Vilniaus universitetas

Ataskaitinė informatikos krypties doktorantų konferencija 2022-09-30

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

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. kovo mėn. 25 d. – 2022 m. rugsėjo 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)			
II (2021/2022) Antras pusmetis	1	2					
II (2022/2023) Pirmas pusmetis	1						
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**
Informatikos ir informatikos inžinerijos tyrimo metodai ir metodika	Išlaikyta: Informatikos ir informatikos inžinerijos tyrimo metodai ir metodika, 2022-06-28, pažymys 7. Natūralios kalbos apdorojimas, 2022-09-26, pažymys 8.	-	-	-	-

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.	Užbaigta mokslinės literatūros apžvalga. Suformuluotas tyrimo tikslas.
	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.	

Visų mokslinių tyrimų ir disertacijos rengimo etapai

	Darbo pavadinimas	Atlikimo terminai	Pastabos
3.	Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų ir kt.) parengimas: 3.1. Tikslų, uždavinių, tyrimo metodikos, ginamųjų teiginių patikslinimas. 3.2. Analitinės disertacijos dalies parengimas. 3.3. Teorinės disertacijos dalies parengimas. 3.4. Eksperimentinės disertacijos dalies parengimas. 3.5. Bendrųjų išvadų formulavimas.	2025 m. balandžio mėn. – 2025 m. rugsėjo mėn.	
4.	Daktaro disertacijos parengimas ir svarstymas padalinyje	2025 m. birželio mėn.	
5.	Daktaro disertacijos gynimas	2025 m. rugsėjo mėn.	

Tyrimo objektas, tikslas ir uždaviniai

„Turinio atpažinimas suskaitmenintuose struktūrizuotuose dokumentuose“

- **Dokumentai**
- **Struktūrizuoti dokumentai**
- **Suskaitmeninti struktūrizuoti dokumentai**
- **Turinio atpažinimas**

Dokumentai

- Dokumentai neberašomi ranka
- Spausdinami arba siunčiami el. paštu
- Dokumento duomenys suvedami į gavėjo sistemas

Automatinis apsikeitimas duomenis – ar tai sprendimas?

- ~1970 – EDI
- ~2000 – cXML, ebXML, … XML
- Šiuo metu vėl bandomi kurti nauji standartai – PDF/A-3, ZUGFeRD, The European standard on eInvoicing (EN 16931)

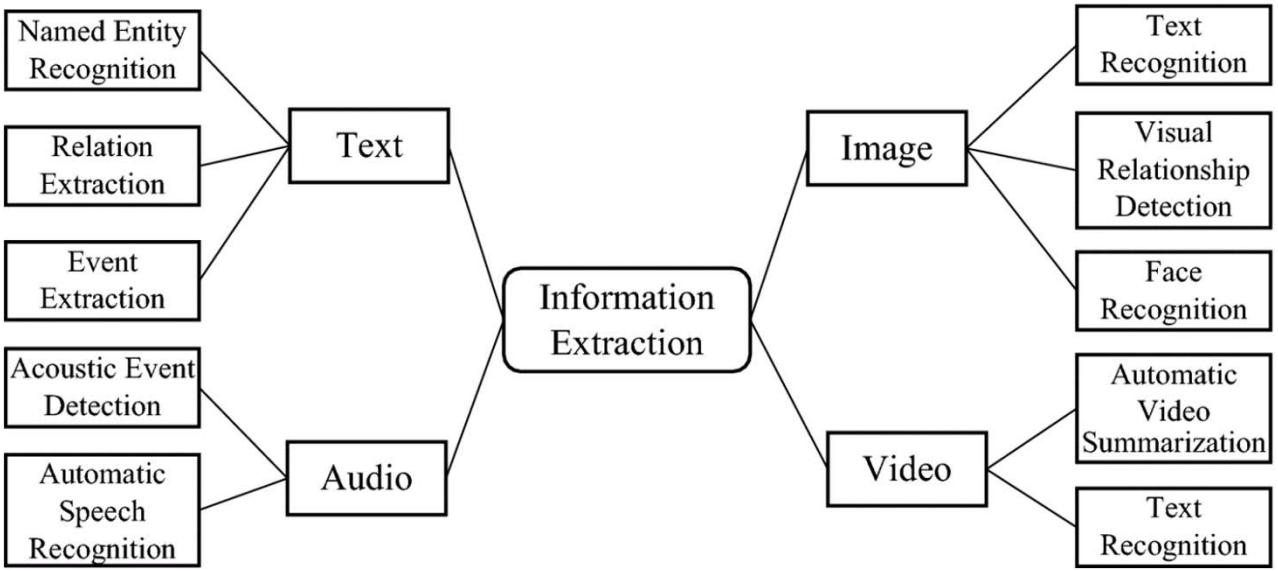
Struktūrizuoti dokumentai

- Griežtai struktūrizuoti
  - Pasas, tapatybės kortelė, CMR važtaraštis
- Iš dalies struktūrizuoti
  - Sąskaita-faktūra, kasos čekis
- Laisvos formos
  - Sutartys, Curriculum Vitae

Suskaitmeninti struktūrizuoti dokumentai

- El. paštu gaunami iš karto skaitmeniniu pavidalu
- Skenuojami, jei gauti popieriniu pavidalu
- Teksto atpažinimas (OCR), jei skenuoti
- Galiausiai turime dokumento tekstą +
  - Koordinates (ne visada)
  - Paveiksliukus (ne visada)

Turinio atpažinimas



Informacijos išgavimas

Adnan, K., Akbar, R. An analytical study of information extraction from unstructured and multidimensional big data. *J Big Data* 6, 91 (2019). https://doi.org/10.1186/s40537-019-0254-8

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**

- **Uždavinių, skirtų tyrimo tikslui pasiekti suformulavimas**
- **Kitų studijų metų rudens pusmetis**

Trumpas per pusmetį gautų mokslinių rezultatų pristatymas

- Detalizuotas disertacijos tyrimo objektas – tekstiniai dokumentai – sąskaitos
- Suformuluotas tyrimo tikslas - atpažinti ir ištraukti tolesniam apdorojimui sąskaitos duomenis, reikšmingus dokumento teisėtumui, apskaitai, sandorio vykdymui
- Atrinkta programinė įranga, generuojanti empirinius duomenis – sąskaitas, atlikti kodo pakeitimai jos generuojamų sąskaitų sulietuvinimui
- Atliktas pirminis tyrimas įvardintų esybių atpažinimo bazinei kokybei įvertinti naudojant standartinę Python natūralios kalbos apdorojimo biblioteką spaCy
- Sudaryta ir nuolat pildoma svarbiausių publikacijų preliminaria disertacijos tematika bazė. Straipsniai yra rūšiuojami, atliekama jų analitinė apžvalga

Kito pusmečio darbo planas.

1. Uždavinių, skirtų tyrimo tikslui pasiekti, suformulavimas
2. Tyrimo metodikos išsikeltiems uždaviniams spręsti parinkimas
3. Teorinio ir empirinio tyrimų suplanavimas pagal pasirinktą metodiką
4. Mokslinių tyrimų disertacijos tema analitinės apžvalgos pildymas naujai atsirandančiais straipsniais
5. Išlaikyti privalomojo dalyko „Fundamentalieji informatikos ir informatikos inžinerijos metodai“ egzaminą

Vilniaus
universitetas

Ačiū už dėmesį

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