Current Trends in IR

TREC — Text REtrieval Conference
(annually in Nov. Gaithersburg, MD)

- Most important development in IR since 1970s
- Before TREC, IR experiments were small and results not directly comparable
- 1991: US DoD launch TIPSTER initiative:
  - Issued several GB of data on CD-ROM
  - Unstructured, standard general texts, e.g. Wall St. Journal, AP Newswire, . . .
  - set of queries for documents
  - Set of relevance judgments
    — which documents relevant to which query.

TREC

- DoD invites researchers to test systems on this data each year
- Results are announced to participants at TREC
- Has been phenomenal success, 1999 will see TREC-8
- Now at TREC-8, as well as the original, "ad-hoc" retrieval task, there are many additional tracks or tasks:
  - Chinese retrieval
  - Cross-Language retrieval
  - Speech Retrieval

RUG Catalogue vs. Altavista

Contrast

- RUG uses controlled subject hierarchy,
  Altavista is a free-text system
- RUG records are structured in a variety of fields,
  Altavista assumes all web pages unstructured.
- RUG records are manually constructed,
  Altavista uses fully automatic indexing techniques
- RUG allows for exact or partial matching,
  Altavista goes for exact match as default.

IR researchers avoid the web as testing ground because experiments are difficult to organise.

Hyperlink Indexing

- Modern IR research concentrates on the TREC collections e.g. Wall Street Journal
  - Documents have minimal internal structure
  - Documents considered independent of one another
- WWW search engines treat web pages *similarly*
  BUT
  web pages contain a rich hyperlink structure
- How do we exploit this?

Standard Search Engines

Problems with Standard Search Engines

- Not all websites equally reputable
- No quality control on the web
- Many sites artificially boost search ratings
- Ambiguity and polysemy still big problems

Possible Solutions

- Manually Assembled Catalogues
- Semantic Networks
- Citation Analysis
- Dynamic Analysis

Manually Assembled Catalogues

Yahoo

- Add only hand-selected pages to catalogue
- Generates good-quality results
- Need human intervention both to maintain the catalogue (choose new keywords) and to select pages to add
- Cannot keep up with expanding web
  — one million new pages join web every day!

Semantic Networks

WordNet

- Defines "concepts"
- Links concepts in a network
- Similar concepts grouped together
- Traverse network to get group of linked concepts for retrieval
- Network hand-built and -tuned

```
Sense 1
gangster, mobster -- (a criminal who is a member of gang)
   => criminal, felon, crook, outlaw, malefactor --
       => wrongdoer, offender -- (person who transgresses ...
           => bad person -- (person who does harm to others)
               => person, individual, someone, somebody, ...
                   => life form, organism, being, ...
                       => entity, something
                   => causal agent, cause, ...
                       => entity, something --
```

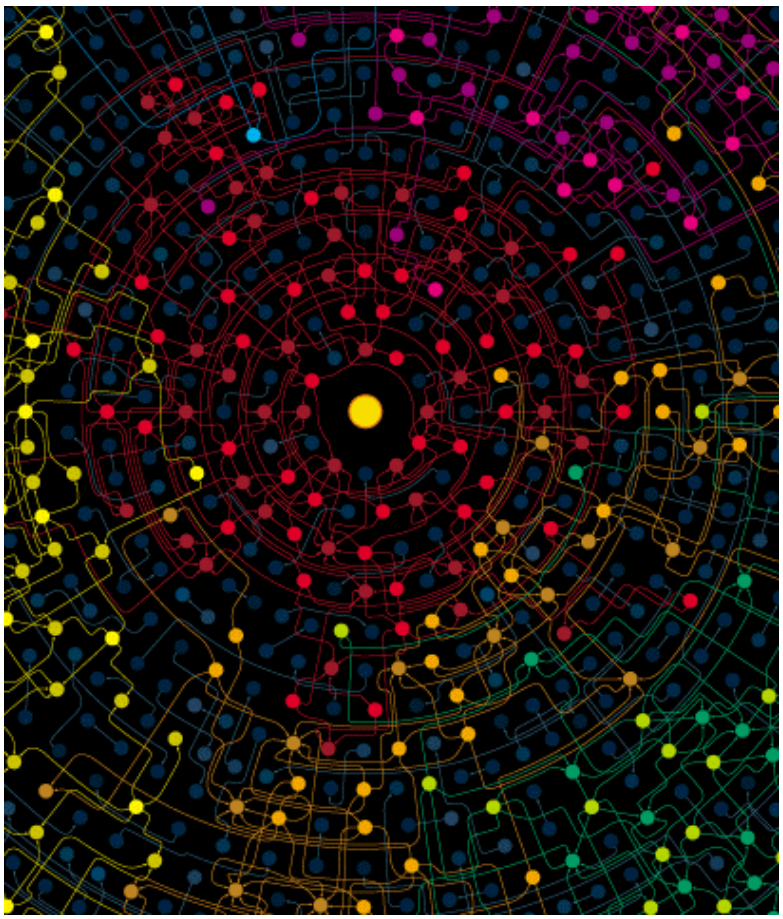
Google (Stanford)

Citation Analysis

- Garfield: importance of journal article proportional to the number of citations it receives (arrowheads)
- $\Rightarrow$ Web pages: good sites are linked to by many others
- Google randomly traverses the web building a list of frequently encountered sites
- Finds universally popular sites, e.g. New York Times
- Favors pages on these sites in ranking search results

Dynamic Analysis

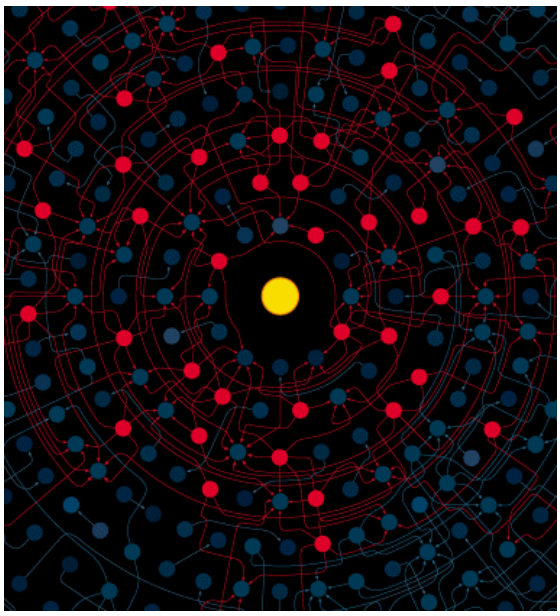
Clustered Links indicate "Web Communities"



- sites which (mostly) point to each other
  - oil spills in Japan
  - resources for Turks living in US
  - fire fighting in Australia

Dynamic Analysis

IBM Clever System



- Like Google but distinguishes between:
  - Hub pages: lists of links (red)
  - Authority pages: sites with content (blue)
    —worth pointing to
- A good hub points to many good authorities, and vice versa
- "Circular definition" utilised by an iterative algorithm to rank results of standard search
- Good hubs and authorities are near the top

Conclusion

- Current IR is stable and reliable
- IR need the query enhancement techniques described in lecture 2
- Research in new areas such as CLIR continues
- The hyperlink indexing systems seem promising, but still in experimental stage
- $\Rightarrow$ We will continue to use existing WWW search engines for a while (maybe longer)