Text Categorization: Overview and applications

4ICT2: Information Management

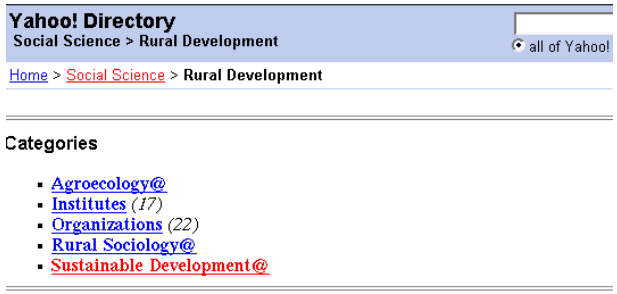
Saturnino Luz
mailto:luzs@cs.tcd.ie

Trinity College, Department of Computer Science
10 Jan 2003

Notes

Background

- Automatic filtering of information
- Improvements on IR systems
- An example of manually built content hierarchies



- Problem:
  - Manually built catalogues are costly to build and maintain

Notes

Related techniques

- From Information Retrieval
  - word indexing (inverted indices etc)
  - corpus statistics (type and token counts, frequency tables etc)
  - feature selection
- From machine learning:
  - automated knowledge acquisition (copying with the knowledge acquisition bottleneck)
  - classifier induction (different techiniques have been applied, yielding various levels of efficacy)

Notes

3-1

Real-world Applications of TC

- Document categorization:
  - Classification of newspaper articles under the appropriate sections (e.g. Science, Politics, etc)
    - Reuters news corpora have been used for training and testing TC systems
  - Classified ads, product catalogues, hierarchical categorisation of web pages...
- Automatic indexing for boolean information retrieval
  - Assigning keywords (and key phrases) of a controlled dictionary to documents
  - Automated metadata creation

Notes

Other instances of potential applications of automatic document categorisation include web product catalogues, mailing list filtering etc.

Examples of controlled dictionaries include the ACM classification system, the MESH thesaurus for medicine, the NASA thesaurus etc.

4-1

Controlled vocabulary: example

MeSH Tree Structures – 2002

Return to Entry Page

1. Anatomy [A]
2. Organisms [B]
   - Invertebrates [B01] +
   - Vertebrates [B02] +
   - Bacteria [B03] +
   - Viruses [B04] +
   - Algae and Fungi [B05] +
   - Plants [B06] +
   - Archaea [B07] +
3. Diseases [C]
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Te
6. Psychiatry and Psychology [F]
7. Biological Sciences [G]
8. Physical Sciences [H]
9. Anthropology, Education, Sociology and Sc
10. Technology and Food and Beverages [J]
11. Humanities [K]
12. Information Science [L]
13. Persons [M]
14. Health Care [N]
15. Geographic Locations [Z]

Vertebrates [B02]
Reptiles [B02.833]
Alligators and Crocodiles [B02.833.100]
▶ Dinosaurs [B02.833.150]
Lizards [B02.833.393] +
Snakes [B02.833.672] +

MeSH Heading	Dinosaurs
Tree Number	B02.833.150
Scope Note	General name for two extinct orders of reptiles from the Mesozoic era: Saurischia and Ornithischia.
Allowable Qualifiers	AB AH BL CF CL EM GD GE IM IN ME MI PH PS SU UR VI
Previous Indexing	Fossils (1975–2001)
Previous Indexing	Reptiles (1966–2001)
History Note	2002
Unique ID	D025061

- From http://www.nlm.nih.gov/mesh/2002/MBrowser.html

Notes

5-1

Applications of TC: Web catalogues

- Hierarchical classification of web pages:
  - Internet portals, recommendation sites etc
  - Manually created catalogues are hard to create and maintain
  - Multiple documents for each category
  - (and possibly) Multiple categories per document
  - Rich sources of extra-linguistic information:
    - links, hypertext structure
    - hierarchical structure of the category set

Notes

6-1

Applications of TC in Computational Linguistics

- Word sense disambiguation
  - Task description: given an ambiguous word, e.g. *bank* as in:
    - *"The banks are selling Euro notes."* vs.
    - *"Dublin grew up at a fording point on the banks of the Liffey."*

    find the sense of each particular occurrence
- *Word contexts* can be seen as *documents*, and
- *Word senses* can be seen as *categories*
- Uses in systems for: context-sensitive spelling, prepositional phrase attachment, machine translation

Notes

Example of PP attachment ambiguity:

- "The children ate the cake with a spoon."
- "The children ate the cake with frosting."

Applications of TC: Text filtering

- "(...) the activity of classifying a stream of incoming documents dispached in an asynchronous way from an information producer to an information consumer"
- Example in [Hayes and Weinstein, 1990]:
  - Producer: a news agency
  - Consumer: a newspaper
- Another example: personalised news web sites
- User interface issues
  - *Adaptive filtering*: user provides (explicit or implicit) feedback and the system builds *user profiles*
  - *routing*: no profile is available and texts are ranked by relevance
  - *batch filtering*: boolean (accept/reject) classification; no user profile is built

Notes

Evaluating TC systems

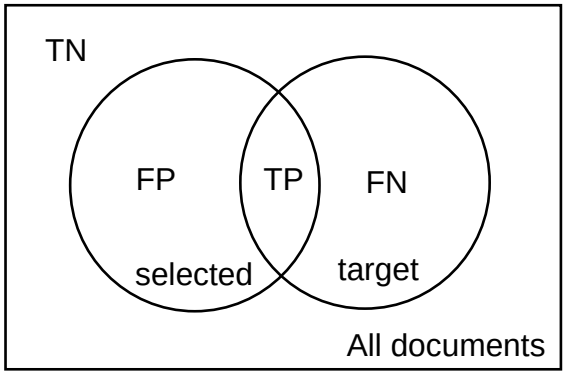
- The most commonly used measures are *Precision* and *recall*:
- Recall: how good the system is at finding relevant documents for a given category (ρ):

$$\rho = \frac{\texttt{true\_positives}}{\texttt{true\_positives} + \texttt{false\_negatives}} \quad (1)$$

- Precision: the "quality" of the classified data (π):

$$\pi = \frac{\texttt{true\_positives}}{\texttt{true\_positives} + \texttt{false\_positives}} \quad (2)$$

Notes



A formal definition of TC

- Notation:
  - $\mathcal{D}$: a set (domain) of documents
    $$\mathcal{D} = \{d_1, ..., d_{|\mathcal{D}|}\}$$
  - $\mathcal{C}$: a set of categories
    $$\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$$
- Text categorisation is the task of assigning a boolean value to each pair $< d_i, c_j >$, s.t.:

$$\begin{cases} < d_i, c_j >= T & \text{if } d_i \text{ is filed under } c_j \\ < d_i, c_j >= F & \text{if } d_i \text{ is NOT filed under } c_j \end{cases} \quad (3)$$

Notes

Target and classifier functions

- The goal of TC is to approximate the (unknown) *target function*:

$$\breve{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow \{T, F\}$$

  which classifies (as defined in (3)) each document *correctly* into one or more categories in $\mathcal{C}$.

- This is done by means of a *classifier function*

$$\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{T, F\}$$

- For evaluation purposes, a target function could be assumed to describe, for instance, a set of hand-annotated documents

References

Philip J. Hayes and Steven P. Weinstein. Construe-TIS: A system for content-based indexing of a database of news stories. In Alain Rappaport and Reid Smith, editors, *Proceedings of the IAAI-90 Conference on Innovative Applications of Artificial Intelligence*, pages 49–66. MIT Press, 1990. ISBN 0-262-68068-8.

Notes