cis20.2
design and implementation of software applications II
spring 2008
session # II.1
information models and systems

topics:

- what is information systems?
- what is information?
- knowledge representation
- information retrieval

what is information systems?

- the field of *information systems (IS)* comprises the following:
  - a number of types of computer-based information systems
  - objectives
  - risks
  - planning and project management
  - organization
  - IS development life cycle
  - tools, techniques and methodologies
  - social effects
  - integrative models

types of information systems

- informal
  - evolve from patterns of human behavior (can be complex)
  - not formalized (i.e., designed)
  - rely on "word of mouth" ("the grapevine")
- manual
  - formalized but not computer based
  - historical handling of information in organizations, before computers (i.e., human "clerks" did all the work)
  - some organizations still use aspects of manual IS (e.g., because computer systems are expensive or don't exist to relace specialized human skills)
- computer-based
  - automated, technology-based systems
  - typically run by an "IT" (information technology) department within a company or organization (e.g., ITS at BC)

computer-based information systems

- data processing systems (e.g., accounting, personnel, production)
- office automation systems (e.g., document preparation and management, database systems, email, scheduling systems, spreadsheets)
- management information systems (MIS) (e.g., produce *information* from *data*, data analysis and reporting)
- decision support systems (DSS) (e.g., extension of MIS, often with some intelligence, allow prediction, posing of "what if" questions)
- executive information systems (e.g., extension of DSS, contain strategic modeling capabilities, data abstraction, support high-level decision making and reporting, often have fancy graphics for executives to use for reporting to non-technical/non-specialized audiences)

why do organizations have information systems?

- to make operations efficient
- for effective management
- to gain a competitive advantage
- to support an organization's long-term goals

IS development life cycle

- feasibility study
- systems investigation
- systems analysis
- systems design
- implementation
- review and maintenance

social effects of IS

- change management
- broad implementation (not just about software)
- education and training
- skill change
- societal and cultural change

integrative models

- computers in society
- the internet revolution (internet 2, web 2.0)
- "big brother"
- ubiquitous computing

what is information?

- definition comprises ideas from philosophy, psychology, signal processing, physics...
- OED:
  - information = "informing, telling; thing told, knowledge, items of knowledge, news"
  - knowledge = "knowing familiarity gained by experience; person's range of information; a theoretical or practical understanding of; the sum of what is known"
- other ideas:
  - relating data to *context*
  - must be *recorded*
  - has potential to become knowledge
- what is the relationship between *data* and *information* and *knowledge* and *intelligence*???

types of information

- can be differented by:
  - form
  - content
  - quality
  - associated information
- properties
  - can be communcated electronically (methods: broadcasting, networking)
  - can be duplicated and shared (issues: ownership, control, maintenance, correction)

intuitive notion of information (from Losee, 1997)

- information must be something, although its exact nature is not clear
- information must be "new" (repeating something old isn't considered "information"... or is it?)
- information must be true (i.e., not "mis-information")
- information must be about something
- note human-centered definition that emphasizes meaning and message

human perspective

- cognitive processing
  - perception, observation, attention
  - reasoning, assimilating, interpreting, inferring
  - communicating
- knowledge, belief
- belief = "an idea held on some support; an internally accepted statement, result of inductive processes combining observed facts with a reasoning process"
- does "information" require a human mind?

meaning versus form

- is the form of information the information itself? or another kind of information?
- is the meaning of a signal or message the signal or message itself?
- representation (from Norman 1993)
  - why do we write things down?
    - Socrates thought writing would obliterate serious thought
    - sound and gestures fade away
  - artifacts help us reason
  - anything not present in a representation can be ignored (do you agree with that?)
  - things left out of a representation are often those things that are hard to represent, or we don't know how to represent them

The Library of Babel, by Jorge Luis Borges (1941)

- a story about a universe comprised of an indefinite (possibly infinite) number of hexagonal rooms, each containing walls of bookshelves that contain books which, in turn contain all possible combinations of letters
- is this information? data? knowledge? intelligence?
- how is the internet like (or unlike) the library of babel?

information theory

- Claude Shannon, 1940's, IBM
- studied communication and ways to measure information
- *communication* = producing the same message at its destination as at its source
- problem: *noise* can distort the message
- message is *encoded* between source (transmitter) and destination (receiver)

communication theory

- many disciplines: mass communication, media, literacy, rhetoric, sociology, psychology, linguistics, law, cognitive science, information science, engineering, medicine...
- *human communication theory:*
  do you understand what I mean when I say something?
- what does it mean to say a message is received? is received the same as understood?
- the *conduit metaphor*
- meaning: syntactic versus semantic

information theory today

- total annual information production including print, film, media, etc is between 1-2 Exabytes (10^{18}) per year
- how to we organize this???
- and remember, it accumulates!
- information hierarchy:
  data → information → knowledge → intelligence

information retrieval

- information *organization* versus *retrieval*
- organization:
  categorizing and describing information objects in ways that people can use them who need to use them
- retrieval:
  being able to find the information objects you need when you need them
- two key concepts:
  - *precision*: did I find what I wanted?
  - *recall*: how quickly did I find it?
- ideally, we want to maximize both precision and recall—this is the primary goal of the field of *information retrieval (IR)*

IR assumptions

- information remains static
- query remains static
- the value of an IR solution is in how good the retrieved information meets the needs of the retriever
- are these good assumptions?
  - in general, information does not stay static; especially the internet
  - people learn how to make better queries
- problems with standard model on the internet:
  - "answer" is a list of hyperlinks that then need to be searched
  - answer list is apparently disorganized

IR process

- IR is iterative
- IR doesn't end with the first answer (unless you're "feeling lucky"...)
- because humans can recognize a partially useful answer; automated systems cannot always do that
- because human's queries change as their understanding improves by the results of previous queries
- because sometimes humans get an answer that is "good enough" to satisfy them, even if initial goals of IR aren't met

"berry-picking" model (from Bates 1989)

- interesting information is scattered like berries in bushes
- the eye of the searcher is continually moving
- new information may trigger new ideas about where to search
- searching is generally not satisfied by one answer

information seeking behavior

- two parts of a process:
  - search and retrieval
  - analysis and synthesis of search results
- search tactics and strategies
  - tactics ⇒ short-term goals, single actions, single operators
  - strategies ⇒ long-term goals, complex actions, combinations of operators (macros)
- need to keep search on track by *monitoring* search
  - check: compare next move with current "state"
  - weigh: evaluate cost/benefit of next move/direction
  - pattern: recognize common actions
  - correct: fix mistakes
  - record: keep track of where you've been (even wrong directions)
- search tactics
  - specify: be as specific as possible in terms you are looking for

  - exhaust: use all possible elements in a query
  - reduce: subtract irrelevant elements from a query
  - parallel: use synonyms ("term" tactics)
  - pinpoint: focus query
  - block: reject terms
- relevance — how can a retrieved document be considered relevant?
  - it can answer original question exactly and completely
  - it can partially answer the question
  - it can suggest another source for more information
  - it can provide background information for answering the question
  - it can trigger the user to remember other information that will help answer the question and/or retrieve more information about the question

parametric search

- most documents have "text" and "meta-data", organized in "fields"
- in parametric search, we can associate search terms with specific fields
- example: search for apartments in a certain **geographic neighborhood** within a certain **price range** of a certain **size**
- the data set can be organized using *indexes* to support parametric search

zone search

- a "zone" is an identified region within a document
- typically the document is "marked up" before you search
- content of a zone is free text (unlike parametric fields)
- zones can also be indexed
- example: search for a book with certain keyword in the title, last name in author and topic in body of document
- does this make the web a database? not really (which you'll see when we get into database definitions next week)

scoring and ranking

- search results can either be **Boolean** (match or not) or **scored**
- scored results attempt to assign a quantitative value to how good the result is
- some web searches can return a **ranked** list of answers, ranked according to their score
- some scoring methods:
  - linear combination of zones (or fields)
  - incidence matrices

linear combination of zones

- assign a weight to each zone (or field) and evaluate:

  $score = 0.6 * (Brooklyn \in neighborhood) + 0.5 * (3 \in bedrooms) + 0.4 * (1000 = price)$

- problem:
  it is frequently hard for a user to assign a weighting that adequately or accurately reflects their needs/desires

incidence matrices

- *recall* = document (or a zone or field in the document) is a binary vector $X \in \{0,1\}^v$
- *query* is a vector
- *score* is overlap measure: $|X \cap Y|$
- example:

	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	0	0	0	1
Brutus	1	0	1	0	0
Caesar	1	0	1	1	1
Calpurnia	1	0	0	0	0
Cleopatra	0	0	0	0	0

score is sum of entries row (or column, depending on what the query is)

- problem: *overlap measure* doesn't consider:
  - term frequency (how often does a term occur in a document)
  - term scarcity in collection (how infrequently does the term occur in all documents in the colletion)
  - length of documents searched
- what about **density**?
  if a document talks about a term more, then shouldn't it be a better match?
- what if we have more than one term?
  this leads to *term weighting*

term weighing

- in previous matrix, instead of 0 or 1 in each entry, put the *number of occurrences* of each term in a document
- this is called the "bag of words" (multiset) model
- problem:
  - score is based on syntactic count but not on semantic count
  - e.g.: *The Red Sox are better than the Yankees.*
    is the same as
    The Yankees are better than the Red Sox.
    (well, only in this example...)
- *count* versus *frequency*
  - search for documents containing "ides of march"
  - Julius Caesar has 5 occurrences of "ides"
  - No other play has "ides"
  - "march" occurs in over a dozen plays
  - All the plays contain "of"

  - By this scoring measure, the top-scoring play is likely to be the one with the most "of"s — is this what we want?
- NOTE that in the IR literature, "frequency" typically means "count" (not really "frequency" in the engineering sense, which would be count normalized by document length...)
- *term frequency (tf)*
  - somehow we want to account for the length of the documents we are comparing
- *collection frequency (cf)*
  - the number of occurrences of a term in a collection (also called *corpus*)
- *document frequency (df)*
  - the number of documents in a collection (corpus) containing the term
- tf x idf or tf.idf
  - tf = term frequency
  - idf = inverse document frequency; could be $1/df$, but more commonly computed as:

$$idf_i = log\left(\frac{n}{df_i}\right)$$

  - "weight" of term i occurring in document d ($w_{i,d}$) is then:
    $w_{i,d} = tf_{i,d} \times idf_i$
    $= tf_{i,d} \times log(n/df_i)$
    where
    $tf_{i,d}$ = frequency of term i in document d
    n = total number of documents in collection
    df_i = number of documents in collection that contain term i
  - weight increases with the number of occurrences within a document
  - weight increases with the rarity of the term across the whole collection
- so now we recompute the matrix using the $w_{i,d}$ formula for each entry in the matrix, and then we can do our ranking with a query

references

- Dr Phil Trinder, Heriot-Watt University, UK
  F29IF1 Database and Information Systems (Fall 2007)
  `http://www.macs.hw.ac.uk/~trinder/DbInfSystems/`

- Prof Ray Larson and Prof Marc Davis, UC Berkeley
  IS 202 Information Organization and Retrieval (Fall 2003)
  `http://www.sims.berkeley.edu/academics/courses/is202/f03/`

- Prof Christopher Manning and Prof Prabhakar Raghavan, Stanford University
  CS276 / LING 286 Information Retrieval and Web Mining (Fall 2006)
  `http://www.stanford.edu/class/cs276/`