UNIVERSITÀ DEGLI STUDI DI BARI
Facoltà di Scienze Matematiche, Fisiche e Naturali
Dipartimento di Informatica

Accesso intelligente all'informazione

Corso di
Metodi per il Ritrovamento dell'Informazione

UNIVERSITÀ DEGLI STUDI DI BARI
Facoltà di Scienze Matematiche, Fisiche e Naturali
Dipartimento di Informatica

Document/Text Mining: From Text to Knowledge

Corso di
Metodi per il Ritrovamento dell'Informazione

Credits

- Ricardo Baeza-Yates
- Berthier Ribeiro-Neto
- Marco de Gemmis
- Giovanni Semeraro
- Pasquale Lops

Gestire la conoscenza

… significa:

- Raccogliere la conoscenza
- Organizzarla (strutturarla, classificarla)
- Distribuirla
- Renderla accessibile a chi ne ha bisogno (nel momento e nel posto in cui serve)

…al fine di:

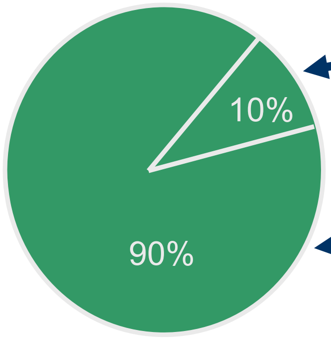
- Risparmiare tempo
- Migliorare la qualità dei servizi
- Ridurre i tempi di accesso all'informazione ed alla fruizione dei servizi

Dati-Informazione-Conoscenza

La conoscenza è un capitale:

- intangibile
- volatile
- difficile da concretizzare e conservare

Circa il 90% dei dati presenti nei database del mondo è in forma non strutturata



Dati in forma strutturata (numerica o codificata)

Dati in forma non strutturata

Automatic Knowledge Management

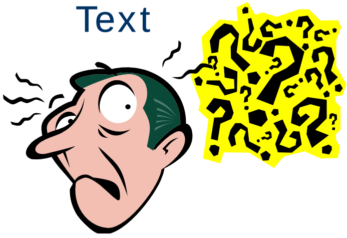
- Obiettivi
  - ✓ Costruzione di sistemi in grado di processare documenti in linguaggio naturale
  - ✓ Acquisizione / Ritrovamento di conoscenza da basi di dati in forma testuale

Acquisizione della Conoscenza

"From Text to Knowledge"

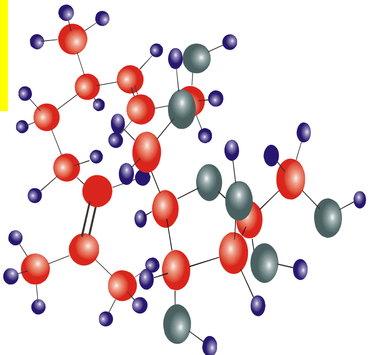
ONTOLOGY
CONSTRUCTION

Text



- Document classification
- Information Extraction
- Text mining

Document-based
ontology definition

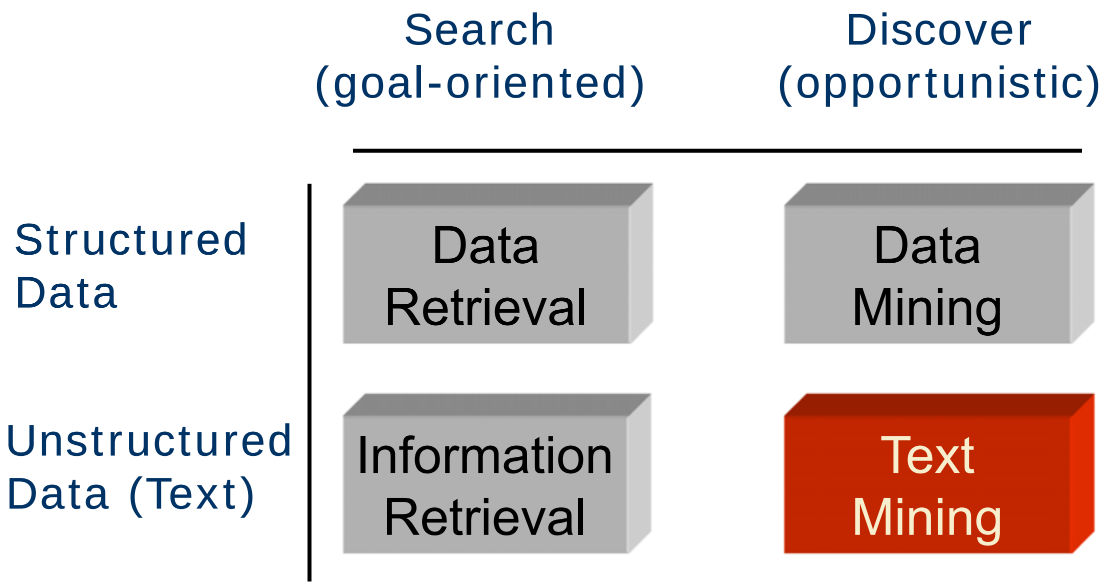


knowledge

Outline

- Knowledge Discovery from Text: Text Mining
  - ✓ Definizione
  - ✓ Data mining vs. Text mining
  - ✓ Perchè Text mining?

“Search” versus “Discover”



	Search (goal-oriented)	Discover (opportunistic)
Structured Data	Data Retrieval	Data Mining
Unstructured Data (Text)	Information Retrieval	Text Mining

Data Retrieval

- Ritrovamento di record in un database strutturato.

Database Type	Structured
Search Mode	Goal-driven
Atomic entity	Data Record
Example Information Need	“Find a Japanese restaurant in Boston that serves vegetarian food.”
Example Query	“SELECT * FROM restaurants WHERE city = boston AND type = japanese AND has_veg = true”

Information Retrieval

- Cerca informazione rilevante in una sorgente di dati non strutturati (tipicamente in formato testo)

Database Type	Unstructured
Search Mode	Goal-driven
Atomic entity	Document
Example Information Need	“Find a Japanese restaurant in Boston that serves vegetarian food.”
Example Query	“Japanese restaurant Boston” or Boston->Restaurants->Japanese

Data Mining

- Scopre nuova conoscenza attraverso l’analisi di dati

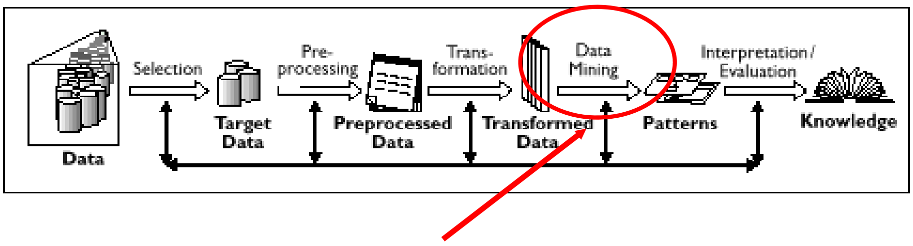
Database Type	Structured
Search Mode	Opportunistic
Atomic entity	Numbers
Example Information Need	“Show trend over time in # of visits to Japanese restaurants in Boston ”

The KDD Process

Knowledge Discovery from Databases

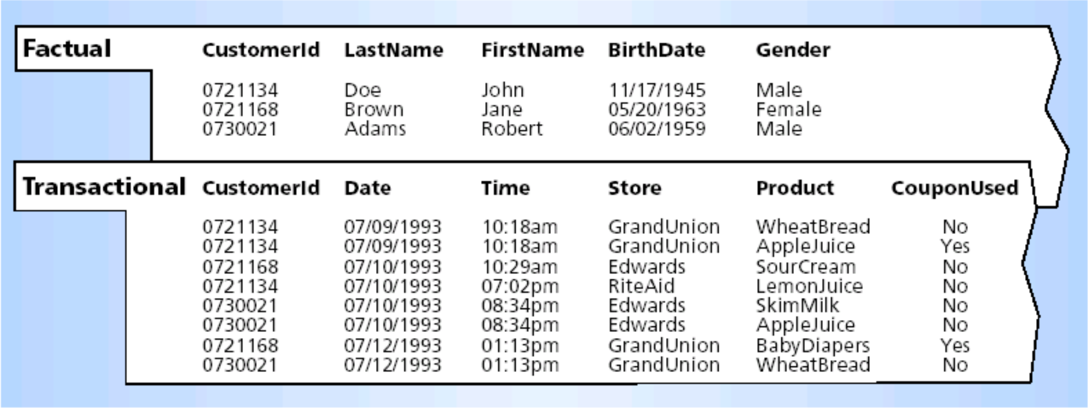
"The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"

Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In Proceedings of The Second Int. Conference on Knowledge Discovery and Data Mining, pages 82—88.



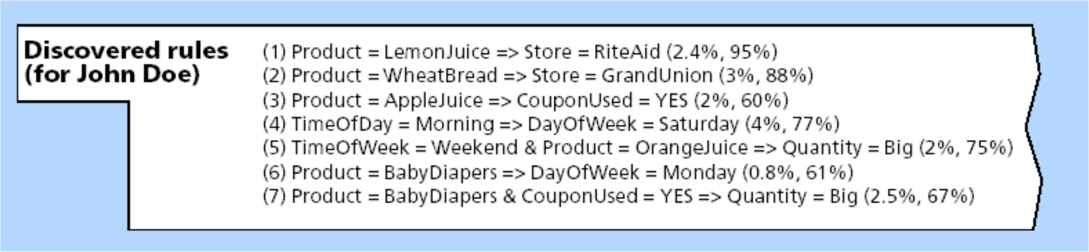
Note: data mining is just one step in the process

Data Mining@work



Factual	CustomerId	LastName	FirstName	BirthDate	Gender
	0721134	Doe	John	11/17/1945	Male
	0721168	Brown	Jane	05/20/1963	Female
	0730021	Adams	Robert	06/02/1959	Male

Transactional	CustomerId	Date	Time	Store	Product	CouponUsed
	0721134	07/09/1993	10:18am	GrandUnion	WheatBread	No
	0721134	07/09/1993	10:18am	GrandUnion	AppleJuice	Yes
	0721168	07/10/1993	10:29am	Edwards	SourCream	No
	0721134	07/10/1993	07:02pm	RiteAid	LemonJuice	No
	0730021	07/10/1993	08:34pm	Edwards	SkimMilk	No
	0730021	07/10/1993	08:34pm	Edwards	AppleJuice	No
	0721168	07/12/1993	01:13pm	GrandUnion	BabyDiapers	Yes
	0730021	07/12/1993	01:13pm	GrandUnion	WheatBread	No



Discovered rules (for John Doe)

(1) Product = LemonJuice => Store = RiteAid (2.4%, 95%)
(2) Product = WheatBread => Store = GrandUnion (3%, 88%)
(3) Product = AppleJuice => CouponUsed = YES (2%, 60%)
(4) TimeOfDay = Morning => DayOfWeek = Saturday (4%, 77%)
(5) TimeOfWeek = Weekend & Product = OrangeJuice => Quantity = Big (2%, 75%)
(6) Product = BabyDiapers => DayOfWeek = Monday (0.8%, 61%)
(7) Product = BabyDiapers & CouponUsed = YES => Quantity = Big (2.5%, 67%)

From Data Mining to Text Mining

- Text Mining, Text Data Mining, Knowledge Discovery from Text, Knowledge Discovery in Textual Data(bases)

“…nontrivial extraction of implicit, previously unknown, and potentially useful information from (large amounts of) textual data”

Text Mining
=
Data Mining (applied to text data)
+
basic linguistics

R. Feldman and I. Dagan, 1995.
Knowledge Discovery in Textual Databases (KDT). In Proceedings of the 1st International Conference on Knowledge Discovery (KDD-95), pp. 112-117, Montreal.

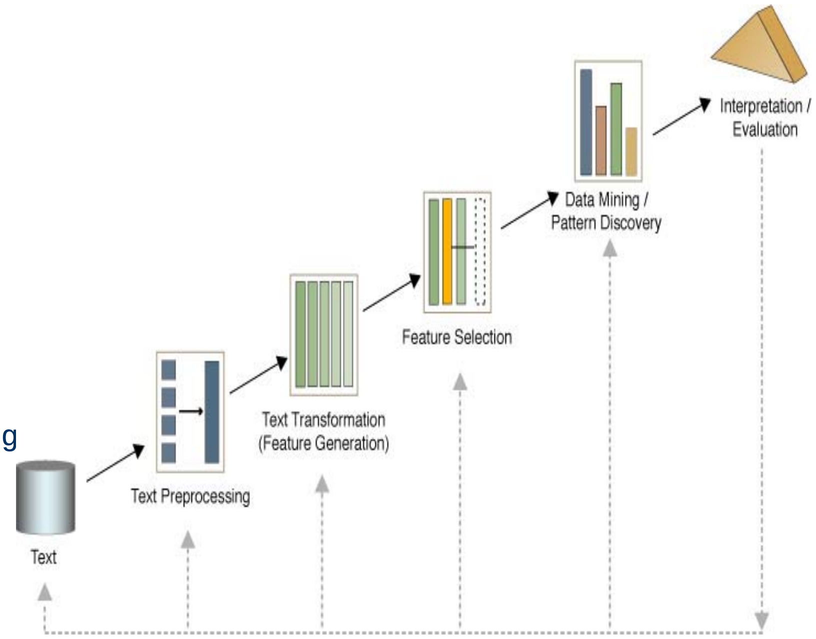
Text Mining

- Discover new knowledge through analysis of text

Database Type	Unstructured
Search Mode	Opportunistic
Atomic entity	Language feature or concept
Example Information Need	“Find the types of food poisoning most often associated with Japanese restaurants”
Example Query	Rank **diseases** found associated with “Japanese restaurants”

Text mining process

- Text preprocessing
  - ✓ Syntactic/Semantic text analysis
- Features Generation
  - ✓ Bag of words
- Features Selection
  - ✓ Simple counting
  - ✓ Statistics
- Text/Data Mining
  - ✓ Classification- Supervised learning
  - ✓ Clustering- Unsupervised learning
- Analyzing results



Text Mining

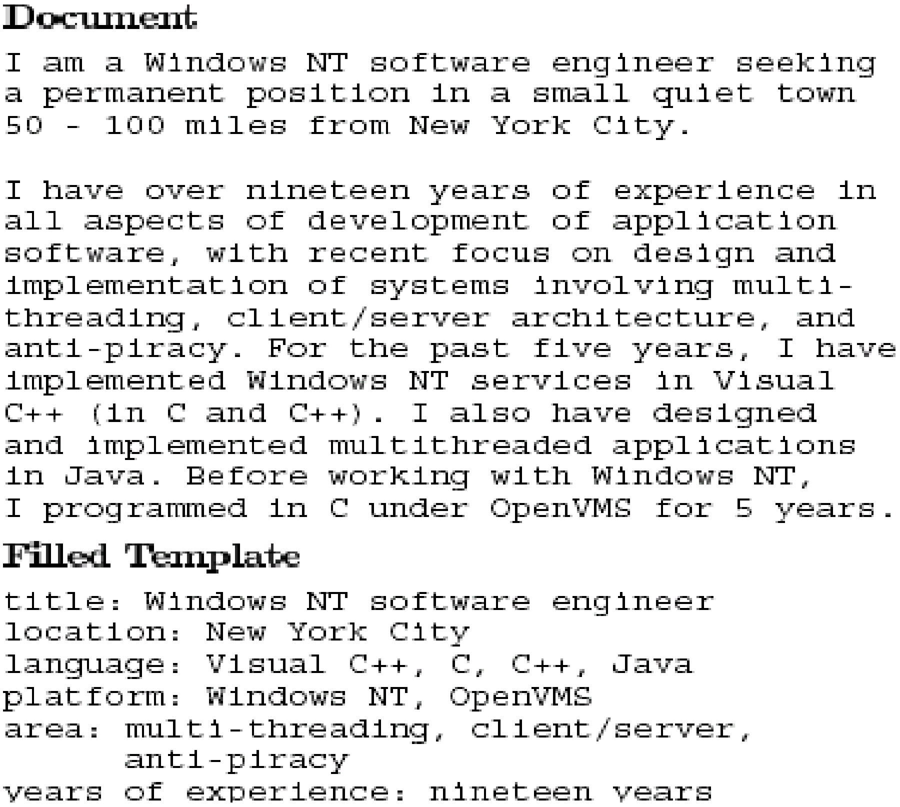
Discover useful and previously unknown “gems” of information in **large text collections**

Patterns

Trends

Associations

Text Mining@work



Document

```
I am a Windows NT software engineer seeking
a permanent position in a small quiet town
50 - 100 miles from New York City.

I have over nineteen years of experience in
all aspects of development of application
software, with recent focus on design and
implementation of systems involving multi-
threading, client/server architecture, and
anti-piracy. For the past five years, I have
implemented Windows NT services in Visual
C++ (in C and C++). I also have designed
and implemented multithreaded applications
in Java. Before working with Windows NT,
I programmed in C under OpenVMS for 5 years.
```

Filled Template

```
title: Windows NT software engineer
location: New York City
language: Visual C++, C, C++, Java
platform: Windows NT, OpenVMS
area: multi-threading, client/server,
      anti-piracy
years of experience: nineteen years
```

Text Mining@work

Information Extraction

- Input
  - ✓ Natural language documents (newspaper article, email message etc.)
  - ✓ Pre specified entities, templates
- Output
  - ✓ Specific substrings/parts of document which match the template.

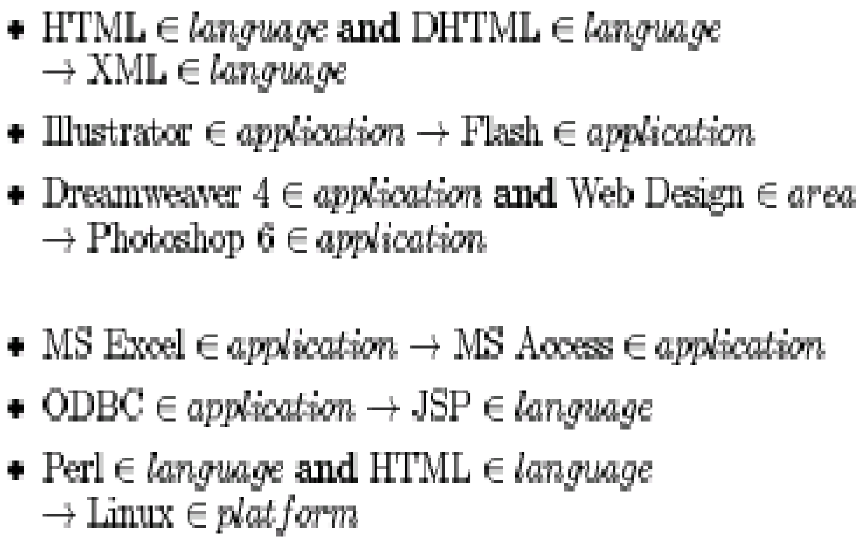
```
Posting from Newsgroup
Telecommunications. Solaris Systems
Administrator. 55-60K. Immediate
need.

3P is a leading telecommunications
firm in need of a energetic
individual to fill the following
position in the Atlanta office:

SOLARIS SYSTEM ADMINISTRATOR
Salary: 50-60K with full benefits
Location: Atlanta, Georgia no
relocation assistance provided
```

FILLED TEMPLATE
job title: SOLARIS SYSTEM ADMINISTRATOR
salary: 55-60K
city: Atlanta
state: Georgia
platform: SOLARIS
area: Telecommunications

Text Mining@work



- HTML $\in$ *language* and DHTML $\in$ *language* $\rightarrow$ XML $\in$ *language*
- Illustrator $\in$ *application* $\rightarrow$ Flash $\in$ *application*
- Dreamweaver 4 $\in$ *application* and Web Design $\in$ *area* $\rightarrow$ Photoshop 6 $\in$ *application*
- MS Excel $\in$ *application* $\rightarrow$ MS Access $\in$ *application*
- ODBC $\in$ *application* $\rightarrow$ JSP $\in$ *language*
- Perl $\in$ *language* and HTML $\in$ *language* $\rightarrow$ Linux $\in$ *platform*

Text Mining nell’Impresa

“Il processo di estrazione di conoscenza, precedentemente sconosciuta, da fonti testuali (agenzie stampa, transazioni, siti Web, e-mail, forum, mailing list…) utilizzabile per prendere decisioni aziendali”

Permette di organizzare/ categorizzare

- *scoprendo tendenze*
- *apprendendo concetti*

Text Mining nell'Impresa

- Perché è necessario…
  - ✓ scoprire quali sono le opinioni, le idee, le tendenze, i gusti degli utenti (clienti) sta diventando sempre più impegnativo:troppi i dati a disposizione e, troppo rapidi i cambi di tendenza
- …Le fonti da analizzare
  - ✓ e-mail, newsgroup, forum, mailing list, lettere, articoli, …
- …L'obiettivo perseguito
  - ✓ analizzare migliaia di testi in pochi secondi, *raggruppandoli* in funzione del loro *contenuto*, estraendo opinioni, tendenze, idee… *degli autori* (analisi delle lettere di lamentela degli utenti)

Text Mining: aree di ricerca correlate

- Information Retrieval
- Text Categorization
- Information Extraction
- Natural Language Processing
- Data Mining

M. Grobelnik, D. Mladenic, and N. Milic-Frayling, 2000.

"Text Mining as Integration of Several Related Research Areas: Report on KDD'2000 Workshop on Text Mining

Intelligent Information Retrieval

Information Retrieval (IR)

- IR deals with the representation, storage, organization of, and access to information items
  - ✓ Types of information items: documents, Web pages, online catalogs, structured records, multimedia objects
- Early goals of the IR area: indexing text and searching for useful documents in a collection
- Searching for pages on the World Wide Web is the most recent “killer app.”
- Concerned firstly with retrieving *relevant* documents to a query.
- Concerned secondly with retrieving from *large* sets of documents *efficiently*.

Typical IR Task

- Given:
  - A corpus of textual natural-language documents.
  - A user query in the form of a textual string.
- Find:
  - A ranked set of documents that are relevant to the query.
- Example of complex information need

Find all documents that address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK)

Typical IR Task

- This full description of the user information need is not necessarily a good query to be submitted to the IR system
- Instead, the user might want to first translate this information need into a query
- This translation process yields a set of *keywords*, or *index terms*, which summarize the user information need
- Given the user query, the key goal of the IR system is to retrieve information that is useful or relevant to the user
- That is, the IR system must rank the information items according to a degree of relevance to the user query

How People Search

- User interaction with search interfaces differs depending on
  - the type of task
  - the domain expertise of the information seeker
  - the amount of time and effort available to invest in the process
- Distinction between **information lookup** and **exploratory search**
- **Information lookup** tasks
  - are akin to fact retrieval or question answering
  - can be satisfied by discrete pieces of information: numbers, dates, names, or Web sites
  - can work well for standard Web search interactions

How People Search

- **Exploratory search** is divided into **learning** and **investigating tasks**
- **Learning search**
  - requires more than single query-response pairs
  - requires the searcher to spend time
    - scanning and reading multiple information items
    - synthesizing content to form new understanding
- **Investigating** refers to a longer-term process which
  - involves multiple iterations that take place over perhaps very long periods of time
  - may return results that are critically assessed before being integrated into personal and professional knowledge bases
  - may be concerned with finding a large proportion of the relevant information available

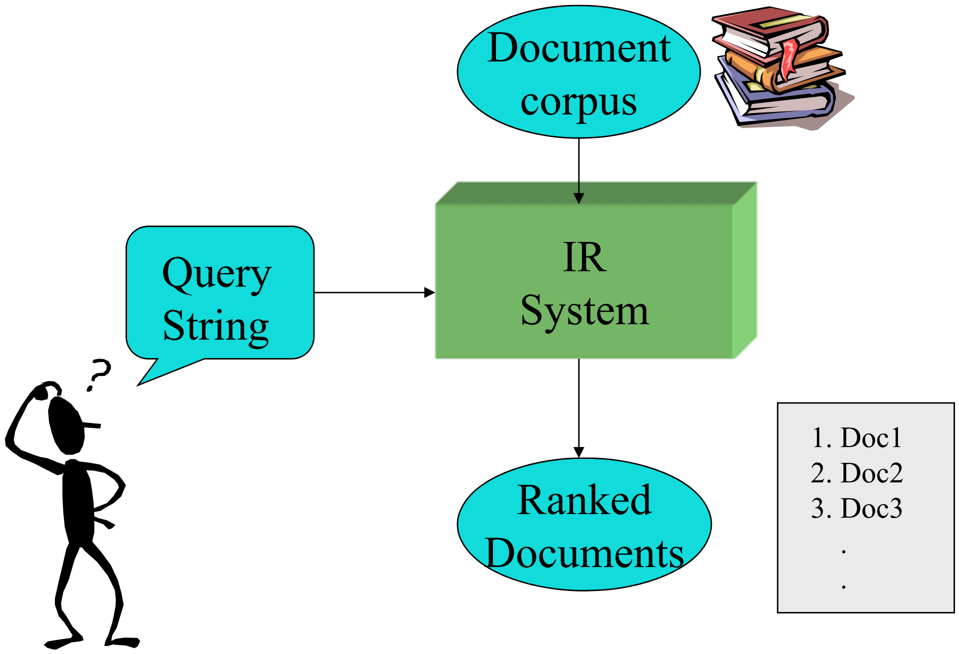
How People Search

- Information seeking can be seen as being part of a larger process referred to as *sensemaking*
- **Sensemaking** is an iterative process of formulating a conceptual representation from a large collection
- Most of the effort in sensemaking goes towards the synthesis of a good representation
- Some sensemaking activities interweave search throughout, while others consist of doing a batch of search followed by a batch of analysis and synthesis

How People Search

- Examples of deep analysis tasks that require sensemaking (in addition to search)
  - ✓ the legal discovery process
  - ✓ epidemiology (disease tracking)
  - ✓ studying customer complaints to improve service
  - ✓ obtaining business intelligence.

IR System



Document corpus

Query String

IR System

Ranked Documents

1. Doc1
2. Doc2
3. Doc3
.
.

Relevance

- Relevance is a subjective judgment and may include:
  - ✓ Being on the proper subject.
  - ✓ Being timely (recent information).
  - ✓ Satisfying the goals of the user and his/her intended use of the information (*information need*).

Keyword Search

- Simplest notion of relevance is that the query string appears verbatim in the document.
- Slightly less strict notion is that the words in the query appear frequently in the document, in any order (*bag of words*).

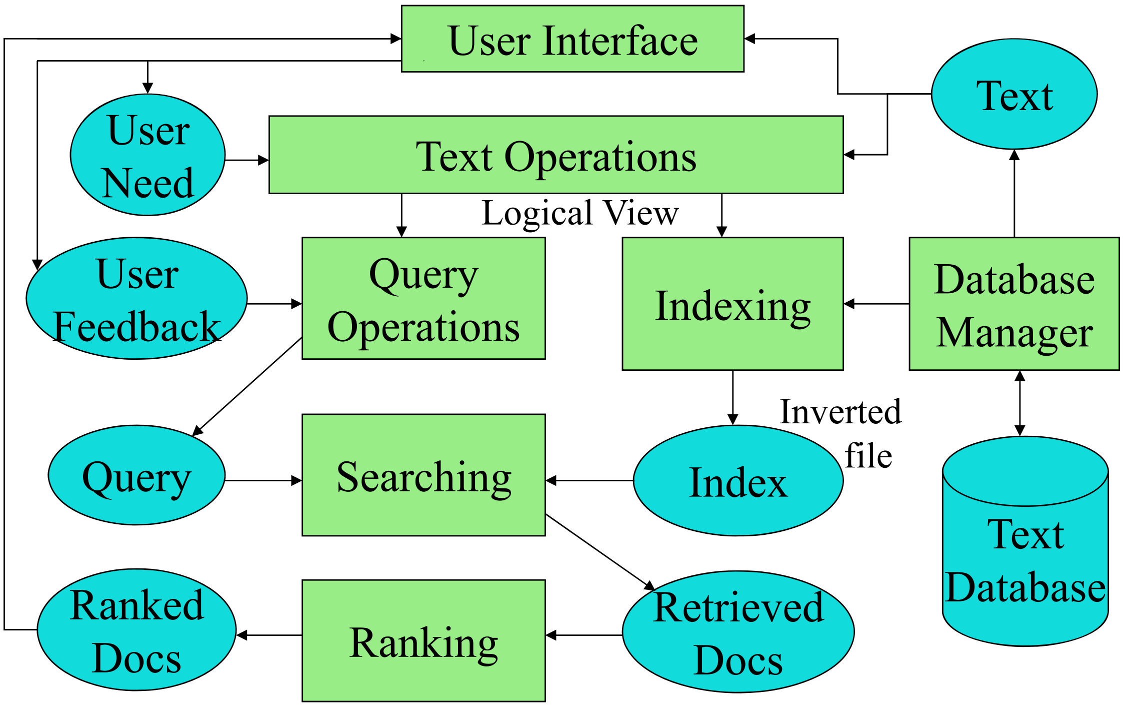
Problems with Keywords

- May not retrieve relevant documents that include *synonymous* terms.
  - ✓ “restaurant” vs. “café”
  - ✓ “PRC” vs. “China”
- May retrieve irrelevant documents that include ambiguous terms (*polysemy*).
  - ✓ “bat” (baseball vs. mammal)
  - ✓ “Apple” (company vs. fruit)

Intelligent IR

- Taking into account the *meaning* of the words used.
- Taking into account the *order* of words in the query.
- Adapting to the user based on direct or indirect feedback (*relevance feedback*): collects feedback, generates new query, repeat retrieval.

IR System Architecture



User Interface

Text

User Need

Text Operations

Logical View

User Feedback

Query Operations

Indexing

Database Manager

Inverted file

Query

Searching

Index

Text Database

Ranked Docs

Ranking

Retrieved Docs

IR System Components

- ***Text Operations*** forms index words (*tokens*).
  - Stopword removal
  - Stemming (reducing words to roots, removing prefix and suffix)
- ***Indexing*** constructs an *inverted index* of word to document pointers.
- ***Searching*** retrieves documents that contain a given query token from the inverted index.
- ***Ranking*** scores all retrieved documents according to a relevance metric. It may also perform *grouping*, i.e. finding commonalities and presenting group of documents.

Intelligent IR?

- Research areas
  - Natural Language Processing
  - Machine Learning

Natural Language Processing

- Focused on the syntactic, semantic, and pragmatic analysis of natural language text and discourse.
- Ability to analyze syntax (phrase structure) and semantics could allow retrieval based on *meaning* rather than keywords.

Natural Lang. Proc: IR Directions

- Methods for determining the sense of an ambiguous word based on context (*word sense disambiguation*).
- Methods for identifying specific pieces of information in a document (*information extraction*).
- Methods for answering specific NL questions from document corpora.

Machine Learning

- Focused on the development of computational systems that improve their performance with experience.
- Automated classification of examples based on learning concepts from labeled training examples (*supervised learning*).
- Automated methods for clustering unlabeled examples into meaningful groups (*unsupervised learning*).