Classic Information Retrieval continued

1

continue defining model for text document search

Last time

- *Document* is
  - Set of terms
  - Bag of terms
  - Sequence of terms
- Each choice has consequences
- "term" is used instead of "word" to signal more general possibilities: serial numbers, nonsense, etc.

2

Modeling: "*query*"

Continue with

- *Query*
  - Basic query is one term
  - Multi-term query is
    - List of terms
      - OR model: *some* terms
      - AND model: *all* terms
    - Boolean combination of terms
    - Other constraints?

3

Modeling: "*satisfying*"

- What determines if document satisfies query?
- That depends ….
  - Document model
  - Query model
- *START SIMPLE*
  - *better understanding*
  - *Use components of simple model later*

4

(pure) Boolean Model of IR

- Document: *set* of terms
- Query: boolean expression over terms
- Satisfying:
  - Doc. evaluates to "true" on single-term query if contains term
  - Evaluate doc. on expression query as you would any Boolean expression
  - doc satisfies query if evals to true on query

5

Boolean Model example

Doc 1: "Computers have brought the world to our fingertips. We will try to understand at a basic level the science -- old and new -- underlying this new Computational Universe. Our quest takes us on a broad sweep of scientific knowledge and related technologies… Ultimately, this study makes us look anew at ourselves -- our genome; language; music; "knowledge"; and, above all, the mystery of our intelligence. (cos 116 description)

Doc 2: "An introduction to computer science in the context of scientific, engineering, and commercial applications. The goal of the course is to teach basic principles and practical issues, while at the same time preparing students to use computers effectively for applications in computer science …" (cos 126 description)

Query:

(principles AND knowledge) OR (science AND engineering)

6

Boolean Model example

Doc 1: "Computers have brought the world to our fingertips. We will try to understand at a basic level the science -- old and new -- underlying this new Computational Universe. Our quest takes us on a broad sweep of scientific knowledge and related technologies… Ultimately, this study makes us look anew at ourselves -- our genome; language; music; "knowledge"; and, above all, the mystery of our intelligence. (cos 116 description)

Doc 2: "An introduction to computer science in the context of scientific, engineering, and commercial applications. The goal of the course is to teach basic principles and practical issues, while at the same time preparing students to use computers effectively for applications in computer science …" (cos 126 description)

Query:

(principles AND knowledge) OR (science AND engineering)

0 1 1 0

Doc 1: FALSE

Boolean Model example

Doc 1: "Computers have brought the world to our fingertips. We will try to understand at a basic level the science -- old and new -- underlying this new Computational Universe. Our quest takes us on a broad sweep of scientific knowledge and related technologies… Ultimately, this study makes us look anew at ourselves -- our genome; language; music; "knowledge"; and, above all, the mystery of our intelligence. (cos 116 description)

Doc 2: "An introduction to computer science in the context of scientific, engineering, and commercial applications. The goal of the course is to teach basic principles and practical issues, while at the same time preparing students to use computers effectively for applications in computer science …" (cos 126 description)

Query:

(principles AND knowledge) OR (science AND engineering)

1 0 1 1

Doc 2: TRUE

Boolean Model example 2

Doc 1: "Computers have brought the world to our fingertips. We will try to understand at a basic level the science -- old and new -- underlying this new Computational Universe. Our quest takes us on a broad sweep of scientific knowledge and related technologies… Ultimately, this study makes us look anew at ourselves -- our genome; language; music; "knowledge"; and, above all, the mystery of our intelligence. (cos 116 description)

Doc 2: "An introduction to computer science in the context of scientific, engineering, and commercial applications. The goal of the course is to teach basic principles and practical issues, while at the same time preparing students to use computers effectively for applications in computer science …" (cos 126 description)

Query: (principles OR knowledge) AND
(science AND NOT(engineering))

Boolean Model example 2

Doc 1: "Computers have brought the world to our fingertips. We will try to understand at a basic level the science -- old and new -- underlying this new Computational Universe. Our quest takes us on a broad sweep of scientific knowledge and related technologies… Ultimately, this study makes us look anew at ourselves -- our genome; language; music; "knowledge"; and, above all, the mystery of our intelligence. (cos 116 description)

Doc 2: "An introduction to computer science in the context of scientific, engineering, and commercial applications. The goal of the course is to teach basic principles and practical issues, while at the same time preparing students to use computers effectively for applications in computer science …" (cos 126 description)

Query: (principles OR knowledge) AND
(science AND NOT(engineering))

Doc 1: (0 OR 1) AND (1 AND NOT(0)) TRUE

Boolean Model example 2

Doc 1: "Computers have brought the world to our fingertips. We will try to understand at a basic level the science -- old and new -- underlying this new Computational Universe. Our quest takes us on a broad sweep of scientific knowledge and related technologies… Ultimately, this study makes us look anew at ourselves -- our genome; language; music; "knowledge"; and, above all, the mystery of our intelligence. (cos 116 description)

Doc 2: "An introduction to computer science in the context of scientific, engineering, and commercial applications. The goal of the course is to teach basic principles and practical issues, while at the same time preparing students to use computers effectively for applications in computer science …" (cos 126 description)

Query: (principles OR knowledge) AND
(science AND NOT(engineering))

Doc 2: (1 OR 0) AND (1 AND NOT(1)) FALSE

(pure) Boolean Model of IR: address "*present in useful form*"

- can mean user interface
- more basic: give list
- meaning of order of list? => RANKING?

- There is no ranking algorithm in pure Boolean model
  - Ideas for making one without changing semantics of "satisfy"?

Doc 1: "Computers have brought the world to our fingertips. We will try to understand at a basic level the science -- old and new -- underlying this new Computational Universe. Our quest takes us on a broad sweep of scientific knowledge and related technologies… Ultimately, this study makes us look anew at ourselves -- our genome; language; music; "knowledge"; and, above all, the mystery of our intelligence. (cos 116 description)

Doc 2: "An introduction to computer science in the context of scientific, engineering, and commercial applications. The goal of the course is to teach basic principles and practical issues, while at the same time preparing students to use computers effectively for applications in computer science …" (cos 126 description)

Query:

(principles OR knowledge) AND (science OR engineering)

Doc 1:	0	1	0	TRUE
Doc 2:	1	0	1	TRUE

RANK?

13

Next Model: Vector Model

- Document: *bag* of terms
- Query: list of terms
- Satisfying:
  - Each document is scored as to the degree it satisfies query (non-negative real number)
  - doc satisfies query if its score is >0
  - Documents are returned in sorted list decreasing by score:
    - Include only non-zero scores
    - Include only highest *n* documents, some *n*

14

How compute score?

1. There is a *dictionary* (aka *lexicon*) of all terms, numbering *t* in all
   - Number the terms 1, …, *t*
2. Change the model of a document (temporarily):
   - A document is a *t*-dimensional vector
   - The i^{th} entry of the vector is the *weight* (importance of) term *i* in the document
3. Change the model of a query (temporarily):
   - A query is a *t*-dimensional vector
   - The i^{th} entry of the vector is the *weight* (importance of) term *i* in the query

15

How compute score, continued

4. Calculate a vector function of the document vector and the query vector to get the score of the document with respect to the query.

Choices:

1. Measure the distance between the vectors:

$$\text{Dist}(\boldsymbol{d},\boldsymbol{q}) = \sqrt{(\Sigma_{i=1}^{t}(\boldsymbol{d}_i - \boldsymbol{q}_i)^2)}$$

- Is *dissimilarity* measure
- Not normalized: Dist ranges [0, inf.)
- Fix: use $e^{-\text{Dist}}$ with range (0,1]
- Is it the right sense of difference?

16

How compute score, continued

2. Measure the angle between the vectors:

Dot product: $\boldsymbol{d}\bullet\boldsymbol{q} = \Sigma_{i=1}^{t}(\boldsymbol{d}_i * \boldsymbol{q}_i)$

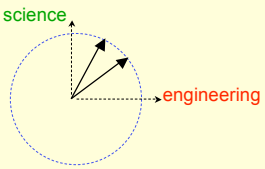
- Is *similarity* measure
- Not normalized: dot product ranges (-inf., inf.)
- Fix: use normalized dot product, range [-1,1]

  $(\boldsymbol{d}\bullet\boldsymbol{q}) / (|\boldsymbol{d}|*|\boldsymbol{q}|)$ where $|\boldsymbol{v}| = \sqrt{\Sigma_{i=1}^{t}(\boldsymbol{v}_i^2)}$
- In practice vector components are non-negative so range is [0,1]
- This most commonly used function for score

17

Normalizing vectors

- If use unit vectors, $\boldsymbol{d} / |\boldsymbol{d}|$ and $\boldsymbol{v} / |\boldsymbol{v}|$ some of issues go away



18

How compute weights d_i and q_i?

First:

observations about this model?

19

Vector model: Observations

- Have matrix of terms by documents
  ⇒Can use linear algebra
- Queries and documents are the same
  ⇒Can compare documents same way
  - Clustering documents
- Document with only some of query terms can score higher than document with all query terms

20

How compute weights

- Vector model *could* have weights assigned by human intervention
  - may add meta-information
  - User setting query weights might make sense
    - User decides importance of terms in own search
  - Humans setting document weights?
    - Who? Billions+ of documents
- Return to model of documents as bag of words – calculate weights
  - Function mapping bag of words to vector

21

Calculations on board

22

Summary weight calculation

- General notation:
  - w_{jd} is the weight of term *j* in document *d*
  - $freq_{jd}$ is the # of times term *j* appears in doc *d*
  - n_j = # docs containing term j
  - *N* = number of docs in collection
- Classic *tf-idf* definition of weight:
  $w_{jd} = freq_{jd} * log(N/ n_j)$

23

Weight of query components?

- Set (list) of terms, ***some choices***:
  1. w_{jq} = 0 or 1
  2. $w_{jq} = freq_{jq} * log(N/ n_j)$
     $= 0 \text{ or } log(N/ n_j)$
- Bag of terms
  - Analyze like document
    Some queries are prose expressions of ***information need***

Do we want idf term in both document weights and query weights?

24

Vector Model example

Doc 1: "Computers have brought the world to our fingertips. We will try to understand at a basic level the science -- old and new -- underlying this new Computational Universe. Our quest takes us on a broad sweep of scientific knowledge and related technologies… Ultimately, this study makes us look anew at ourselves -- our genome; language; music; "knowledge"; and, above all, the mystery of our intelligence. (cos 116 description)

Frequencies:

science 1; knowledge 2; principles 0; engineering 0

Doc 2: "An introduction to computer science in the context of scientific, engineering, and commercial applications. The goal of the course is to teach basic principles and practical issues, while at the same time preparing students to use computers effectively for applications in computer science …" (cos 126 description)

Frequencies:

science 2; knowledge 0; principles 1; engineering 1

25

Vector model example cont.

- Consider the 5 100-level and 200-level COS courses as the collection (109, 217, 226)
- Only other appearance of our 4 words is "science" once in 109 description.
- idf:
  science ln(5/3) = .51
  engineering, principles, knowledge:
  ln(5/1) = 1.6

26

Term by Doc. Table: $freq_{jd}$ * $log(N/ n_j)$

	Doc 1	Doc 2
science	.51	1.02
engineering		1.6
principles		1.6
knowledge	3.2	

27

Unnormalized dot product for query:
science, engineering, knowledge, principles
using 0/1 query vector

- Doc 1: 3.71
- Doc 2: 4.22

- If documents have about same vector length, this right ratio for normalized (cosine) score

28

Additional ways to calculate document weights

- Dampen frequency effect:
  $w_{jd} = 1 + log\ (freq_{jd})$ *if* $freq_{jd} > 0$; *0 otherwise*
- *Use smoothing term to dampen effect:*
  $W_{jd} = a + (1-a)\ freq_{jd} / max_p\ (freq_{pd})$
  - *a is typically .4 or .5*
  - *Can multiply second term by idf*
- *Effects for long documents (Section 6.4.4)*

29

Where get dictionary of *t* terms?

- Pre-determined dictionary.
  – How sure get all terms?
- Build lexicon when collect documents
  – What if collection dynamic: add terms?

30

Query models advantages

- Boolean
  - No ranking in pure
  + Get power of Boolean Algebra:
    expressiveness
    optimization of query forms
- Vector
  + Query and document look the same
  + Power of linear algebra
  - No requirement all terms present in pure

31

Classic IR models - Taxonomy

- Boolean
- Vector
  – bag of words
- Probabilistic

32