Information Retrieval

Vannevar Bush

- Director of the Office of Scientific Research and Development (1941-1947)


Vannevar Bush, 1890-1974

- End of WW2 - what next big challenge for scientists?

Historic Vision

"A memex is a device in which an an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

Vannevar Bush, As we may think, *Atlantic Monthly*, July 1945.

Prophetic: Hypertext

- "associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing."

Vannevar Bush, As we may think, *Atlantic Monthly*, July 1945

Prophetic: Wikipedia et al

- "Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified."

Vannevar Bush, As we may think, *Atlantic Monthly*, July 1945

Vision

" This is a much larger matter than merely the extraction of data for the purposes of scientific research; it involves the entire process by which man profits by his inheritance of acquired knowledge"

Vannevar Bush, As we may think, *Atlantic Monthly*, July 1945

"The applications of science have built man a well-supplied house, and are teaching him to live healthily therein. They have enabled him to throw masses of people against one another with cruel weapons. They may yet allow him truly to encompass the great record and to grow in the wisdom of race experience. He may perish in conflict before he learns to wield that record for his true good. Yet, in the application of science to the needs and desires of man, it would seem to be a singularly unfortunate stage at which to terminate the process, or to lose hope as to the outcome."

" final paragraph "As we may think"

Historic Goals

"Google's mission is to organize the world's information and make it universally accessible and useful" Larry Page, Sergey Brin, Google's mission statement, ~ 1998.

"A memex is a device in which an an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory." Vannevar Bush, As we may think, *Atlantic Monthly*, July 1945.

Information Retrieval Problem

- User wants information from a collection of "objects": information need
- User formulates need as a "query"
  - Language of information retrieval system
- System finds objects that "satisfy" query
- System presents objects to user in "useful form"
- User determines which objects from among those presented are relevant CRITICAL NOTION

➢ Define each of the words in quotes
➢ Develop algorithms

9

Think first about text documents

Although search has changed, classic techniques still provide foundations
– our starting point

- Early digital searches – digital card catalog:
  - subject classifications, keywords
- "Full text" : words + natural language syntax
  - No "meta-structure"
- Classic study
  - Gerald Salton SMART project 1960's

10

Scaling

- What are attributes changing from 1960's to online searches of today?
- How do they change problem?

11

Develop models

Begin with document:

How do we view document contents?

12

Modeling: "*query*"

How do we want to express a query?

What does it mean?

13

Modeling: "*query*"

We will consider

- *Query*
  - Basic query is one term
  - Multi-term query is (choose one):
    - Set of terms
    - Sequence of terms
      - multiplicity?
      - Other constraints?
    - Boolean combination of terms

14

Modeling: "*satisfying*"

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

15

Present results in *"useful form"*

- most basic: give list of results
- meaning of order of list? => RANKING
- Goals of ranking
  - Order documents that satisfy a query by how well match the query
  - Capture relevance to user by algorithmic method of ordering

16

(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

17

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

18

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

19

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

20

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 OR 0) AND (1 AND NOT(1)) FALSE

21

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

- most basic: give list of results
- meaning of order of list? => RANKING?
- There is no sense of ranking in pure Boolean model
  - need idea in addition to "satisfying documents": generalize model

22

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?

23

Restrict Boolean Model

- AND model: query is the AND of a set of query terms: term_1 AND term_2 AND…
  - just need specify set of terms
  - This model used by current search engines
- OR model: query is the OR of a set of query terms: term_1 OR term_2 OR …
  - just need specify set of terms
  - This original model for IR development
    - why?

24