**Course: Advance Bio Informatics**

**Module Title: Document Classification**

**Module No: 100**

**Document Classification**

**Given:**

A description of instance, x ∈ X, where X is the instance language or instance space. e.g. how to represent text documents.

A fixed set of categories C = {c1, c2,…, cn}

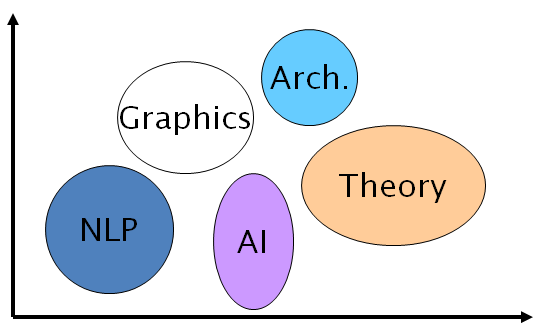
**Determine**

The category of

x: c(x) ∈ C,

Where c(x) is a categorization function whose domain is X and whose range is C.

**A Graphical View of Text Classification**



**Text classification – Example**

This concerns you as a patient.   
  
Our medical records indicate you have had a history of illness. We are now encouraging all our patients to use this highly effective and safe solution.   
  
Proven worldwide, feel free to read the many reports on our site from the BBC & ABC News.   
  
We highly recommend you try this Anti-Microbial Peptide as soon as possible since its world supply is limited. The results will show quickly.

**Classification Methods**

**Manual classification:**

* Used by Yahoo, Medline.
* Very accurate when job is done by experts.
* Consistent when problem size and team is small.
* Difficult and expensive to scale.

**Automatic document classification:**

* Hand-coded rule-based systems.
* Reuters, CIA, … Commercial systems have complex query languages (everything in IR query languages + accumulators)

**Supervised learning** of document-label assignment function:

Autonomy, Kana, MSN, Verity, …

Naive Bayes (simple common method).

k-Nearest Neighbors (simple, powerful).

Support-vector machines (new, more powerful).

**Examples**

LABELS=BINARY “spam” / “not spam”

LABELS=TOPICS finance / sports / Asia

LABELS=OPINION “like” / “hate” / neutral

LABELS=AUTHOR Shakespeare/ Marlowe/ “Ben Jonson”