Hodnocení vedoucího závěrečné práce

České vysoké učení technické v Praze **Fakulta informačních technologií**

Student: Bc. Anna Kučerová
Vedoucí práce: Ing. Luboš Krčál
Název práce: Approximate Pattern Matching In Sparse Multidimensional Arrays Using Machine Learning Based Methods
Obor: Znalostní inženýrství

Datum vytvoření: 5. 6. 2017

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 5:*
1. Náročnost a další komentář k zadání	*1=mimořádně náročné zadání,* ***2=náročnější zadání,*** *3=průměrně náročné zadání,* *4=lehčí, ale ještě dostatečně náročné zadání,* *5=nedostatečně náročné zadání*

Popis kritéria:
Podrobněji charakterizujte diplomovou (bakalářskou) práci a její případné návaznosti na předchozí nebo běžící projekty. Dále posuďte, čím je zadání této ZP náročné. (U obtížnější ZP lze dále tolerovat některé nedostatky, které by u ZP standardní obtížnosti tolerovány nebyly; a naopak u jednoduché ZP mohou být zjištěné nedostatky hodnoceny přísněji.)

Komentář:
Jedná se o zadání, které je aktuálně předmětem výzkumu. Existující algoritmy pro přibližné vyhledávání byly k datu práce publikovány pouze na teoretické úrovni, nad řídkými poli pak vůbec. Velká šířka oboru zpracování multidimenzionálních polí pak přidává na složitosti zejména ve fázi rešerše, kde je treba vytvořit si rozsáhlý obraz problematiky, vyžaduje porozumnění mnoha algoritmů i aplikací.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
2. Splnění zadání	*1=zadání splněno,* ***2=zadání splněno s menšími výhradami,*** *3=zadání splněno s většími výhradami,* *4=zadání nesplněno*

Popis kritéria:
Posuďte, zda předložená ZP splňuje zadání. V komentáři uveďte body zadání, které nebyly zcela splněny, případně rozšíření ZP oproti původnímu zadání. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Komentář:
Rešeršní část práce byla splněna nad rámec očekávání, stejně jako návrh algoritmu kombinující teoreticko-informatický problém se znalostním inženýrstvím. Menší výhrady se týkají pouze implementace a testování navržených algoritmů, kde v některých okrajových případech neodpovídaly výsledky teoretickým očekáváním.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
3. Rozsah písemné zprávy	***1=splňuje požadavky,*** *2=splňuje požadavky s menšími výhradami,* *3=splňuje požadavky s většími výhradami,* *4=nesplňuje požadavky*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části.

Komentář:
Rozhas naprosto vyhovující. Část rešerše, která je vůči hlavnímu tématu okrajově relevantní, byla přesunuta do apendixu.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Věcná a logická úroveň práce	*80 (B)*

Popis kritéria:
Posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti. Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře.

Komentář:
Logická úroveň a organizace práce naprosto bez výtek. Práce je organizovaná podobně jako vědecké články v oboru, s mnohem rozsáhlejší rešeršní částí a podrobnějším popisem implementace a měření.

Výtky k pochopitelnosti práce jsou, zejména v části měření, kde není kompletně diskutována návaznost na předchozí teoretické výsledky v oboru, očekávané a skutečné výsledky implementace, a popis implementačních překážek, na které studentka narazila.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*75 (C)*

Popis kritéria:
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 14/2015, článek 3.

Komentář:
Typografická úprava je na dobré úrovni. Práce je slabší po jazykové (anglické) stránce. Celý text je pochopitelný, nicméně spousta formulací je "kostrbatá" a příliš připomíná překlad slovo od slova. Potřeba nativního korektora je značná.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*85 (B)*

Popis kritéria:
Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení ZP. Charakterizujte výběr studijních pramenů. Posuďte, zda student využil všechny relevantní zdroje nebo zda se pokoušel řešit již vyřešené problémy. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Komentář:
Zdroje pokrývají problematiku zcela a do rozumné hloubky. Odlišení od vlastních výsledků je dobře popsané a jasně dané strukturou práce.

Po formální stránce mám několik výtek. Pořadí referencí v seznamu referencí je mi záhadou. Navíc, některé reference v seznamu se nevyskytují v textu.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*90 (A)*

Popis kritéria:
Vyjádřete se k úrovni dosažených hlavních výsledků ZP, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, apod. Případně také zhodnoťte, zda software nebo zdrojové texty, které nevytvořil sám student, byly v ZP použity v souladu s licenčními podmínkami a autorským právem. Popište případnou publikační činnost a získaná ocenění související s řešením této ZP.

Komentář:
Dosažené výsledky kompletně pokrývají zadání práce. Rešeršní část poté svou hloubkou toto i překračuje.

Navržené algoritmy pro přibližné vyhledávání vzorků v multidimenzionálních polích fungují, byť v některých okrajových případech ne optimálně. Implementace a meření předchozích, do této doby pouze teoretických algoritmů je bezproblémová.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*
8. Komentář o využitelnosti výsledků	

Popis kritéria:
Uveďte, zda hlavní výsledky ZP rozšiřují již publikované známé výsledky a/nebo přinášející zcela nové poznatky. Uveďte možnosti využití výsledků ZP v praxi.

Komentář:
Práce má vysoce nadprůměrný publikační potenciál na následujíccíh urovních:

- Implementace a detailní měření doposud pouze teoretických algoritmů nad skutečnými a umělými daty, a s datovým modelem multidimenzionálních polí, ktery je běžně používaný v moderních databázích.
- Nové algoritmy využívající array kernely, similarity hashování a indexování pro výlepšení filtrovacích fází doposud znamých algoritmů.

V prvním případě je potřeba více rozvinout měření, více diskutovat vliv datového modelu polí a implementační detaily. V druhém případě je treba rozvinout navržené algoritmy více do hloubky, a to zejména po teoretické stránce.

Studentce bylo navrženo pokračovat i po odevzdání diplomové práce alespoň na první variantě.

Následující není výtka, pouze poznámka. Nebyly prozkoumány, navrženy a naimplementovány metody s předpočítaným indexem, které mohly použité algoritmy urychlit až řádově, a více tak rozšírit publikační potenciál.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 5:*
9. Aktivita a samostatnost studenta v průběhu řešení	9a: ***1=výborná aktivita,*** *2=velmi dobrá aktivita,* *3=průměrná aktivita,* *4=slabší, ale ještě dostatečná aktivita,* *5=nedostatečná aktivita* 9b: *1=výborná samostatnost,* ***2=velmi dobrá samostatnost,*** *3=průměrná samostatnost,* *4=slabší, ale ještě dostatečná samostatnost,* *5=nedostatečná samostatnost*

Popis kritéria:
Posuďte, zda byl student během řešení aktivní, zda dodržoval dohodnuté termíny, jestli své řešení průběžně konzultoval a zda byl na konzultace dostatečně připraven (9a). Posuďte schopnost studenta samostatné tvůrčí práce (9b).

Komentář:
Dobrá komunikace i aktivita. Místy bylo třeba studentku motivovat do více kreativních a odvážnějších kroků, zejména ve fázi návrhu algoritmu a implementace.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*90 (A)*

Popis kritéria:
Shrňte stránky ZP studenta, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení **nemusí** být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích 1 až 9.

Text hodnocení:
Zadání diplomové práce je náročné a z velké části postavené na aktuálním výzkumu v oboru vícedimenzionálních polí a přibližného vyhledávání, které je komplikované, neobsahuje ucelené rešeršní publikace, a teprve v posledních pár letech se objevují praktické implementace v podobě array databází.

Zadání bylo splněno kompletně. Rešeršní část svojí kvalitou značně předčila očekávání. Implementace existujících, do této doby pouze teoretických algoritmů se podařila bez výhrad. Návrh vlastních algoritmů studentky s použitím array kernelů a similarity hashování pro vylepšení stávajících algoritmů je ucelený a smysluplný. Jejich implementace funguje. V některých okrajových případech se projevují implementační nedostatky a malé nedostatky v měření, jejich závažnost je vzhledem ke složitosti a rozsahu práce marginální.

Oceňuji také aktivitu studentky při komunikaci s autory předchozího výzkumu v oboru.

Práce má vysoce nadprůměrný publikační potenciál na následujících úrovních:
- Implementace a detailní měření doposud pouze teoretických algoritmů nad skutečnými a umělými daty, a s datovým modelem multidimenzionálních polí, ktery je běžně používaný v moderních databazích.
- Nové algoritmy využívající array kernely, similarity hashování a indexování pro výlepšení filtrovacích fází doposud znamých algoritmů.

Doporučují studentce investovat čas a snahu navíc do publikace svých výsledků i po ukončení svého studia. Minimálné první publikační potenciál by byla škoda nevyužit.

Doplňující otázky:
- Similairty hashování nad array kernely: Co je přesně ve navrženém algoritmu array kernel? Proč je pouze jednodimenzionální? Mohl by být vícedimenzionální? Kde přesně se kernel používá a porovnává s patternem, nad jakou podmnožinou polí?
- Vliv chunkování na časovou a paměťovou složitost algoritmu: Jak se liší chunkování pomoci hyperkrychlí a chunkování podle jedné z dimenzí? Proč se ve SciDB používá pouze uniformní chunkování (ne hyperkrychle, ale uživatelem definované hyperkvádry)?
- Proč časová složitost některých algoritmů neodpovídá očekavané teoretické složitosti? Například Fig 5.4, Error=16, LSB. Jak opravit implementaci nebo datový model polí?

Podpis vedoucího práce: