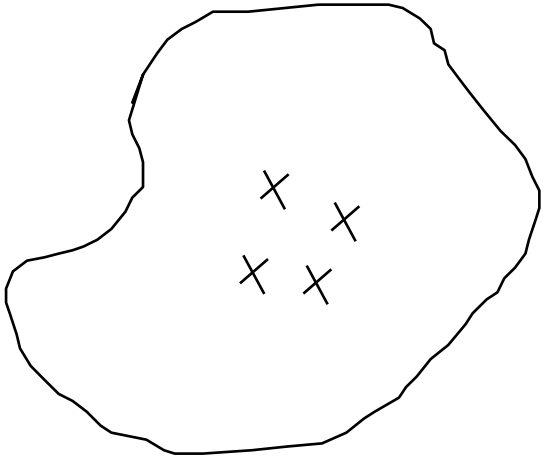
A Virtual Document Space



Original State



After Assignment of good discriminator



After Assignment of poor discriminator

Good Term Assignment

- When a term is assigned to the documents of a collection, the few items to which the term is assigned will be distinguished from the rest of the collection.
- This should increase the average distance between the items in the collection and hence produce a document space less dense than before.

Poor Term Assignment

- A high frequency term is assigned that does not discriminate between the items of a collection.
- Its assignment will render the document more similar.
- This is reflected in an increase in document space density.

Term Discrimination Value

- definition

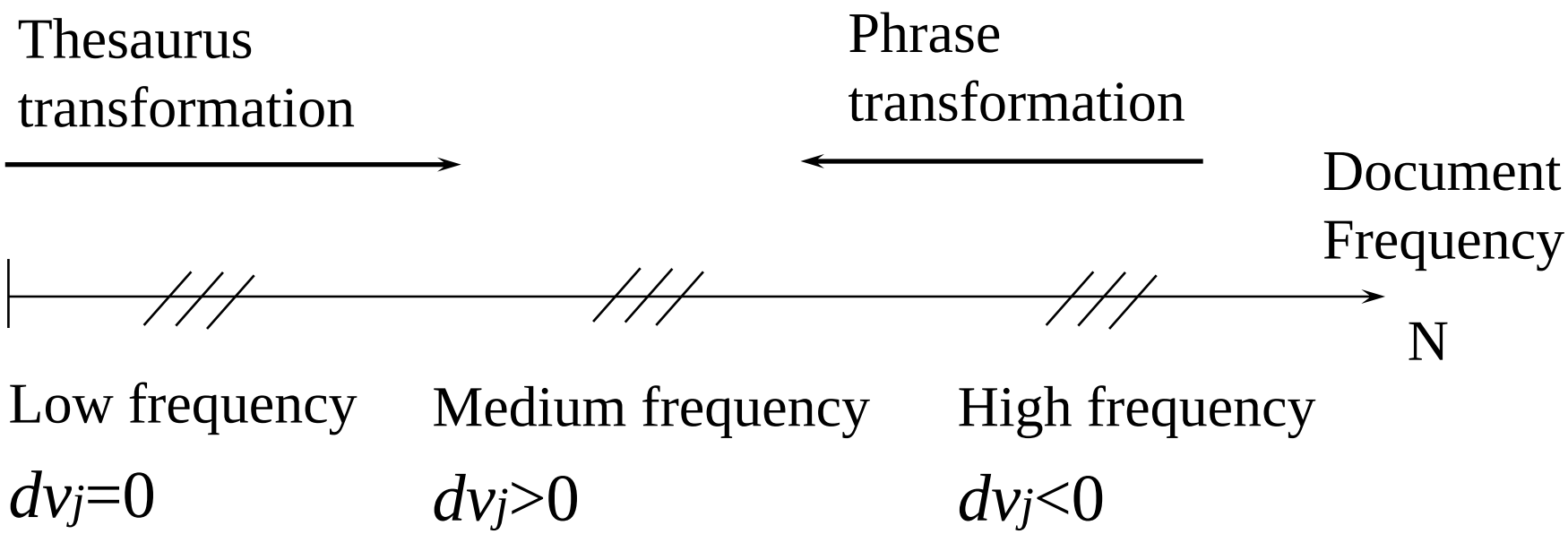
  $dv_j = Q - Q_j$

  where Q and Qj are space densities before and after the assignments of term Tj.

$$Q = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} sim(D_i, D_k)$$

- $dv_j > 0$, T_j is a good term;
  $dv_j < 0$, T_j is a poor term.

Variations of Term-Discrimination Value with Document Frequency



Another Term Weighting

- $w_{ij} = tf_{ij}$ x dv_j
- compared with $w_{ij} = tf_{ij} \times \log \frac{N}{df_j}$
  - $\frac{N}{df_j}$: decrease steadily with increasing document frequency
  - dv_j: increase from zero to positive as the document frequency of the term increase, decrease shapely as the document frequency becomes still larger.

Term Relationships in Indexing

- Single-term indexing
  - Single terms are often ambiguous.
  - Many single terms are either too specific or too broad to be useful.
- Complex text identifiers
  - subject experts and trained indexers
  - linguistic analysis algorithms, e.g., NP chunker
  - term-grouping or term clustering methods

Term Classification (Clustering)

$$\begin{array}{c|cccc} & T_1 & T_2 & T_3 & T_t \\ D_1 & d_{11} & d_{12} & \cdots & d_{1t} \\ D_2 & d_{21} & d_{22} & \cdots & d_{2t} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ D_n & d_{n1} & d_{n2} & \cdots & d_{nt} \end{array}$$

Term Classification (Clustering)

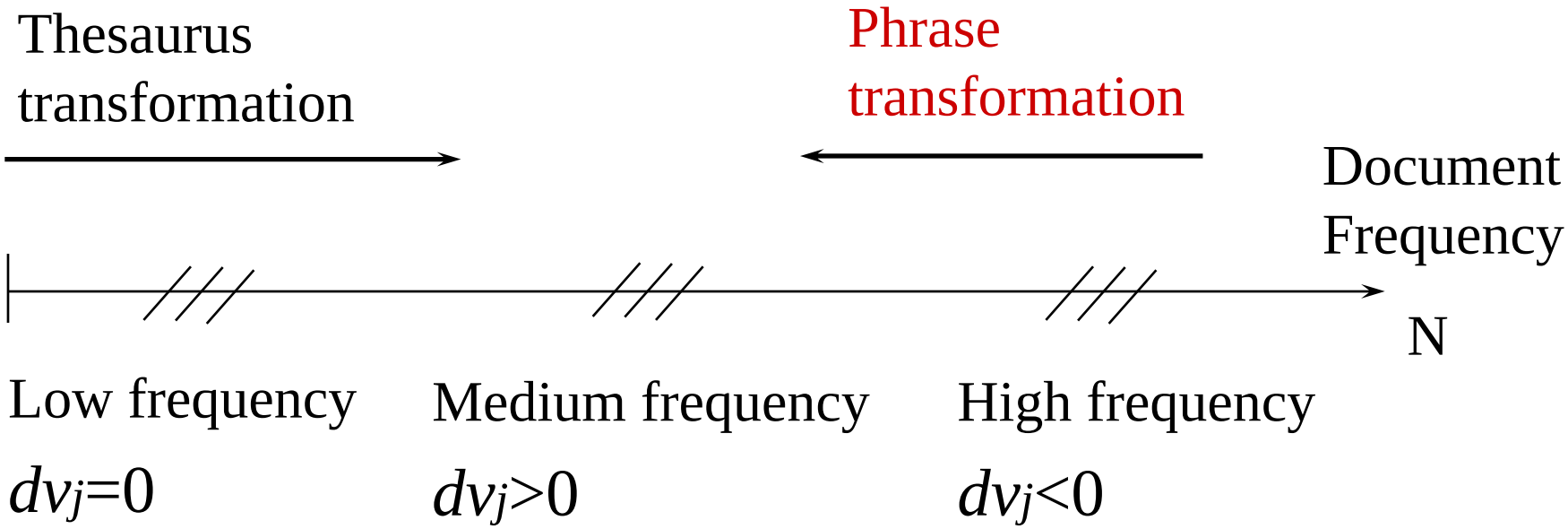
- Column part
  Group terms whose corresponding column representation reveal similar assignments to the documents of the collection.
- Row part
  Group documents that exhibit sufficiently similar term assignment.

Linguistic Methodologies

- Indexing phrases:
  nominal constructions including adjectives and nouns
  - Assign syntactic class indicators (i.e., part of speech) to the words occurring in document texts.
  - Construct word phrases from sequences of words exhibiting certain allowed syntactic markers (noun-noun and adjective-noun sequences).

Term-Phrase Formation

- Term Phrase
  a sequence of related text words carry a more specific meaning than the single terms
  e.g., “computer science” vs. computer;



Simple Phrase-Formation Process

- the principal phrase component (phrase head)
  a term with a document frequency exceeding a stated threshold, or exhibiting a negative discriminator value
- the other components of the phrase
  medium- or low- frequency terms with stated co-occurrence relationships with the phrase head
- common function words
  not used in the phrase-formation process

An Example

- *Effective retrieval systems are essential for people in need of information.*
  - "are", "for", "in" and "of": common function words
  - "system", "people", and "information": phrase heads

The Formatted Term-Phrases

effective retrieval systems essential people need information

Phrase Heads and Components Must Be Adjacent	Phrase Heads and Components Co-occur in Sentence
1. retrieval system*	6. effective systems
2. systems essential	7. systems need
3. essential people	8. effective people
4. people need	9. retrieval people
5. need information*	10. effective information*
	11. retrieval information*
	12. essential information*
2/5	5/12

*: phrases assumed to be useful for content identification

The Problems

- A phrase-formation process controlled only by word co-occurrences and the document frequencies of certain words is not likely to generate a large number of high-quality phrases.
- Additional syntactic criteria for phrase heads and phrase components may provide further control in phrase formation.

Additional Term-Phrase Formation Steps

- Syntactic class indicator are assigned to the terms, and phrase formation is limited to sequences of specified syntactic markers, such as adjective-noun and noun-noun sequences.

  Adverb-adjective × adverb-noun ×

- The phrase elements are all chosen from within the same syntactic unit, such as subject phrase, object phrase, and verb phrase.

Consider Syntactic Unit

- *effective retrieval systems are essential for people in the need of information*
- subject phrase
  - effective retrieval systems
- verb phrase
  - are essential
- object phrase
  - people in need of information

Phrases within Syntactic Components

[$_{subj}$ *effective retrieval systems*] [$_{vp}$ *are essential*]
for [$_{obj}$ *people need information*]

- Adjacent phrase heads and components within syntactic components
  - retrieval systems*
  - people need 2/3
  - need information*
- Phrase heads and components co-occur within syntactic components
  - effective systems

Problems

- More stringent phrase formation criteria produce fewer phrases, both good and bad, than less stringent methodologies.
- Prepositional phrase attachment, e.g.,

  The man saw the girl with the telescope.
- Anaphora resolution

  He dropped the plate on his foot and broke it.

Problems (*Continued*)

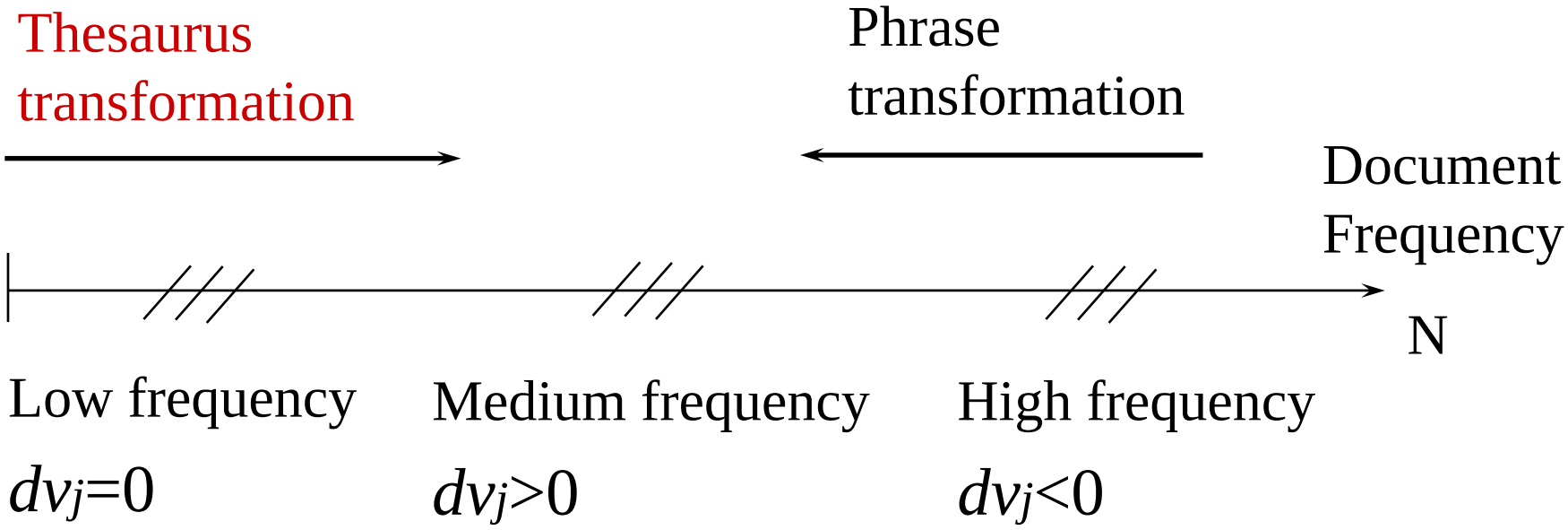
- Any phrase matching system must be able to deal with the problems of
  - synonym recognition
  - differing word orders
  - intervening extraneous word
- Example
  - retrieval of information vs. information retrieval

Equivalent Phrase Formulation

- Base form: text analysis system
- Variants:
  - system analyzes the text
  - text is analyzed by the system
  - system carries out text analysis
  - text is subjected to system analysis
- Related term substitution
  - text: documents, information items
  - analysis: processing, transformation, manipulation
  - system: program, process

Thesaurus-Group Generation

- Thesaurus transformation
  - broadens index terms whose scope is too narrow to be useful in retrieval
  - a thesaurus must assemble groups of related specific terms under more general, higher-level class indicators



Sample Classes of Roget's Thesaurus

Class Indicator	Entry	Class Indicator	Entry
760	permission	763	offer
	leave		presentation
	sanction		tender
	allowance		overture
	tolerance		advance
	authorization		submission
761	prohibition		proposal
	veto		proposition
	disallowance		invitation
	injunction	764	refusal
	ban		declining
	taboo		noncompliance
762	consent		rejection
	acquiescence		denial
	compliance		
	agreement		
	acceptance		

同義詞詞林

- 12 large categories
- 94 middle categories
- 1,428 small categories
- 3,925 word clusters

A	People		
Aa	a collective name		
01	Human being	The people	Everybody
02	I	We	
03	You	You	
04	He/She	They	
05	Myself	Others	Someone
06	Who		
Ab	people of all ages and both sexes		
01	A Man	A Woman	Men and Women
02	An Old Person	An Adult	The old and the young
03	A Teenager		
04	An Infant	A Child	
Ac	posture		
01	A Tall Person	A Dwarf	
02	A Fat Person	A Thin Person	
03	A Beautiful Woman	A Handsome Man	

A. PERSON (人): **Aa**. general name (泛稱), **Ab**. people of all ages and both sexes (男女老少), **Ac**. posture (體態), **Ad**. nationality/citizenship (籍屬), **Ae**. occupation (職業), **Af**. identity (身分), **Ag**. situation (狀況), **Ah**. relative/family dependents (親人/眷屬), **Ai**. rank in the family (輩次), **Aj**. relationship (關係), **Ak**. morality (品行), **Al**. ability and insight (才識), **Am**. religion (信仰), **An**. comic/clown type (丑類)
B. THING (物): **Ba**. generally called (統稱), **Bb**. (擬狀物), **Bc**. part of an object (物體的部分), **Bd**. a celestial body (天體), **Be**. terrian features (地貌), **Bf**. meteorological phonomena (氣象), **Bg**. natural substance (自然物), **Bh**. plant (植物), **Bi**. animals (動物), **Bj**. micro-organism (微生物), **Bk**. the whole body (全身), **Bl**. secretions/excretions (排泄物/分泌物), **Bm**. Material (材料), **Bn**. Building (建築物), **Bo**. machines and tools (機具), **Bp**. appliances (用品), **Bq**. Clothing (衣物), **Br**. edibles/medicines/drugs (食品/藥物/毒品)
C. TIME AND SPACE (時間與空間): **Ca**. time (時間), **Cb**. space (空間)
D. ABSTRACT THINGS (抽象事物): **Da**. event/circumstances (事情/情況), **Db**. reason/logic (事理), **Dc**. looks (外貌), **Dd**. functions/properties (性能), **De**. character/ability (性格/才能), **Df**. conscious (意識), **Dg**. analogical thing (比喻物), **Dh**. imaginary things (臆想物), **Di**. society/politics (社會/政法), **Dj**. economy (經濟), **Dk**. culture and education (文教), **Dl**. disease (疾病), **Dm**. Organization (機構), **Dn**. quantity/unit (數量/單位)
E. CHARATERISTICS (特徵): **Ea**. external form (外形), **Eb**. surface looks/seeming (表象), **Ec**. color/taste (顏色/味道), **Ed**. Property (性質), **Ee**. virtue and ability (德才), **Ef**. Circumstances (境況)
F. MOTION (動作). **Fa**. motion of upper limbs (hands) (上肢動作), **Fb**. motion of lower limbs (legs) (下肢動作), **Fc**. motion of head (頭部動作), **Fd**. motion of the whole body (全身動作)
G. PSYCHOLOGICAL ACTIVITY (心理活動): **Ga**. state of mind (心理狀態), **Gb**. activity of mind (心理活動), **Gc**. capability and willingness (能/願)
H. ACTIVITY (活動): **Ha**. political activity (政治活動), **Hb**. military activity (軍事活動), **Hc**. administrative management (行政管理), **Hd**. Production (生產), **He**. economical activity (經濟活動), **Hf**. communications and transportation (交通運輸), **Hg**. education and hygiene scientific research (教衛科研), **Hh**. recreational and sports activities (文體活動), **Hi**. social contact (社交), **Hj**. Life (生活), **Hk**. religionary activity (宗教活動), **Hl**. superstitious belief activity (迷信活動), **Hm**. public security and judicature (公安/司法), **Hn**. wicked behavior (惡行)
I. PHENOMENON AND CONDITION (現象與狀態): **Ia**. natural phenomena (自然現象), **Ib**. physiology phenomena (生理現象), **Ic**. facial expression (表情), **Id**. object status (物體狀態), **Ie**. Situation (事態), **If**. circumstances (mostly unlucky) (境遇), **Ig**. the beginning and the end (始末), **Ih**. Change (變化)
J. TO BE RELATED (關聯): **Ja**. association (聯繫), **Jb**. similarities and dissimilarities (異同), **Jc**. to operate in coordination (配合), **Jd**. existence (存在), **Je**. Influence (影響)
K. AUXILIARY PHRASE (助語): **Ka**. quantitative modifier (疏狀), **Kb**. preposition (中介), **Kc**. conjunction (聯接), **Kd**. auxiliary (輔助), **Ke**. interjection (呼嘆), **Kf**. Onomatopoeia (擬聲)
L. GREETINGS (敬語)

The Indexing Prescription (1)

- Identify the individual words in the document collection.
- Use a stop list to delete from the texts the function words.
- Use an suffix-stripping routine to reduce each remaining word to word-stem form.
- For each remaining word stem T_j in document D_i, compute w_{ij}.
- Represent each document D_i by

$$D_i=(T_1, w_{i1}; T_2, w_{i2}; \ldots, T_t, w_{it})$$

Word Stemming

- effectiveness --> effective --> effect
- picnicking --> picnic
- king -\-> k

Some Morphological Rules

- Restore a silent e after suffix removal from certain words to produce “hope” from “hoping” rather than “hop”
- Delete certain doubled consonants after suffix removal, so as to generate “hop” from “hopping” rather than “hopp”.
- Use a final y for an i in forms such as “easier”, so as to generate “easy” instead of “easi”.

The Indexing Prescription (2)

- Identify individual text words.
- Use stop list to delete common function words.
- Use automatic suffix stripping to produce word stems.
- Compute term-discrimination value for all word stems.
- Use thesaurus class replacement for all low-frequency terms with discrimination values near zero.
- Use phrase-formation process for all high-frequency terms with negative discrimination values.
- Compute weighting factors for complex indexing units.
- Assign to each document single term weights, term phrases, and thesaurus classes with weights.

Query vs. Document

- Differences
  - Query texts are short.
  - Fewer terms are assigned to queries.
  - The occurrence of query terms rarely exceeds 1.

$Q=(w_{q1}, w_{q2}, \ldots, w_{qt})$ where w_{qj}: inverse document frequency
$D_i=(d_{i1}, d_{i2}, \ldots, d_{it})$
where d_{ij}: term frequency*inverse document frequency

$$sim(Q,D) = \sum_{j=1}^{t} w_{qj} \bullet d_{ij}$$

Query vs. Document

- When non-normalized documents are used, the longer documents with more assigned terms have a greater chance of matching particular query terms than do the shorter document vectors.

$$sim(Q, Di) = \frac{\sum_{j=1}^{t} w_{qj} \bullet d_{ij}}{\sqrt{\sum_{j=1}^{t} (d_{ij})^2}}$$

or

$$sim(Q, Di) = \frac{\sum_{j=1}^{t} w_{qj} \bullet d_{ij}}{\sqrt{\sum_{j=1}^{t} (d_{ij})^2 \bullet \sum_{j=1}^{t} (w_{qj})^2}}$$

Relevance Feedback

- Terms present in previously retrieved documents that have been identified as relevant to the user's query are added to the original formulations.
- The weights of the original query terms are altered by replacing the inverse document frequency portion of the weights with term-relevance weights obtained by using the occurrence characteristics of the terms in the previous retrieved relevant and nonrelevant documents of the collection.

Relevance Feedback

- $Q = (w_{q1}, w_{q2}, ..., w_{qt})$
- $D_i = (d_{i1}, d_{i2}, ..., d_{it})$
- New query may be the following form
  $Q' = \alpha\{w_{q1}, w_{q2}, ..., w_{qt}\}+\beta\{w'_{qt+1}, w'_{qt+2}, ..., w'_{qt+m}\}$
- The weights of the newly added terms T_{t+1} to T_{t+m} may consist of a combined term-frequency and term-relevance weight.

Final Indexing

(1) Identify individual text words.
(2) Use a stop list to delete common words.
(3) Use suffix stripping to produce word stems.
(4) Replace low-frequency terms with thesaurus classes.
(5) Replace high-frequency terms with phrases.
(6) Compute term weights for all single terms, phrases, and thesaurus classes.
(7) Compare query statements with document vectors.
(8) Identify some retrieved documents as relevant and some as nonrelevant to the query.

Final Indexing

(9) Compute term-relevance factors based on available relevance assessments.

(10) Construct new queries with added terms from relevant documents and term weights based on combined frequency and term-relevance weight.

(11) Return to step (7).
Compare query statements with document vectors ……..

Summary of expected effectiveness of automatic indexing

- Basic single-term automatic indexing -
- Use of thesaurus to group related terms in the given topic area +10% to +20%
- Use of automatically derived term associations obtained from joint term assignments found in sample document collections 0% to -10%
- Use of automatically derived term phrases obtained by using co-occurring terms found in the texts of sample collections +5% to +10%
- Use of one iteration of relevant feedback to add new query terms extracted from previously retrieved relevant documents +30% to +60%