Title: Document classification

URL: https://en.wikipedia.org/wiki/Document_classification

PageID: 1331441

Categories: Category:Data mining, Category:Information science, Category:Knowledge representation, Category:Machine learning, Category:Natural language processing

---

Document classification or document categorization is a problem in library science , information science and computer science . The task is to assign a document to one or more classes or categories . This may be done "manually" (or "intellectually") or algorithmically . The intellectual classification of documents has mostly been the province of library science, while the algorithmic classification of documents is mainly in information science and computer science. The problems are overlapping, however, and there is therefore interdisciplinary research on document classification.

The documents to be classified may be texts, images, music, etc. Each kind of document possesses its special classification problems. When not otherwise specified, text classification is implied.

Documents may be classified according to their subjects or according to other attributes (such as document type, author, printing year etc.). In the rest of this article only subject classification is considered. There are two main philosophies of subject classification of documents: the content-based approach and the request-based approach.

"Content-based" versus "request-based" classification

Content-based classification is classification in which the weight given to particular subjects in a document determines the class to which the document is assigned. It is, for example, a common rule for classification in libraries, that at least 20% of the content of a book should be about the class to which the book is assigned. In automatic classification it could be the number of times given words appears in a document.

Request-oriented classification (or -indexing) is classification in which the anticipated request from users is influencing how documents are being classified. The classifier asks themself: “Under which descriptors should this entity be found?” and “think of all the possible queries and decide for which ones the entity at hand is relevant” (Soergel, 1985, p. 230).

Request-oriented classification may be classification that is targeted towards a particular audience or user group. For example, a library or a database for feminist studies may classify/index documents differently when compared to a historical library. It is probably better, however, to understand request-oriented classification as policy-based classification : The classification is done according to some ideals and reflects the purpose of the library or database doing the classification. In this way it is not necessarily a kind of classification or indexing based on user studies. Only if empirical data about use or users are applied should request-oriented classification be regarded as a user-based approach.

Classification versus indexing

Sometimes a distinction is made between assigning documents to classes ("classification") versus assigning subjects to documents (" subject indexing ") but as Frederick Wilfrid Lancaster has argued, this distinction is not fruitful. "These terminological distinctions,” he writes, “are quite meaningless and only serve to cause confusion” (Lancaster, 2003, p. 21). The view that this distinction is purely superficial is also supported by the fact that a classification system may be transformed into a thesaurus and vice versa (cf., Aitchison, 1986, 2004; Broughton, 2008; Riesthuis & Bliedung, 1991). Therefore, assigning a subject term to a document in an index is equivalent to assigning that document to the class of documents indexed by that term (all documents indexed or classified as X belong to the same class of documents).

Automatic document classification (ADC)

Automatic document classification tasks can be divided into three sorts: supervised document classification where some external mechanism (such as human feedback) provides information on the correct classification for documents, unsupervised document classification (also known as document clustering), where the classification must be done entirely without reference to external information, and semi-supervised document classification , where parts of the documents are labeled by the external mechanism. There are several software products under various license models available.

Techniques

Automatic document classification techniques include:

Artificial neural network

Concept Mining

Decision trees such as ID3 or C4.5

Expectation maximization (EM)

Instantaneously trained neural networks

Latent semantic indexing

Multiple-instance learning

Naive Bayes classifier

Natural language processing approaches

Rough set -based classifier

Soft set -based classifier

Support vector machines (SVM)

K-nearest neighbour algorithms

tf–idf

Applications

Classification techniques have been applied to

spam filtering , a process which tries to discern E-mail spam messages from legitimate emails

email routing , sending an email sent to a general address to a specific address or mailbox depending on topic

language identification , automatically determining the language of a text

genre classification, automatically determining the genre of a text

readability assessment , automatically determining the degree of readability of a text, either to find suitable materials for different age groups or reader types or as part of a larger text simplification system

sentiment analysis , determining the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.

health-related classification using social media in public health surveillance

article triage, selecting articles that are relevant for manual literature curation, for example as is being done as the first step to generate manually curated annotation databases in biology

See also

Classification

Compound-term processing

Concept-based image indexing

Document

Document retrieval

Information retrieval

Knowledge organization

Knowledge organization system

Library classification

Subject (documents)

Subject indexing

References

Further reading

Fabrizio Sebastiani. Machine learning in automated text categorization . ACM Computing Surveys, 34(1):1–47, 2002.

Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. Information Retrieval: Implementing and Evaluating Search Engines Archived 2020-10-05 at the Wayback Machine . MIT Press, 2010.

External links

Introduction to document classification

Bibliography on Automated Text Categorization Archived 2019-09-26 at the Wayback Machine

Bibliography on Query Classification Archived 2019-10-02 at the Wayback Machine

Text Classification analysis page

Learning to Classify Text - Chap. 6 of the book Natural Language Processing with Python (available online)

TechTC - Technion Repository of Text Categorization Datasets Archived 2020-02-14 at the Wayback Machine

David D. Lewis's Datasets

BioCreative III ACT (article classification task) dataset [usurped]

v

t

e

AI-complete

Bag-of-words

n -gram Bigram Trigram

Bigram

Trigram

Computational linguistics

Natural language understanding

Stop words

Text processing

Argument mining
Collocation extraction
Concept mining
Coreference resolution
Deep linguistic processing
Distant reading
Information extraction
Named-entity recognition
Ontology learning
Parsing Semantic parsing Syntactic parsing
Semantic parsing
Syntactic parsing
Part-of-speech tagging
Semantic analysis
Semantic role labeling
Semantic decomposition
Semantic similarity
Sentiment analysis
Terminology extraction
Text mining
Textual entailment
Truecasing
Word-sense disambiguation
Word-sense induction
Compound-term processing
Lemmatisation
Lexical analysis
Text chunking
Stemming
Sentence segmentation
Word segmentation
Multi-document summarization
Sentence extraction
Text simplification
Computer-assisted
Example-based
Rule-based
Statistical

Transfer-based

Neural

BERT

Document-term matrix

Explicit semantic analysis

fastText

GloVe

Language model (large)

Latent semantic analysis

Seq2seq

Word embedding

Word2vec

Corpus linguistics

Lexical resource

Linguistic Linked Open Data

Machine-readable dictionary

Parallel text

PropBank

Semantic network

Simple Knowledge Organization System

Speech corpus

Text corpus

Thesaurus (information retrieval)

Treebank

Universal Dependencies

BabelNet

Bank of English

DBpedia

FrameNet

Google Ngram Viewer

UBY

WordNet

Wikidata

Speech recognition

Speech segmentation

Speech synthesis

Natural language generation

Optical character recognition

Document classification
Latent Dirichlet allocation
Pachinko allocation
Automated essay scoring
Concordancer
Grammar checker
Predictive text
Pronunciation assessment
Spell checker
Chatbot
Interactive fiction
Question answering
Virtual assistant
Voice user interface
Formal semantics
Hallucination
Natural Language Toolkit
spaCy