Modern Information Retrieval

Ricardo Baeza-Yates and Berthier Ribeiro-Neto

Addison Wesley, 2006

Modern Information Retrieval

Chapter 1

Introduction

Motivation

- IR: representation, storage, organization of, and access to information items

- Focus is on the user information need

- User information need:
  - Find all docs containing information on college tennis teams which: (1) are maintained by a USA university and (2) participate in the NCAA tournament

- Emphasis is on the retrieval of information (not data)

Motivation

- Data retrieval
  - which docs contain a set of keywords?
  - well defined semantics
  - a single erroneous object implies failure!
- Information retrieval
  - information about a subject or topic
  - semantics is frequently loose
  - small errors are tolerated
- IR system:
  - generates a ranking which reflects relevance
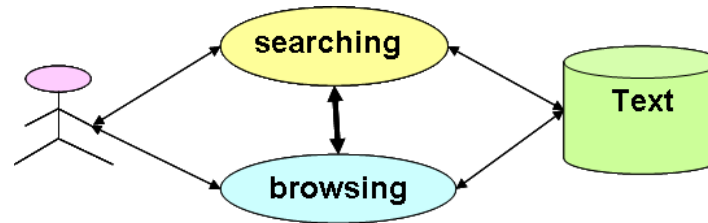 - notion of relevance is most important

Motivation

- IR at the center of the stage
  - In the last 20 years:
    - classification and categorization
    - systems and languages
    - user interfaces and visualization
    - Area was seen as of narrow interest
  - With the Web
    - search is a key application
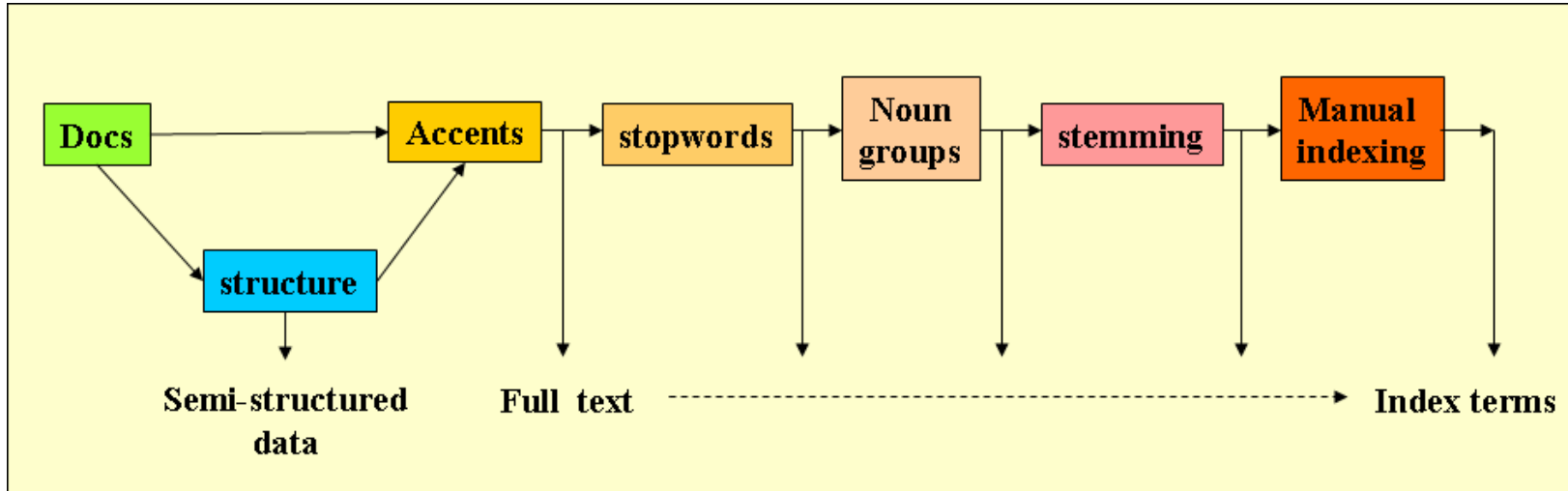 - IR is a fundamental science in the Web

Basic Concepts

- The User Task

- Retrieval
  - information or data
  - more specific purpose
- Browsing
  - glancing around
  - More generic information: F1, cars, France, tourism

Basic Concepts

Docs → Accents → stopwords → Noun groups → stemming → Manual indexing

Docs → structure → Semi-structured data

Full text - - - - - - - - - - - - - - - - - - → Index terms

- Document representation can be viewed as a continuum, in which the logical view of the docs shifts from full text (Web) to customized index terms (Libraries)

The Retrieval Process