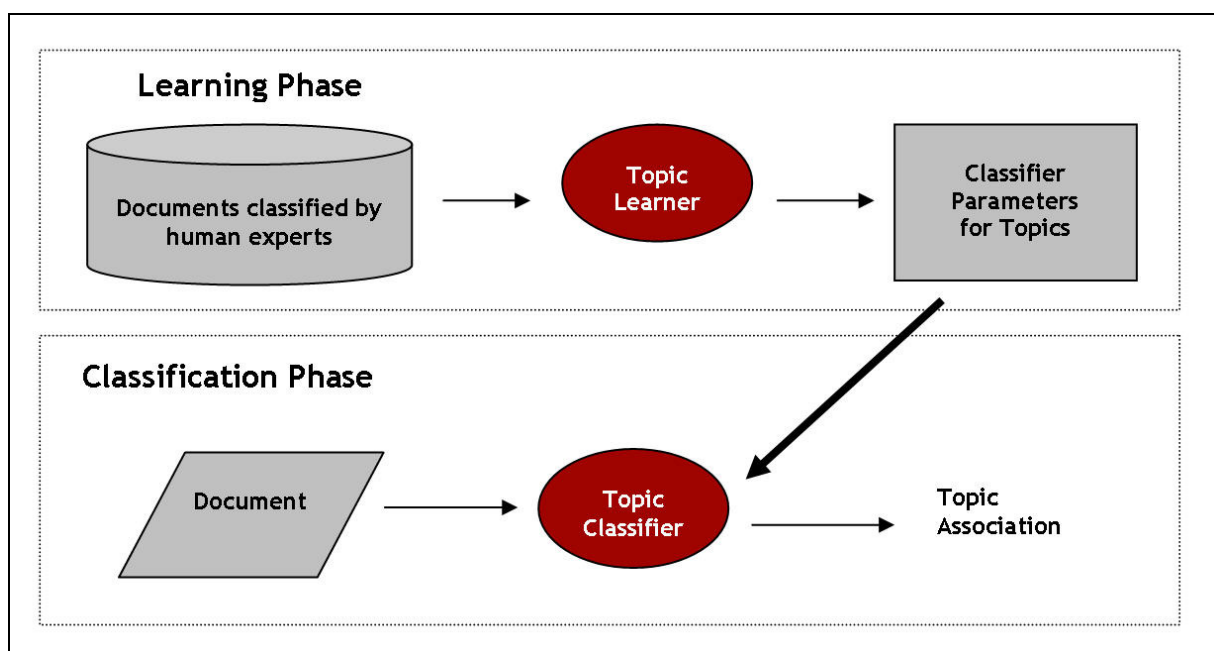


## FACT SHEET

### Fully automatic text classification with TopicFinder within the REEGLE knowledge portal

Intrafind's TopicFinder is a product for fully automatic text classification. Basically text classification means *content-based assignment* of one or more *predefined topics (categories)* to text. TopicFinder works in two phases, the *learning phase* and the subsequent *classification or application phase*. In the *learning phase* users define topics by providing *sample documents (training examples)* for each of these topics. In the *classification phase* new (previously unseen) documents can be given to the topic classifier which returns a topic association (a rating or classification for each topic).



Intrafind's TopicFinder combines advanced linguistic preprocessing (baseform and compound word analysis, phrase detection) and latest machine learning technology (support vector machines) to achieve outstanding classification accuracy.

#### Use in the project

Reegle is a search engine for environmental information. Besides full-text search, search can be restricted to special topics. The user may state his query (e.g. a person name) and restrict search to documents about one of the predefined topics such as „renewable energy / wind energy“. Topic associations are delivered by Intrafind's TopicFinder. Thus no manual indexing and categorization is necessary.

## Defined hierarchies

Currently two independent topic hierarchies are used. There are 33 topics in the „sectors“ hierarchy

- CLIMATE-PROTECTION
  - CLIMATE-PROTECTION-CARBON-FINANCING
  - CLIMATE-PROTECTION-CLIMATE-CHANGE
  - CLIMATE-PROTECTION-EMISSIONS-TRADING
  - CLIMATE-PROTECTION-MISC
- COGENERATION
- DISTRICT-HEATING-SYSTEMS
- ENERGY
- ENERGY-EFFICIENCY
  - ENERGY-EFFICIENCY-APPLIANCES
  - ENERGY-EFFICIENCY-BUILDINGS
  - ENERGY-EFFICIENCY-COOLING
  - ENERGY-EFFICIENCY-ENERGY-SERVICES
  - ENERGY-EFFICIENCY-HEATING
  - ENERGY-EFFICIENCY-INDUSTRIAL-APPLICATIONS
  - ENERGY-EFFICIENCY-LIGHTING
  - ENERGY-EFFICIENCY-MISC
  - ENERGY-EFFICIENCY-TRANSMISSION-DISTRIBUTION
  - ENERGY-EFFICIENCY-TRANSPORT
- RENEWABLE-ENERGY
  - RENEWABLE-ENERGY-BIOFUELS
  - RENEWABLE-ENERGY-BIOGAS
  - RENEWABLE-ENERGY-BIOMASS
  - RENEWABLE-ENERGY-GEOTHERMAL-ENERGY
  - RENEWABLE-ENERGY-HYDRO-POWER-LARGE-SCALE
  - RENEWABLE-ENERGY-HYDRO-POWER-SMALL-SCALE
  - RENEWABLE-ENERGY-HYDROGEN-FUEL-CELLS
  - RENEWABLE-ENERGY-MARINE-(WAVE-TIDAL)-ENERGY
  - RENEWABLE-ENERGY-MISC
  - RENEWABLE-ENERGY-PHOTOVOLTAICS
  - RENEWABLE-ENERGY-SOLAR-THERMAL-ENERGY
  - RENEWABLE-ENERGY-WIND-ENERGY
- RURAL-ELECTRIFICATION

and 12 topics in the „type of information“ hierarchy.

- DEVELOPMENTS-IN-RE-EE-FINANCING
- DEVELOPMENTS-IN-RE-EE-POLICY
- ECONOMIC-DATA-OCOR
- ENERGY-CONSUMPTION-AND-DEMAND-DATA-BCOR
- ENERGY-SUPPLY-AND-PRICING-DATA-BCOR
- FINANCING-MECHANISMS-BM&RMS
- LAWS-AND-REGULATIONS
- NATIONAL-OR-REGIONAL-P&S
- NEWS-AND-ANNOUNCEMENTS
- REPORTS-OR-STUDIES-ON-RE-EE-F&I
- REPORTS-OR-STUDIES-ON-RE-EE-P&R
- SOURCES-OF-FINANCE-FOR-RE-AND-EE

## Training phase

About 4000 documents were annotated manually in order to supply input for the training phase. These data has been supplied by Reegle. Classification accuracy (average recall and precision) is about 80% for the sectors hierarchy and 70% for the „type of information“ hierarchy. Note that these numbers have been computed on independent test sets (5-fold cross validation).

TopicFinder offers a very intuitive training and testing GUI. It allows to identify possible inconsistencies in the manual topic associations of training and test data (e.g. very similar documents with differing topic associations) which should be reviewed by experts. Usually experts agree to change manual topic association or at least accept additional topic associations and after retraining on these modified training data classification accuracy is

increased. Since this step has not yet been done in the Reegle project we expect to achieve even higher classification accuracy in the near future.

### **User feedback & automatic re-training**

Furthermore, TopicFinder and the Reegle Web-site offer users the possibility to give feedback concerning wrong or missing topic associations of documents. Furthermore TopicFinder automatically marks new documents for which topic associations are unclear for later verification by experts. Both kinds of documents can be included into the training data and due to regular retraining the system constantly improves.

### **Technical integration – Application Phase**

The classification tool TopicFinder is fully integrated with the search engine iFinder. Classification is done before indexing the newly acquired content. The generated topics are stored additionally to the original content within the iFinder index. Thus it's possible to combine fulltext queries and topic queries. The user will be able to get e.g. *all* new documents within the topic "*Renewable Energy*" or to find all documents matching the entered query "*Ethanol*" within a selected subtopic "*Biofuels*".

### **Benefits**

- Time and cost savings by the automatic topic assignment instead of manual processing by editors.
- Improved usability of the research process by topic filters.
- More reliable results of higher quality.

---

### **Company - IntraFind Software AG**

Finding, filtering, presenting, and structuring relevant textual information in an intelligent way - that's what our text-mining and retrieval products have been created for. Our products are based on a sophisticated combination of linguistic, semantic and statistic / information-theoretic techniques.

- linguistic text analysis
- thesaurus-based search
- similarity search
- associativ-semantic search
- text classification for topic based search
- multi- and crosslingual search
- information extraction, Named Entity Recognition
- clustering

Using the IntraFind products provides clear advantage for our customers: High-quality search solutions with high efficiency and an excellent price/performance ratio.

### **Contact**

IntraFind Software AG, Fraunhofer Strasse 15, 82152 Martinsried ,Germany

fon: 0049-89-8906-9700, web: [www.intrafind.de](http://www.intrafind.de), online democenter: [www.intrafind.org](http://www.intrafind.org)