Document classification

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.

Contents

"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.[1] 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[2]).

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[3]). 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,[4] 2004;[5] Broughton, 2008;[6] Riesthuis & Bliedung, 1991[7]). Therefore, is the act of labeling a document (say by assigning a term from a controlled vocabulary to a document) at the same time to assign 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**,[8] where parts of the documents are labeled by the external mechanism. There are several software products under various license models available.[9][10][11][12][13]

Techniques

Automatic document classification techniques include:

- Expectation maximization (EM)
- Naive Bayes classifier
- tf–idf
- Instantaneously trained neural networks
- Latent semantic indexing
- Support vector machines (SVM)
- Artificial neural network
- K-nearest neighbour algorithms
- Decision trees such as ID3 or C4.5
- Concept Mining
- Rough set-based classifier
- Soft set-based classifier
- Multiple-instance learning
- Natural language processing approaches

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[14]
- language identification, automatically determining the language of a text
- genre classification, automatically determining the genre of a text[15]
- 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 [16]
- 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.[17]

See also

- Categorization

- Classification (disambiguation)
- Compound term processing
- Concept-based image indexing
- Content-based image retrieval
- Document
- Supervised learning, unsupervised learning
- Document retrieval
- Document clustering
- Information retrieval
- Knowledge organization
- Knowledge Organization System
- Library classification
- Machine learning
- Native Language Identification
- String metrics
- Subject (documents)
- Subject indexing
- Text mining, web mining, concept mining

Further reading

- Fabrizio Sebastiani. Machine learning in automated text categorization (https://arxiv.org/pdf/cs.ir/0110053). 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 (http://www.ir.uwaterloo.ca/book/). MIT Press, 2010.

References

1. Library of Congress (2008). The subject headings manual. Washington, DC.: Library of Congress, Policy and Standards Division. (Sheet H 180: "Assign headings only for topics that comprise at least 20% of the work.")
2. Soergel, Dagobert (1985). Organizing information: Principles of data base and retrieval systems (https://books.google.com/books?id=cHbNCgAAQBAJ&printsec=frontcover#v=onepage&q&f=false). Orlando, FL: Academic Press.
3. Lancaster, F. W. (2003). Indexing and abstracting in theory and practice. Library Association, London.
4. Aitchison, J. (1986). "A classification as a source for thesaurus: The Bibliographic Classification of H. E. Bliss as a source of thesaurus terms and structure." Journal of Documentation, Vol. 42 No. 3, pp. 160-181.
5. Aitchison, J. (2004). "Thesauri from BC2: Problems and possibilities revealed in an experimental thesaurus derived from the Bliss Music schedule." Bliss Classification Bulletin, Vol. 46, pp. 20-26.
6. Broughton, V. (2008). "A faceted classification as the basis of a faceted terminology: Conversion of a classified structure to thesaurus format in the Bliss Bibliographic Classification (https://link.springer.com/article/10.1007/s10516-007-9027-7) (2nd Ed.).]" Axiomathes, Vol. 18 No.2, pp. 193-210.
7. Riesthuis, G. J. A., & Bliedung, St. (1991). "Thesaurification of the UDC." Tools for knowledge organization and the human interface, Vol. 2, pp. 109-117. Index Verlag, Frankfurt.
8. Rossi, R. G., Lopes, A. d. A., and Rezende, S. O. (2016). Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts (https://www.sciencedirect.com/science/article/pii/S0306457315000990). Information Processing & Management, 52(2):217–257.
9. An Interactive Automatic Document Classification Prototype (https://pdfs.semanticscholar.org/bea4/a204239556a29228decc9e029c326e4900b7.pdf)
10. Interactive Automatic Document Classification Prototype (https://seer.lcc.ufmg.br/index.php/jidm/article/download/43/41An) Archived (https://web.archive.org/web/20150424122349/https://seer.lcc.ufmg.br/index.php/jidm/article/download/43/41An) April 24, 2015, at the Wayback Machine
11. Document Classification - Artsyl (https://archive.is/20141208063727/http://www.artsyltech.com/da_classification.htmlAutomatic)
12. ABBYY FineReader Engine 11 for Windows (http://www.abbyy.com/ocr_sdk_windows/what_is_new/classification/)
13. Classifier - Antidot (http://www.antidot.net/classifier/)

14. Stephan Busemann, Sven Schmeier and Roman G. Arens (2000). Message classification in the call center (https://arxiv.org/pdf/cs/0003060). In Sergei Nirenburg, Douglas Appelt, Fabio Ciravegna and Robert Dale, eds., Proc. 6th Applied Natural Language Processing Conf. (ANLP'00), pp. 158-165, ACL.
15. Santini, Marina; Rosso, Mark (2008), *Testing a Genre-Enabled Application: A Preliminary Assessment* (http://www.bcs.org/upload/pdf/ewic_fd08_paper7.pdf) (PDF), BCS IRSG Symposium: Future Directions in Information Access, London, UK, pp. 54–63
16. X. Dai, M. Bikdash and B. Meyer, "From social media to public health surveillance: Word embedding based clustering method for twitter classification," SoutheastCon 2017, Charlotte, NC, 2017, pp. 1-7. doi: 10.1109/SECON.2017.7925400, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7925400&isnumber=7925258
17. Krallinger, M; Leitner, F; Rodriguez-Penagos, C; Valencia, A (2008). "Overview of the protein-protein interaction annotation extraction task of Bio *Creative* II" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2559988). *Genome Biology*. 9 Suppl 2: S4. doi:10.1186/gb-2008-9-s2-s4 (https://doi.org/10.1186%2Fgb-2008-9-s2-s4). PMC 2559988 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2559988). PMID 18834495 (https://www.ncbi.nlm.nih.gov/pubmed/18834495).

External links

- Introduction to document classification (http://isp.imm.dtu.dk/thor/projects/multimedia/textmining/node11.html)
- Bibliography on Automated Text Categorization (http://www.cs.technion.ac.il/~gabr/resources/atc/atcbib.html)
- Bibliography on Query Classification (http://liinwww.ira.uka.de/bibliography/Ai/query-classification.html)
- Text Classification (http://www.gabormelli.com/RKB/Text_Classification_Task) analysis page
- Learning to Classify Text - Chap. 6 of the book Natural Language Processing with Python (http://www.nltk.org/book/ch06.html) (available online)
- TechTC - Technion Repository of Text Categorization Datasets (http://techtc.cs.technion.ac.il)
- David D. Lewis's Datasets (http://www.daviddlewis.com/resources/testcollections/)
- BioCreative III ACT (article classification task) dataset (http://www.biocreative.org/tasks/biocreative-iii/ppi/)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Document_classification&oldid=858163401"

This page was last edited on 5 September 2018, at 14:03 (UTC).