

Active Learning For Hierarchical Text Classification

Web search engines have been adopted by most universities for searching webpages in their own domains. Basically, a user sends keywords to the search engine and the search engine returns a flat ranked list of webpages. However, in university search, user queries are usually related to topics. Simple keyword queries are often insufficient to express topics as keywords. On the other hand, most E-commerce sites allow users to browse and search products in various hierarchies. It would be ideal if hierarchical browsing and keyword search can be seamlessly combined for university search engines. The main difficulty is to automatically classify and rank a massive number of webpages into the topic hierarchies for universities. In this thesis, we use machine learning and data mining techniques to build a novel hybrid search engine with integrated hierarchies for universities, called SEEU (Search Engine with Hierarchy for Universities). Firstly, we study the problem of effective hierarchical webpage classification. We develop a parallel webpage classification system based on Support Vector Machines. With extensive experiments on the well-known ODP (Open Directory Project) dataset, we empirically demonstrate that our hierarchical classification system is very effective and outperforms the traditional flat classification approaches significantly. Secondly, we study the problem of integrating hierarchical classification into the ranking system of keywords-based search engines. We propose a novel ranking framework, called ERIC (Enhanced Ranking by Hierarchical Classification), for search engines with hierarchies. Experimental results on four large-scale TREC (Text REtrieval Conference) web search datasets show that our ranking system with hierarchical classification outperforms the traditional flat keywords-based search methods significantly. Thirdly, we propose a novel active learning framework to improve the performance of hierarchical classification, which is important for ranking webpages in hierarchies. From our experiments on the benchmark text datasets, we find that our active learning framework can achieve good classification performance yet save a considerable number of labeling effort compared with the state-of-the-art active learning methods for hierarchical text classification. Fourthly, based on the proposed classification and ranking methods, we present a novel hierarchical classification framework for mining academic topics from university webpages. We build an academic topic hierarchy based on the commonly accepted Wikipedia academic disciplines. Based on this hierarchy, we train a hierarchical classifier and apply it to mine academic topics. According to our comprehensive analysis, the academic topics mined by our method are reasonable and consistent with the real-world topic distribution in universities. Finally, we combine all the proposed techniques together and implement the SEEU search engine. According to two usability studies conducted in the ECE and the CS departments at our university, SEEU is favored by the majority of participants. To conclude, the main contribution of this thesis is a novel search engine, called SEEU, for universities. We discuss the challenges toward building SEEU and propose effective machine learning and data mining methods to tackle them. With extensive experiments on well-known benchmark datasets and real-world university webpage data sets, we demonstrate that our system is very effective. In addition, two usability studies of SEEU in our university show that SEEU has a great promise for university search.

This book constitutes the refereed proceedings of the 16th European Conference on Machine Learning, ECML 2005, jointly held with PKDD 2005 in Porto, Portugal, in October 2005. The 40 revised full papers and 32 revised short papers presented together with abstracts of 6 invited talks were carefully reviewed and selected from 335 papers submitted to ECML and 30 papers submitted to both, ECML and PKDD. The papers present a wealth of new results in the area and address all current issues in machine learning.

"This reference brings together an impressive array of research on the development of Science, Technology, Engineering, and Mathematics curricula at all educational levels"--Provided by publisher.

This book constitutes the refereed conference proceedings of the 12th International Conference on Multi-disciplinary Trends in Artificial Intelligence, MIWAI 2018, held in Hanoi, Vietnam, in November 2018. The 16 full papers presented together with 9 short papers were carefully reviewed and selected from 65 submissions. They are organized in the following topical sections: control, planning and scheduling, pattern recognition, knowledge mining, software applications, strategy games and others.

The book Intelligent Systems and Applications - Proceedings of the 2020 Intelligent Systems Conference is a remarkable collection of chapters covering a wider range of topics in areas of intelligent systems and artificial intelligence and their applications to the real world. The Conference attracted a total of 545 submissions from many academic pioneering researchers, scientists, industrial engineers, students from all around the world. These submissions underwent a double-blind peer review process. Of those 545 submissions, 177 submissions have been selected to be included in these proceedings. As intelligent systems continue to replace and sometimes outperform human intelligence in decision-making processes, they have enabled a larger number of problems to be tackled more effectively. This branching out of computational intelligence in several directions and use of intelligent systems in everyday applications have created the need for such an international conference which serves as a venue to report on up-to-the-minute innovations and developments. This book collects both theory and application based chapters on all aspects of artificial intelligence, from classical to intelligent scope. We hope that readers find the volume interesting and valuable; it provides the state of the art intelligent methods and techniques for solving real world problems along with a vision of the future research.

This book constitutes the proceedings of the 18th China National Conference on Computational Linguistics, CCL 2019, held in Kunming, China, in October 2019. The 56 full papers presented in this volume were carefully reviewed and selected from 134 submissions. They were organized in topical sections named: linguistics and cognitive science, fundamental theory and methods of computational linguistics, information retrieval and question answering, text classification and summarization, knowledge graph and information extraction, machine translation and multilingual information processing, minority language processing, language resource and evaluation, social computing and sentiment analysis, NLP applications.

The three volume proceedings LNAI 11051 – 11053 constitutes the refereed proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2018, held in Dublin, Ireland, in September 2018. The total of 131 regular papers presented in part I and part II was carefully

reviewed and selected from 535 submissions; there are 52 papers in the applied data science, nectar and demo track. The contributions were organized in topical sections named as follows: Part I: adversarial learning; anomaly and outlier detection; applications; classification; clustering and unsupervised learning; deep learningensemble methods; and evaluation. Part II: graphs; kernel methods; learning paradigms; matrix and tensor analysis; online and active learning; pattern and sequence mining; probabilistic models and statistical methods; recommender systems; and transfer learning. Part III: ADS data science applications; ADS e-commerce; ADS engineering and design; ADS financial and security; ADS health; ADS sensing and positioning; nectar track; and demo track.

First Published in 2008. Routledge is an imprint of Taylor & Francis, an informa company.

Once considered disruptive to learning, technology has increasingly become an integrated and valued part of the modern classroom. In particular, mobile technologies provide the ability to encourage evocative student learning through new experiences. Promoting Active Learning through the Integration of Mobile and Ubiquitous Technologies showcases the widely varied ways that technology can be applied to enhance classroom learning. Closely examining and critiquing the best methods in assimilating technologies, this publication is a valuable resource for faculty, teachers, administrators, technology staff, directors of learning centers, and other education technology leaders interested in incorporating new technologies within the classroom for engaging student learning.

The two-volume set LNAI 7301 and 7302 constitutes the refereed proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2012, held in Kuala Lumpur, Malaysia, in May 2012. The total of 20 revised full papers and 66 revised short papers were carefully reviewed and selected from 241 submissions. The papers present new ideas, original research results, and practical development experiences from all KDD-related areas. The papers are organized in topical sections on supervised learning: active, ensemble, rare-class and online; unsupervised learning: clustering, probabilistic modeling in the first volume and on pattern mining: networks, graphs, time-series and outlier detection, and data manipulation: pre-processing and dimension reduction in the second volume.

Traditional Pattern Recognition (PR) and Computer Vision (CV) technologies have mainly focused on full automation, even though full automation often proves elusive or unnatural in many applications, where the technology is expected to assist rather than replace the human agents. However, not all the problems can be automatically solved being the human interaction the only way to tackle those applications. Recently, multimodal human interaction has become an important field of increasing interest in the research community. Advanced man-machine interfaces with high cognitive capabilities are a hot research topic that aims at solving challenging problems in image and video applications. Actually, the idea of computer interactive systems was already proposed on the early stages of computer science. Nowadays, the ubiquity of image sensors together with the ever-increasing computing performance has open new and challenging opportunities for research in multimodal human interaction. This book aims to show how existing PR and CV technologies can naturally evolve using this new paradigm. The chapters of this book show different successful case studies of

multimodal interactive technologies for both image and video applications. They cover a wide spectrum of applications, ranging from interactive handwriting transcriptions to human-robot interactions in real environments.

The two-volume set LNAI 7818 + LNAI 7819 constitutes the refereed proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2013, held in Gold Coast, Australia, in April 2013. The total of 98 papers presented in these proceedings was carefully reviewed and selected from 363 submissions. They cover the general fields of data mining and KDD extensively, including pattern mining, classification, graph mining, applications, machine learning, feature selection and dimensionality reduction, multiple information sources mining, social networks, clustering, text mining, text classification, imbalanced data, privacy-preserving data mining, recommendation, multimedia data mining, stream data mining, data preprocessing and representation.

The seven-volume set comprising LNCS volumes 7572-7578 constitutes the refereed proceedings of the 12th European Conference on Computer Vision, ECCV 2012, held in Florence, Italy, in October 2012. The 408 revised papers presented were carefully reviewed and selected from 1437 submissions. The papers are organized in topical sections on geometry, 2D and 3D shapes, 3D reconstruction, visual recognition and classification, visual features and image matching, visual monitoring: action and activities, models, optimisation, learning, visual tracking and image registration, photometry: lighting and colour, and image segmentation.

Here are the refereed proceedings of the First International Conference on Knowledge Science, Engineering and Management, KSEM 2006, held in Guilin, China in August 2006 in conjunction with PRICAI 2006. The book presents 51 revised full papers and 57 revised short papers together with 4 invited talks, reporting a wealth of new ideas and current research results in the broad areas of knowledge science, knowledge engineering, and knowledge management.

There are more than one billion documents on the Web, with the count continually rising at a pace of over one million new documents per day. As information increases, the motivation and interest in data warehousing and mining research and practice remains high in organizational interest. The Encyclopedia of Data Warehousing and Mining, Second Edition, offers thorough exposure to the issues of importance in the rapidly changing field of data warehousing and mining. This essential reference source informs decision makers, problem solvers, and data mining specialists in business, academia, government, and other settings with over 300 entries on theories, methodologies, functionalities, and applications.

The Fourth SIAM International Conference on Data Mining continues the tradition of providing an open forum for the presentation and discussion of innovative algorithms as well as novel applications of data mining. This is reflected in the talks by the four keynote speakers who discuss data usability issues in systems for data mining in science and engineering, issues raised by new technologies that generate biological data, ways to find complex structured patterns in linked data, and advances in Bayesian inference techniques. This proceedings includes 61 research papers.

This three-volume set LNAI 6911, LNAI 6912, and LNAI 6913 constitutes the refereed proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2011, held in Athens, Greece, in September

2011. The 121 revised full papers presented together with 10 invited talks and 11 demos in the three volumes, were carefully reviewed and selected from about 600 paper submissions. The papers address all areas related to machine learning and knowledge discovery in databases as well as other innovative application domains such as supervised and unsupervised learning with some innovative contributions in fundamental issues; dimensionality reduction, distance and similarity learning, model learning and matrix/tensor analysis; graph mining, graphical models, hidden markov models, kernel methods, active and ensemble learning, semi-supervised and transductive learning, mining sparse representations, model learning, inductive logic programming, and statistical learning. a significant part of the papers covers novel and timely applications of data mining and machine learning in industrial domains.

The 30-volume set, comprising the LNCS books 12346 until 12375, constitutes the refereed proceedings of the 16th European Conference on Computer Vision, ECCV 2020, which was planned to be held in Glasgow, UK, during August 23-28, 2020. The conference was held virtually due to the COVID-19 pandemic. The 1360 revised papers presented in these proceedings were carefully reviewed and selected from a total of 5025 submissions. The papers deal with topics such as computer vision; machine learning; deep neural networks; reinforcement learning; object recognition; image classification; image processing; object detection; semantic segmentation; human pose estimation; 3d reconstruction; stereo vision; computational photography; neural networks; image coding; image reconstruction; object recognition; motion estimation.

This book constitutes the refereed proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, held in Paris, France, in June 2018. The 18 full papers, 26 short papers, and 9 poster papers presented were carefully reviewed and selected from 99 submissions. The papers are organized in the following topical sections: Opinion Mining and Sentiment Analysis in Social Media; Semantics-Based Models and Applications; Neural Networks Based Approaches; Ontology Engineering; NLP; Text Similarities and Plagiarism Detection; Text Classification; Information Mining; Recommendation Systems; Translation and Foreign Language Querying; Software Requirement and Checking.

The eight-volume set comprising LNCS volumes 9905-9912 constitutes the refereed proceedings of the 14th European Conference on Computer Vision, ECCV 2016, held in Amsterdam, The Netherlands, in October 2016. The 415 revised papers presented were carefully reviewed and selected from 1480 submissions. The papers cover all aspects of computer vision and pattern recognition such as 3D computer vision; computational photography, sensing and display; face and gesture; low-level vision and image processing; motion and tracking; optimization methods; physics-based vision, photometry and shape-from-X; recognition: detection, categorization, indexing, matching; segmentation, grouping and shape representation; statistical methods and learning; video: events, activities and surveillance; applications. They are organized in topical sections on detection, recognition and retrieval; scene understanding; optimization; image and video processing; learning; action, activity and tracking; 3D; and 9 poster sessions.

This volume constitutes the refereed proceedings of the 14th International Conference on Hybrid Artificial Intelligent Systems, HAIS 2019, held in León, Spain, in September 2019. The 64 full papers published in this volume were carefully reviewed and selected

from 134 submissions. They are organized in the following topical sections: data mining, knowledge discovery and big data; bio-inspired models and evolutionary computation; learning algorithms; visual analysis and advanced data processing techniques; data mining applications; and hybrid intelligent applications.

This edition of this handbook updates and expands its review of the research, theory, issues and methodology that constitute the field of educational communications and technology. Organized into seven sectors, it profiles and integrates the following elements of this rapidly changing field.

This two-volume set (CCIS 1367-1368) constitutes reviewed and selected papers from the 10th International Advanced Computing Conference, IACC 2020, held in December 2020. The 65 full papers and 2 short papers presented in two volumes were thoroughly reviewed and selected from 286 submissions. The papers are organized in the following topical sections: Application of Artificial Intelligence and Machine Learning in Healthcare; Using Natural Language Processing for Solving Text and Language related Applications; Using Different Neural Network Architectures for Interesting applications; ?Using AI for Plant and Animal related Applications.- Applications of Blockchain and IoT.- Use of Data Science for Building Intelligence Applications; Innovations in Advanced Network Systems; Advanced Algorithms for Miscellaneous Domains; New Approaches in Software Engineering.

This book constitutes the refereed proceedings of the 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, held in Montréal, QC, Canada, in May 2014. The 22 regular papers and 18 short papers presented together with 3 invited talks were carefully reviewed and selected from 94 submissions. The papers cover a variety of topics within AI, such as: agent systems; AI applications; automated reasoning; bioinformatics and BioNLP; case-based reasoning; cognitive models; constraint satisfaction; data mining; E-commerce; evolutionary computation; games; information retrieval; knowledge representation; machine learning; multi-media processing; natural language processing; neural nets; planning; privacy-preserving data mining; robotics; search; smart graphics; uncertainty; user modeling; web applications.

Comprehensive Coverage of the Entire Area of Classification Research on the problem of classification tends to be fragmented across such areas as pattern recognition, database, data mining, and machine learning. Addressing the work of these different communities in a unified way, *Data Classification: Algorithms and Applications* explores the underlying algorithms of classification as well as applications of classification in a variety of problem domains, including text, multimedia, social network, and biological data. This comprehensive book focuses on three primary aspects of data classification: Methods-The book first describes common techniques used for classification, including probabilistic methods, decision trees, rule-based methods, instance-based methods, support vector machine methods, and neural networks. Domains-The book then examines specific methods used for data domains such as multimedia, text, time-series, network, discrete sequence, and uncertain data. It also covers large data sets and data streams due to the recent importance of the big data paradigm. Variations-The book concludes with insight on variations of the classification process. It discusses ensembles, rare-class learning, distance function learning,

active learning, visual learning, transfer learning, and semi-supervised learning as well as evaluation aspects of classifiers.

This book constitutes the proceedings of the 23rd International Conference on Discovery Science, DS 2020, which took place during October 19-21, 2020. The conference was planned to take place in Thessaloniki, Greece, but had to change to an online format due to the COVID-19 pandemic. The 26 full and 19 short papers presented in this volume were carefully reviewed and selected from 76 submissions. The contributions were organized in topical sections named: classification; clustering; data and knowledge representation; data streams; distributed processing; ensembles; explainable and interpretable machine learning; graph and network mining; multi-target models; neural networks and deep learning; and spatial, temporal and spatiotemporal data.

The key idea behind active learning is that a machine learning algorithm can perform better with less training if it is allowed to choose the data from which it learns. An active learner may pose "queries," usually in the form of unlabeled data instances to be labeled by an "oracle" (e.g., a human annotator) that already understands the nature of the problem. This sort of approach is well-motivated in many modern machine learning and data mining applications, where unlabeled data may be abundant or easy to come by, but training labels are difficult, time-consuming, or expensive to obtain. This book is a general introduction to active learning. It outlines several scenarios in which queries might be formulated, and details many query selection algorithms which have been organized into four broad categories, or "query selection frameworks." We also touch on some of the theoretical foundations of active learning, and conclude with an overview of the strengths and weaknesses of these approaches in practice, including a summary of ongoing work to address these open challenges and opportunities. Table of Contents: Automating Inquiry / Uncertainty Sampling / Searching Through the Hypothesis Space / Minimizing Expected Error and Variance / Exploiting Structure in Data / Theory / Practical Considerations

Text classification is becoming a crucial task to analysts in different areas. In the last few decades, the production of textual documents in digital form has increased exponentially. Their applications range from web pages to scientific documents, including emails, news and books. Despite the widespread use of digital texts, handling them is inherently difficult - the large amount of data necessary to represent them and the subjectivity of classification complicate matters. This book gives a concise view on how to use kernel approaches for inductive inference in large scale text classification; it presents a series of new techniques to enhance, scale and distribute text classification tasks. It is not intended to be a comprehensive survey of the state-of-the-art of the whole field of text classification. Its purpose is less ambitious and more practical: to explain and illustrate some of the important methods used in this field, in particular kernel approaches and techniques.

The World Wide Web has become an extremely popular way of publishing and

distributing electronic resources. Though the Web is rich with information, collecting and making sense of this data is difficult because it is rather unorganized. Building an Intelligent Web introduces students and professionals to the state-of-the-art development of Web Intelligence techniques and teaches how to apply these techniques to develop the next generation of intelligent Web sites. Each chapter contains theoretical bases, which are also illustrated with the help of simple numeric examples, followed by practical implementation. Students will find Building an Intelligent Web to be an active and exciting introduction to advanced Web mining topics. Topics covered include Web Intelligence, Information Retrieval, Semantic Web, Classification and Association Rules, SQL, Database Theory, Applications to e-commerce and Bioinformatics, Clustering, Modeling Web Topology, and much more!

The book presents the proceedings of four conferences: The 26th International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'20), The 18th International Conference on Scientific Computing (CSC'20); The 17th International Conference on Modeling, Simulation and Visualization Methods (MSV'20); and The 16th International Conference on Grid, Cloud, and Cluster Computing (GCC'20). The conferences took place in Las Vegas, NV, USA, July 27-30, 2020. The conferences are part of the larger 2020 World Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE'20), which features 20 major tracks. Authors include academics, researchers, professionals, and students. Presents the proceedings of four conferences as part of the 2020 World Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE'20); Includes the research tracks Parallel and Distributed Processing, Scientific Computing, Modeling, Simulation and Visualization, and Grid, Cloud, and Cluster Computing; Features papers from PDPTA'20, CSC'20, MSV'20, and GCC'20.

The 3-volume set LNAI 12712-12714 constitutes the proceedings of the 25th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2021, which was held during May 11-14, 2021. The 157 papers included in the proceedings were carefully reviewed and selected from a total of 628 submissions. They were organized in topical sections as follows: Part I: Applications of knowledge discovery and data mining of specialized data; Part II: Classical data mining; data mining theory and principles; recommender systems; and text analytics; Part III: Representation learning and embedding, and learning from data.

Over the last two decades, researchers are looking at imbalanced data learning as a prominent research area. Many critical real-world application areas like finance, health, network, news, online advertisement, social network media, and weather have imbalanced data, which emphasizes the research necessity for real-time implications of precise fraud/defaulter detection, rare disease/reaction prediction, network intrusion detection, fake news detection, fraud advertisement detection, cyber bullying identification, disaster events prediction, and more.

Machine learning algorithms are based on the heuristic of equally-distributed balanced data and provide the biased result towards the majority data class, which is not acceptable considering imbalanced data is omnipresent in real-life scenarios and is forcing us to learn from imbalanced data for foolproof application design. Imbalanced data is multifaceted and demands a new perception using the novelty at sampling approach of data preprocessing, an active learning approach, and a cost perceptive approach to resolve data imbalance. *Data Preprocessing, Active Learning, and Cost Perceptive Approaches for Resolving Data Imbalance* offers new aspects for imbalanced data learning by providing the advancements of the traditional methods, with respect to big data, through case studies and research from experts in academia, engineering, and industry. The chapters provide theoretical frameworks and the latest empirical research findings that help to improve the understanding of the impact of imbalanced data and its resolving techniques based on data preprocessing, active learning, and cost perceptive approaches. This book is ideal for data scientists, data analysts, engineers, practitioners, researchers, academicians, and students looking for more information on imbalanced data characteristics and solutions using varied approaches.

Asia Information Retrieval Symposium (AIRS) 2008 was the fourth AIRS conference in the series established in 2004. The first AIRS was held in Beijing, China, the second in Jeju, Korea, and the third in Singapore. The AIRS conferences trace their roots to the successful Information Retrieval with Asian Languages (IRAL) workshops, which started in 1996. The AIRS series aims to bring together international researchers and developers to exchange new ideas and the latest results in information retrieval. The scope of the conference encompasses the theory and practice of all aspects of information retrieval in text, audio, image, video, and multimedia data. We are pleased to report that AIRS 2006 received a large number of 144 submissions. Submissions came from all continents: Asia, Europe, North America, South America and Africa. We accepted 39 submissions as regular papers (27%) and 45 as short papers (31%). All submissions underwent double-blind reviewing. We are grateful to all the area Co-chairs who managed the review process of their respective area efficiently, as well as to all the Program Committee members and additional reviewers for their efforts to get reviews in on time despite the tight time schedule. We are pleased that the proceedings are published by Springer as part of their Lecture Notes in Computer Science (LNCS) series and that the papers are EI-indexed.

Organizes major concepts, theories, methodologies, trends, challenges and applications of data mining (DM) and knowledge discovery in databases (KDD). This book provides algorithmic descriptions of classic methods, and also suitable for professionals in fields such as computing applications, information systems management, and more.

Mining social networks has now becoming a very popular research area not only for data mining and web mining but also social network analysis. Data mining is a technique that has the ability to process and analyze large amount of data and by this to discover valuable information from the data. In recent year, due to the growth of social communications and social networking websites, data mining becomes a very

