

TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika

Mia Marta Heikla

**Ettekannete automaatne sessioonideks jagamine
teaduskonverentside jaoks**

Bakalaureusetöö (9 EAP)

Juhendaja(d): Meelis Kull

Tartu 2024

Ettekannete automaatne sessioonideks jagamine teaduskonverentside jaoks

Lühikokkuvõte:

Selle töö eesmärk on luua keelemudeli abil meetod, mis suudaks jagada teaduskonverentsile vastuvõetud artiklid teemade kaupa sessioonideks, et osalejad saaksid kuulata järjest sarnase teemaga ettekandeid. Meetodi väljatöötamisel on vajalik eelnevalt teada konverentsi ajagraafiku probleemi, hea sessiooni pealkirja loomise printsiipe ja juturoboti promptide loomise põhimõtteid. Meetodile antakse ette teadusartiklite pealkirjad ja abstraktid. Nende andmete põhjal esitatakse keelemudelile päringud võimalike sessioonide pealkirjade genereerimiseks, millele järgneb järelpäringute kasutamine pealkirjade üldistamiseks. Ühekordselt esinevad pealkirjad vaadatakse üle, et tagada nende sobivus masinõppe konverentsi kontekstis, eemaldades madala väärtusega pealkirjad. Allesjäänud pealkirju hinnatakse keelemudeliga vastavalt iga ettekande sisule ja pealkirjale, millele järgneb sessioonideks jagamine lineaarse täisarvulise optimeerimise algoritmi abil. Protsessi lõpetab Levenshteini kauguse kasutamine sessioonide jaotamise sarnasuse hindamiseks.

Võtmesõnad:

Konverentsi ajagraafiku probleem, keelemudelid, prompt, sessioonide jaotamine, teaduskonverents

CERCS: P176 Tehisintellekt

Automatic division of presentations into sessions for scientific conferences

Abstract:

The aim of this work is to create a method, using a language model, that can divide the articles accepted for a scientific conference into sessions by topic, so that participants can listen to a succession of presentations on a similar topic. In developing the method, it is necessary to know in advance the problem of scheduling a conference, the principles of creating a good session title, and the principles of creating story prompts. The titles and abstracts of scientific articles will be provided for the method. Based on this data, the language model is queried to generate possible session titles, followed by the use of prompts to generalise the titles. Occurring titles are reviewed to ensure their suitability in the context

of a machine learning conference by removing low-value titles. The remaining titles are evaluated with a language model according to the content and title of each presentation, followed by a segmentation into sessions using a linear integer optimization algorithm. The process is completed by using Levenshtein distance to estimate the similarity of the segmentation of the sessions.

Keywords:

Conference Timetable Problem, Language Models, Prompt, Division of Sessions, Scientific Conference

CERCS: P176 Artificial intelligence

Sisukord

| | |
|---|----|
| Sissejuhatus | 5 |
| 1. Hea konverentsi ajagraafiku printsiibid | 6 |
| 1.1 Pealkirjade olulisus | 6 |
| 1.2 Pealkirjade kujundamise põhimõtted | 6 |
| 1.3 Hea prompti printsiibid | 9 |
| 2. Varasemad konverentsi ajagraafiku probleemi lahendused | 10 |
| 2.1 Kõnelejate populaarsus atraktiivsus koefitsientidest lähtudes loodud ajagraafik | 10 |
| 2.2 Sisupõhise lähenemisega konverentsi ajagraafik | 10 |
| 2.3 Personaalne lähenemine konverentsi ajakavale | 11 |
| 2.4 Esineja põhise perspektiivi kaudu loodud ajaplaan | 11 |
| 2.5 Analüüs | 12 |
| 3. Koostatud meetodi kirjeldus | 14 |
| 3.1 Meetodi lühikirjeldus | 14 |
| 3.2 Meetodis kasutatavate teekide kirjeldus | 14 |
| 3.3 Sessioonide optimeerimise ülesande esitamine täisarvulise lineaarse planeerimise ülesandena | 16 |
| 3.4 Meetodi detailne tehniline kirjeldus | 18 |
| 3.5 Meetodide parameetrite võrdlus | 27 |
| 3.6 Meetodide täpsuse analüüs | 28 |
| 3.7 Meetodi potentsiaal tulevikuks | 33 |
| 3.8 Meetodi kood | 33 |
| 4. Loodud mudeli kasulikkus ECML PKDD teaduskonverentsile | 34 |
| Kokkuvõte | 35 |
| Viidatud kirjandus | 36 |

Sissejuhatus

Uurimustöö keskmes seisab loodava automaatse sessioonideks jagamise meetodi arendamist ja rakendamist ECML PKDD (Euroopa masinõppe konverents ja teadmiste avastamise põhimõtted andmebaasides) konverentsi kontekstis. Konverents ise on iga-aastane sündmus masinõppe ja andmebaaside valdkonnas. Lähtepunktiks on 2022. aasta ECML PKDD konverentsi andmed, mis pakuvad alusmaterjali meetodi testimiseks ja täiustamiseks. Konverentsi ajakava on struktureeritud mitmepäevase üritusena, kus igal päeval on mitu paralleelsessiooni, mis toimuvad erinevates ruumides. Päeva jooksul on mitu sessiooni, mis kestavad tavaliselt 1-2 tundi. Iga sessioon koosneb mitmest ettekandest, mis on grupeeritud teemade või teadusvaldkondade järgi. Oluline on, et igal artiklil on kindel kuuluvus ühte sessiooni ja kõikides sessioonides oleks võrdne arv artikleid. Traditsiooniliselt on ajakava loomine olnud ekspertide ülesanne, kes hindavad ja kategoriseerivad sisu käsitsi.

Käesoleva uurimustöö peamine eesmärk on välja töötada automaatne meetod, kasutades juturoboti tehnoloogiat, et optimeerida andmeanalüüsi ja sessioonide organiseerimist suuremahuliste konverentside jaoks. Meetod hõlmab sessiooni pealkirjade genereerimist sisu põhjal ja artiklite jagamist sessioonideks vastavalt nende sisule, eesmärgiga muuta konverentsi struktuuri efektiivsemaks ja tõhusamaks.

Enne kui asume mudeli loomise ja rakendamise maailma on hädavajalik mõista konverentsi ajagraafiku probleemi, hea sessiooni pealkirja loomise printsiipe ja juturoboti promptide loomise põhimõtteid. Konverentsi ajagraafiku probleemi mõistmine võimaldab arendada meetodit, mis on sarnane artiklite sessioonideks jaotamise protsessile. Sessioonideks jaotamine on nimetatud probleemi üks alamülesandeid. Esialgu on vaja mõista, et millise suurusega konverentsiga on tegemist – olgu selleks osalevate artiklite koguarv või sessioonide arv, mis tuleb hinnata ja jaotada. Lisaks on oluline mõista konverentsi teemat, kuna see määrab andmete struktuuri ja nende tõlgendamise keerukuse. Seega on töö eesmärgiks mitte üksnes luua mudel sessioonide loomiseks, vaid ka lahendada konkreetset selle konverentsi analüüsiga seotud väljakutseid, olles samal ajal suuteline mõistma erinevate valdkondade keerukust ja eripärasid. Terve lõputöö kirjaliku osa koostamise vältel kasutasin sõnastuse parendamiseks toetava tööriistana ChatGPT 4 [1].

1. Hea konverentsi ajagraafiku printsiibid

Konverentsi korraldamisel on oluline tagada ajagraafiku selgus, jälgitavus ja vastavus osalejate ootustele. Kvaliteetne ajagraafik aitab optimeerida osalejate kogemust ja maksimeerida õppimisvõimalusi. Hea ajagraafiku kujundamisel on oluline sessiooni pealkirjade valik. Need pealkirjad ei tohi olla liiga üldised, kuid samas peavad kokkuvõtlikult edastama sessiooni sisu. Konverentsidel on sageli paralleelsessioonid, mis tähendab, et erinevad sessioonid toimuvad samal ajal mitmes kohas. Kuna osaleja saab korraga viibida ainult ühes sessioonis, on oluline, et iga paralleelsessioon sisaldaks mitu ettekannet, mis osalejale huvi pakuvad. Seetõttu on temaatiliselt sarnaste ettekannete ühte sessiooni koondamine mõistlik lähenemine. Järgnevalt vaatleme lähemalt, kuidas luua tõhusaid pealkirju, mis aitavad saavutada eduka konverentsi ajaplaani.

1.1 Pealkirjade olulisus

Sessiooni pealkirjad omavad konverentsi ajagraafikus võtmerolli, sest need on esimesed elemendid, mis köidavad potentsiaalsete osalejate tähelepanu.

Hea sessiooni pealkiri:

- **Kajastab korrektselt artikli sisu:** Sessiooni pealkiri peab lühidalt, kuid täpselt kirjeldama abstrakt sisu ja tagama eesmärgi, et osalejad saavad kiiresti aru sessiooni olulisusest ja kontekstist;
- **Annab selge ülevaate:** Pealkiri peab edastama, mida osalejad sessioonilt oodata võivad;
- **Erineb teistest:** Hea pealkiri aitab sessioonil silma paista teiste hulgast, mis on eriti oluline suurtel konverentsidel.

1.2 Pealkirjade kujundamise põhimõtted

Pealkirjad peaksid olema võimalikult konkreetsed ja asjakohased. Vältima peaks liialt üldiseid väljendeid, mis ei anna edasi sessiooni unikaalsust ega fookust. Näiteks pealkiri "Masinõpe" võib tunduda liiga lai ja üldine. Selle asemel võiks kasutada "Masinõppe kasutus meditsiinis", mis on spetsiifilisem ja informatiivsem. Selgemate võtmesõnade kasutamine aitab potentsiaalsetel osalejatel mõista, millised teemasid või tehnoloogiaid sessioonis

käsitletakse. Pealkirja valikul tuleks mõelda sellele, kuidas sessiooni pealkiri kajastab korrektselt abstrakti sisu. Eduka konverentsi ajagraafiku koostamisel on sessioonide pealkirjade valikul tähtis roll. Hea pealkiri on konkreetne, lühike, sisaldab võtmesõnu ja äratavad osalejate huvi.

| Sessiooni pealkiri: | Artikli pealkiri: | Autor(id): |
|----------------------------------|---|--|
| Active, Semi-Supervised Learning | A Stopping Criterion for Transductive Active Learning | Daniel Kottke (University of Kassel)*; Christoph Sandrock (University of Kassel); Georg Kreml (Utrecht University); Bernhard Sick (University of Kassel) |
| | Deep Active Learning for Detection of Mercury's Bow Shock and Magnetopause Crossings | Sahib Julka (University of Passau)*; Nikolas Kirschstein (University of Passau); Michael Granitzer (University of Passau); Alexander Lavrukhin (M.V.Lomonosov Moscow State University); Ute V Amerstorfer (Space Research Institute, Austrian Academy of Sciences) |
| | Online Active Classification via Margin-based and Feature-based Label Queries | Tingting Zhai |
| | Stream-Based Active Learning for Sliding Windows Under Verification Latency | Tuan Minh Pham |
| | Semi-supervised Latent Block Model with pairwise constraints | Paul Riverain |
| Computer Vision | Charge Own Job: Saliency Map and Visual Word Encoder for Image-Level Semantic Segmentation | Yuhui Guo (Renmin University of China)*; Xun Liang (Renmin University of China); hui tang (Renmin University of China); Xiangping Zheng (Renmin University of China); Bo Wu (Renmin University of China); Xuan Zhang (Renmin University of China) |
| | A Novel Data Augmentation Technique for Out-of-Distribution Sample Detection using Compounded Corruptions | Ramya S Hebbalaguppe (Indian Institute of Technology, Delhi)*; Soumya Suvra Ghosal (University of Wisconsin-Madison); Jatin Prakash (Indian Institute of Technology, Delhi); Harshad Khadilkar (IIT Mumbai); Chetan Arora (Indian Institute of Technology Delhi) |

| | | |
|--|--|---|
| | Understanding Adversarial Robustness of Vision Transformers via Cauchy Problem | Zheng Wang (Exeter University)*; Wenjie Ruan (University of Exeter) |
| | Supervised Contrastive Learning for Few-Shot Action Classification | Hongfeng Han (Renmin University of China); Nanyi Fei (Renmin University of China); Zhiwu Lu (Renmin University of China)*; Ji-Rong Wen (Renmin University of China) |
| | Wavelet-Packets for Deepfake Image Analysis and Detection | Moritz Wolter |
| Meta-Learning Neural Architecture Search | MRF-UNets: Searching UNet with Markov Random Fields | Zifu Wang (KU Leuven)*; Matthew B. Blaschko (KU Leuven) |
| | Efficient Automated Deep Learning for Time Series Forecasting | Difan Deng (Leibniz Universität Hannover)*; Florian M Karl (Fraunhofer-Institut für Integrierte Schaltungen IIS); Frank Hutter (University of Freiburg); Bernd Bischl (LMU Munich); Marius Lindauer (Leibniz University Hannover) |
| | Adversarial Projections to Tackle Support-Query Shifts in Few-Shot Meta-Learning | Aroof Aimen (Indian Institute of Technology, Ropar)*; Bharat Ladrecha (Indian Institute of Technology Ropar); Narayanan C Krishnan (IIT Palakkad) |
| | Context Abstraction to Improve Decentralized Machine Learning in Structured Sensing Environments | Massinissa Hamidi (Laboratoire LIPN-UMR CNRS 7030, PRES Sorbonne Paris Cité)*; Aomar Osmani (Laboratoire LIPN-UMR CNRS 7030, PRES Sorbonne Paris Cité) |
| | Discovering wiring patterns influencing neural network performance | Aleksandra I Nowak (Jagiellonian University)*; Romuald Janik (Jagiellonian University) |
| | Automatic Feature Engineering through Monte Carlo Tree Search | Yiran Huang (Karlsruhe Institute of Technology)*; Yexu Zhou (KIT); Michael Hefenbrock (TECO); Till Riedel (KIT); Likun Fang (KIT); Michael Beigl |

Tabel 1 Näide 2022. aasta ECML PKDD konverentsi sessioonidest [2]

Pealkirjade hoolikas planeerimine aitab tagada konverentsi sisu selguse ning meelitab ligi õige sihtrühma.

1.3 Hea prompti printsiibid

Prompt (prompt - “keelemudelilt tavakeeles küsitud küsimus või päring, (kirjalikult) edastatud teade (nt vestlusrobotiga suheldes) [3].”) on oluline vahend suhtluses keelemudeliga, sest see määrab, millist ja kui kvaliteetset sisu mudel väljastab. Hea prompt peab olema selge, täpne ja piisavalt üksikasjalik, et suunata mudelit genereerima soovitud tüüpi vastuseid, mis on oluline nii teabe täpsuse kui ka relevantsuse tagamiseks. Konverentsi ajakava planeerimisel aitab efektiivne prompt keelemudelil mõista ülesande nõudmisi ja genereerida asjakohaseid sessiooni pealkirju, mis vastavad konkreetsetel masinõppe konverentsi teemadele. Selleks, et juturobot vastaks parima väljundiga, pidasin prompti loomisel silmas firma OpenAI [4] poolt loodud heatavstid:

- Asjakohasemate vastuste saamiseks täpsustada oma küsimust;
- Anda keelemudelile ülesanne võtta mingi tegelase roll töö ülesande lahendamiseks;
- Sisu erinevate osade vahel kasutada eristamiseks selgeid eraldajaid;
- Loetleda ülesande lõpetamiseks vajalikud sammud või formaat;
- Tuua näiteid;
- Määretleda kui kaua mudel väljundi esitadamiseks aeg võib kulutada.

Lisaks OpenAI poolt määratletud põhimõtetele promptide loomiseks kasutan ka järgnevaid tehnikaid:

- Enne mudeli rakendamist on vajalik täpselt kirjeldada probleemi, mida juturobot lahendada hakkab. Tuleb selgelt välja tuua konverentsi suurus, hinnatavate artiklite koguarv, sessioonide jaotuse vajadus ning konverentsi temaatika;
- Tähtis on kasutada samasuguseid sümboleid. Näiteks, kui kasutada jutumärke, tuleb vaadata, et ei oleks segamini erinevad jutumärgi sümboolid ‘ “ ja “ ’;
- Sümbolite arvu vähendamiseks on soovitatav pealkirjad eraldada reavahetusega (\n) selle asemel, et kasutada komasid ja jutumärke. Selline lähenemine lisab teksti ainult kaks sümbolit, mitte kolm (“,”).

Testimise käigus selgus, et ChatGPT 3.5 mudel suudab korraga hinnata maksimaalselt 150 - 170 sessiooni pealkirja ja ChatGPT 4 korraga kuni 250 pealkirja (kokku umbes 1000 tokenit). Kuigi keelemudelid on võimelised töötleva rohkem tookeneid, on eeltoodud arv pealkirju ChatGPT optimaalne töömaht üheks sisendiks.

2. Varasemad konverentsi ajagraafiku probleemi lahendused

Järgnevad alapeatükid annavad ülevaate olemasolevatest konverentsi ajakava optimeerimis võimalustest ja lahendustest. Viimane alapeatükk koosneb ajagraafiku probleemi lahenduste analüüsist ja sobivusest artiklite sessioonideks jaotamiseks.

2.1 Kõnelejate populaarsus atraktiivsus koefitsientidest lähtudes loodud ajagraafik

Teschi bakalaureusetöö [5] tutvustab põhjalikult, kuidas loodi meetod suure konverentsi ajakava loomiseks. Tema järgi on ajagraafiku loomisel esmalt oluline jagada konverentsi ettekannete teemad klastriteks. Konkreetse konverentsi puhul jagab iga registreerunud esineja ise valitud teema vastavasse klassifikatsiooni. Sessiooni korraldamisel tuleb samasse klastrisse kuuluvad ettekanded paigutada samasse ruumi, soodustades paremat orienteerumist ja vältimaks osalisteta istungeid, tagades osalejate püsiva motivatsiooni olla ruumis terve sessiooni vältel. Ruumi mahutavuse piirangud nõuavad strateegilist planeerimist, eelistades atraktiivsemaid istungeid suuremates ruumides, et vältida ülerahvastatust. Ühtlase atraktiivsuse jaotumise tagamine kogu programmi vältel võimaldab igal ajahetkel pakkuda võrdselt huvipakkuvaid istungeid. Meetodi tulemuslikkus sõltub atraktiivsuskoeffitsientidest, mis arvutatakse tõmbeindeksite abil. Tesch [5] rõhutab ka kõneleja populaarsuse mõju, soovitades koguda spetsiifilist huvipakkuvat teavet igaks sessiooniks tulevaste konverentside planeerimiseks.

2.2 Sisupõhise lähenemisega konverentsi ajagraafik

Castaño jt [6] loodud teadustöö käsitleb konverentsi ajakava sisupõhise optimeerimise keerukust, võttes arvesse sisu sarnasust ja kattuvust ning nendest tulenevaid piiranguid. Graafiku planeerimine hõlmab programmi poolt ajakava hoolikat koostamist, mis võtab arvesse nii klastreid kui ka ajalist jaotust. Nende uuringus pakutakse välja kaheetapiline lahendusmeetod, kus kõigepealt liigitatakse ettekanded metoodiliselt sisu alusel gruppideks. Seejärel koostatakse ajakava, eelistades sisu sarnasust, et vältida sarnase materjaliga artiklite samaaegset planeerimist. Castaño jt [6] kasutavad eelkõige algoritmilist meetodit, milleks on GASP-algoritm (metaheuristiline algoritm, mida tavaliselt rakendatakse kombinatoorsete optimeerimis probleemide puhul). Algoritmi kasutatakse selleks, et kiirendada optimaalsete

lahenduste leidmist piiratud arvutus aja jooksul. (Heuristiline algoritm - Matemaatilises programmeerimises kasutatakse heuristilisi tehnikaid või lihtsalt heuristikat selleks, et leida optimeerimise probleemidele head või peaaegu optimaalsed lahendused. Heuristika, mis algselt arendati välja diskreetse meetodina on mõeldud konkreetsete optimeerimis probleemide tõhusaks lahendamiseks [7].)

2.3 Personaalne lähenemine konverentsi ajakavale

Konverentsi planeerimise probleemi üldine eesmärk on maksimeerida kogu osavõtjate arvu, võttes arvesse esinejate eelistusi. Vangerven jt [8] artiklis süvenetakse konverentsi ajakava praktilisse teostamisse, võttes kasutusele kolmefaasilise strateegia:

- esmalt ettekanded jaotatakse ruumide ja ajavahemike kaupa ära;
- teiseks ettekanded rühmitatakse sessioonideks;
- viimaks määratakse iga osaleja jaoks ideaalne ajakava vastavalt tema eelistusele.

Seda on rakendatud konverentside MathSport 2013, MAPSP 2015 ja ORBEL 2017 puhul. Kuigi meetodil võib esineda raskusi sidusate sessioonide loomisel ainult osalejate eelistuste alusel, leiti uuringus, et loodud sessioonide puhul ilmnes suhteline sidusus, mis tulenes teemade lähedusest osalejate profiilides. Tähelepanuväärne aspekt on olemasoleva meetodi juures osalejate ebavõrdne kohtlemine, mille puhul paljude ettekannete valik võib avaldada ebaproportsionaalset mõju kogu konverentsi päevakavale.

2.4 Esineja põhise perspektiivi kaudu loodud ajaplaan

Thompson [9] on uurinud traditsioonilist esineja-keskset perspektiivi konverentside planeerimisel ning püüab kõrvaldada aja konflikte, viies need kokku esinejate eelistustega. Graafiku loomisel kasutatakse süstemaatilise meetodi heuristilist algoritmi, mis on loodud spetsiaalselt konverentside planeerimise jaoks. Optimeerimismeetod alustab ilma planeeritud istungiteta ja lisab neid järk-järgult, määrates osalejaid sõltuvalt nende huvidest. Kui osaleja on varem määratud huvi alusel sessioonile, määratakse ta uuesti uuele sessioonile, et hinata, kas uut ajakava tuleks rakendada, peale mida muudab algoritm iteratiivselt esialgset ajakava, kustutades ja lisades uuesti sessioone.

2.5 Analüüs

Kuigi on olemas erinevaid optimeerimise ja klastertamise lahendusi (klaster - “kogum ühesuguseid või omavahel seotud objekte, nähtusi vms [10].”), esineb neis teatavaid puudujääke. Näiteks Teschi bakalaureusetöös [5] uuritakse konverentsi ajakava alternatiivseid lähenemisviise, mis hõlmavad osalejate arvul põhinevat rühmitamist ja ruumilist jaotust, võttes arvesse kõnelejate populaarsust ja atraktiivsuse koefitsienti. Teschi meetodi peamine nõrkuseks on, et kõnelejate populaarsus-andmed ei ole usaldusväärsed. Kui andmed populaarsuse kohta kogutakse registreerunud osalejate käest, siis on võimalik saada võimalikult tõene tulemus, kuid vastasel juhul võib populaarsus tegelikkusest erineda.

Castaño jt. [6] uurivad konverentsi ajakava sisupõhist optimeerimist, käsitledes sisu sarnasuse ja katvusega seotud keerukusi, pakkudes kaheetapilist lahendusmeetodit. See hõlmab ettekannete grupeerimist sisu järgi ja ajakava koostamist, eelistades sisu sarnasust. Castaño jt. loodu meetodi teeb positiivseks see, et lahendus arvestab sisulist sarnasust sessioonideks jagamisel. Kuid puudujäägiks on, et programmi läbijooksutamiseks jäetud piiratud aeg, mistõttu ei ole uuritud, kuidas mõjutab aeg ideaalse graafiku loomist.

Vangerven jt. [8] artiklis käsitletakse konverentsi ajakava koostamist kolmefaasilise strateegiaga, mille eesmärk on maksimeerida osalejate arvu. Esmalt rühmitatakse ettekanded ruumide ja ajavahemike järgi, seejärel grupeeritakse need sessioonideks ning viimaks kohandatakse ajakava iga osaleja eelistustele vastavalt. Vangerven jt. [8] endi sõnul, isegi kui meetodil on raskusi sidusate sessioonide loomisega pelgalt osalejate eelistuste alusel, on teema lähedus osalejate profiilides aidanud saavutada sessioonide suhtelise sidususe. Seevastu tekitanud probleeme ebavõrdse kohtlemise näol, mis mõjutab kogu konverentsi päevakava. Seega meetodi puudujääk on, et mõne kõneleja sessioonidel on vähe osalejaid kuulamas võrreldes teiste sessioonidega.

Thompson [9] uurib konverentside planeerimisel traditsioonilist esineja-keskset perspektiivi, püüdes kõrvaldada ajakonflikte. Selleks viiakse need kokku esinejate eelistustega, kasutades spetsiaalselt selleks otstarbeks loodud heuristilist algoritmi. Algoritmi optimeerimismeetod alustab planeerimata istungitega ja lisab need järk-järgult, kohandades osalejaid vastavalt

nende huvidele. Iteratiivselt muudetakse ajakava sessioone kustutades ja uuesti lisades, et saavutada võimalikult hea tulemus. Thompsoni loodud meetodi nõrkuseks on ebavõrdsete olukordade loomise võimalus, kus mõned kõnelejad saavad ideaalsed ajakavad, kuid teised peavad rääkima vähem soodsatel aegadel (näiteks kui publik on juba väsinud). Lisaks võib esineda ajavahemikke, mis ei ole ühegi kõneleja eelistuseks, kuid määratakse siiski neile, mis võib mõjutada esinemise kvaliteeti ja osalejate rahulolu.

Töö käigus välja töötatav meetod artiklite automaatseks sessioonideks jagamiseks teaduskonverentside jaoks eristub kõigist senistest lahendustest, kuna meetod kasutab lisa abivahendina ChatGPT keelemudeli poolt antavate sessiooni pealkirja sobivuse hindamist. Teschi bakalaureusetöö [5] optimeerimine on konkreetsetes kontekstis asjakohane, sest osalejate registreerumise põhjal tehtav rühmitamine on optimaalne lahendus. Kuid loodava meetodi jaoks on puudu osalejate soovide/eelistuste andmestik. Lisaks tuleb meetodi loomisel genereerida potentsiaalsete klasterite pealkirjad, mis suurendab rakenduse keerukust. Seega kuuluvad loodava meetodi ettekanded mitme potentsiaalse teemabloki alla ning tuleb leida täiendav optimeerimine klasterite jaoks. Teine puudus seisneb selles, et kui luua ise andmed kõneleja populaarsuse kohta, võtab see palju aega ja võib olla ebatäpne. Tehes seda ChatGPT abil, siis endiselt ei saa täpset tulemust, sest mudelil puuduvad andmed iga teadlase tuntuse kohta, mis muudab sobiva populaarsuse koefitsiendi leidmise keeruliseks ja ajakulukaks. Castaño ja tema kaasautorite [6] esitatud lähenemine sobib siinkohal, kuna see rõhutab sisu sarnasust, mis on kooskõlas töö käigus loodava mudeli eesmärkidega. Vangerven ja tema kaasautorite [8] ettepanekud on ebapiisavad, kuna puuduvad täpsed andmed konverentsil osalejate profiilide kohta, muutes osalejate teemade huvide eeldused liialt üldiseks. Thompsoni [9] esitatud lähenemine on osaliselt piisav, sest selle abil konstrueeritakse kõnelejate soovide alusel ajakavad. Siiski piirab kõnelejate eelistuste andmete puudumine selle kohandamist konkreetsele meetodile.

Teaduskonverentsi andmete põhjal loodava automaatse sessioonide jaotamise meetodi loomisel tuleb põhiliselt kasuks Castaño ja tema kaasautorite [6] poolt loodud sisupõhist optimeerimist, sest keelemudel suudab hinnata sisu sarnasuse põhjal sessiooni sobivust.

3. Koostatud meetodi kirjeldus

Järgnevates alapeatükkides kirjeldan meetodiga kaasnevaid teeke, prompti koostamise põhimõtteid ja meetodi enda olemust. Loodud kood on kirjutatud Python 3.11.5 versioonis. Jupyter Notebook, mis on osa Anaconda 3 platvormist, on kasutusel liidesena. Jupyter Notebooki kasutus võimaldab koodi tõhusalt vormistada ning tagab, et keelemudeliga ühenduse katkemise korral on võimalik teatud koodiosad kergesti uuesti käivitada.

3.1 Meetodi lühikirjeldus

Meetodi esimeses etapis loetakse andmed sisse. Andmed koosnevad teadusartiklite pealkirjadest ja abstraktidest. Artiklite pealkirjade ja abstraktide põhjal esitatakse keelemudelile päring, et see genereeriks võimalikke sessioonide pealkirju. Järelpäringuga korrigeeritakse loodud pealkirju, muutes neid üldisemaks. Pealkirjad, mis esinevad vaid korra, hinnatakse üle, keskendudes nende sobivusele masinõppe konverentsi kontekstis. Seejärel eemaldatakse madala väärtusega sessioonipealkirjad algselt loodud tabelist. Alles jäänud sessiooni pealkirju hinnatakse, võttes arvesse nende sobivust iga ettekande artikli pealkirja ja sisu põhjal. Järgmisena toimub sessioonideks jagamine lineaarse optimeerimisülesande lahendamise teel. Lõpuks hinnatakse Levenshteini kauguse (Levenshteini kaugus - Levenshteini kaugus on sõnade erinevuste mõõdik, mis kvantifitseerib kahe jada erinevuse, arvutades kõige vähem redigeerimistoiminguid (nt tähemärgi sisestamine, kustutamine ja asendamine), mida on vaja ühe sõne teiseks muutmiseks [11].) abil üks-ühele suhet sessiooni pealkirjade sarnasust, võrreldes käsitsitehtud sessioonide ja meetodi loodud sessioonide jaotamise erinevust. Järgmises peatükis on meetodid esitatud üksikasjalikumalt.

3.2 Meetodis kasutatavate teekide kirjeldus

Azure OpenAI on ettevõtetele suunatud, kasutaja poolt täielikult hallatav pakett, mis pakub tarbijatele täiustatud juurdepääsu OpenAI võimalustele. Microsoft annab ettevõttele võimaluse kasutada selle rubriigi raames suuremahulisi mudeleid [12].

Pythoni pakett **Pyscipopt** pakub Pythoni liidest SCIP (ingl *Solving Constraint Integer Programs*), mis on raamistik kitsendustega matemaatiliste optimeerimisülesannete

lahendamiseks [13]. SCIP on üks kiireimaid mitteärilisi lahendusi segatud täisarvulise mittelineaarse planeerimise (ingl MINLP - *mixed integer nonlinear programming*) ja segatud täisarvulise planeerimise (ingl MIP - *mixed integer programming*) jaoks [14]. Põhikomponent, mida kasutati pycipoptist, on Model-klass, mis võimaldab defineerida optimeerimise ülesandeid ja lahendada neid SCIPi algoritmide abil [15]. Funktsioon quicksum on pycipoptis sisalduv utiliit, mis on mõeldud väljendite jada tõhusaks summeerimiseks, mida on sageli vaja optimeerimisülesannete formuleerimisel [16].

Loodav sessioonideks jaotamise meetod kasutab täisarvulist lineaarset planeerimist (edaspidi ingl ILP), mis on optimeerimisülesannete tüüp, kus sihtfunktsioon ja piirangud on lineaarsed ning kõik otsustusmuutujad on täisarvulised [13]. ILP eesmärk on leida otsustusmuutujatele täisarvulised väärtused, mis optimeerivad antud lineaarset sihtfunktsiooni, samal ajal rahuldades lineaarseid piiranguid [17]. Selles kontekstis on oluline märkida, et SCIP teek, mida kasutatakse meetodi arendamisel, toetab ILP lahendamise võimekust, pakkudes tugevat alust keeruka sessioonideks jaotamise ülesande tõhusaks lahendamiseks [18].

Täisarvulise lineaarse planeerimise ülesande definitsioon [19]:

Maksimeerida: $C_1X_1 + C_2X_2 + \dots + C_nX_n$, kus $X_1, \dots, X_n \in \mathbb{Z}$, $C \in \mathbb{R}$

Kitsendused: $A_{11}X_1 + \dots + A_{1n}X_n \leq B_1$, $A \in \mathbb{R}$, $C \in \mathbb{R}$

...

$A_{m1}X_1 + \dots + A_{mn}X_n \leq B_n$, kus $X_1, \dots, X_n \geq 0$

JSON on avatud standardne failivormingu ja andmevahetusformaad, mis salvestab ja edastab andmeobjekte, mis koosnevad massiividest ja atribuutide-väärtuste paaridest, kasutades inimestele arusaadavat teksti [20]. JSON on laialdaselt kasutatav andmevorming, millel on mitmesuguseid rakendusi elektroonilises andmevahetuses, näiteks serveritega suhtlevad veebirakendused [20].

Pandas on Pythoni pakett, mis pakub suure jõudlusega hõlpsasti kasutatavaid andmestruktuure ja andmeanalüüsi vahendeid. Kõige sagedamini kasutatav andmestruktuur pandases on DataFrame ehk eesti keeles andme raamistik, mida võib pidada relatsiooniliseks

andmetabeliks, millel on read ja veerud [21].

NumPy on vektoriseerimise, indekseerimise ja ülekandmise funktsioonid, mis on kiired ja kohandatavad. NumPy abil on saadaval ulatuslikud matemaatilised funktsioonid, juhusliku arvu generaatorid, lineaaralgebra operatsioonid, Fourier' teisendused ja muud funktsioonid [22].

Defaultdict Pythoni **Collection** moodul pakub vaikeväärtusi olematutele võtmetele [23]. Meetod on sisseehitatud 'dict' klassi alamklassi, kuid ühe keerdkäiguga: teek tühistab ühe meetodi ja lisab teise kirjutatava instantsi muutuja juurde [23]. Konstruktor võtab valikulise argumendi 'default_factory', seda tehase funktsiooni kasutatakse puuduvate väärtuste andmiseks, kui kasutatakse võtmeid, mida ei ole olemas. Vaikimisi, kui ei anda 'default_factory' ette, käitub see nagu tavaline sõnastik [23].

3.3 Sessioonide optimeerimise ülesande esitamine täisarvulise lineaarse planeerimise ülesandena

Järgnevalt on kirjeldatud, kuidas täisarvulist lineaarsete kitsendustega lineaarset planeerimist saab kasutada konverentsi artiklite sessioonidesse sobivuse maksimeerimiseks.

Definitsiooni põhjal tähistavad A, B, C parameetreid ja x_1, \dots, x_n optimeerimistulemusi. Toon sisse järgmised binaarsed muutujad:

X_{ij} - on binaarne muutuja, kus i indekseerib artikleid ja j indekseerib seansse. X_{ij} on osa otsustusmuutujate $x_1 \dots x_n$ komplektist

Y_j - on binaarne muutuja, mida indekseerib j, mis tähistab seansse. Y_j on osa otsustusmuutujate $x_1 \dots x_n$ komplektist

Nendel muutujatel peavad kehtima järgmised kitsendused:

- Iga artikkel peab kuuluma täpselt ühte sessiooni;
- Igas sessioonis peab olema täpselt 5 artiklit või 0 artiklit;
- Valitud sessioone on kokku täpselt 57. Sessioonide koguarv arvutakse sessioonide arv jagatud mitu artiklit on ühes sessioonis, ehk $285/5 = 57$.

Neid kitsendusi saab esitada lineaarsete võrranditega järgmiselt:

Kui i on artikliel indeks ja j sessiooni indeks, siis $\forall i, \forall j$

$$1. X_{i1} + \dots + X_{in} = 1$$

$$2. X_{ij} + \dots + X_{mj} = S \times Y_j$$

$$\text{Kui } Y_j = 0, \text{ siis } X_{1j} + \dots + X_{mj} = 0$$

$$\text{Kui } Y_j = 1, \text{ siis } X_{1j} + \dots + X_{mj} = S$$

$$3. Y_1 + \dots + Y_n = N_{\text{sessiooni}}$$

Loodud meetodis tuleb maksimeerida sessioonide jaotamisel saadud sessiooni pealkirjade skooride summat, et leida parimad tulemused sessiooni pealkirjade, skooride ja artiklile kuuluvuse vahel. Selleks, et vähendada maksimeerimija jaoks minevat arvutus mahtu, on optimeerimise protsessile kehtestatud kolm kitsendust, et tagada lahenduse struktuurne korrektsus ja praktiline rakendatavus. Esimene reegel nõuab, et igale artiklile määrataks jaotuses täpselt üks sessioon, tagades sellega artiklite ühtlase jaotuse üle kõikide sessioonide. Teine reegel sätestab, et igasse sessiooni peab kuuluma täpselt viis artiklit, mis kindlustab sessioonide sisulise tiheduse ja asjakohasuse. Kus B sümboliseerib artiklite arvu, mis konkreetse sessiooni raames esitatakse ja kas sessioon osutus valituks. Kolmas ja viimane reegel piirab sessioonide koguarvu, milleks on antud juhul M artiklit jagatuna viiega, mis tähendab, et meie näites moodustub 57 sessiooni. Kolme reegliline struktureeritud lähenemine mitte ainult ei optimeeri konverentsi ajakava, vaid ka tagab, et iga sessioon on sisult rikas ja mitmekesine, vastates osalejate huvidele ja ootustele.

Järgneval joonisel on näha sessioonide jagamise koodi ja töökäigu kirjeldusi kommentaaridena:

```

# Sessiooni gruppi suurus
group_size = 5

# Loeb admed failist
df_scores = pd.read_csv('df_scores.csv', sep=',')
article_ids = df_scores['id_name'].tolist()
session_titles = df_scores.columns[1:].tolist()

# Loob uue mudeli
model = Model("ArticleSessionAssignment")

# Lisab iga artikli-seansi kombinatsiooni jaoks binaarsed otsustusmuutujad
x = {}
for article_id in article_ids:
    for session_title in session_titles:
        x[(article_id, session_title)] = model.addVar(vtype="B", name=f"x_{article_id}_{session_title}")
y = {}
for session_title in session_titles:
    y[session_title] = model.addVar(vtype="B", name=f"y_{session_title}")

# Eesmärk: Sobivusskooride summa maksimeerimine
model.setObjective(
    quicksum(x[(article_id, session_title)] * df_scores.loc[df_scores['id_name'] == article_id, session_title].values[0]
    for article_id in article_ids for session_title in session_titles), "maximize")

# Piirang: iga artikkel on määratud täpselt ühele seansile
for article_id in article_ids:
    model.addCons(quicksum(x[(article_id, session_title)] for session_title in session_titles) == 1)

# Piirang: kui seansi pealkiri on valid, on sellel täpselt 'group_size' artiklid; muidu pole sellel artikleid
for session_title in session_titles:
    model.addCons(quicksum(x[(article_id, session_title)] for article_id in article_ids) == group_size * y[session_title])

# Piirang: validus on ainult piiratud arv seansi pealkirju
total_sessions = len(article_ids) // group_size
model.addCons(quicksum(y[session_title] for session_title in session_titles) == total_sessions)

# Lahendav probleemi
model.optimize()

```

Joonis 1. Sessioonideks jagamise kood.

Käesolev kood realiseerib matemaatilise optimeerimise seatud eesmärgid jaotada artikleid erinevatele sessioonidele, lähtudes nende sobivuse hinnangutest. Algfaasis loetakse andmed CSV-failist, mille järel luuakse iga artikli ja sessiooni kombinatsiooni kohta binaarsed otsustusmuutujad. Mudeli eesmärk on maksimeerida kõigi sobivuse skooride summasid. Kehtestatud piirangud tagavad, et iga artikkel on määratud täpselt ühele sessioonile ja iga sessioon sisaldab etteantud arvu artikleid, mis on määratletud muutujaga `group_size`, või ei sisalda üldse artikleid. Samuti piiratakse validavate sessioonide arvu, lähtudes artiklite koguarvust. Peale optimeerimise probleemi lahendamist esitab programm artiklite jaotuse sessioonide vahel koos optimaalse eesmärgi väärtusega, kui see on lahenduv.

3.4 Meetodi detailne tehniline kirjeldus

Loodud sai kaks meetodit, mille mudelite põhiline erinevus seisneb selles, et kas on kasutatud teatud osadel ChatGPT 3.5 turbot või ChatGPT 4.

Andmetöötlus:

Alustuseks on iga andmetöötlusprogrammi protsessi osaks andmete sisselugemine ja töötlus. Iga sessiooni ID-d, artiklite pealkirjad ja abstraktid tuleb hoolikalt failist sisse lugeda. Mõista andmete olulisust ja vajadusel parandada või korrastada andmeid. Eelkõige oli antud töös vajalik etteantud pealkirjade ja sessiooni pealkirjade lühendite pikalt välja kirjutamine, et hiljem saaks hinnata loodud sessiooni pealkirja professionaali loodud pealkirjadega. Näiteks kui pealkirjas sisaldab “Meta-Learning NAS” siis pikalt väljakirjutatuna on “Meta-Learning Neural Architecture Search”. Ainult nii saame kindlustada, et meie analüüs põhineb terviklikkel andmetel ja toob esile uurimistöö sisulise tähenduse.

| | SessionTitles | ID | TITLE | ABSTRACT |
|-----|--------------------------------|------|---|---|
| 0 | Active, semi-sup. learning (1) | 190 | SemiTE: Semi-supervised Individual Treatment ... | Recent years have witnessed a surge of interes... |
| 2 | Active, semi-sup. learning (1) | 442 | Exploring Latent Sparse Graph for Large-Scale ... | We focus on developing a novel scalable graph-... |
| 4 | Active, semi-sup. learning (1) | 891 | SMFM4L: Multi-typed Objects Multi-view Multi-i... | Multi-typed objects Multi-view Multi-instance ... |
| 6 | Active, semi-sup. learning (1) | 926 | Consistent and Tractable Algorithm for Markov ... | Markov network (MN) structured output classifi... |
| 8 | Active, semi-sup. learning (1) | 952 | Multi-Task Adversarial Learning for Semi-Super... | Trajectory-User Linking (TUL), which aims to l... |
| ... | ... | ... | ... | ... |
| 560 | Transfer and Multitask | 916 | Fast and Accurate Importance Weighting for Cor... | Bias in datasets can be very detrimental for a... |
| 562 | Transfer and Multitask | 1171 | Newer is not always better: Rethinking transfe... | Fine-tuning of large pre-trained image and lan... |
| 564 | Transfer and Multitask | 1255 | Learning to Teach Fairness-aware Deep Multi-Ta... | Fairness-aware learning mainly focuses on sing... |
| 566 | Transfer and Multitask | 1288 | Overcoming Catastrophic Forgetting via Directi... | This paper studies a new design of the optimiz... |
| 568 | Transfer and Multitask | 1371 | On the relationship between disentanglement an... | One of the main arguments behind studying dise... |

285 rows × 4 columns

Joonis 2. Sisseloetud andmed näidis.

Sessiooni pealkirjade loomine juturobotiga:

Mudel kasutab teeki Azure OpenAI, et suhelda Juturobotiga. Juturobot suudab ennustada potentsiaalseid sessiooni pealkirju artiklite sisu põhjal. Formuleeritud küsimus, mis antakse juturobotile on järgnev:

| | |
|-----------------------------|------------------------------------|
| Originaalne prompt: (ingl) | Tõlgitud lühendatud prompt: (est) |
|-----------------------------|------------------------------------|

| | |
|---|---|
| <p>We are organising a machine learning / data mining conference. Each of the 200+ accepted articles will make a presentation at the conference.</p> <p>We now need to start the work of arranging the article presentations into many parallel sessions, each with a thematic title.</p> <p>Example session titles are the following: Computer Vision Deep Learning Graph Neural Networks Interpretability and Explainability Knowledge Graphs Networks and Graphs NLP and Text Mining</p> <p>As the first step, we need to generate several potential session titles for each of the articles. After that, we will gather frequently occurring session titles and find a way to arrange the presentations into sessions with these titles.</p> <p>Please go through the list of articles provided below. For each article, generate 6 potential session titles. A good session title must be concise and suitable for at least 2% and at most 10% of overall papers accepted to the conference.</p> <p>This means that the title should not be overly specific, but also not overly broad. Whenever possible, prefer a shorter title, but do not use acronyms. Make the first 3 of the titles a bit more specific and the last 3 a bit more general.</p> <p>Do not provide any other output than just the results in JSON format as the following example shows:</p> <pre>{ "ID1": ["sessionTitle1", "sessionTitle2", "sessionTitle3", "sessionTitle4", "sessionTitle5", "sessionTitle6"], "ID2": ["sessionTitle1", "sessionTitle2", "sessionTitle3", "sessionTitle4", "sessionTitle5", "sessionTitle6"] }</pre> <p>Here's the list of article IDs, titles and</p> | <p>Iga artikli jaoks tuleb kõigepealt välja mõelda mitu võimalikku sessiooni pealkirja. Seejärel koostame nimekirja sageli esinevatest sessiooni pealkirjadest ja mõtleme välja, kuidas rühmitada ettekanded nendele pealkirjadele vastavateks sessioonideks. Palun vaadake allpool esitatud artiklite nimekirja. Leidke iga artikli kohta kuus võimalikku sessiooni pealkirja. Hea sessiooni pealkiri peaks olema lühike ja sobima mitte rohkem kui 2% ja mitte rohkem kui 10% konverentsile vastuvõetud ettekannete koguarvust. See tähendab, et pealkiri ei tohiks olla liiga üldine ega ka liiga konkreetne. Võimalusel valige lühem pealkiri, kuid vältige akronüümide kasutamist. Kolm viimast pealkirja võiksid olla veidi üldisemalt sõnastatud ja 3 esimest veidi täpsemalt.</p> |
|---|---|

| | |
|------------|--|
| abstracts. | |
|------------|--|

Tabel 2 Formuleeritud küsimus, mis antakse juturobotile

Selleks, et saavutada keelemudeli tulemustes suurem täpsus, määrasin osaliselt subjektiivsed kriteeriumid sobiva pealkirja saavutamiseks. Määratluse kohaselt peaks sessiooni pealkiri sobima vähemalt 2% ja mitte rohkem kui 10% konverentsile vastuvõetud ettekannete koguarvust. Tähendades, et pealkiri ei tohiks olla liiga üldine ega liiga spetsiifiline. Pikad pealkirjad võivad osalejate tähelepanu hajutada või segadusse ajada. Ideaalne pealkiri on tavaliselt kuni 5 sõna, mis suudab edastada sessiooni peamist ideed ilma üleliigsete sõnadeta.

Selline prompt aitab saavutada seda, et juturobot genereeriks sessioonipealkirjad, mis oleksid üldised. Erinevate promptitega testides mõistsin/õppisin, et kui pealkirja üldisust väga rangelt mitte piiritleda, siis suudab juturobot kõrvale hiilida ja ennustada väga spetsiifilisi pealkirju. Prompti ette andmisel on vaja silmas pidada tokenite arvu, kui tokenite arv ületab lubatud määra (ChatGPT 3.5 on 4096 tokenit [24]) siis ei võtta juturobot vastu prompti. Samas on tähtis, kui mitu pärimist tehakse, sest iga token maksab ("Token ehk üksikmärk, eksemplar on märgi esitis üksikeksemplarina, üksiku olemasoleva eseme või sündmusena." [25]). Selleks, et vähendada mudeli hinda päringute rohkuse tõttu, annan juturobotile ette 10 artiklit koos nende pealkirjadega, sel juhul jääb sisendi maht koos artikli sisuga alla lubatud maksimaalmäära.

Meetod küsib juturobotilt järel päringuga küsimuse, et sessiooni pealkirjad parandada üldisemaks. Kasulik on teha keelemudelile järelpäring, sest kui mudel on pealkirja genereerinud, mis on enamjaolt üldised aga lisanud sõna lõppu liiga spetsiifiline osa, siis väikese muudatusega saab juba ideaalse sessiooni pealkirja. Näiteks kui on sessioon pealkiri "Supervised Learning from Observational Data" aga lõppu lisatud sõna "from Observational Data", mis küll on sarnaste sisutega artiklite jaoks üldine aga konverentsi raames mittepiisavalt üldine, siis parem oleks "Supervised Learning".

| | |
|---------------------------------|--------------------------------|
| Järel päringu küsimus: (ingl.) | Järel päringu küsimus: (est.) |
|---------------------------------|--------------------------------|

| | |
|---|--|
| Go through your output and provide an improved set of session titles in the same format. Make titles shorter and more general. Do not use acronyms. For each word, consider whether you can drop it because the essence of the title would remain the same. | Käige oma väljund läbi ja esitage täiustatud seansi pealkirjad samas formaadis. Tehke pealkirjad lühemaks ja üldisemaks. Ärge kasutage akronüüme. Iga sõna puhul kaaluge, kas võite selle välja jätta, sest pealkirja sisu jääks samaks. |
|---|--|

Tabel 3 Järel päringu prompt

| | |
|---|--|
| Keelemudeli esimese päringu järgne tulemus: | { "190": ["Treatment Effect Estimation", "Supervised Learning from Observational Data", "Disagreement-based Co-training", "Predictive Modelling", "Data Mining and Analysis", "Machine Learning Techniques"], "442": ["Scalable Semi-supervised Learning", "Latent Sparse Graph Construction", "Optimization in Large-Scale Data", "Large Scale Data Analysis", "Modern Machine Learning Methods", "Semi-supervised Learning"], "891": ["Multi-view Multi-instance Multi-label Learning", "Interconnected Multi-typed Objects", "Selective Matrix Factorization", "Machine Learning Optimization", "Data Structure and Representation", "Advanced Learning Methods"], "926": ["Markov Network Learning", "Consistent and Tractable Algorithm", "Loss Minimization in Learning", "Data Supervision Methods", "Advanced Learning Systems", "Machine Learning and Deep Learning Application"], "952": ["Trajectory-User Linking", "Kalman filtering and seq2seq model", "Multi-task adversarial learning", "Privacy and Security in Machine Learning", "Advanced Techniques in Data Mining", "Semi-supervised Learning"], "1128": ["Near Out-of-distribution Detection", "Radar Targets Detection", "Contrastive Learning Insights", "Sensor Data Processing", "Deep Learning and Non-Deep Methods Comparison", "Data Mining and Machine Learning Approach"], "375": ["Class-Imbalanced Node Classification", "Reinforcement Mixup", "Self-supervised Context Prediction", "AI in Graph Networks", "Learning Mechanisms in Neural Networks", "Machine Learning Applications in Node Classification"], "481": ["Multi-domain Active Learning", "Semi-supervised Anomaly Detection", "Multi-armed Bandits Method", "Active Learning and its Applications", "Anomaly Detection and its Methods", "Learning Techniques for Multi-domain Data"], "531": ["Class-Mixed Data Generation", "Out-Of-Distribution Detection", "Contrastive Learning", "Data Augmentation and Class Augmentation", "Machine Learning for Data Generation", "Advanced Learning Techniques for Data Detection"], "1054": ["Transductive Active Learning", "Stopping Criterion", "Probabilistic Active Learning Framework", "Techniques in Active Learning", "Deep Learning for Data Labeling", "Advanced ML Techniques for Data Classification"] } |
| Järelpäringu tulemus: | { "190": ["Treatment Estimation", "Supervised Learning", "Co-training", "Predictive Modelling", "Data Mining", "Machine Learning"], "442": ["Scalable Learning", "Sparse Graph Construction", "Large-Scale Optimization", "Data Analysis", "Learning Methods", "Semi-supervised |

| | |
|--|--|
| | Learning"], "891": ["Multi-view Learning", "Interconnected Objects", "Matrix Factorization", "Optimization", "Data Representation", "Learning Methods"], "926": ["Markov Network", "Tractable Algorithm", "Loss Minimization", "Data Supervision", "Learning Systems", "Deep Learning"], "952": ["Trajectory Linking", "Kalman Filtering", "Adversarial Learning", "Data Security", "Data Mining", "Semi-supervised Learning"], "1128": ["Distribution Detection", "Radar Detection", "Contrastive Learning", "Data Processing", "Deep Learning", "Data Mining"], "375": ["Node Classification", "Reinforcement Mixup", "Context Prediction", "Graph Networks", "Neural Networks", "Machine Learning"], "481": ["Active Learning", "Anomaly Detection", "Bandits Method", "Learning Applications", "Anomaly Methods", "Learning Techniques"], "531": ["Data Generation", "Distribution Detection", "Contrastive Learning", "Data Augmentation", "Data Generation", "Data Detection"], "1054": ["Transductive Learning", "Stopping Criterion", "Active Learning", "Learning Techniques", "Data Labeling", "Data Classification"]} |
|--|--|

Tabel 4 Esimese päringu ja järelpäringute tulemused

Sellegipoolest ei ole garanteeritud, et juturobot genereeriks alati ühepalju sessioonide pealkirju. Kui ChatGPT annab vähem kui 6 sessiooni pealkirja, siis lisame tühiväärtused, et kõik pealkirjad saaks teisenda sessiooni pealkirjade põhjal transpose() ja one-hot-vectori meetodeid kasutades andmestikuks sessioonide sobivuse hindamise jaoks.

Üleliigsete sessiooni pealkirjade vähendamine:

Keelemudel on nüüdseks genereerinud arvukalt pealkirju, millest paljud ei sobi vähemalt viiele artiklile sessiooni pealkirjaks. Selleks on vajalik kogu sessiooni pealkirjade arvu vähendada. Üks viis, kuidas vähendada loodud sessiooni pealkirjade arvu, on uuesti analüüsida pealkirjade sobivust. Esiteks, kui tabelis on mõni genereeritud sessiooni pealkiri esinenud mitu korda, saame arvestada, et need sessiooni pealkirjad sobivad vähemalt kahele artiklile ja ei vaja ülehindamist. Teiseks saab lasta keelemudelil hinnata üle pealkirjad, mis on esinenud vaid korra. Uus päring keelemudelile peaks hindama, kas iga pealkiri sobib üleüldiselt 0-st kuni 1-ni sessiooni pealkirjaks masinõppe konverentsi raames. Keelemudeli tulemusena saadud hinnangu väärtustest saan keskmise väärtuse, mille väärtuse põhjal välistan madalama skoori väärtusega pealkirjad. Eesmärk on saada sessiooni pealkirjade kogumaht kuni 700 pealkirja peale. Selleks on vaja kontrollida, kas uus vähendatud pealkirjade arv on lubatud mahu piires või on vaja vähendada tulemust veelgi. Selleks, et tulemust veelgi vähendada, on vaja lisada keskmisele väärtusele 0,1 punkti. Uue tekkinud

keskmise väärtusega saab välistada rohkem pealkirju. Seejärel kontrollida, kas pealkirjade arv mahub soovitud mahu sisse. Sel viisil toimides on saavutatud mõistlik maht sessiooni pealkirju, et hinnata iga sessiooni pealkirja sobivust konkreetsele artiklile.

Äsja tehtud optimeerimise tulemusel on võimalik vähendada järgmiste päringute arvu ChatGPT 4-le 3000 päringult umbes 500 päringule. Vähendamine on vajalik, kuna keelemudelid saavad korraga hinnata ainult 250 pealkirja, isegi kui tokenite arv võimaldaks suuremat mahtu.

Sessiooni pealkirjade hindamine artiklile põhjal:

Seejärel on mudelil vaja hinnata kõikide sessiooni tiitlite sobivust igale artiklile. Sessioonide hindamise promptiga on tähtis juturobotile anda konkreetne piirang, et mudel ei tohiks uusi pealkirju genereerida ega parandada. Sellest olenemata võib hindamise järel juturobot lisada mõne üleliigse jutumärgi või tühiku, mis tuleb eemaldada ja ühendada ilma lisanditeta sessiooni pealkirjaga. Kui ChatGPT on kogemata genereerinud lisa pealkirjasid, siis kustutame duplikatsioonid.

| Päringu küsimus: (ingl) | Päringu küsimus: (est) |
|--|--|
| <p>We are organising a machine learning / data mining conference. Each of the 200+ accepted articles will make a presentation at the conference. We now need to start the work of arranging the article presentations into many parallel sessions, each with a thematic title. We have already identified a list of candidate session titles. Your task is to consider the given article and, for each session title, evaluate whether the article can be in that session. This evaluation should reflect how well the topic of each session aligns with the themes, methods, and objectives discussed in the abstract.</p> <p>Please provide your assessment using a scale from 0 to 1, where 0 indicates that the article does not suit to be in that session, and 1 signifies perfect relevance. The output should be formatted in JSON, with each session title serving as a key and its</p> | <p>Me korraldame masinõppe/andmekaave konverentsi. Iga üle 200 vastuvõetud artiklist teeb konverentsil ettekande. Nüüd peame alustama tööd, et korraldada artiklite esitlused mitmeteks paralleelseteks sessioonideks, millest igaühel on oma teema. Me oleme juba kindlaks teinud nimekirja kandidaat istungite pealkirjadest. Teie ülesanne on kaaluda antud artiklit ja hinnata iga sessiooni pealkirja puhul, kas artikkel võib kuuluda sellesse sessiooni. See hinnang peaks kajastama, kui hästi vastab iga istungi teema kokkuvõttes käsitletud teemadele, meetoditele ja eesmärkidele.</p> <p>Palun andke oma hinnang, kasutades skaalat 0-1, kus 0 tähendab, et artikkel ei sobi antud sessioonile, ja 1 tähendab, et see sobib ideaalselt sessioonile. Tulemus tuleks vormistada JSON-vormingus, kus iga sessiooni pealkiri on võtmeks ja vastav</p> |

| | |
|--|--|
| <p>corresponding relevance score as the value. Use the exact format shown below for your response:</p> <pre>{ "SessionTitle1": score1, "SessionTitle2": score2, ... }</pre> <p>The article to be evaluated is:</p> | <p>asjakohasuse hinne on väärtuseks. Kasutage oma vastuses täpselt allpool esitatud formaati:</p> <pre>{ "SessionTitle1": score1, "SessionTitle2": score2, ... }</pre> <p>Hinnatav artikkel on:</p> |
| <p>Päringu tulemus ID 1038 kohta:</p> | <pre>{"AI Applications": 0.8, "AI Optimization": 0.9, "AI Techniques": 0.6, "AI for Feature Selection": 1, "AI for Information Retrieval": 0, "AI for Stream Analysis": 0, "Active Learning": 0.2, "Active Learning Advancements": 0.2, "Active Learning Evolution": 0.2, ... "Advances in Feature Engineering": 0.7, "Algorithm Advances": 0.8, "Algorithm Analysis": 0.8, "Algorithm Configuration": 0.7, "Algorithm Design": 0.8, "Algorithm Design in Meta-learning": 0.5, "Algorithm Improvement": 0.8, "Algorithm Methods": 0.8, "Algorithm Optimization": 0.9, "Algorithm Performance": 0.9, "Algorithmic Enhancements": 0.9, "Algorithmic Improvements": 0.8, "Algorithms Task Sequencing": 0.2}</pre> |

Tabel 5 Sessiooni pealkirjade vähendamise prompt ja vastus

Sessiooni pealkirjade lisa optimeerimine ja duplikatsioonide tegemine:

Selleks, et vähendada viimasele sessioonide jaotamisele mõeldud arvutuse osa ajakulu, on vajalik vähendada sessiooni pealkirjade arvu andmestikus. Summa leidmise järel kustutame need sessiooni pealkirjad, mille summa väärtus on alla 1.3, sest alla selle kokku ei tule ühte tervik sessiooni, kuhu kuulub 5 artiklit. Summa väärtus võiks olla suurem, kuid konkreetse meetodi puhul pole tehtud ühtegi testi, et uurida, kas tulemus halveneb, kui välistada rohkem sessiooni pealkirju.

Hindamistabelisse on vaja veel lisada tagasi duplikatsioonid tingimusel, et sessiooni pealkirja summa on suure väärtusega, sellisel juhul võiks sama sessiooni tabelis esineda rohkem. Seega kui summa on üle väärtuse 20 siis meetod teeb ühe duplikatsiooni, kui 40 siis kaks duplikatsiooni jne. Sarnast lähenemist on kasutatud ECML PKDD konverentsil.

Sessioonideks jagamine ja optimeerimine:

Sessioonideks jagamise protsessis kasutasin OpenAI ChatGPT 4 tehnoloogiat, et paremini mõista ja rakendada olemasolevaid meetodeid. Alustuseks pöördusin keelemudeli poole, et uurida, millised teekid ja lahendused sessioonide jaotamiseks on juba välja töötatud. Esimeses sammus formuleerisin prompti, määratledes konverentsi suuruse ja eesmärgid, mis aitasid suunata päringut. Seejärel võrdlesin keelemudeli poolt genereeritud lahendusi standardsete näidisülesannete ja oma käsitsi koostatud lahendustega. Selline lähenemine võimaldas mul hinnata ja optimeerida mudeli tõhusust reaalsetes stsenaariumides, tagades et sessioonide jaotamine toetub asjakohastele ja efektiivsetele meetoditele. Lühendatud esimene prompt tõlgitud:

| Prompti sisend: | Prompti väljund lühendatud: |
|--|---|
| Ma korraldan masinõppe konverentsi, tahan jagada 240 vastuvõetud tööd 40 sessioonile, millest igaühes on 6 tööd. Olen juba kindlaks teinud 100 potentsiaalset sessiooni pealkirja ja iga ettekande puhul kindlaks teinud, millised sessiooni pealkirjad on selle jaoks suurepärased (skoor 1) ja millised on enam-vähem ok (skoor 0,5). Millise tarkvara abil saaksin korraldada ettekanded sessioonidesse nii, et ükski ettekanne ei oleks sessioonis, mille pealkiri ei sobi, ja et üldine keskmine tulemus oleks maksimaalne. | 1. Modelleerimine täisarvulise lineaarse programmeerimise (ILP) probleemina: Võite kasutada ILP-lahendajaid, nagu IBMi CPLEX, Gurobi või avatud lähtekoodiga SCIP-lahendaja. 2. Metaheuristilised lähenemisviisid: Geneetilised algoritmid, simuleeritud lõõmutamine või sipelgapesade optimeerimine võiksid olla selle probleemi lahendamiseks sobivad. Nende meetodite rakendamiseks võib kasutada tarkvara/tööriistu nagu DEAP (Distributed Evolutionary Algorithms in Python) või Optuna. 3. ... |

Tabel 6 Täpsem suhtluse ajalugu keelemudeliga [18]

SPIC mudeli (Sessiooni Põhinev Indeks ja Klasterdaja) abil jagame Artiklid parimateks sessioonideks. SPIC mudel on tööriist, mida kasutatakse andmeanalüüsis, eriti suurte andmemahutude korral, et paremini mõista andmekogumite struktuuri ja leida olulisi mustreid. SPIC mudel keskendub andmekogumite jagamisele sessioonideks, kus sessioon võib olla

näiteks kasutaja interaktsioon veebilehega või ajavahemik teatud sündmuste vahel. Seejärel analüüsib SPIC teek iga sessiooni sisu, et mõista selle olulisust või teha järeldusi vastavalt eelnevalt seatud kriteeriumitele. Teek võimaldab leida olulisi mustreid või seoseid andmetes, mis omakorda võib aidata tuvastada olulisi sessioone või artikleid. Täisarvuline lineaarne programmeerimise abil saab lahendada sessioonideks jagamise probleemi. Selleks määratakse otsustusmuutujad, kas artikkel kuulub sessiooni või mitte ja seejärel saab maksimeerida üldist sobivust skoori. Lisades piirangud, et iga ettekanne kuuluks ainult ühte sessiooni, iga sessioonis oleks täpselt 5 artiklit ning ükski ettekanne ei kuuluks sessiooni, mille sessiooni pealkiri ei ole sobiv.

Pealiskaudne analüüs:

Pärast tulemuste saamist viime läbi täiendava analüüsi, kus kontrollime Levenshteini kauguse abil, kui sarnased on originaalsete sessioonide pealkirjad ja uute sessioonide pealkirjad. Levenshteini kaugus võimaldab hinnata pealkirjade sarnasust, võttes arvesse tähemärkide erinevusi, ja aitab tuvastada, kas uued sessioonid on piisavalt sarnased originaalsetega.

Tegelikult ei ole võimalik mõõta meetodi täpsust pelgalt selle põhjal, kui sarnased on loodud pealkirjad 2022 ECML PKDD konverentsi sessioonide pealkirjadega. Selle põhjuseks on asjaolu, et valitud sessiooni pealkiri võib erinevate ekspertide hinnangul olla sobimatu. Seetõttu tuleb mudeli täpsuse hindamiseks võrrelda meetodi genereeritud tulemusi originaalidega ja lasta mitmel eksperdil hinnata, kas teatud artikkel sobib antud sessiooni pealkirja alla kümne palli süsteemil.

3.5 Meetodide parameetrite võrdlus

Meetodite parameetrite võrdluse käigus on ilmnenu, et ChatGPT 4 mudel on märkimisväärselt aeglasem kui ChatGPT 3.5 Turbo- seda ligikaudu 2 korda. Kahjuks ei ole Tartu Ülikoolil juurdepääsu ChatGPT 4 Turbo versioonile, mis võib olla kiirem. Kiirus on oleneb samuti arvuti võimekusest. ChatGPT 4 mudeli täpsus, mõõdetuna Levenshteini kauguse järgi, on 34/258, samas kui ChatGPT 3.5 mudeli täpsus, arvutatuna üks-ühele suhtena, on 35/258. Mudel ChatGPT 4 maksumus oli 80 eur ja see hõlmas kokku 1865055 tokenit, ühtlasi kui mudel ChatGPT 3.5 oli kordades odavam hinnaga 32 eur, sisaldades 1675635 tokenit.

| Models | Context | Input (Per 1,000 tokens) | Output (Per 1,000 tokens) |
|------------------------|---------|--------------------------|---------------------------|
| GPT-3.5-Turbo-0125 | 16K | €0.0005 | €0.0014 |
| GPT-3.5-Turbo-Instruct | 4K | €0.0014 | €0.002 |
| GPT-4-Turbo | 128K | €0.010 | €0.028 |
| GPT-4-Turbo-Vision | 128K | €0.010 | €0.028 |
| GPT-4 | 8K | €0.028 | €0.056 |
| GPT-4 | 32K | €0.056 | €0.111 |

Joonis 3. Toonenit maksu- ja mahumäär [24].

| | accuracy and fine-grained interpretability | active learning | anomaly detection | causal inference | classification tasks | deep neural networks | electronic health records analysis | few-shot learning for anomaly detection | graph convolutional neural networks | knowledge graphs | ... | exploratory analysis of hierarchical datasets (2) | novel architecture (2) | reinforcement learning (2) | transfer learning (2) | anomaly detection (3) | anomaly detection with autoencoders on data streams (3) |
|---|--|-----------------|-------------------|------------------|----------------------|----------------------|------------------------------------|---|-------------------------------------|------------------|-----|---|------------------------|----------------------------|-----------------------|-----------------------|---|
| 0 | 346 | 1054 | 1012 | 1357 | 1018 | 435 | 1333 | 1289 | 1287 | 354 | ... | 366 | 1189 | 1074 | 1171 | 1117 | 1083 |
| 1 | 477 | 1177 | 578 | 190 | 1317 | 599 | 1340 | 1383 | 486 | 587 | ... | 821 | 1358 | 394 | 519 | 713 | 1288 |
| 2 | 707 | 1368 | 916 | 642 | 347 | J15 | 212 | 146 | 542 | 604 | ... | 972 | 526 | 459 | 612 | 802 | 1392 |
| 3 | 880 | 481 | 923 | J22 | 445 | J40 | 249 | 849 | 634 | 696 | ... | 988 | 672 | 538 | 671 | J18 | 518 |
| 4 | J14 | J28 | J25 | J41 | J3 | J42 | 586 | 891 | J31 | J10 | ... | J6 | 873 | 869 | 792 | J36 | 831 |

5 rows × 57 columns

Joonis 4. Loodud sessiooni jaotus, ChatGPT 3.5 versiooniga.

| | ai algorithms | ai for feature selection | active learning techniques | advanced time-series learning methods | adversarial learning | algorithm performance | anomaly detection | approaches to node classification | artificial intelligence | challenges in multi-agent learning | ... | graph machine learning (1) | machine learning (1) | self-supervised learning (1) | ai applications (2) | deep learning (2) | deep learning techniques (2) | graph neural networks (2) |
|---|---------------|--------------------------|----------------------------|---------------------------------------|----------------------|-----------------------|-------------------|-----------------------------------|-------------------------|------------------------------------|-----|----------------------------|----------------------|------------------------------|---------------------|-------------------|------------------------------|---------------------------|
| 0 | 1280 | 1038 | 1054 | 1298 | 548 | 526 | 1128 | 764 | 1210 | 1242 | ... | 1184 | 407 | 1000 | 1093 | 1333 | 1293 | 1022 |
| 1 | 531 | 1340 | 1177 | 755 | 795 | 543 | 1334 | 848 | 519 | 1297 | ... | 1383 | 643 | 216 | 586 | 245 | 191 | 1027 |
| 2 | 712 | 1367 | 1368 | 759 | 852 | 665 | 146 | 872 | 828 | 1379 | ... | 626 | 668 | 792 | 713 | 311 | 932 | 1052 |
| 3 | 904 | 408 | 481 | 953 | 926 | J20 | 158 | 973 | 928 | 869 | ... | 802 | J11 | 927 | 860 | 381 | 944 | 1256 |
| 4 | J42 | J49 | J28 | J3 | 952 | J30 | 578 | J10 | J9 | 883 | ... | J50 | J56 | J46 | J6 | J43 | J14 | 666 |

5 rows × 38 columns

Joonis 5. Loodud sessiooni jaotus tabel 2, ChatGPT 4 versiooniga.

Sellised tõendus põhised võrdlused aitavad mõista erinevate mudelite jõudlus omadusi ja on hädavajalikud optimaalse valiku tegemiseks.

3.6 Meetodide täpsuse analüüs

Selleks, et määrata kindlaks, kumb mudel — kasutades kas ChatGPT 4 või ChatGPT 3.5 teeki — pakub paremaid tulemusi sessiooni pealkirjade jaotamisel, oli vaja võrrelda loodud mudelite ja algsete käsitsi sessiooni jaotuste pealkirjade täpsuse sobivust. Hindamise protsessis osalesid masinõppe doktorandid, kes analüüsisid kahte erinevat aspekti. Esiteks

hinnati, kui hästi sobib iga mudeli genereeritud sessiooni pealkiri konkreetsele artiklile. Selleks paluti neil anda hinnang vahemikus 0 (ei sobi üldse) kuni 10 (sobib ideaalselt).

Teiseks vaadeldi sessiooni pealkirja üldist sobivust, kus valikud hõlmasid järgmisi kategooriaid: sobiv, peaaegu sobiv(kuid vajab lühendamist), liiga üldine, liiga spetsiifiline või muudel põhjustel sobimatu. Näiteks juhul, kui sessiooni pealkiri oli "Problems in semi-supervised learning", leiti, et kuigi pealkiri põhimõtteliselt sobib, ei anna sõnaühend "Problems in" olulist lisaväärtust. Seega oleks eelistatum ja asjakohasem pealkiri lühemalt "Semi-supervised learning". Sellised tagasiside dokumentatsioonid aitavad paremini mõista, millised pealkirjad on tõhusad ja kuidas neid edaspidi optimeerida, et tagada sessiooni pealkirjade suurem relevantsus.

Hindamise täpsuse tagamiseks arendasin koodi, mis valis juhuslikult 30 artiklit 258 artiklist koosnevast kogust. Juhuslikult leitud 30 artiklit võimaldas mul hinnata kokku 90 sessiooni pealkirja sobivust, et määrata kolme meetodi täpsust üksikasjalikult. Esialgse hinnangu andmisel kasutasin kõiki 90 sessiooni pealkirja, et saada ülevaade iga artikli ja selle vastava sessiooni pealkirja sobivusest. Teise hinnangu jaoks optimeerisin hindamisprotsessi, eemaldades korduvad sessiooni pealkirjad, mis viis hindamiseks jäänud unikaalsete sessiooni pealkirjade arvu 51-ni.

| | | | | |
|-------|--|------|--|---|
| Mudel | Palun hinnata, kas artikkel sobib selle pealkirjaga sessiooni, hinda 0st (üldse ei sobi) 10ni (sobib ideaalselt) | | Please rate whether the article suits into the session with the given title, scale from 0 (not suitable at all) to 10 (perfectly suitable) | |
| Score | Sessiooni pealkiri: | ID: | Artikli pealkiri: | Abstrakt: |
| 2 | model optimization | 1308 | Improved Regret Bounds for Online Kernel Selection under Bandit Feedback | <p>In this paper, we improve the regret bound for online kernel selection under bandit feedback. Previous algorithm enjoys a $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound for Lipschitz loss functions. We prove two types of regret bounds improving the previous bound. For smooth loss functions, we propose an algorithm with a $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound where $LT^{(H)}$ is the cumulative losses of optimal hypothesis in $H = \{f \in H: \frac{1}{\epsilon} \frac{1}{\sqrt{t}} \leq H(f)\}$. The data-dependent bound keeps the previous worst-case bound and is smaller if most of candidate kernels match well with the data. For Lipschitz loss functions, we propose an algorithm with a $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound asymptotically improving the previous bound. We apply the two algorithms to online kernel selection with time constraint and prove new regret bounds matching or improving the previous $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound where R is the time budget. Finally, we empirically verify our algorithms on online regression and classification tasks.</p> |
| 1 | time constraints | 1308 | Improved Regret Bounds for Online Kernel Selection under Bandit Feedback | <p>In this paper, we improve the regret bound for online kernel selection under bandit feedback. Previous algorithm enjoys a $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound for Lipschitz loss functions. We prove two types of regret bounds improving the previous bound. For smooth loss functions, we propose an algorithm with a $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound where $LT^{(H)}$ is the cumulative losses of optimal hypothesis in $H = \{f \in H: \frac{1}{\epsilon} \frac{1}{\sqrt{t}} \leq H(f)\}$. The data-dependent bound keeps the previous worst-case bound and is smaller if most of candidate kernels match well with the data. For Lipschitz loss functions, we propose an algorithm with a $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound asymptotically improving the previous bound. We apply the two algorithms to online kernel selection with time constraint and prove new regret bounds matching or improving the previous $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound where R is the time budget. Finally, we empirically verify our algorithms on online regression and classification tasks.</p> |
| 8 | bandits and online learning | 1308 | Improved Regret Bounds for Online Kernel Selection under Bandit Feedback | <p>In this paper, we improve the regret bound for online kernel selection under bandit feedback. Previous algorithm enjoys a $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound for Lipschitz loss functions. We prove two types of regret bounds improving the previous bound. For smooth loss functions, we propose an algorithm with a $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound where $LT^{(H)}$ is the cumulative losses of optimal hypothesis in $H = \{f \in H: \frac{1}{\epsilon} \frac{1}{\sqrt{t}} \leq H(f)\}$. The data-dependent bound keeps the previous worst-case bound and is smaller if most of candidate kernels match well with the data. For Lipschitz loss functions, we propose an algorithm with a $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound asymptotically improving the previous bound. We apply the two algorithms to online kernel selection with time constraint and prove new regret bounds matching or improving the previous $O(\frac{1}{\epsilon} \frac{1}{\sqrt{t}} \frac{1}{2H} + 1) K^{1/3} T^{2/3}$ expected bound where R is the time budget. Finally, we empirically verify our algorithms on online regression and classification tasks.</p> |

Joonis 6. Näide doktorantide hinnatud tabelist 1.

| | | |
|---|---|--|
| Example about a case that could be "almost suitable, but should be shortened": Session title: 'Problems in semi-supervised learning' - while in principle it is ok, the part 'Problems in' does not add any value, and a preferable title would be 'Semi-supervised learning'. | suitable almost suitable, but should be shortened not suitable, because too general not suitable, because too specific not suitable, for some other reason | If you respond 'not suitable, for some other reason', then please specify the reason in the comment. |
| Sessioni pealkiri: | suitability | comment |
| model optimization | suitable | |
| time constraints | not suitable, for some other reason | |
| bandits and online learning | suitable | |
| electronic health records analysis | suitable | |
| natural language processing and text-mining | not suitable, because too general | |
| model performance | not suitable, because too general | |

Joonis 6. Näide doktorantide hinnatud tabelist 2.

Täpsuse hindamiseks viin esimeses etapis läbi kahe peamise analüüsi: a) määran iga meetodi protsentuaalse keskmise täpsuse, kasutades kümne-palli süsteemi, mis võimaldab kvantifitseerida iga mudeli efektiivsust suhteliselt ühtsel skaalal; b) kui võtan algse käsitsi loodud sessioonide jaotuse punktist a saadud täpsuse ja oletan, et see tulemus on ideaalseks tulemuseks, siis saan võrrelda selle väärtuse suhet teiste meetoditega, et teha kindlaks, kui tõhusad on kaks loodud meetodit.

Teises hindamisetapis keskendun üldisele sobivusele, analüüsides, kui hästi sessiooni pealkirjad valikusse sobivad. Analüüs hõlmab sessiooni pealkirjade sobivuse hindamist erinevate kategooriate lõikes — sobiv, peaaegu sobiv (kuid vajab lühendamist), liiga üldine, liiga spetsiifiline või muudel põhjustel sobimatu. Hindamise käigus leian suhted ja kvantiteedid, mis näitavad, kui palju igas kategooriasse ühekordseid erilisi ja kõikide artiklite sessiooni pealkirju kuulus.

Hindamise tulemused:

Esimese hinnangu põhjal jagunes meetodide protsentuaalne keskmine täpsus järgnevalt:

- Juba loodud originaalse teaduskonverentsi sessioonide hinnanguline täpsus kümne-palli süsteemis on 62%;
- Meetod mis kasutab teeki ChatGPT 3.5 versiooni abi hinnanguline täpsus kümne-palli süsteemis on 45.7%;
- Meetod mis kasutab teeki ChatGPT 4 versiooni hinnanguline täpsus kümne-palli süsteemis on 58.7%.

Kui nüüd võtta arvesse, et originaalne sessioonide jaotus on parim lahendus siis ülejäänud mudelite tulemus on järgnev:

- Meetod mis kasutab teeki ChatGPT 3.5 versiooni täpsus on 73.7%;
- Meetod mis kasutab teeki ChatGPT 4 versiooni täpsus on 94.7%.

| Sessiooni pealkirja sobivus: | Esines kokku unikaalseid sessiooni pealkirju: | Käsitsi loodud sessiooni pealkirjade osakaal artiklitest: | Meetodi ChatGPT 3.5 sessiooni pealkirjade osakaal artiklitest: | Meetodi ChatGPT 4 sessiooni pealkirjade osakaal artiklitest: |
|---|---|---|--|--|
| sobivad | 25 | 15/24 | 16/30 | 13/30 |
| peaaegu sobivad, kuid seda tuleks lühendada | 0 | 0/24 | 0/30 | 0/30 |
| sobimatu, sest liiga üldine | 12 | 6/24 | 4/30 | 15/30 |
| sobimatu, sest liiga spetsiifiline | 7 | 0/24 | 7/30 | 0/30 |
| sobimatu, sest mingil muul | 6 | 3/24 | 3/30 | 2/30 |

| | | | | |
|----------|--|--|--|--|
| põhjused | | | | |
|----------|--|--|--|--|

Tabel 7 Sessiooni pealkirjade sobivus üldiselt ehk teine uurimus tulemus.

Esimeses hinnangus selgus, et doktorandid hindasid originaalseid sessioonide jaotusi madalaks, jäädes vahemikku veidi üle poole kuni alla kolmveerandi protsendi. Uuringu andmed näitavad, et käsitsi loodud jaotused ning ChatGPT 4 abil loodud jaotused on suhteliselt võrdsed, erinevused nende täpsuses jäävad vaid umbes 3% piiresse. Hoolimata sellest, et ChatGPT 3.5 abil loodud meetodi tulemused jäid kõige nõrgemaks, mis võib olla põhjustatud vähemate treening andmete kasutamisest mudeli treenimisel, ei jää see meetod siiski oluliselt maha, saavutades peaaegu kolmveerandi täpsuse.

Arvestades, et käsitsi loodud sessioonide jaotus peetakse täielikult tõseks, näitavad tulemused, et ChatGPT 4 abil loodud jaotused on ideaalsele tulemusele üllatavalt lähedal, mis tõestab loodud meetodi edukust täpse sessioonide jaotamise lahenduse loomisel. Samuti, kuigi ChatGPT 3.5 meetod genereeris kõige rohkem liiga spetsiifilisi pealkirju, ei jää see täpsuselt palju maha, hoolimata sellest, et ChatGPT 4 mudel kipub genereerima liiga üldiseid pealkirju, mis võib järelpärimisel veelgi süveneda.

Teises hinnangus ilmnes hindamisprotsessi nõrkus, kuna ei esinenud ühtegi sessiooni pealkirja, mis olid "peaaegu sobiv". Põhjus võib olla kuna doktorandid võivad kalduda valima kategooriaid "sobiv" või "sobimatu", vältides keerulisemaid otsuseid, nagu "peaaegu sobiv". Hinnangus oli sessiooni pealkiri "Application" sattunud kaks korda, mis tõi tulemuse sisse vastuolulise hinnangu: ühel juhul peeti seda liiga üldiseks ja teisel juhul sobivaks. Seega "Application" sessiooni pealkiri on välja jäetud lõplikust arvestusest, kuna tulemus on vastuoluline. Tabelist 7 on näha, et kõige enam sobivaid pealkirju pakkus välja teeki ChatGPT 3.5 meetodi abil loodud sessiooni pealkirju, mis aga oli ainuke meetod, mis genereeris liiga spetsiifilisi pealkirju. Samas kui meetod, mis kasutas ChatGPT 4 teeki, tootis kõige rohkem liiga üldisi pealkirju, mis võis tulla sellest, et kui teha päringut ChatGPT 4 -le, siis on vaja veel täpsemalt sõnastada prompt.

3.7 Meetodi potentsiaal tulevikuks

Meetodit saab muuta efektiivsemaks, kui pärast juturoboti genereeritud pealkirjasid, hindab erialaspetsialist need üle. Spetsialist eemaldab üleliigsed või sobimatud pealkirjad ning vajadusel lisab uusi paremaid. Inimese poolt läbi viidav optimeerimisprotsess vähendab hindamiseks vajalike pealkirjade arvu, muutes mudeli jooksutamise kuluefektiivsemaks.

Tulevikus peitub loodud mudeli väärtus selle võimes vähendada ajagraafikute loomise kulusid mitme konverentsi jaoks. Kasutades seda meetodit, on võimalik märkimisväärselt kokku hoida rahalisi vahendeid, kuna kulutused piirduvad ainult mudeli käitamisega. Samuti väheneb vajadus kulutada töötunde, mis tavaliselt kuluksid manuaalsele artiklite sessioonideks jagamiseks. Meetodi kasutamise efektiivsus ei säästa ainult ressursse, vaid suurendab korraldusprotsessi kiirust ja täpsust, muutes konverentside korraldamise oluliselt lihtsamaks ja ökonoomsemaks.

3.8 Meetodi kood

Kogu kood on lisades:

Lisa 1 meetod, mis kasutab ChatGPT 3.5 teeki: Konverentsi sessioonide jaotamine-3.5.ipynb

Lisa 2 meetod, mis kasutab ChatGPT 4 teeki: Konverentsi sessioonide jaotamine-4.ipynb

Lisa 3 sisend andmed: headers.xlsx

4. Loodud mudeli kasulikkus ECML PKDD teaduskonverentsile

ChatGPT 4 loodud mudel on osutunud tõhusaks just selle konkreetse konverentsi raames, kuna konverentsi korraldaja on nõustunud eelnevalt hindama ja kinnitama sobivad pealkirjad. Tänu selle eeltööle on meetod suuteline märkimisväärselt aega kokku hoidma, kuna sessioonide jaotamine on juba teostatud. Samal ajal kui mudel töötab ja päringuid teostab, saab korraldaja tegeleda teiste ülesannetega, mis aitab kokku hoida väärtuslikku aega. Lõplik ja põhjalik kvaliteedi hindamine seisab professionaalil endiselt veel ees. Mistõttu on tõenäoline, et ekspert peab sessioonide jaotuse üle vaatama ja tegema vajalikke parandusi, et tagada optimeerimise protsessi tõhusus ja täpsus. Ajagraafiku loomiseks on nüüd vajalik vaid määrata sessioonidele sobivad ruumid ja ajad. Aasta 2022 andmete põhjal arvutatud meetodi jooksumise maksumus on 80 eurot, samas kui manuaalse ajagraafiku koostamiseks kulub kahe ööpäeva jagu töötunde. Kulu/aja määr näitab, et automaatne mudel mitte ainult ei säästa aega, vaid on ka kuluefektiivne, muutes konverentsi korraldamise protsessi ökonoomsemaks ja sujuvamaks.

Kokkuvõte

Käesolev uurimus keskendub kahe meetodi väljatöötamisele, mis erinevad peamiselt kasutatud keelemudelite versioonides, hõlmates nii ChatGPT 3.5 Turbo kui ka ChatGPT 4 mudelit. Meetod kasutab Azure OpenAI teeki suhtlemiseks juturobotiga, mis ennustab potentsiaalseid sessiooni pealkirju vastavalt artikli sisule ja hindab artikli sessiooni kuuluvuse sobivust. Oluline on formuleerida küsimus juturobotile nii, et see genereeriks üldiseid sessiooni pealkirju. Testides erinevaid prompte selgus, et ilma range piiranguta pealkirjade üldisusele võib juturobot kõrvale hiilida ja ennustada väga spetsiifilisi pealkirju. Lisaks sellele tehakse juturobotile järelpäring, et muuta sessiooni pealkirjad üldisemaks. Järelpäring on kasulik, sest mudel võib genereerida pealkirja, mis on üldine, kuid mittepiisav konkreetse konverentsi kontekstis. Vajadusel tehakse täiendavaid päringuid, et vähendada sessiooni pealkirjade arvu, tagades et need on piisavalt üldised ja asjakohased. Tööriistana kasutatakse SPIC mudelit artiklite jagamiseks parimateks sessioonideks, mis aitab mõista andmekogumite struktuuri ja tuvastada olulisi mustreid. Viimaks hinnatakse sessioonide pealkirjade sobivust Levenshteini kauguse abil, et tagada nende sarnasus originaalidega, ning mitme eksperdi hinnangul, mis kindlustab pealkirjade asjakohasuse vastavalt kümne palli süsteemile.

Loodud meetod on väärtuslik abivahend ajagraafikute loomisel, sest see võimaldab aega kokku hoida tänu eelnevalt tehtud sessioonide jaotusele. Mudeli efektiivsust suurendab oluliselt inimese sekkumine, mis seisneb vahepealsetes hindamistes, et kinnitada sessioonide pealkirjade asjakohasust. Oluline on keelemudeli väljundit kontrollida, sest muidu juhtub, et mudel annab väljundiks poolikuid või puudulike tulemusi. Lisaks on võimalik meetodi jooksutamise mahtu vähendada, tehes keelemudelile lisapäringuid, mis aitavad hinnata, kas sessiooni pealkiri sobib spetsiifiliselt masinõppe konverentsile. Seega toob loodud meetodi kasutamine kaasa ajalise ja rahalise kokkuhoiu, suurendades konverentsi korraldamise tõhusust.

Viidatud kirjandus

- [1]OpenAI (2022). ChatGPT 4: <https://chat.openai.com>.
- [2]ecmlpkdd, 2022. ecmlpkdd.org/2022/. (04.02.2024)
- [3]Sõnaveeb: EKI ühendsõnastik, 2024. [Otsing \(sonaveeb.ee\)](https://otsing.sonaveeb.ee).
- [4]OpenAI Platform s.a. <https://platform.openai.com/docs/guides/prompt-engineering>. (28.04.2024)
- [5]Tesch A. .Optimization of Large-Scale Conference Schedules. Berliini tehnikaülikool matemaatika osakonna Bakalaaurusetöö. 2013. <https://www.semanticscholar.org/paper/Optimization-of-Large-Scale-Conference-Schedules-Tesch/c86cd10fdceee48d03198d9243af38971cd6fb41>. (04.02.2024)
- [6] Fabián Castaño, Jimmy Carvajal, Nubia Velasco - Content-Based Conference Scheduling Optimization. *IEEE Journals & Magazine | IEEE Xplore*. 2019. <https://ieeexplore.ieee.org/abstract/document/8891884>. (04.02.2024)
- [7]H. A. Eiselt ja C. -l. Sandblom - Heuristic Algorithms. *Springer eBooks*, Jan. 01, 2000. https://link.springer.com/chapter/10.1007/978-3-662-04197-0_11. (06.02.2024)
- [8]Vangerven B, Ficker A, Goossens D, Passchyn W, Spieksma F, Woeginger GJ. - Conference scheduling - A personalized approach. *Omega*. 2018, lk 38–47. [Conference scheduling — A personalized approach - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0167636918300000). (06.02.2024)
- [9]Thompson GM. - Improving conferences through session scheduling. *Cornell Hotel and Restaurant Administration Quarterly*, 2002, lk 71–76. [Improving Conferences Through Session Scheduling - Gary M. Thompson, 2002 \(sagepub.com\)](https://www.sagepub.com/journalsPermissions.nav?path=/journals/cornell-hotel-and-restaurant-administration-quarterly/vol22-no4/200212). (10.02.2024)
- [10]Sõnaveeb: EKI ühendsõnastik, 2024. [Otsing \(sonaveeb.ee\)](https://otsing.sonaveeb.ee).
- [11]Z. Wang and M. Noda. Alarm Data Analysis for Safe Plant Operations: Case Study of Ethylene Plant. *Computer-aided chemical engineering*, Jan. 01, 2018. <https://www.sciencedirect.com/science/article/abs/pii/B9780444642417503803>. (06.02.2024)
- [12]Tc T. All you need to know about Microsoft and OpenAI’s collaboration so far. *Techcircle*. 2023.

<https://www.techcircle.in/2023/01/09/how-has-the-investment-in-openai-benefited-microsoft#:~:text=Azure%20OpenAI%20service%20is%20a%20fully>. (28.04.2024)

[13] SCIP. s.a. <https://scipopt.org/#scipoptsuite>. (28.04.2024)

[14] SCIP Doxygen Documentation: First Steps Walkthrough. s.a. <https://www.scipopt.org/doc/html/GETTINGSTARTED.php>. (28.04.2024)

[15] Maher SJ, Miltenberger M, Pedroso JP, Rehfeldt D, Schwarz R, Serrano F. PySCIPOpt: Mathematical Programming in Python with the SCIP Optimization Suite. Loengukonspektid arvutiteaduses, 2016, lk 301–307. [PySCIPOpt: Mathematical Programming in Python with the SCIP Optimization Suite | SpringerLink](#).

[16] The Leader in Decision Intelligence Technology - Gurobi Optimization. Gurobi Optimization 2024. [Gurobi Optimizer Release Notes v11.0.1 - Gurobi Optimization](#). (28.04.2024)

[17] Lin P, Cai S, Zou M, Lin J. Local Search for Integer Linear Programming. 2023. <https://www.semanticscholar.org/paper/Local-Search-for-Integer-Linear-Programming-Lin-Cai/2135ecb0eb61d8b76cac030f486ee3b5916ef97a>.

[18] OpenAI Platform s.a. [Session Assignment Optimization Methods \(openai.com\)](#).

[19] Integer linear programming - introduction. Loengukonspektid arvutiteaduses, Colorado ülikool. 2013. <https://home.cs.colorado.edu/~srirams/courses/csci5654-fall13/ilpLectures.pdf>.

[20] JSON s.a. <https://www.json.org/json-en.html>. (28.04.2024)

[21] Getting started — pandas 2.2.2 documentation s.a. https://pandas.pydata.org/docs/getting_started/index.html#getting-started. (28.04.2024)

[22] NumPy - s.a. <https://numpy.org/>. (28.04.2024)

[23] collections — Container datatypes. Python Documentation s.a. <https://docs.python.org/3/library/collections.html>. (28.04.2024)

[24] OpenAI Platform s.a. [Azure OpenAI Service - Pricing | Microsoft Azure](#).

[25] Sõnaveeb: Semiootika terminibaas. s.a [Otsing \(sonaveeb.ee\)](#).

I. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, _____ Mia Marta Heikla _____,

(autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

_____ Ettekannete automaatne sessioonideks jagamine teaduskonverentside
jaoks _____,

(lõputöö pealkiri)

mille juhendaja on _____ Meelis Kull _____,

(juhendaja nimi)

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commons'i litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Mia Marta Heikla

15.05.2024