Metódy strojovej organizácie kolekcie dokumentov (informačný prieskum)

Martin Záborský

Abstrakt

Obrovské množstvo dát v rôznej forme je v súčasnosti nemožné spracovávať „ručne“, teda bez výpočtovej podpory a rôznych modelov, ktoré proces spracovania značne urýchľujú a umožňujú analytickými metódami identifikovať relevantné vzťahy, prípadne opakujúce sa vzory predstavujúce nové informácie, resp. znalosti. Výpočtové teórie a nástroje využívajú poznatky z rôznych odborov a ich vhodnou synergiou sa snažia uľahčiť spracovanie veľkého množstva dát tak, aby uspokojili informačné potreby používateľov.

Kľúčové slová: digitálne humanitné vedy, získavanie znalostí, strojové učenie, text mining

Abstract

A huge amount of data in various forms is currently impossible to process "manually", i.e. without computational support and different models, which speed up the whole process and enable analytical methods to identify relevant relationships and repeating patterns representing new information or knowledge. Computational theories and tools use insights from different disciplines and, by their appropriate synergies, seek to facilitate the processing of large amounts of data to meet the information needs of users.

Keywords: digital humanities, knowledge discovery, machine learning, text mining

Úvod

Súčasné výpočtové prostriedky a algoritmy umožňujú relatívne rýchle vyhľadávanie aj vo veľkom archíve, avšak potenciálne obrovské množstvo dokumentov získané z procesu vyhľadávania môže pre používateľa predstavovať problém, nakoľko nedokáže efektívne posúdiť ich relevanciu v súvislosti s hľadanou témou alebo frázou. Existujú rôzne metódy, ako sa s týmto problémom vysporiadať a ponúknuť užívateľovi výsledky zoradené podľa vypočítanej relevancie, napr. Okapi BM25 (Robertson a Zaragoza 2009) alebo PageRank (Page et al. 1998), ale používateľ musí pri vyhľadávaní vedieť formulovať, čo vlastne hľadá. Z uvedeného vyplýva, že nie je problém vyhľadať relevantné dokumenty pre zadanú frázu, problém je efektívne uspokojiť informačné potreby používateľa.

Práve v tomto rozkole vzniká priestor pre výpočtové metódy na organizovanie, vyhľadávanie a pochopenie súvislostí vo veľkých korpusoch, keďže používateľ často ani nevie, čo konkrétne hľadá (v zmysle kľúčových slov) alebo pozná iba okruh svojho záujmu – pre hľadanú frázu môže systém poskytnúť relevantné výsledky, ale pokiaľ je zadaná fráza nesprávne, resp. príliš všeobecne formulovaná, môžu byť prezentované výsledky zoradené podľa vypočítanej relevancie s ohľadom na informačné potreby používateľa nepodstatné. Tento problém v minulosti riešilo ručné posúdenie obsahu dokumentov. Katalogizovanie dokumentov do predefinovaných kategórií a ich zaradenie do vhodne navrhnutej adresárovej štruktúry umožnilo používateľovi celú kolekciu prehľadávať na rôznych úrovniach všeobecnosti (tzv. *browsing*), až kým v nej nenašiel požadovanú informáciu. Keďže však nie je v ľudských silách prečítať alebo aspoň posúdiť obrovské množstvo dokumentov, ktoré je v súčasnosti k dispozícii a ktoré sa každý deň rozrastá rýchlym tempom, nepredstavuje opísaný prístup schodnú cestu pri efektívnej organizácii veľkej kolekcie dokumentov.

S rastúcim množstvom teda rastie potreba kolekciu dokumentov aspoň nejakým spôsobom organizovať, čo je s ohľadom na efektivitu bez strojovej podpory prakticky nereálne. Hoci súčasná technológia a metódy neumožňujú strojom porozumieť významu textu a ľudský faktor bude stále zohrávať dôležitú úlohu pri posudzovaní jeho obsahu, niektoré úlohy získavania znalostí z textov sa s rôznou mierou úspešnosti dajú riešiť metódami strojového učenia, ktoré pri riešení problému využívajú pravdepodobnostné a štatistické princípy s cieľom odhaliť vzory skryté v textoch a na základe nich identifikovať nové znalosti o kolekcii, resp. o jednotlivých dokumentoch. Tieto metódy prinášajú dobré výsledky napriek tomu, že sa sémantikou textu ani hlbšou lingvistickou analýzou nezaoberajú a na dokument často nazerajú iba ako na multimnožinu slov v ňom obsiahnutých (tzv. model *bag-of-words*).

Úloha (automatickej alebo poloautomatickej) organizácie dokumentov predstavuje problém strojového učenia s cieľom usporiadania a zaradenia dokumentov do tried alebo kategórií, ktoré majú spoločné charakteristiky z určitého uhla pohľadu. Tento pohľad môže zahŕňať rôzne indikátory a úrovne, ktoré sa líšia v závislosti od aplikácie. Podľa toho, či sú k dispozícii vopred definované kategórie a k nim zaradené vzorové dokumenty, ktoré slúžia ako trénovacie dáta, hovoríme o *strojovom učení kontrolovanom* (učenie s učiteľom, supervised learning), resp. *nekontrolovanom* (učenie bez učiteľa, unsupervised learning). Príkladom kontrolovaného učenia je problém kategorizácie dokumentov, príkladom nekontrolovaného učenia je zhlukovanie dokumentov.

Problematikou získavania znalostí z textov sa zaoberáme z perspektívy tzv. *digitálnych humanitných vied* (známych tiež pod anglickým pojmom digital humanities), ktoré predstavujú snahu o využívanie digitálnych technológií a výpočtových prostriedkov v humanitných odboroch (Drucker et al. 2013).

Prehľad úloh a metód

Výskumom v oblasti metód strojového učenia pre objavovanie znalostí z textov sa s ohľadom na množstvo dostupných dát v neštruktúrovanej podobe a potrebu ich spracovania zaoberá pomerne široké spektrum odborníkov, ktorí sa na problém dívajú z rôznych uhlov pohľadu alebo aplikačných domén. Ich snaha smeruje k efektívnemu opisu prvkov v kolekcii dokumentov a vzťahov medzi nimi v rozličných podmienkach a situáciách. Metódy objavovania znalostí z textov principiálne vychádzajú z algoritmov hĺbkovej analýzy dát, ktoré sa postupne vyvíjali niekoľko rokov a matematický aparát použitý v metódach je často oveľa starší ako samotná metóda na riešenie konkrétnej úlohy. S ohľadom na multidisciplinaritu a rozsiahly záber tejto problematiky si nerobíme nárok na jej vyčerpávajúce pokrytie. Uvedieme diela opisujúce principiálne fungovanie základných metód, z ktorých sú častokrát odvodené rôzne modifikácie, resp. vylepšenia a spomenieme vybrané nástroje poskytujúce výpočtové prostriedky pre aplikáciu metód text miningu.

Modelovanie tém

Metódy modelovania tém sa postupne zdokonaľovali od LSA (Deerwester et al. 1988), cez PLSA (Hoffman 1999) až k hojne používanej LDA (Blei, Ng a Jordan 2003) tvoriacej ideový základ pre ďalšie metódy, ktoré z nej vychádzajú. Model hLDA (Blei et al. 2004) rozširuje LDA o možnosť odvodenia hierarchie tém. Podobne model PAM (Li a McCallum 2006) umožňuje zachytiť koreláciu medzi témami a modelovať rôzne vzťahy medzi nimi. Pre spomínané metódy je potrebné definovať počet modelovaných tém. Teh et al. (2006) navrhli zmiešaný model využívajúci hierarchický Dirichletovský proces, ktorý predpokladá, že počet tém je vopred neznámy a odvodí sa z pozorovaných dát.

Varianty modelov tém vznikli pre úlohy kontrolovaného strojového učenia, napr. sLDA (Blei a McAuliffe 2008), MedLDA (Zhu, Ahmed a Xing 2009) alebo Labeled LDA (Ramage et al. 2009), resp. Labeled PAM (Bakalov et al. 2012), ktoré na rozdiel napr. od sLDA miesto jedinej uvažujú s niekoľkými kategóriami priradenými dokumentu. S hierarchiou tém so zakomponovaným kontrolovaným učením uvažuje model HLLDA (Petinot et al. 2011) alebo SHLDA (Nguyen, Boyd-Graber a Resnik 2013). Prehľad modelov tém a ich funkčných rozšírení uvádzajú Liu et al. (2016).

Zhluková analýza dokumentov

Pre úlohu nehierarchického zhlukovania predstavuje základný prístup relatívne stará metóda k-means (MacQueen 1967). Arthur a Vassilvitskii (2007) navrhli metódu k-means++, ktorá podľa prezentovaných experimentov oproti k-means prináša vyššiu rýchlosť a presnosť zhlukovania. Výkonnosť a efektívnosť k-means pre text mining sa snažili zlepšiť Mahdavi a Abolhassani (2009). Huangová (2008) porovnala rôzne metriky pre zhlukovanie dokumentov. Hotho, Staab a Stumme (2003) zahrnuli do procesu zhlukovania dokumentov ontológie, ktoré pomáhajú riešiť problém ignorovania vzťahov medzi termami modelu bag-of-words. Ontológie

použili aj Yueová et al. (2015), ktorí skúmali ich použitie pri fuzzy zhlukovaní. Singh, Tiwari a Garg (2011) experimentálne skúmali algoritmy zhlukovania dokumentov a ukázali, že fuzzy zhlukovanie v rôznych prípadoch prináša lepšie výsledky ako k-means. Zlepšovaním metód fuzzy zhlukovania textov sa zaoberali Deng et al. (2010).

Okrem variantov metódy k-means sa pre úlohu zhlukovania skúmali aj iné prístupy. Zhlukovaním založenom na hustote sa zaoberali napr. Ester et al. (1996), ktorý navrhli algoritmus DBSCAN. McLachlan a Basford (1988) ukázali, že zhlukovanie je možné riešiť prostredníctvom pravdepodobnostných zmiešaných modelov, čo vedie k použitiu modelov tém. Zhong a Ghosh (2005) empiricky testovali generatívne prístupy zhlukovania a porovnali ich s diskriminačnými metódami. Zhao a Karypis (2005) experimentálne overili výkonnosť algoritmov pre hierarchické zhlukovanie dokumentov a navrhli algoritmus pre hierarchické zhlukovanie kombinujúci aglomeratívny a deliaci prístup. Hellerová a Ghahramani (2005) uviedli algoritmus aglomeratívneho hierarchického zhlukovania založený na pravdepodobnosti, resp. bayesovskej štatistike a porovnali ho s algoritmami založenými na vzdialenosti. Prehľadný opis prístupov a algoritmov zhlukovania spolu s aplikáciami v rôznych oblastiach uvádzajú Aggarwal a Reddy (2014), resp. Jain (2010).

Kategorizácia dokumentov

Potreba katalogizovať a triediť akékoľvek dáta dala podnet pre vznik viacerých metód kategorizácie založených na rôznych princípoch. Jednoduchý princíp diskriminačnej klasifikačnej metódy kNN (Cover a Hart 1967) dosahuje pri kategorizácii textových dokumentov napriek svojej jednoduchosti pomerne uspokojivé výsledky. Opis problémov kNN a prehľad rôznych zlepšení pre text mining rozoberá Barigou (2016). Na dôležitosť nastavenia parametrov pre výkonnosť kNN upozornil Lim (2004). Erkan et al. (2008) uvádzajú rôzne varianty metódy kNN, porovnávajú ich s ostatnými metódami klasifikácie textov a zaoberajú sa možnosťami využitia prvkov jazykových modelov v kNN. Jiang et al. (2012) sa pokúsili zlepšiť výkonnosť kNN zakomponovaním algoritmu zhlukovania.

Medzi najefektívnejšie diskriminačné metódy klasifikácie patrí metóda SVM (Boser, Guyon a Vapnik 1992). Cortesová a Vapnik (1995) uvažovali s voľným okrajom umožňujúcim chyby klasifikácie, čím rozšírili možnosti použitia SVM na neseparovateľné trénovacie dáta. Joachims (1999) navrhol modifikáciu TSVM a ukázal, že pre účel klasifikácie textov je vhodnejšia. Podrobný opis princípu, vlastností a rozšírení SVM vo všeobecnosti uvádza napr. Steinwart a Christmann (2008). Hamel (2009) preskúmal SVM z perspektívy získavania znalostí z dát.

Pre pravdepodobnostné modely klasifikácie predstavuje základné východisko metóda Naive Bayes (McCallum a Nigam 1998) založená na aplikácii Bayesovej vety. Schneider (2005) a Kim et al. (2006) poukázali na jej nedostatky a navrhli jej úpravy pre klasifikáciu textov. Rôznymi prístupmi a modifikáciami Naive Bayes sa zaoberali Calders a Verwer (2010). Peng, Schuurmans

a Wang (2004) uvažovali o zakomponovaní jazykových modelov. Využitie zmiešaných pravdepodobnostných modelov pre kategorizáciu dokumentov skúmali napr. Novovičová a Malík (2003) alebo McCallum (1999). Na úlohu kategorizácie dokumentov sa dajú využiť tiež modely tém spomínané vyššie.

Nezanedbateľný vplyv a zastúpenie metód pri riešení úloh získavania znalostí z dát a textov majú neurónové siete, ktoré predstavujú dôležitý nástroj umelej inteligencie schopnej riešiť pestrú škálu problémov. Ich použitie pre kategorizáciu textu skúmali napr. Ruiz a Srinivasan (2002), Cho a Lee (2003), ktorí využili kombináciu viacerých neurónových sietí, prípadne Johnson a Zhang (2015) alebo Moraes, Valiati a Gavião Neto (2013), ktorí uviedli empirické porovnanie efektívnosti klasifikácie dokumentov pri nasadení neurónových sietí a použití metódy SVM. Prehľad metód strojového učenia využiteľných pre klasifikáciu textových dokumentov uvádzajú Khan et al. (2010).

Získavanie znalostí z textov

Okrem uvedených zdrojov existuje celý rad ucelených zborníkov a kníh, ktoré sa problematike získavania znalostí z textov venujú metodicky komplexnejšie, uvádzajú typické úlohy, rôzne príklady a venujú priestor viacerým metódam či už kontrolovaného alebo nekontrolovaného strojového učenia, napr. Aggarwal a Zhai (2012), Miner et al. (2012) alebo Zhai a Massung (2016). Manning, Raghavan a Schütze (2008) sa primárne venujú problematike vyhľadávania informácií, ale veľká časť obsahu je aktuálna aj pre text mining. Podobne sú využiteľné mnohé princípy spomínané v publikáciách zameraných na opis metód strojového učenia pre data mining, napr. Han, Kamberová a Pei (2011) alebo Witten et al. (2016).

Aplikačné nástroje pre text mining

Prehľad komerčných aj voľne dostupných aplikačných nástrojov, ktoré implementujú pokročilé metódy pre riešenie úloh strojového učenia vo vzťahu k získavaniu znalostí z textov, uvádza napr. Kaur a Chopra (2016) alebo portál KDnuggets (2017). Okrem nástrojov strojového učenia pôvodne navrhnutých pre data mining, ktoré boli s ohľadom na potreby rozšírené o možnosti spracovania textu, napr. Weka (Witten et al. 2016) alebo RapidMiner (Hofmann a Klinkenberg 2013), existujú prostredia vyvíjané špeciálne pre potreby získavania znalostí z textov, napr. knižnica Gensim (Řehůřek a Sojka 2010) a platforma NLTK (Bird, Klein a Loper 2014) pre programovací jazyk Python, toolkit MeTA (Massung, Geigle a Zhai 2016) vytvorený v jazyku C++ alebo balík aplikácií Mallet (McCallum 2016) vyvíjaný v jazyku Java. Tieto aplikácie obsahujú pokročilé funkcie pre štatistickú analýzu textových dokumentov, rozbor ich štruktúry, spracovanie jazykových modelov a implementujú rôzne algoritmy modelovania tém, klasifikácie dokumentov, zhlukovania a ďalších úloh text miningu.

Veľký potenciál pre nasadenie štatistických metód strojového učenia ponúka vývojové prostredie a jazyk pre štatistické výpočty R (2016). Pomerne jednoduchá rozšíriteľnosť o nové funkcie, veľké množstvo dostupných rozšírení a rozsiahla komunita vývojárov promptne reagujúca na aktuálne trendy ponúkajú možnosti pre implementáciu sofistikovaných skriptov kombinujúcich rôzne metódy s nástrojmi na vizualizáciu výsledkov.

Zoznam bibliografických zdrojov

Aggarwal, C. C. a Reddy, C. K. (eds.). 2014. *Data Clustering: Algorithms and Applications*. Boca Raton: CRC Press. ISBN 978-1-4665-5822-9.

Aggarwal, C. C. a Zhai, C. (eds.). 2012. *Mining Text Data*. New York: Springer. ISBN 978-1-4614-3222-7.

Arthur, D. a Vassilvitskii, S. 2007. K-means++: The Advantages of Careful Seeding. In: *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*. Philadelphia: Society for Industrial and Applied Mathematics. ISBN 978-0-898716-24-5.

Bakalov, A. et al. 2012. Topic Models for Taxonomies. In: *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. New York: Association for Computing Machinery. ISBN 978-1-4503-1154-0.

Barigou, F. 2016. Improving K-nearest neighbor efficiency for text categorization. In: *Neural Network World*. Praha: České vysoké učení technické. 2016, vol. 26, no. 1. ISSN 2336-4335.

Bird, S., Klein, E. a Loper, E. 2014. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit* [online]. NLTK Project. [cit. 2017-02-24]. Dostupné na: http://www.nltk.org/book

Blei, D. a McAuliffe, J. 2008. Supervised topic models. In: *Advances in Neural Information Processing Systems 20*. Red Hook: Curran Associates. ISBN 978-1-60560-352-0.

Blei, D. et al. 2004. Hierarchical topic models and the nested Chinese restaurant process. In: *Advances in Neural Information Processing Systems 16*. Cambridge: MIT Press. ISBN 978-0262201520.

Blei, D., Ng, A. a Jordan, M. 2003. Latent Dirichlet Allocation. In: *Journal of Machine Learning Research*. Cambridge: MIT Press. 2003, vol. 3. ISSN 1532-4435.

Boser, B. E., Guyon, I. M. a Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York: Association for Computing Machinery. ISBN 0-89791-497-X.

Calders, T. a Verwer, T. 2010. Three naive Bayes approaches for discrimination-free classification. In: *Data Mining and Knowledge Discovery*. Hingham: Kluwer Academic Publishers. 2010, vol. 21, no. 2. ISSN 1573-756X.

Cortes, C. a Vapnik, V. 1995. Support-vector networks. In: *Machine Learning*. Hingham: Kluwer Academic Publishers. 1995, vol. 20, no. 3. ISSN 1573-0565.

Cover, T. a Hart, P. 1967. Nearest Neighbor Pattern Classification. In: *IEEE Transactions Information Theory*. Piscataway: IEEE Press. 1967, vol. 13, no. 1. ISSN 1557-9654.

Deerwester, S. et al. 1988. Improving Information Retrieval with Latent Semantic Indexing. In: *Proceedings of the 51st Annual Meeting of the American Society for Information Science*. Medford: Learned Information. ISBN 978-0938734291.

Deng, J. et al. 2010. An Improved Fuzzy Clustering Method for Text Mining. In: *Proceedings of the Second International Conference on Networks Security, Wireless Communications and Trusted Computing*. Washington: IEEE Computer Society. ISBN 978-1-4244-6598-9.

Drucker, J. et al. 2013. *Introduction to Digital Humanities* [online]. Los Angeles: UCLA Center for Digital Humanities. [cit. 2017-01-26]. Dostupné na: http://dh101.humanities.ucla.edu

Erkan, G. et al. 2008. *Improved Nearest Neighbor Methods For Text Classification With Language Modeling and Harmonic Functions* [online]. Ann Arbor: University of Michigan. [cit. 2017-02-24]. Dostupné na: http://clair.si.umich.edu/~radev/papers/tc.pdf

Ester, M. et al. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Palo Alto: AAAI Press. ISBN 1-57735-004-9.

Hamel, L. 2009. *Knowledge Discovery with Support Vector Machines*. New York: Wiley. ISBN 978-0470371923.

Han, J., Kamber, M. a Pei, J. 2011. *Data Mining Concepts and Techniques (Third Edition)*. Waltham: Morgan Kaufmann. ISBN 978-0-12-381479-1.

Heller, K. A. a Ghahramani, Z. 2005. Bayesian Hierarchical Clustering. In: *Proceedings of the 22nd International Conference on Machine Learning*. New York: Association for Computing Machinery. ISBN 1-59593-180-5.

Hofmann, M. a Klinkenberg, R. (eds.). 2013. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Boca Raton: CRC Press. ISBN 978-1-4822-0549-7.

Hofmann, T. 1999. Probabilistic Latent Semantic Analysis. In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. New York: Association for Computing Machinery. ISBN 1-55860-614-9.

Hotho, A., Staab, S. a Stumme, G. 2003. Ontologies Improve Text Document Clustering. In: *Proceedings of the Third IEEE International Conference on Data Mining*. Washington: IEEE Computer Society. ISBN 0-7695-1978-4.

Huang, A. 2008. Similarity measures for text document clustering. In: *Proceedings of the Sixth New Zealand Computer Science Research Student Conference*. Hamilton: University of Waikato.

Cho, S. a Lee, J. 2003. Learning Neural Network Ensemble for Practical Text Classification. In: *Intelligent Data Engineering and Automated Learning*. Berlin: Springer. ISBN 978-3540-45080-1.

Jain, A. K. 2010. Data clustering: 50 years beyond K-means. In: *Pattern Recognition Letters*. New York: Elsevier. 2010, vol. 31, no. 8. ISSN 0167-8655.

Jiang, S. et al. 2012. An Improved k-Nearest Neighbor Algorithm for Text Categorization. In: *Expert Systems with Applications*. Tarrytown: Pergamon Press. 2012, vol. 39, no. 1. ISSN 0957-4174.

Joachims, T. 1999. Transductive Inference for Text Classication using Support Vector Machines. In: *Proceedings of the Sixteenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann Publishers. ISBN 1-55860-612-2.

Johnson, R. a Zhang, T. 2015. Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding. In: *Advances in Neural Information Processing Systems 28*. Cambridge: MIT Press. ISBN 978-1-5108-2502-4.

Kaur, A. a Chopra, D. 2016. Comparison of Text Mining Tools. In: *5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*. New York: Institute of Electrical and Electronics Engineers, 2016. ISBN 978-1-5090-1489-7.

KDnuggets, 2017. *Text Analysis, Text Mining, and Information Retrieval Software* [online]. KDnuggets. [cit. 2017-02-21]. Dostupné na: http://www.kdnuggets.com/software/text.html

Khan et al. 2010. A Review of Machine Learning Algorithms for Text-Documents Classification. In: *Journal of Advances in Information Technology*. Oulu: Academy Publisher. 2010, vol. 1, no. 1. ISSN 1798-2340.

Kim, S. et al. 2006. Some Effective Techniques for Naive Bayes Text Classification. In: *IEEE Transactions on Knowledge and Data Engineering*. New York: Institute of Electrical and Electronics Engineers. 2006, vol. 18, no. 11. ISSN 1041-4347.

Li, W. a McCallum, A. 2006. Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. In: *Proceedings of the 23rd international conference on Machine learning*. New York: ACM Press. ISBN 1-59593-383-2.

Lim, H. S. 2004. Improving kNN Based Text Classification with Well Estimated Parameters. In: *Neural Information Processing: Lecture Notes in Computer Science, vol. 3316*. Berlin: Springer. ISBN 978-3-540-30499-9.

Liu, L. et al. 2016. An overview of topic modeling and its current applications in bioinformatics. In: *SpringerPlus*. Berlin: Springer. 2016, vol. 5, no. 1. ISSN 2193-1801.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press. 1967, vol. 1. ISSN 0097-0433.

Mahdavi, M. a Abolhassani, H. 2009. Harmony K-means algorithm for document clustering. In: *Data Mining and Knowledge Discovery*. Hingham: Kluwer Academic Publishers. 2009, vol. 18, no. 3. ISSN 1573-756X.

Manning, C. D., Raghavan, P. a Schütze, H. 2008. *Introduction to Information Retrieval*. New York: Cambridge University Press. ISBN 978-0521865715.

Massung, S., Geigle, C. a Zhai, C. 2016. MeTA: A Unified Toolkit for Text Retrieval and Analysis. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – System Demonstrations*. Stroudsburg: Association for Computational Linguistics. ISBN 978-1945626036.

McCallum, A. 1999. *Multi-Label Text Classification with a Mixture Model Trained by EM* [online]. Orlando: AAAI-99 Workshop on Text Learning. [cit. 2017-02-18]. Dostupné na: https://people.cs.umass.edu/~mccallum/papers/multilabel-nips99s.ps

McCallum, A. 2016. *MALLET: A Machine Learning for Language Toolkit* [online]. Amherst: University of Massachusetts Amherst. [cit. 2017-02-16]. Dostupné na: http://mallet.cs.umass.edu

McCallum, A. a Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In: *Papers from the AAAI-98 Workshop on Learning for Text Categorization, Technical Report WS-98-05*. Palo Alto: AAAI Press. ISBN 978-1-57735-058-3.

McLachlan, G. J. a Basford, K. E. 1988. *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker. ISBN 0824776917.

Miner, G. et al. 2012. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Waltham: Academic Press. ISBN 978-0123870117.

Moraes, R., Valiati, J. F. a Gavião Neto, W. P. 2013. Document-level Sentiment Classification: An Empirical Comparison between SVM and ANN. In: *Expert Systems with Applications*. Tarrytown: Pergamon Press. 2013, vol. 40, no. 2. ISSN 0957-4174.

Nguyen, V., Boyd-Graber, J. a Resnik, P. 2013. Lexical and Hierarchical Topic Regression. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates. ISBN 978-1-62748-003-1.

Novovičová, J. a Malík, A. 2003. Application of Multinomial Mixture Model to Text Classification. In: *Pattern Recognition and Image Analysis: Lecture Notes in Computer Science, vol. 2652*. Berlin: Springer. ISBN 978-3-540-44871-6.

Page, L. et al. 1998. *The PageRank Citation Ranking: Bringing Order to the Web* [online]. Stanford: Stanford University. [cit. 2017-02-08]. Dostupné na: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Peng, F., Schuurmans, D. a Wang, S. 2004. Augmenting Naive Bayes Classifiers with Statistical Language Models. In: *Information Retrieval*. Hingham: Kluwer Academic Publishers. 2004, vol. 7, no. 3. ISSN 1573-7659.

Petinot, Y., McKeown, K. a Thadani, K. 2011. A Hierarchical Model of Web Summaries. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics. ISBN 978-1-932432-88-6.

R Core Team. 2016. *R: A language and environment for statistical computing* [online]. R Foundation for Statistical Computing. [cit. 2016-12-16]. Dostupné na: https://www.r-project.org

Ramage, D. et al. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics. ISBN 978-1-932432-59-6.

Robertson, S. a Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. In: *Foundations and Trends in Information Retrieval*. Hanover: Now Publishers. 2009, vol. 3, no. 4. ISSN 1554-0677.

Ruiz, M. a Srinivasan, P. 2002. Hierarchical Text Categorization Using Neural Networks. In: *Information Retrieval*. Hingham: Kluwer Academic Publishers. 2002, vol. 5, no. 1. ISSN 1386-4564.

Řehůřek, R. a Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. Msida: University of Malta. ISBN 2-9517408-6-7.

Schneider, K. 2005. Techniques for Improving the Performance of Naive Bayes for Text Classification. In: *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*. Berlin: Springer. ISBN 978-3-540-24523-0.

Singh, V. K., Tiwari, N. a Garg, S. 2011. Document Clustering Using K-Means, Heuristic K-Means and Fuzzy C-Means. In: *Proceedings of the 2011 International Conference on Computational Intelligence and Communication Networks*. Washington: IEEE Computer Society. ISBN 978-0-7695-4587-5.

Steinwart, I. a Christmann, A. 2008. *Support Vector Machines (Information Science and Statistics)*. New York: Springer. ISBN 978-0-387-77241-7.

Teh, Y. W. et al. 2006. Hierarchical Dirichlet Processes. In: *Journal of the American Statistical Association*. Alexandria: American Statistical Association. 2006, vol. 101, no. 476. ISSN 1537-274X.

Witten, I. H. et al. 2016. *Data Mining: Practical Machine Learning Tools and Techniques (Fourth Edition)*. Cambridge: Morgan Kaufmann. ISBN 978-0128042915.

Yue, L. et al. 2015. A Fuzzy Document Clustering Approach Based on Domain-specified Ontology. In: *Data & Knowledge Engineering*. Amsterdam: Elsevier. 2015, vol. 100, part A. ISSN 0169-023X.

Zhai, C. a Massung, S. 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. New York: Association for Computing Machinery. ISBN 978-1-97000-117-4.

Zhao, Y. a Karypis, G. 2005. Hierarchical Clustering Algorithms for Document Datasets. In: *Data Mining and Knowledge Discovery*. Hingham: Kluwer Academic Publishers. 2005, vol. 10, no. 2. ISSN 1573-756X.

Zhong, S. a Ghosh, J. 2005. Generative model-based document clustering: a comparative study. In: *Knowledge and Information Systems*. Berlin: Springer. 2005, vol. 8, no. 3. ISSN 0219-3116.

Zhu, J., Ahmed, A. a Xing, E. P. 2009. MedLDA: Maximum Margin Supervised Topic Models. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. New York: Association for Computing Machinery. ISBN 978-1-60558-516-1.