Remarks

3: cytowanie

6: Wstęp jest za długi. Wstęp ma mieć 2-3 strony. Wstęp i zakończenie piszemy na końcu, gdy cała reszta jest już napisana.

10: state of the art dla aplikacji sieciowych

11: Dlaczego aplikacja webowa?

III: brak autorów

Silesian
University
of Technology

SILESIAN UNIVERSITY OF TECHNOLOGY

FACULTY OF AUTOMATIC CONTROL, ELECTRONICS AND COMPUTER SCIENCE

PROGRAMME: INFORMATICS

Final Project

Title of engineer thesis

author: Name Surname

supervisor: Name Surname, DSc PhD

Gliwice, January 2022

Contents

Abstract

The text of the abstract should be copied into a respective field in the APD system. The Abstract with keywords should not exceed one page.

Keywords: 2-5 keywords, separated by commas

Chapter 1

Introduction

Technological advancement being observed in the past years fundamentally changed the relevance of data in today's digitalized world. Information became an innovation stimulus in the area of research and development.

The quantity of data that organizations produce, process, store and share is at a continuous growth. An enormous amount of 1.8 zettabytes ($1.8 \cdot 10^{21}$ bytes) of new data was produced only in the 2011, and every two consecutive years this number was doubling [19]. The ever increasing volume of data being collected globally is further excited by improvements to the overall population's accessibility to the digital technology. Cisco Systems estimates within its annual report [cytowanie] that 66 percent of the world population will have an access to the web by the 2023, compared to 51 percent in 2018, whereas the number of devices that are connected to the web will reach a staggering value of three times as many as the entire population size – demonstrating a total of 60 percent expansion when compared to 2018. Even the area of mobile connection, which was established long ago, is still sustaining a growth – by 2023, a mobile connectivity will be a privilege for 70 percent of the world's population, compared to 66 percent in 2018. The global average mobile network speeds will be tripled through a rapid increase from 13.2 Mbps in 2018 to 43.9 Mbps in 2023 [3].

After decades of observed technological advancement and innovation, the global internet traffic finally entered the zettabyte-era, as it had reached a magnitude of one zettabyte in 2016, and in the calender month being as early as September [4].

The vast quantities of processed information allowed for brand new research fields to form, e.g. big data analytics, which is also considered as a field that will entirely alter all parts of civilization such as businesses or society [12]. The applications of big data analytics and its impact can be observed already today in a wide spectrum of domains.

Organizations collect and process the information to derive new insights and enhance the services they provide to the customers through statistical analysis or newly developed computer science processes such as data mining. The utility of delivered services is increased at a lower cost and improved efficiency through the insights derived from the collected information about how the service is consumed [1].

Information that was deidentified may flow from system to a system. Various types of organizations collect data to take advantage from the insights derived out of the data. Leading technology companies, such as Google and Facebook to name a few, sell anonymised collected user data access to their partner advertisers [12]. Large Hadron Collider (LHC) located in CERN, being the largest physical experiment, annually produce approximately 30 petabytes of data. LHC takes an advantage of light sensors that monitor the collisions of hundreds of millions of particles accelerated nearly to the speed of light. The collisions create enormous amount of data to be processed by computer algorithms in the hope of discovering new particles, e.g. a Nobel prize awarding discovery of Higgs boson had taken place in 2012 [2].

Enterprise stores such as Amazon or SAP Commerce Cloud collect the information regarding the way of how the visitors browse and interact with these stores. Collected information may involve the behavioral data related to customer engagement, such as the pages we visit, event clicks, or the way we scroll the page. The insights derived from the collected information enable making future improvements of these services - for example by improving the digital marketing or performance improvements based on the metrics [11]. The customer experience is also improved as based on the collected data the advertisements or item recommendations can be taylored to the specific user's preferences. The recommendation engine may also attempt to match your profile data to people of similiar profile to provide better recommendations. Services attempt to analyze the behavioural patterns such as

time of day we browse the store or what circumstances caused our last visit to finish. Even the details such as the exact neighbourhood location we live in, combined with its estimated wealth, organizations may attempt to guess our potential income level [12]. These data analytics are performed to increase the change of buying yet another item.

Services of enterprise organizations are running on all the collected data which can take various forms such as database entries, metrics, logs or outgoing messages. New data streaming technologies working at a large scale needed to be engineered to handle the continuous flow of data between systems and databases. An example of such technology includes Apache Kafka which generates even more than a trillion of messages per day for individual large enterprise organizations taking advantage of it [10][13].

Następne do zrobienia, w uproszczeniu i w podanej kolejności:

- Jak widać, dane są używane wszędzie. Przeciętna osoba sobie nawet nie zdaje sprawy.
- Istota prywatności danych, ochrona danych osobowych.
- Prawo wolno reaguje na uregulowanie ochrony danych osobowych: GDPR dopiero w 2018. Prawo różnie działa w różnych krajach.
- W czasach narastających danych, dane muszą być zanominizowane, bo...
- ...ale:

It is not possible to predetermine – in a general fashion and without an extra context – the data masking techniques that need to be combined together in order to achieve an actual anonymisation of the previously identifiable information. Instead, a context consisting of:

- exact representation form of the data being processed *(What is the data?)*
- data processors who use the data *(Who will use the data?)*
- processing purpose, e.g. research objective *(How will the data be used?)*

is always required when considering an effective way to anonymise data under that context [9].

Wniosek: generyczny sposób anonimizacji - nie ma takiego.

- Mogą istnieć dobre algorytmy anonimizacji, ale są optymalale jedynie w określonych kontekstach
- **Cel pracy**: stworzenie generycznego rozwiązania do anonimizacji baz danych w dowolnej formie danych (nieważne co dane reprezentują).
- Kontroler danych zna swoje dane i sam dostosowuje optymalny sposób anonimizacji.
- Rozwiązanie może być dostarczone jako free-access (research), B2C lub nawet B2B.
- Zakres pracy, opis rozdziałów - na samym końcu, po zrobieniu innych rozdziałów.

[Wstęp jest za długi. Wstęp ma mieć 2-3 strony. Wstęp i zakończenie piszemy na końcu, gdy cała reszta jest już napisana.]

- introduction into the problem domain
- settling of the problem in the domain
- objective of the thesis
- scope of the thesis
- short description of chapters
- clear description of contribution of the thesis's author – in case of more authors table with enumeration of contribution of authors

Chapter 2

Data anonymisation

Personal data is... Anonymisation is... what, why.

2.1 Data protection

Data protection: what is, why, goals. Other definitions: data subject, data controller, data processor.

2.1.1 Consent

Majority of data protection regulations are based on a consent of an individual, i.e. information can be used for secondary purposes only if an individual explicitly acknowledge their consent for that. This may sound easier said than done due to the unobvious difficulties data controllers would be facing when obtainining such a broad authorization consent for all possible secondary purpose usages. Consider a patient entering medical facility for an ordinary appointment. The patient could find it unusual, disturbing or even shocking if upon his entrance to the facility he was to receive an overwhelming form including dozens of independent consent authorization requests. The consents could give the impression of being seemingly unrelated to his visit in the first place, e.g. a constent to share the data with researchers of an university located on another continent. In the end that could destroy the data subject's trust – in this case patient's trust.

This theoretical scenario may not easily be implemented in the real world counterpart, as it could be even impossible to know or predict all possible secondary purpose usages in the first place, and consent based authorization is all about knowing the usages.

Consider a newly discovered purpose to process personal information of an already existing database. Getting a consent after the data had already been collected, i.e. backwards in time, would be impossible to accomplish as the data controller would need to contact potentially hundreds of thousands of people for their explicit consent. New purpose can be discovered years after the data collection.

Having that in mind, no consent is required when processing the data that is already anonymised. A data that was stripped from personal identifying or identifiable information data can be used in any way and can be shared with third parties without previously agreed consent. Data controllers now face a realistic to solve problem of information anonymisation rather than an unrealistic problem of consents collection [9].

2.1.2 General Data Protection Regulation

Protection of personal data is a fundamental right and freedom of people as defined under the Recital 1 of General Data Protection Regulation (GDPR) and the Charter of Fundamental Rights of the European Union [6][5]. Newly discovered challenges for the protection of personal data arise due to the ongoing globalization and quick development of digital technologies. This in turn vastly had increased both the scope of the gathering of the data, and the sharing of thereof. General Data Protection Regulation (GDPR) is a data protection law that was introduced on May 25, 2018 to addresses these data privacy related issues [8].

One of the primary objectives of GDPR is personal data privacy protection. Implementation of this law had immediately increased the significance of data anonymisation as an information sanitization process in today's world [14]. Anonymisation being a specific form of data masking suddenly became more relevant in today's world for the reason that the strict regulations defined in GDPR do not apply to the anonymized information. Data protection principles covered throughout GDPR concern only the processing of information that is identified to a natural

person, or that is yet to be identified. Given the fact that anonymised information is by definition not relating to a person and hence it can be altogether exempted from the requirements of the GDPR, that applies only to personal data, as stated under Recital 26 [7]:

> The principles of data protection should apply to any information concerning an identified or identifiable natural person. [...] The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

2.2 Anonymisation and pseudonymisation

What is pseudonymization? How does it differ?

2.3 Data masking techniques

Generalisation, suppression, etc.

2.4 Deidentification

Nawet jeśli dane personall identifiable information PII są usunięte, to można połączyć osobe z pomocą istniejących nieusuniętych danych np. płeć oraz external informacjach np. social media. Komputery są coraz szybsze, źródeł danych coraz więcej - napastnikom coraz łatwiej.

Przykłady takich ataków (Netflix Prize, IMDB).

// todo Conclusion: Anonymisation = false sense of security! All the known examples of this type of identification are from the research world — no commercial or malicious uses have yet come to light — but they prove that anonymization is not an absolute protection.

Why you can't really anonymize the data:
https://web.archive.org/web/20140109052803/http://strata.oreilly.com/2011/05/anonymize-data-limits.html

2.5 Existing solutions

List of existing solutions.

- definicja anonimizacji
- rodzaje animonimizacji (od razu odnośniki do literatury (bibliografia))
- metody, narzędzia
- problem analysis
- state of the art, problem statement
- literature research (all sources in the thesis have to be referenced [16, 15, 17, 18])
- description of existing solutions (also scientific ones, if the problem is scientifically researched), algorithms, location of the thesis in the scientific domain

[state of the art dla aplikacji sieciowych]

Chapter 3

Requirements and tools

[Dlaczego aplikacja webowa?]

- functional and nonfunctional requirements
- use cases (UML diagrams)
- description of tools
- methodology of design and implementation

Chapter 4

External specification

- hardware and software requirements
- installation procedure
- activation procedure
- types of users
- user manual
- system administration
- security issues
- example of usage
- working scenarios (with screenshots or output files)

Figure 4.1: Figure caption (below the figure).

Chapter 5

Internal specification

- concept of the system
- system architecture
- description of data structures (and data bases)
- components, modules, libraries, resume of important classes (if used)
- resume of important algorithms (if used)
- details of implementation of selected parts
- applied design patterns
- UML diagrams

Use special environment for inline code, eg **descriptor** or **descriptor_gaussian**. Longer parts of code put in the figure environment, eg. code in Fig. 5.1. Very long listings–move to an appendix.

```
1  class descriptor_gaussian : virtual public descriptor
2  {
3     protected:
4        /** core of the gaussian fuzzy set */
5        double _mean;
6        /** fuzzyfication of the gaussian fuzzy set */
7        double _stddev;
8
9     public:
10       /** @param mean core of the set
11           @param stddev standard deviation */
12       descriptor_gaussian (double mean, double stddev);
13       descriptor_gaussian (const descriptor_gaussian & w
            );
14       virtual ~descriptor_gaussian();
15       virtual descriptor * clone () const;
16
17       /** The method elaborates membership to the
            gaussian fuzzy set. */
18       virtual double getMembership (double x) const;
19
20 };
```

Figure 5.1: The **descriptor_gaussian** class.

Chapter 6

Verification and validation

- testing paradigm (eg V model)
- test cases, testing scope (full / partial)
- detected and fixed bugs
- results of experiments (optional)

Chapter 7

Conclusions

```
1  public class ColumnShuffleFacade extends AnonymisationFacade {
2
3    AnonymisationService<ColumnShuffle> anonymisationService = new
     ↪  ColumnShuffleService();
4
5    @Override
6    List<Pair<String, String>> getAnonymisedRows(ColumnOperations oper,
     ↪  List<Pair<String, String>> rows) {
7      ColumnShuffle columnShuffle = oper.getColumnShuffle();
8      if (columnShuffle == null) {
9        return null;
10     }
11     return anonymisationService.anonymise(rows, columnShuffle);
12   }
13
14   @Override
15   protected AnonymisationService<ColumnShuffle> getAnonymisationService() {
16     return anonymisationService;
17   }
18 }
```

Figure 7.1: Java code often uses long names.

- achieved results with regard to objectives of the thesis and requirements
- path of further development (eg functional extension . . .)
- encountered difficulties and problems

Table 7.1: A caption of a table is **above** it.

	method						
			alg. 3			alg. 4, $\gamma = 2$	
ζ	alg. 1	alg. 2	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 3$	$\beta = 0.1$	$\beta = -0.1$
0	8.3250	1.45305	7.5791	14.8517	20.0028	1.16396	1.1365
5	0.6111	2.27126	6.9952	13.8560	18.6064	1.18659	1.1630
10	11.6126	2.69218	6.2520	12.5202	16.8278	1.23180	1.2045
15	0.5665	2.95046	5.7753	11.4588	15.4837	1.25131	1.2614
20	15.8728	3.07225	5.3071	10.3935	13.8738	1.25307	1.2217
25	0.9791	3.19034	5.4575	9.9533	13.0721	1.27104	1.2640
30	2.0228	3.27474	5.7461	9.7164	12.2637	1.33404	1.3209
35	13.4210	3.36086	6.6735	10.0442	12.0270	1.35385	1.3059
40	13.2226	3.36420	7.7248	10.4495	12.0379	1.34919	1.2768
45	12.8445	3.47436	8.5539	10.8552	12.2773	1.42303	1.4362
50	12.9245	3.58228	9.2702	11.2183	12.3990	1.40922	1.3724

Bibliography

[1] Luk Arbuckle and Khaled El Emam. *Building an Anonymization Pipeline: Creating Safe Data*. O'Reilly Media, Inc, 2020. ISBN: 978-1492053439.

[2] [brak autorów]. 'Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC'. In: *Physics Letter B* 716.1 (2012), pp. 1–29.

[3] Inc. Cisco Systems. *Cisco Annual Internet Report (2018–2023)*. Tech. rep. URL: `https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf` (visited on 21/01/2022).

[4] Inc. Cisco Systems. *The Zettabyte Era Officially Begins*. URL: `https://blogs.cisco.com/sp/the-zettabyte-era-officially-begins-how-much-is-that` (visited on 21/01/2022).

[5] European Comission. *Charter of Fundamental Rights of the European Union*. URL: `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT&from=EN` (visited on 21/01/2022).

[6] European Comission. *General Data Protection Regulation - Recital 1*. URL: `https://gdpr-info.eu/recitals/no-1` (visited on 21/01/2022).

[7] European Comission. *General Data Protection Regulation - Recital 26*. URL: `https://www.privacy-regulation.eu/en/recital-26-GDPR.htm` (visited on 21/01/2022).

[8] European Comission. *General Data Protection Regulation - Recital 6*. URL: `https://gdpr-info.eu/recitals/no-6` (visited on 21/01/2022).

[9] Khaled El Emam and Luk Arbuckle. *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O'Reilly Media, Inc, 2020. ISBN: 978-1449363079.

[10] Apache Software Foundation. *Apache Kafka – uses*. URL: `https://kafka.apache.org/uses` (visited on 21/01/2022).

[11] Chris Lee. *Integrating Web Analytics with your SAP Commerce Cloud Storefront - An Overview*. URL: `https://www.sap.com/cxworks/article/2589633935/integrating_web_analytics_with_your_sap_commerce_cloud_storefront_an_overview` (visited on 21/01/2022).

[12] Bernard Marr. *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. John Wiley and Sons Ltd, 2016. ISBN: 978-1119231387.

[13] Neha Narkhede, Gwen Shapira and Todd Palino. *Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale*. O'Reilly Media, Inc., 2017. ISBN: 978-1491936160.

[14] Petter Ødegård. 'Data Anonymization for Research'. MA thesis. Norwegian University of Science and Technology, 2019.

[15] Name Surname and Name Surname. *Title of a book*. Hong Kong: Publisher, 2017. ISBN: 83-204-3229-9-434.

[16] Name Surname and Name Surname. 'Title of an article in a journal'. In: *Journal Title* 157.8 (2016), pp. 1092–1113.

[17] Name Surname, Name Surname and N. Surname. 'Title of a conference article'. In: *Conference title*. 2006, pp. 5346–5349.

[18] Name Surname, Name Surname and N. Surname. *Title of a web page*. 2021. URL: `http://somewhere/on/the/internet.html` (visited on 30/09/2021).

[19] Colin Tankard. *Big data security*. URL: `https://www.sciencedirect.com/science/article/abs/pii/S1353485812700636` (visited on 21/01/2022).

Appendices

Index of abbreviations and symbols

DNA	deoxyribonucleic acid
MVC	model–view–controller
N	cardinality of data set
μ	membership function of a fuzzy set
$\mathbb{E}$	set of edges of a graph
$\mathcal{L}$	Laplace transformation

Listings

(Put long listings in the appendix.)

```
1  partition fcm_possibilistic::doPartition
2                                  (const dataset & ds)
3  {
4     try
5     {
6        if (_nClusters < 1)
7           throw std::string ("unknown number of clusters"
               );
8        if (_nIterations < 1 and _epsilon < 0)
9           throw std::string ("You should set a maximal
               number of iteration or minimal difference --
               epsilon.");
10       if (_nIterations > 0 and _epsilon > 0)
11          throw std::string ("Both number of iterations
               and minimal epsilon set -- you should set
               either number of iterations or minimal
               epsilon.");
12
13       auto mX = ds.getMatrix();
14       std::size_t nAttr = ds.getNumberOfAttributes();
15       std::size_t nX    = ds.getNumberOfData();
16       std::vector<std::vector<double>> mV;
17       mU = std::vector<std::vector<double>> (_nClusters)
```

```
        ;
18      for (auto & u : mU)
19          u = std::vector<double> (nX);
20      randomise(mU);
21      normaliseByColumns(mU);
22      calculateEtas(_nClusters, nX, ds);
23      if (_nIterations > 0)
24      {
25          for (int iter = 0; iter < _nIterations; iter++)
26          {
27              mV = calculateClusterCentres(mU, mX);
28              mU = modifyPartitionMatrix (mV, mX);
29          }
30      }
31      else if (_epsilon > 0)
32      {
33          double frob;
34          do
35          {
36              mV = calculateClusterCentres(mU, mX);
37              auto mUnew = modifyPartitionMatrix (mV, mX);
38
39              frob = Frobenius_norm_of_difference (mU,
                    mUnew);
40              mU = mUnew;
41          } while (frob > _epsilon);
42      }
43      mV = calculateClusterCentres(mU, mX);
44      std::vector<std::vector<double>> mS =
            calculateClusterFuzzification(mU, mV, mX);
45
46      partition part;
47      for (int c = 0; c < _nClusters; c++)
```

```
48          {
49              cluster cl;
50              for (std::size_t a = 0; a < nAttr; a++)
51              {
52                  descriptor_gaussian d (mV[c][a], mS[c][a]);
53                  cl.addDescriptor(d);
54              }
55              part.addCluster(cl);
56          }
57          return part;
58      }
59      catch (my_exception & ex)
60      {
61          throw my_exception (__FILE__, __FUNCTION__,
                __LINE__, ex.what());
62      }
63      catch (std::exception & ex)
64      {
65          throw my_exceptionn (__FILE__, __FUNCTION__,
                __LINE__, ex.what());
66      }
67      catch (std::string & ex)
68      {
69          throw my_exception (__FILE__, __FUNCTION__,
                __LINE__, ex);
70      }
71      catch (...)
72      {
73          throw my_exception (__FILE__, __FUNCTION__,
                __LINE__, "unknown␣expection");
74      }
75  }
```

List of additional files in electronic submission (if applicable)

Additional files uploaded to the system include:

- source code of the application,
- test data,
- a video file showing how software or hardware developed for thesis is used,
- etc.

List of Figures

List of Tables