

# Day 1 Notes

## Why info retrieval ?

- Origin in library sciences
- Information overload. We need some way of organizing this
- We need a way to handle unstructured data. It is becoming the majority type of data
- To handle structured data use some db system
- Database system: You first need to define your data and how its structured
- Semantic Web & RTF:

## History

- Popularized by Vannevar Bush (1945) -> described very abstract form of www / http
- Work started in the 50s, got beter through 70s-80s, then web material in 90s, and present day is mostly scalability

## Preforming info retrieval

1. Crawler / indexer
2. Analyser
3. Query parser
4. Ranking

## Core concepts

- Query representation: lexical gap: say vs said
- Semantic gap: ranking vs retrieval
- *finish up this*

# Day 2 Notes

## Modern Search Engine

- High precision understanding of natural language
- Demand of accuracy = what matches your query
- Demand of Efficiency = retrieve information quickly
- Demand of convenience = Organize knowledge (via table or graph summarizing information to save you a click). Can be a bit counter-intuitive in terms of measuring site ranking via click.
- Demand of diversity

## Types of info retrieval

1. Recommendation system.
  - Search engine matches query, recommendation system kind of pushes relevant content
2. Product Search
3. Question answering
4. Document understand, text crawling, and mining
5. Advertisement
6. Desktop search , web search

## IR vs DB

- Info: Unstructured data, subjective semantics

- DB: Structured data, well defined semantics (predefined, exact answer)

–finish up–

## IR and DBs (II)

- IR  $\Rightarrow$  DBs . Approximate search available in DBs
- DBs  $\Rightarrow$  IR . Use information extraction to convert unstructured data to structured data. e.g. knowledge base
- Semi-structured representation: XML data queries

## IR vs NLP

- Info retrieval: Computational approaches, statistical understanding of language, large scale
- NLP : Cognitive , symbolic, and computational approaches . Semantic understanding of language, small scale problems

## IR vs NLP (II)

- IR  $\Rightarrow$  NLP : larger data collections, scalable NLP techniques
- NLP  $\Rightarrow$  IR :

–finish up– info

Bush, as we may think. Chapter 1: boolean retrieval