CSA4020

Multimedia Systems:

Adaptive Hypermedia Systems

Lecture 1: Introduction

About Myself

- Dr. Christopher Staff, Dept. of Computer Science and AI
- cstaff@cs.um.edu.mt
- http://www.cs.um.edu.mt/~cstaff
- Room 402, New Computer Building
- Timetable available from Web site

About the Course

- 2 credits
- Assessment: 100% test
- Web site:
  http://www.cs.um.edu.mt/~cstaff/courses/lectures/csa4020/
- Recommended Texts

  Van Rijsbergen, Keith, Information Retrieval. (http://www.dcs.gla.ac.uk/Keith/Preface.html)

  Baeza-Yates, R., and Ribeiro-Neto, B. Modern Information Retrieval.

  Brusilovsky, *et. al*. (eds). Adaptive Hypertext and Hypermedia.

Introduction to IR

- Why IR?

  Think about your own file collections…

  One day you want to find information that you know is in one of the files…

  You try to remember which file contains it…

  Maybe you gave it a meaningful name…

  Maybe you organised files into directories… once you locate the right directory, you can look at each file…

  If you are relying on your own memory, or you are relying on “clues” that you may have left yourself, and it is hard to locate info, what chance has a stranger?

- Books, libraries, and the Internet

  Books provide information about their contents…

  Libraries too…

  But notice that libraries do not simply re-use the tools provided by books…
  Keywords, subject indices, shelf marks…

  Information Retrieval and the Internet, in particular, the http protocol, are inextricably, but inexactly, linked…

  Information Retrieval on the Internet also comes in two main forms:

  Tools similar to a library's
  except that files can belong to more than one category
  A super-set of (one of the) tools for books

Is all information made equal?

- Data may be:

    - Formatted (structured)
    - Non-formatted (unstructured)

        Textual (articles, reports, papers, books)
        Non-textual (video, graphics, audio)

- In formatted data, “meaning” is exposed and made implicit in the structure of the data (eg, RDBMS)

- In non-formatted data, “meaning” is buried in the “file”/document – how can it be exposed?

- In formatted data, structured query language is used to formulate query

- In non-formatted data… ? Simplest way is by matching keywords against an index

Why is IR a difficult problem?

- Huge amount of data, which effects efficiency, effectiveness, timeliness, and user-friendliness

  Google indexes more than 2 billion web pages (www.google.com/press/guide/reviewguide_3.html), and this is only part of the indexable Web

  It takes time to index pages (though only a portion of it will change since the last time the index was built)

  It takes time to search through the index (users want a response within seconds)

  Index will almost certainly be out-of-date

  Have you encountered a IR system which allows you to ask a complex query in NL and gives you reasonable results?

Why is IR difficult? (cont.)

- Difficult to capture semantics in unstructured information:

  Find all reports about football matches in which Liverpool beat Manchester United

  ```
  select * from Directory where Surname = “Attard”
  ```

- Unrestricted domains:

  Cannot always pre-define category to which the doc belongs
  Word sense ambiguity

- Broad user base: novice to expert

  Search tool and subject matter! If you cannot describe the info (because you’re a novice IR tool user, or because you’re new to the topic), it’s going to be hard to find

Why is IR difficult (cont.)

- Information at the right level of detail:

  User expectations (FAQ vs. Manual)
  User expertise (novice vs. expert)

- Distributed and interlinked:

  Info might be spread over several, linked documents

- Efficiency vs. Effectiveness:

  Accuracy vs. Time to process query
  Usability vs. Accuracy of results
  Timeliness vs. Quality of index
  Quality of index vs. Storage space
  …

Document Retrieval Model