

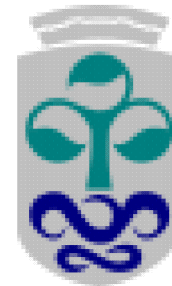
Thesaurus Topic Assignment using Hierarchical Text Categorization

F.J. Ribadas, E. Lloves and **V.M. Darriba**



**Compilers and
Languages**

<http://www.colegroup.org>



Computer Science Dept.
University of Vigo

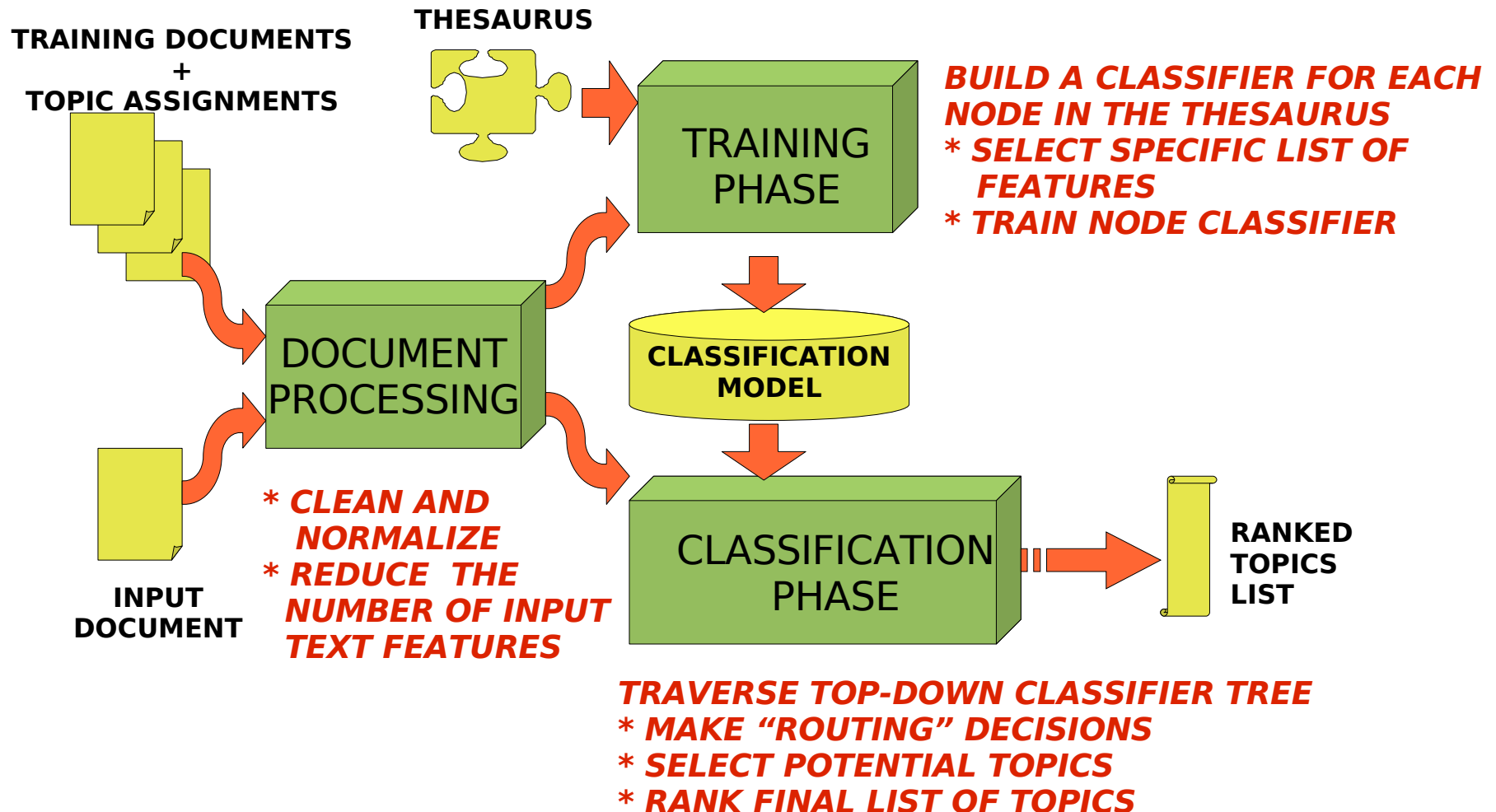
iNEWS 2007

Improving Non English Web Searching
July 27, 2007. Amsterdam

MOTIVATION & BASIC IDEAS

- **Motivation:** Efficient processing in doc. management tasks
 - Use of specialized thesaurus (hierarchically structured concepts)
 - **Our case:** legislative texts in Spanish (subventions, grants, ...)
- **Our approach:** use multiple label text classification over hierarchical categories to automatically assign descriptive topics
- **Method:** Use hierarchy to break the problem into small pieces [Koller&Sahami, 1997]
 - Train a set of partial classifiers associated to each internal node in the topic tree structure
 - Take advantage of topic structure to reduce:
 - features to be taken into account+classes at each classification step
 - More specific decisions → improve overall categorization quality

SYSTEM ARCHITECTURE



FUTURE WORK:

- Fine tuning of parameters and application in new domains and languages
- Integration of feedback and auto-correction features
- Application as a categorization tool on web directories (Open Directory Project)