VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY

UNIVERSITY OF INFORMATION TECHNOLOGY

DS307

SOCIAL MEDIA ANALYSIS

Faculty of Information Science and Engineering

University of Information Technology, VNU-HCM

This Course's Contents

Introduction to Information Retrieval

Why Study IR?

- Many reasons, but if you want a one-word answer:

Google …

- Examines billions of web pages
- Returns results in less than half a second
- Valued at gazillions of dollars by the public market

How Does Google Work?

- Only Google know, but . . .
- Uses hundreds of thousands of machines
- Uses some sophisticated computer science (efficient storage and searching of large datasets)
- Uses an innovative ranking algorithm (based on the hypertext structure of the web)

How Does Google Work?

- Underlying Google is basic IR technology.
- The Web is indexed.
  - an index links terms with pages.
- A user's information need is represented as a query.
- Queries are matched against web pages.
  - Google attempts to return pages which are relevant to the information need.

IR is More Than Web Search

- IR is much older than the Web (1950s –)
- The Web has some unique characteristics which make it a special case.
- IR deals with tasks other than searching:
  - Categorizing Documents.
  - Summarizing Documents.
  - Answering Questions.
  - …

Motivation for IR

- Searching literature databases.
- Web search.
- Volume of information stored electronically is growing at ever faster rates
  - need to search it
  - categorize it
  - filter it
  - translate it
  - summarize it
  - ...

Biomedical Information

- Biomedical literature is growing at a startling rate
  - Around 1,000,000 new articles are added to Medline each year
- Tasks:
  - Literature search
  - Creation and maintenance of biological databases
  - Knowledge discovery from text mining

Document Retrieval

- IR is often used to mean Document Retrieval
- Primary task of an IR system:
  retrieve documents with content that is relevant to a user's information need
- How do we represent content?
- How do we represent information need?
- How do we decide on relevance?

Document Retrieval

- Representation/Indexing
  - Representation of documents and requests:
    - bag of words?
    - stop words, upper/lower case, . . . *
    - query language
  - Storing the documents, building the index
- Searching
  - Is a document relevant to the query?
    - models of IR: Boolean, **vector-space**, **probabilistic**
  - Efficient algorithms for searching large datasets

What IR is Not

- An IR system is not a Database Management System
- A DBMS stores and processes well-defined data
- A search in a DBMS is exact / deterministic
- Search in an IR system is probabilistic
  - inherent uncertainty exists at all stages of IR: information need, formulating query, searching

A Simple Retrieval Model

- Bag of Words approach
  - A document is represented as a bag of words – Word order is ignored
  - Syntactic structure is ignored
  - ...
- Relevance is determined by comparing the words in the document with the words in a query
- Simple approach has been very effective

Vector Space Model

- Provides a ranking of documents with respect to a query
- Documents and queries are vectors in a multi-dimensional information space
- Key questions:
  - What forms the dimensions of the space? ∗ terms, concepts, . . .
  - How are document and query vectors compared?

Coordinate Matching

- Document relevance measured by the number of query terms appearing in a document
- Terms provide the dimensions
  - Large vocabulary ⇒ high dimensional space
- Similarity measure is the dot-product of the query and document vectors

Simple Example

- Term vocabulary: (England, Australia, Pietersen, Hoggard, run, wicket, catch, century, collapse)
- Documents:
  - d1: Australia collapse as Hoggard takes 6 wickets
  - d2: Pietersen's century puts Australia on back foot
- Queries:
  - q1: {Hoggard, Australia, wickets}
- Query, document similarity
  - $q1.d1 = (0,1,0,1,0,1,0,0,0) \cdot (0,1,0,1,0,1,0,0,1) = 3$
  - $q1.d2 = (0,1,0,1,0,1,0,0,0) \cdot (0,1,1,0,0,0,0,1,0) = 1$

Term Frequency (TF)

- Coordinate matching does not consider the frequency of query terms in documents
- Term vocabulary: (England, Australia, Pietersen, Hoggard, run, wicket, catch, century, collapse)
- d1: Australia collapsed as Hoggard took 6 wickets. Flintoff praised Hoggard for his excellent line and length.
- q1: {Hoggard, Australia, wickets}
- $q1 \cdot d1 = (0,1,0,1,0,1,0,0,0) \cdot (0,1,0,2,0,1,0,0,1) = 4$

Term Frequency (TF)

- Coordinate matching does not consider the number of documents query terms appear in
- Term vocabulary: (England, Australia, Pietersen, Hoggard, run, wicket, catch, century, collapse)
- d2: Flintoff took the wicket of Australia's Ponting, to give him 2 wickets for the innings and 5 wickets for the match.
- q1: {Hoggard, Australia, wickets}
- $q1 \cdot d2 = (0,1,0,1,0,1,0,0,0) \cdot (0,1,0,0,0,3,0,0,0) = 4$

Inverse Document Frequency (IDF)

- Assume wicket appears in 100 documents in total, Hoggard appears in 5, and Australia in 10
  (ignoring IDF of other terms)
- d1: Australia collapsed as Hoggard took 6 wickets. Flintoff praised Hoggard for his excellent line and length.
- d2: Flintoff took the wicket of Australia's Ponting, to give him 2 wickets for the innings and 5 wickets for the match.
- q1: {Hoggard, Australia, wickets}
- $q1 \cdot d1 = (0,1,0,1,0,1,0,0,0) \cdot (0,1/10,0,2/5,0,1/100,0,0,1) = 0.411$
- $q1 \cdot d2 = (0,1,0,1,0,1,0,0,0) \cdot (0,1/10,0,0/5,0,3/100,0,0,0) = 0.13$

Document Length

- Terms in documents can have high term frequencies simply because the document is long
- Normalise similarity measure, M, by Euclidean length:

$$\mathrm{M}(\mathrm{Q},\mathrm{D}) = \frac{Q.\mathrm{D}}{|\mathrm{Q}||\mathrm{D}|}$$

Vector Space Similarity

- The terms in the query vector and document vector are weighted: $Q \cdot D = \sum_t w_{Q,t}.w_{D,t}$
- $w_{D,t} = TF \times IDF$
- Vector of weights determines position of document in the information space

Vector Space Similarity

- $M(Q, D) = \frac{Q.D}{|Q||D|} = \frac{1}{|Q||D|} \sum_t w_{Q,t}.w_{D,t} =$ cosine(Q, D)
  - where: $|D| = \sqrt{\sum_t w_{Q,t}^2}$
- Similarity measure is the cosine of the angle between the query and document vectors

Language Understanding?

- Want a system which “understands” documents and query and matches them?
  - use semantic representation and logical inference
- Until recently such technology was not robust / did not scale to large unrestricted text collections
- But:
  - useful for restricted domains
  - now used for some large-scale tasks (QA, IE)
- Is a “deep” approach appropriate for document retrieval?
  - Powerset (Natural Language Search) think so (see www.powerset.com)

Tasks in IR (broadly conceived)

- Document Retrieval (ad-hoc retrieval)
- Document Filtering or Routing
- Document Categorization
- Document Summarizing
- Information Extraction
- Question Answering

Other Topics

- Multimedia IR (images, sound, . . .)
  - but text can be of different types (web pages, e-mails, . . .)
- User-system interaction (HCI)
- Browsing

Evaluation

- IR has largely been treated as an empirical, or engineering, task
- Evaluation has played an important role in the development of IR
- DARPA/NIST Text Retrieval Conference (TREC)
  - began in 1992
  - has many participants
  - uses large text databases
  - considers many tasks in addition to document retrieval

IR in One Sentence

- “Indexing, retrieving and organizing text by probabilistic or statistical techniques that reflect semantics without actually understanding”

  (James Allan, Umass)

Brief History of IR

- 1960s
  - development of basic techniques in automated indexing and searching
- 1970s
  - Development of statistical methods / vector space models
  - Split from NLP/AI
  - Operational Boolean systems
- 1980s
  - Increased computing power
  - Spread of operational systems

Brief History of IR

- 1990s and 2000s
- Large-scale full text IR systems for retrieval and filtering
- Dominance of statistical ranking approaches
- Web search
- Multimedia and multilingual applications
- Question Answering
- TREC evaluations

Brief History of IR

- Course book
  - **Introduction to Information Retrieval**

Manning, Raghavan, Schu¨tze
http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html

Q&A

Thank you!