Day 2 & 3 & 4

Query is not Info Need (they are different)

Salsa -> Need lessons for dance/ Wish to cook Jpn cuisine with salsa sauce/ How can I use up my salsa in two days?

Recipe salsa -> Wish to cook Jpn cuisine with salsa sauce/ How can I use up my salsa in two days?

How to cook salsa -> How can I use up my salsa in two days?/ Wish to cook Jpn cuisine with salsa sauce

Relevance in principle

- The relevance of document D wrt need X is the degree to which D satisfies X, whereas only query Q is submitted to an IR system
- By definition, no relevance of D exists
- IR system predicts the need behind the query

X -> Q -> IR
match↓
Relevance↖↖D (D must be returned to satisfy need X)

Relevance in practice: Two majorities

- **Topical R (relevance)** ≈ subject R (has more than one topics)≈ aboutness ← query
  - Static and superficial match between Q&D (doesn't depend on users but environment?)
  - R (jack-o-lantern, Halloween) is always 1
- Tends to define objectivity relevance

- **User-centered R** ≈ subjective R ← info need
  - Judgment can be different depending on the individual user
  - Situational R

  R(X, clip art) = 1 -> 0 **What is X?**

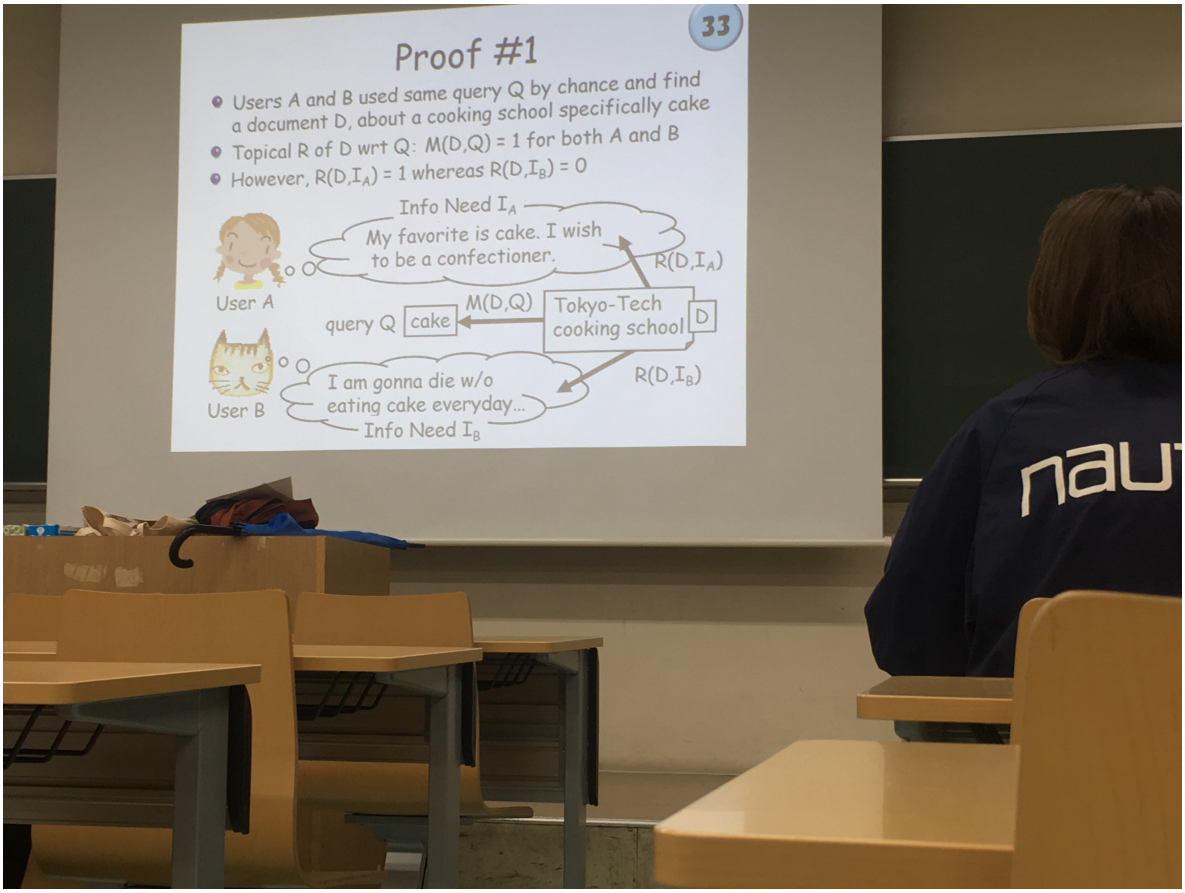
  R(X, lottery) = 0 -> 1 Relevance of X changes as soon as new year begins

Cranfield experiments (1960s)

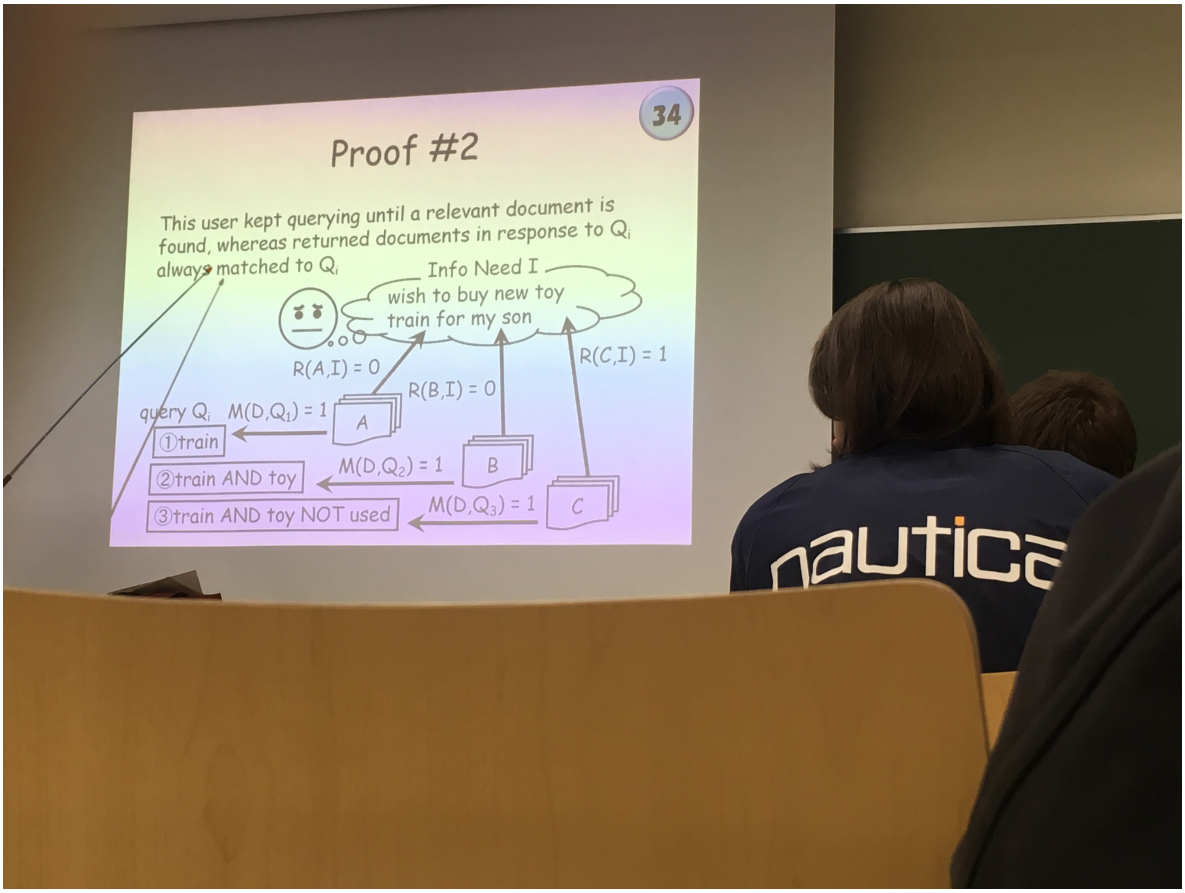
- Goal: benchmark test for indexing systems
  - Topic: description for info need
  - Document collection: which gives rel-docs
  - Relevance judgement: rel-docs for each topic
- criticism: **"topical R" ≈ query ≠ info need**
  - Dynamic nature of users and environment during search is not considered
- However, recent paradigm is a successor of Cranfield experiments and its derivative
  - Example: TREC (text retrieval conference)

Exercise

- Prove the following proposition is false
- The topical R of document D wrt query Q the same as the relevance of D wrt info need
- Proof #1
  - Users A and B used same query Q by chance and find a document D about a cooking school specifically cake
  - Topical R of D wrt Q: M(D,Q) = 1 for both A and B
  - However, R(D,I_a) = 1 whereas R(D, I_b) = 0
    - User A Info Need I_a: my favorite is cake. I wish to be a confectioner.
    - User B Info Need I_b: I am gonna die without eating cake everyday...

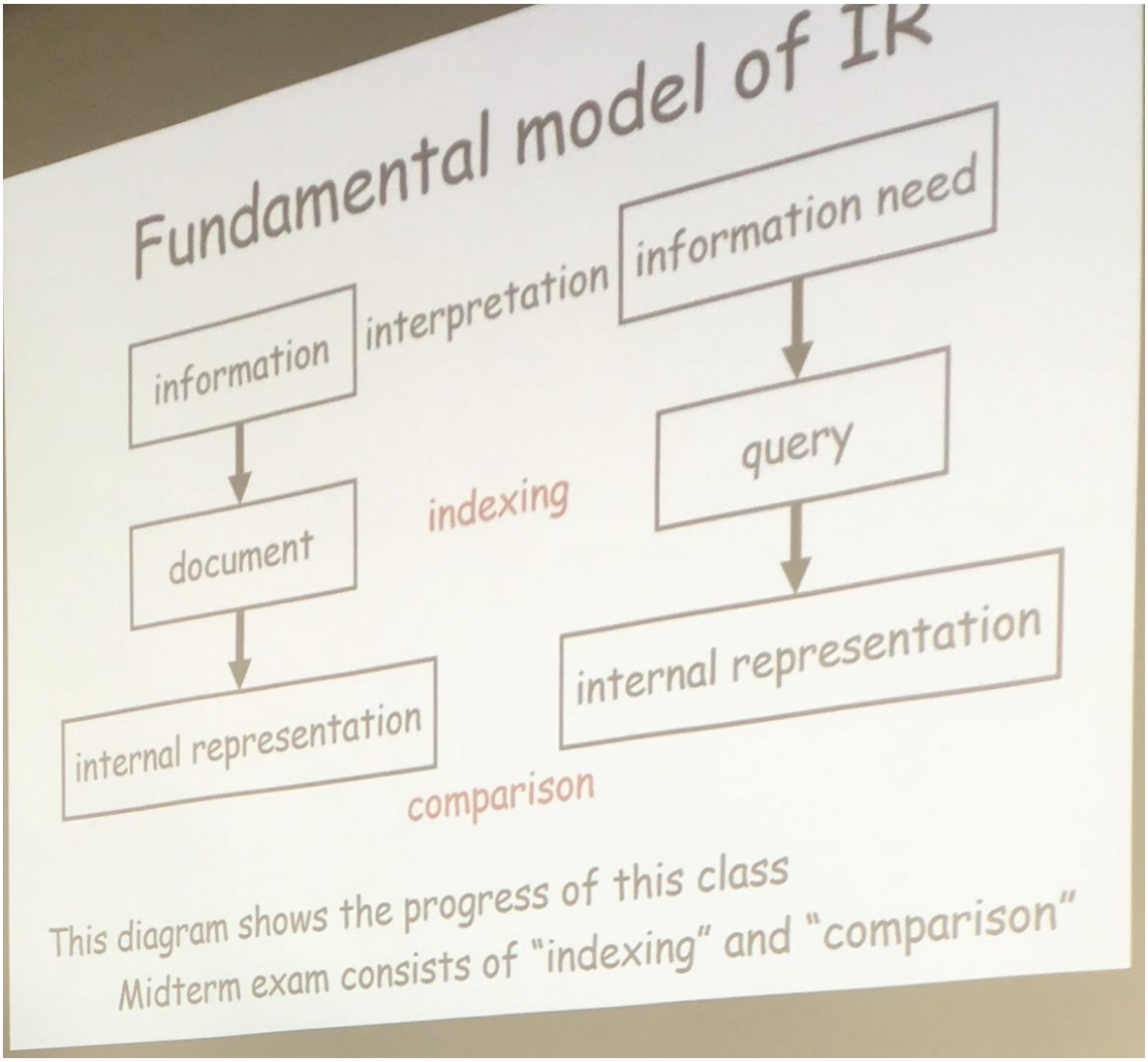


- Proof #2



- Hint
  - Imagine two users coincidently submit the same query driven by different desires
  - Imagine a user repeats modifying and submitting their query for a single investigation

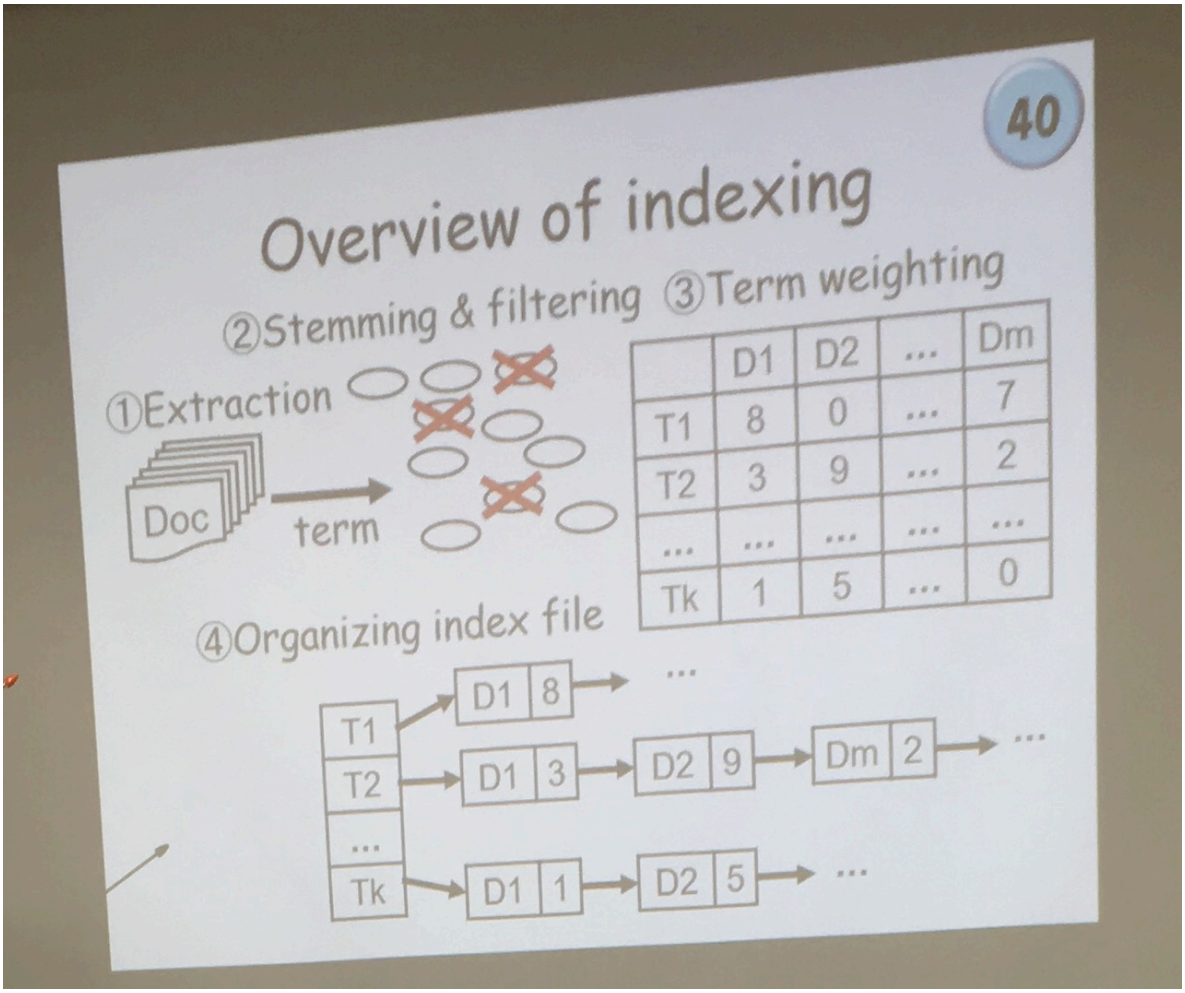
Fundamental model of IR



Indexing

- Producing an index from a set of documents
  - Given a key, index returns its value quickly
  - Book: key = word/phrase value = page
  - IR: key = char/word/phrase (index) term value = docID
- Improve the following two Es
  - Efficiency: time/space complexity
  - Effectiveness: quality of system's output

Overview of indexing



Requirement for IR

- IR systems are required to retrieve documents [(A) **accurately**] and [(B) **exhaustively**]
  - Trade-off
- Indexing is the first step to determine the upperbound for your system
- What are “good” terms?
  - Specificity: restrict # of candidate docs
  - Exhaustiveness: catch as many potentially promising candidate docs as possible

Language-specific problems NLP

- If sentence is space-delimited, as in English, extracting indexing terms is trivial
- In agglutinative languages, NLP is effective
  - A) Each phrase = a content word + one or more suffixes following inflection rule
  - B) Sentence has no lexical segmentation (jp)
    sentence is space-delimited phrases (mn)
  - For A) hand-cragted inflection rule that identifies original forms is the first step
  - For B) “Morphological analysis” that segments input Japanese text into

words

Language-independent unit

- (Character) N-gram, N is the number of objects
  - A sequence of N charactersin documents
  - [Example] we study indexing, N = 2
  - Ans: we es st tu ud dy yi in nd de ex xi in ng
  - we e_ _s st tu ud dy y_ _i in nd de ex xi in ng
  - _w we e_ _s st tu ud dy y_ _i in nd de ex xi in ng_
  - In the 2nd case: nonexper/ non-expert can be matched

What is N-gram good for?

- For agglutinative language: definitely yes
  - N=2: two-kanji words, such as 情报 and 检索 can exhaustively be collected in Chinese and Japanese
- How about for English?
  - Robust against variants of the same word, such as theater/theatre & dialog/ dialogue
  - Similar reasons: typographical errors, simple abbreviations
  - When increasing exhaustivity, accuracy is decreasing

Stemming

- Standardize inflected words in documents and query so that the same word can be matched
- A set of manually produced rules
  - Eg. porter[1980] is a popular one
  - ***sses or ***ies ->remove es
  - Available to the public online: libraries for perl or python (approx. 100 codes)
- Effectiveness of stemming is not clear
  - Expect for highly inflected languages, such as Finnish and Mongolian

Filtering Stopwords

- Unfortunately, not all candidates are useful for IR, on the contrary no use or even worsen
  - They are called **stopword** and filtered out
  - Two properties associated with stopwords
    - **High frequent words**: such as “the”, which is no specificity because appears in most of the documents (but no universal criteria)
    - **Function words**: article, preposition, auxiliary verb, conjunction

    the word itself has little meaning but serves a grammatical function

Exercise

A) what is the side effect of stemming?
We lose some information. Also, some words change meaning in plural: glass/ glasses
Googe -> goos; geese -> gees => not equal but should be????
Meaness, meaning -> mean => false equal????

B) what is the side effect of stop word filtering?
We might block something important by accident.
Eg. Beatles <-> The Beatles

- +Novel titles/ movies - include func words
- Ppl search for entire sentences

Stopwords are no longer used.