Univerzita Pardubice
Fakulta elektrotechniky a informatiky

POSUDEK VEDOUCÍHO DIPLOMOVÉ PRÁCE

Jméno studenta: Bc. Pavel Klečanský
Téma práce: Deduplikace dat a jejich využití
Vedoucí diplomové práce: Ing. Monika Borkovcová, Ph.D.

Cílem práce je podrobný popis algoritmů a problematiky deduplikace spojování záznamů a vytvoření knihovny, která umožní provádět spojování záznamů mezi dvěma zdroji dat a deduplikaci jednoho zdroje dat. Práce bude využívat Jaro-Winklerovu vzdálenost, Levenshteinovu vzdálenost, Damerau-Levensteinovu vzdálenost, Jaccardův index a Q-gram podobnost.

1. **Jaké metody (příslušející navazujícímu magisterskému studiu) diplomant ve své práci uplatnil?**

Řešená oblast se zabývá technikami deduplikace a spojováním záznamů, které rozšiřují studentovi znalosti z oblasti databázových systémů a datových skladů o další techniky čištění dat, metody blokování a indexování a samozřejmě samotné algoritmy používané pro porovnávání atributů a záznamů, tedy algoritmy editační vzdálenosti a algoritmy založené na tokenech.

2. **Co diplomant při vypracování své DP vytvořil?**

Předložená práce shrnuje poznatky z oblasti práce s tzv. nečistými daty, jejich klasifikaci, možnostmi fúze dat a samotnými metodami a algoritmy využívanými při hledání podobnosti v datech. Před tvorbou knihovny, která je v práci nazvaná jako Match4j autor porovnával stávající řešení, kdy se zaměřil na Splink, JedAI, Febrl a ostatní a zhodnotil jejich možnosti využití. Technologický stack vytvořené knihovny spočíval v implementaci v jazyku Java s využitím třech externích knihoven opencsv pro parsování souborů CSV, logovací knihovny logback-classic a smile-core jakožto knihovnu klasifikačních algoritmů a algoritmů strojového učení.

3. **Jak diplomant prokázal správnost navrženého řešení problému?**

Možnosti použití vytvořené knihovny jsou řádně popsané v kapitole 8 včetně uvedení výsledků měření na množinách dat pro testování algoritmů, a to pro algoritmy blokování, podobnosti a klasifikační algoritmy. Navržené řešení je plně funkční a řádně popsané.

4. **Podařilo se diplomantovi splnit cíle práce, které mu byly uloženy?**

Cíle práce byly splněny, výstup práce byl vytvořen podle zadání, a ačkoliv se jednalo o novou oblast, diplomant splnil vše dle zadání a dle domluvy. Při tvorbě byl patrný zájem autora o řešenou problematiku.

5. **Jaká je kvalita textu diplomové práce z hlediska jeho struktury, srozumitelnosti, jazykové a typografické úrovně?**

Práce má logické členění, je srozumitelná a má řádnou jazykovou úroveň. Text je doplněn o ilustrace v podobě tabulek a obrázků, které zvyšují úroveň pochopení psaného textu. Celkově práce splňuje formální náležitosti.

Univerzita Pardubice
Fakulta elektrotechniky a informatiky

6. **Jak byla vyhodnocena kontrola textu DP (případně zdrojových kódů softwaru) pomocí systému pro odhalování plagiátů mezi závěrečnými pracemi?**

Kontrola původnosti práce byla shledána s výsledkem - není plagiát.

7. **Které nejasnosti vyskytující se v DP by měl diplomant objasnit při obhajobě a jaké jsou další připomínky k DP?**

K práci nemám zásadní připomínky, celkově práci pokládám za velmi zdařilou a kvalitně zpracovanou.

8. **V závěru je nutno jednoznačně uvést, zda je práce doporučena či nedoporučena k obhajobě a jakým klasifikačním stupněm je hodnocena?**

Práce splňuje všechny požadavky kladené na diplomovou práci. Práci doporučuji k obhajobě a hodnotím ji známkou A.

Otázky k obhajobě:

S jakými problémy jste se potýkal při řešení praktické části práce?
Jaké jsou možnosti rozšíření Vašeho řešení a jaké kroky byste navrhl pro jeho naplnění?

V Pardubicích dne 29.05.2024

Ing. Monika Borkovcová, Ph.D.