Vilniaus universitetas

Ataskaitinė informatikos krypties doktorantų konferencija 2023-03-21

Rolandas Gricius (VU DMSTI doktorantas, Išmaniųjų technologijų tyrimų grupė)

Preliminari darbo tema.

Turinio atpažinimas suskaitmenintuose struktūrizuotuose dokumentuose.

Recognising the contents in digitised structured documents.

Darbo vadovas.

Prof. dr. Igoris Belovas.

Doktorantūros studijų laikotarpis.

2021 m. spalio mėn. 1 d. – 2025 m. rugsėjo mėn. 30 d.

Ataskaitinis laikotarpis.

2022 m. spalio mėn. 1 d. – 2023 m. kovo mėn. 30 d.

Visų studijų planas ir jo vykdymo suvestinė

Studijų metai	Egzaminai		Dalyvavimas konferencijose		Publikacijos		
	Planas	**Įvykdyta**	**Planas**	**Įvykdyta**	**Planas**	**Įvykdyta**	**Būklė**
I (2021/2022) Pirmas pusmetis	1	1		1 (L)			
I (2021/2022) Antras pusmetis	1	**2**					
II (2022/2023) Pirmas pusmetis	1	1		**1 (T)**			
II (2022/2023) Antras pusmetis	1		1 (T)				
III (2023/2024) Pirmas pusmetis							
III (2023/2024) Antras pusmetis					1 (CA WoS)		
IV (2024/2025) Pirmas pusmetis							
IV (2024/2025) Antras pusmetis			1 (T)		1 (CA WoS)		

Vilniaus universitetas

Ataskaitinių metų darbo planas ir jo vykdymo suvestinė

Egzaminai		Dalyvavimas konferencijose		Publikacijos	
Planas	**Įvykdyta**	**Planas**	**Įvykdyta**	**Planas**	**Įvykdyta**
Fundamentalieji informatikosir informatikos inžinerijos metodai	Išlaikyta: Fundamentalieji informatikosir informatikos inžinerijos metodai, 2023-01-24, pažymys 8.	1(T)	Pranešimas tarptautinėje konferencijoje „International Conference on Pattern Recognition Applications and Methods (ICPRAM) 2023“, Lisabona, Portugalija, 2023-02-22 – 24 d.	-	-

Vilniaus universitetas

Visų mokslinių tyrimų ir disertacijos rengimo etapai

	Darbo pavadinimas	Atlikimo terminai	Pastabos
1.	**Mokslinių tyrimų disertacijos tema apžvalga ir analizė:** 1.1. Analitinės apžvalgos atlikimas. 1.2. Disertacijos tyrimo objekto detalizavimas.	2021 m. spalio mėn. – 2022 m. kovo mėn.	
	1.3. Analitinės apžvalgos užbaigimas. 1.4. Mokslinių problemų susietų su tyrimo objektu identifikavimas ir tyrimo tikslo suformavimas.	2022 m. balandžio mėn. – 2022 m. rugsėjo mėn.	
2.	Mokslinio tyrimo vykdymas:		
	2.1. Tyrimo metodikos sudarymas	2022 m. spalio mėn. – 2023 m. kovo mėn.	
	2.2. Teorinis tyrimas	2023 m. balandžio mėn. – 2023 m. rugsėjo mėn.	
	2.3. Empirinis tyrimas	2023 m. spalio mėn. – 2024 m. rugsėjo mėn.	
	2.4. Gautų rezultatų analizė ir apibendrinimas	2024 m. spalio mėn. – 2025 m. kovo mėn.	

Tyrimo objektas, tikslas ir uždaviniai

- **Tyrimo objektas – tekstas sąskaitose-faktūrose (angl. invoices), gautas tiesiogiai arba po OCR procedūros**
- **Tikslas – naudojant natūralios kalbos apdorojimo metodus, atpažinti ir ištraukti tolesniam apdorojimui sąskaitos duomenis, reikšmingus:**
  - **teisėtumui – privalomus pagal teisės aktus duomenis**
  - **apskaitai – data, pirkėjo ir pardavėjo duomenys, sandorio ir mokesčių sumos**
  - **sandorio vykdymui – pristatymo duomenys, apmokėjimo detalės**
- **Uždaviniai – sudaryti duomenų rinkinį tyrimui, atlikti teorinį tyrimą identifikuojant metodus, empirinį tyrimą palyginant jų veikimą ir modifikuoti pritaikant Lietuvos specifikai ir surinktiems duomenims**

Duomenų rinkinys tyrimui

- Esami sąskaitų rinkiniai nedideli, ne visuomet anotuoti, todėl netinka giliam mokymui
- Viešai prieinamų lietuviškų duomenų rinkinių iš viso nėra
- Dauguma tyrimų naudoja neviešinamus duomenų rinkinius, todėl rezultatus sunku palyginti
- Priimtas sprendimas duomenis tyrimui generuoti

Duomenų generavimas

- Atlikta mokslinės literatūros analizė dokumentų (sąskaitų) generavimo tema
- Atrinkta publikuota programinė įranga skirta sąskaitų generavimui
- Identifikuoti trūkumai (tik anglų ir prancūzų kalbos, nepakankamai tikroviškai generuojami adresai, įmonių pavadinimai)
- Atliktas bandymas (proof of concept) modifikuoti programinę įrangą lietuviškų sąskaitų generavimui

Duomenų generavimas

- **Rezultatai pristatyti konferencijoje „*12th International Conference on Pattern Recognition Applications and Methods*“**

Trumpas per pusmetį gautų mokslinių rezultatų pristatymas

- **Suformuluoti tolesni tyrimo uždaviniai – sukaupti duomenų rinkinį, atlikti teorinį ir empirinį tyrimus, modifikuoti palygintus algoritmus pritaikant Lietuvos specifikai ir duomenų rinkiniui**
- **Duomenų (sąskaitų) generavimo sprendimas 2023-02-23 pristatytas pranešime tarptautinėje konferencijoje ICPRAM(12th International Conference on Pattern Recognition Applications and Methods). Pranešimo pavadinimas – *Generation of Synthetic Invoices for the Training of Machine Learning Models***
- **Nuolat pildoma svarbiausių publikacijų preliminaria disertacijos tematika bazė. Straipsniai yra rūšiuojami, atliekama jų analitinė apžvalga**

Kito pusmečio darbo planas.

1. Metodų naudojamų turinio atpažinimo suskaitmenintuose struktūrizuotuose dokumentuose tyrimas
2. Metodų naudojamų turinio atpažinimo suskaitmenintuose struktūrizuotuose dokumentuose adaptavimas ir modifikavimas, atsižvelgiant į Lietuvos specifiką
3. Turinio atpažinimo suskaitmenintuose struktūrizuotuose dokumentuose algoritmų tyrimas
4. Mokslinių tyrimų disertacijos tema analitinės apžvalgos pildymas naujai atsirandančiais straipsniais
5. Pradėti rengti publikaciją apie sąskaitų generavimo sprendimą Web of Science reitinguojamame leidinyje

Vilniaus
universitetas

Ačiū už dėmesį

Rolandas Gricius
VU DMSTI doktorantas
rolandas.gricius@mif.stud.vu.lt