Processing Content with

LCIndex

You want to know what the most relevant topics in your documents are? You want high-quality indexing of your documents? You want a concept-based search facility for your documents? You want to create an efficient index for your publication?

LCIndex offers these functionalities on a very high level. It allows for easy and quick indexing of large sets of documents and thus creates the basis for efficient search.

Automatic Indexation

LCIndex creates a concept-based representation of a document's content. It extracts and exhibits those concepts that optimally represent the content of a document. Such a representation can then be the basis for a concept-based retrieval of information. So, for a given document, **LCIndex** might tell you that it is about *3D computer aided design* or about *cost reduction strategies for hospitals*. As the representations are meant to be concept-based you would find a document even by searching with *strategies for the reduction of costs in hospitals*.

LCIndex extracts the most relevant key phrases from a text. To that end, linguistic analysis techniques from **LCCore** and **LCTerm** are used and enhanced with statistical methods that weight the relevance of a term. **LCIndex** thus differentiates between relevant and less relevant terms, or concepts, for the representation of the content of a text.

The linguistic terms that qualify for representing the content of a text are first extracted on the basis of linguistic criteria. However, linguistic properties are not sufficient for qualifying as a key phrase. Statistical techniques are used to determine the relevance of the linguistic terms. A standard measure for the determination of the relevance of a term is informativity. This is usually measured as tf-idf (term frequency - inverse document frequency).

A very important feature for calculating tf-idf is that the search terms are reduced to their base form. This may not be that difficult for English, but it is for other languages. The purpose is that inflected forms are not considered independent words but instances of the same word. Variations between British and American English are also mapped onto a common base form.

Thesaurus-based Indexation

If a thesaurus is available this is a very useful resource for indexing. A thesaurus constitutes a conceptual representation of a piece of the world and thus provides a valuable resource for extracting and weighting terms. In addition, knowledge about semantic relations such as the relation between superordinate and subordinate terms is used to represent the content of a document or a piece of a document (e.g. the term *bipolar electrode* infers the term *bipolar device*).

Linguistic processing of thesaurus terms allows for intelligent search by using variants of terms as search terms as in the examples shown on the right.

Features

- thesaurus-based indexing
- free indexing
- statistical weighting of key phrases
- document classification
- word clouds
- document similarity

Languages

- German
- English

Search and Find

cost saving
→ *... a substantial **cost reduction** and thus a competitive price ...*

radiation transport
→ *... collective effects to **transport of radiation** in plasmas.*

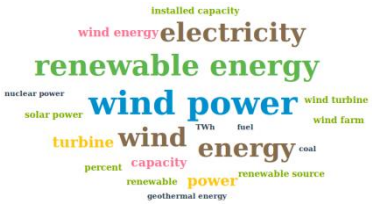
waste gas purification
→ *To **purify waste gas**, use is made of gas scrubbers in which*

abrasive grinder
→ *... hire an **abrasive** floor **grinder** to sort that one out*

signal processing
→ *..., with an emphasis on discrete **signal** and image **processing**.*

Word clouds

Word clouds are collections of semantically similar concepts. They are determined by co-occurrence in large document collections. Word clouds can be used for different purposes. One application scenario is associative search. If a search term does not lead to relevant results semantically similar concepts might help.



Indexation can also take benefit from word clouds in that ambiguities may be resolved. If a term such as *virus* is found to be relevant for a given document it is not yet clear whether a computer virus or an organic virus is meant. By comparing the word clouds of the two readings of *virus* with the prominent terms of the document the ambiguity can be resolved

Classification

In document classification, the documents are assigned a domain label. On a general level such a domain label may be *chemistry* or *engineering*. On a more refined level the labels may be *artificial intelligence* or *semiconductor theory*. We offer various classification systems but we are also happy to use your classification scheme if you have one.

For classification, a classifier is trained with large (classified) document collections. However, it is also possible to establish a classification system without training data.

Document Clustering

Indexing and classification pave the way for clustering documents into groups. The document clustering technique makes a simple assumption, namely, the more similar the indexing of some documents the more similar the documents are.

Similarity of documents may also be used for searching for documents.

Need more Information?

If you have further questions, or would like to know how we can help you with your specific requirements, don't hesitate to contact us.

IAI Linguistic Content AG

Martin-Luther-Str. 14
66111 Saarbrücken
+49 (0)681 38951-0

www.iailc.de
kontakt@iailc.de