Classic Information Retrieval II

1

Modeling: "*document*", "*query*", "*satisfying*"

(ignore "*presenting in useful form*" for now)

Last time

- *Document* is
  - Set of terms
  - Bag of terms
  - Sequence of terms

Continue with

- *Query* is ????

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

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 OR knowledge) AND (science AND NOT(engineering))

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 OR knowledge) AND (science AND NOT(engineering))
Doc 1: 0 1 1 0 TRUE

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 OR knowledge) AND (science AND NOT(engineering))
Doc 2: 1 0 1 1 FALSE

(pure) Boolean Model of IR: come to "*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"?

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 OR knowledge) AND (science OR engineering)

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

RANK?

13

Next Model: Vector Model

- Document: *bag* of terms
- Query: list of terms (could be bag 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

Choices:

2. Measure the angle between the vectors:

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

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

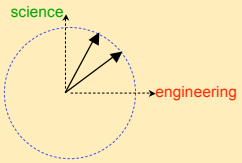
$$(\boldsymbol{d}\cdot\boldsymbol{q}) / (|\boldsymbol{d}|*|\boldsymbol{q}|) \quad \text{where } |\boldsymbol{v}| = \sqrt{\Sigma_{i=1}^{t}(\boldsymbol{v}_i^2)}$$

   - 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: what do you observe 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
- User setting query weights might make sense
  - User decides importance of terms in own search
- Someone setting document weights makes no sense
  - Huge number documents – billions
- Return to model of documents as bag of words – calculate weights

21

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:
  $w_{jd} = freq_{jd} * log(N/ n_j)$

22