Information Systems M
Introduction

Prof. Paolo Ciaccia
http://www-db.deis.unibo.it/courses/SI-M/

Course contents (general)

A set of models, algorithms, and general techniques aiming to make easier the problem of finding the “best” information in large data collections

- There are plenty of modern applications/scenarios in which this is THE problem, e.g.:
  - Web search engines
  - E-commerce sites
  - Multimedia databases
  - Recommender systems
  - ...
Introduction Sistemi Informativi M 3

Scenario 1: “good” e-commerce solutions

- E-commerce web sites usually make available search forms so as to select only those products/items of interest:
  - used cars, apartments, books, hotels, flights, electronic auctions, ...

- In these scenarios:
  - the client aims to find something that matches at best her preferences
  - the system aims to return to the user the best answers in a reasonable amount of time

- A “good” e-commerce site will provide users with “high quality” results without losing in efficiency, and without requiring to users long search sessions
Relaxing a search criterion...

Even here no result, yet…

...the system is anyway able to return "good" answers!

Introduction  Sistemi Informativi M
Scenario 2: “good” multimedia solutions

- Multimedia database systems (MMDBMSs) are nowadays used in many application domains:
  - Design, fashion, medicine, biological sciences, press agencies, video-on-demand, e-learning, ...

- In all these cases a common query type is the so-called Query By Example, i.e., “give me something similar to this” (this image, this song, ...)

- In this case
  - the client aims to find something that (semantically) matches the query object
  - the system aims to deliver to the user the best answers in a reasonable amount of time

- A “good” MMDBMS will provide users with “high quality” results without losing in efficiency

---

Content-based Image Retrieval

This is the query object...

… which is the “best” system?
Scenario 3: Web search engines

- Given a search query, a good search engine will return the “most interesting” Web pages, which are relevant to the user query.

- Even in this case:
  - the client aims to find something that is relevant to her information need
  - the system aims to deliver to the user the best answers in a reasonable amount of time

- Besides the set of pages managed by the search engines, other factors can influence which pages are returned as well as their relative order.

The “Pareto set” query

- Let P be a “multi-criteria” optimization problem, i.e., a problem where we want to minimize, at the same time, the objective functions $F_1, \ldots, F_j, \ldots, F_N$, e.g.:
  
  $F_1(x) = \sqrt{x + 1}$
  
  $F_2(x) = x^2 - 4x + 5 = (x - 2)^2 + 1$

- A solution $A$ for problem $P$ is an assignment of values to the variables in $P$, such that all constraints are satisfied (e.g., $X_i \geq 0$)

- A solution $A$ is Pareto optimal iff (if and only if) there is no other solution $B$ that dominates $A$, that is, $B$ is better than $A$ for at least one $F_i$ and not worse than $A$ on all other $F_j$’s:
  
  $A$ is Pareto optimal $\iff \neg \exists B: (\exists k: F_k(B) < F_k(A) \land \forall j: F_j(B) \leq F_j(A))$

- The Pareto set is the set of all Pareto optimal solutions.
  
  - ...we’ll come back to these concepts later in the course
  
  - Unclear? Let’s find some more explanation on the Web...

\[
\begin{align*}
F_1(x) &= \sqrt{x + 1} \\
F_2(x) &= x^2 - 4x + 5 = (x - 2)^2 + 1
\end{align*}
\]
The “Pareto set” query on Google

The “Pareto set” query on Altavista

Nice result!

Not exactly what we are looking for! This is why!
Scenario 4: suggesting new products...

- Systems that recommend/suggest products to their users have to take into account the information overload phenomenon, where too many alternatives make it hard for users to choose among.

- The available techniques are collectively known as "information filtering", and find major applications in two types of systems:
  - publish & subscribe (pub/sub) systems: "deliver new published events to subscribers who expressed interests in such events" (e.g., news)
  - recommender systems: "suggest products to customers when they come back to visit a site" (e.g., electronic catalogs)

- In both cases:
  - the client aims to be informed on what best matches her preferences (or "profile")
  - the system aims to deliver to the user the best answers in a reasonable amount of time

Amazon.com recommendations

Introduction Sistemi Informativi M 13
Amazon.com recommendations: why?

Why was I recommended this?

We recommended...

Refactoring
by Martin Fowler (Author), et al
List Price: $49.95
Price: $29.99
Leave it (Not interested)

Because you purchased or rated...

Transaction Processing
by Jim Gray, Andreas Reuter

...a possible explanation...

Recap

Finding the most valuable information is challenging
- Large volume of data
- Imprecision of query specification
- Sometimes not clear what to look for
- User- and even context-dependent relevance of results

We will study a variety of approaches that share the following objectives:
- Maximize user satisfaction (effectiveness)
- Minimize system overhead (efficiency)
- Minimize user effort (usability)

The details will vary depending on a variety of factors, including:
- system architecture (e.g., local vs distributed)
- data model (e.g., relational vs unstructured)
- system resources (e.g., indexes)
Course contents (more detailed)

- Top-k queries
  - How to efficiently return only the “best” k answers

- Skyline and more general preference queries
  - How to determine the “best” results without using numbers

- Information retrieval
  - How to find the “best” documents or web pages

- Multimedia information retrieval
  - How to find the “best” images, videos, graphs, time series, ...

- Recommender systems
  - How to suggest the “best” products to users