FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Student:	Bc. Lukáš Renc
Oponent práce:	Ing. Ondřej Guth, Ph.D.
Název práce:	Automata Approach to Approximate Tree Pattern Matching
Obor:	Teoretická informatika

Datum vytvoření: 6. 6. 2020

Hodnotící kritérium:	*Způsob hodnocení – následující škálou 1 až 4:*
1. 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 dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posuďte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Komentář:
Zadání patří mezi obtížnější, neboť vyžaduje krok do neznáma: návrh vlastního algoritmu. Zadání konstatuji jako splněné.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
2. Písemná část práce	*55 (E)*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále 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. Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 26/2017, článek 3. Posuďte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Komentář:
Rozsah práce je na počet stran přiměřený. Text obsahuje části, které jsou celkem zbytečné, a to především v kapitole 2 (Related work). Vyloženě nerelevantní je několik stran o UML diagram version: porovnávání diagramů; odkazovaný článek se netýká předmětu práce, navíc student se v popisu nezaměřuje na algoritmický aspekt porovnání orientovaných acyklických grafů, ale pouze na aplikaci (vizuální podání rozdílu mezi dvěma diagramy). Celá podkapitola 2.2 (vyhledávání v řetězcích) je rovněž nerelevantní (dávala by trochu smysl, kdyby obsahovala informace o vyhledávání s Levenshteinovou/editační vzdáleností). I přes tyto zbytečné části rešerše je tato část nedostatečně zpracovaná a působí dojmem, že tam je, jen aby tam něco bylo. V práci tedy naprosto chybí ucelený přehled o přibližném vyhledávání ve stromech, vzdálenostech mezi stromy a algoritmech na jejich výpočet. Výjimku tvoří Selkowova vzdálenost (i tak je algoritmus na její výpočet popsán velmi nepochopitelně).

Po věcné stránce text obsahuje řadu chyb a nepřesností takového významu, že činí tuto teoretickou práci těžko pochopitelnou. V popisu konstrukce konečného automatu na vyhledávání s Hammingovou vzdáleností chybí vyhledávací smyčka; nejde o opomenutí, student jinde v textu píše, že tato smyčka tam není záměrně; bez ní ale automat nemůže vyhledávat. Pseudokód v algoritmu 2 není formálně správně: mnoho vlastností výsledného zásobníkového automatu není definováno (např. množina koncových stavů, zásobníková abeceda, počáteční zásobníkový symbol); přitom další text a algoritmy tyto vlastnosti využívají; tento pseudokód, příklad 3.7 a důkaz teorému 3.3 si ani vzájemně neodpovídají (někde se mluví o odstranění symbolu Z0 ze zásobníku, což se ale neděje). Bylo by vhodné si pseudokódy před odevzdáním alespoň jednou odkrokovat.

Navržená vzdálenost dvou stromů, především část "block type matching", je velmi zajímavá a dává smysl. Nápad na tuto vzdálenost považuji za největší přínos této práce. O její definici 3.1 to ale říci nelze: kvůli formálním chybám (otočené "menší než", které se vyskytuje na mnoha místech v práci, ale hlavně "Ti je nejpravějším potomkem Ti") nejde pochopit, v čem vzdálenost spočítá, bez přečtení dalších částí práce. Z textu není jasné, jak má fungovat algoritmus vyhledávání se studentem navrženou vzdáleností. Text obsahuje pouze algoritmus na konstrukci nedeterministického zásobníkového automatu (velmi nedůsledně specifikovaný, čtenář si musí mnoho domyslet). Dále je obsažen důkaz časové složitosti vyhledávání nad tímto automatem, ze kterého není poznat, jakým způsobem je nedeterministické vyhledávání realizováno. Až z implementace (a ne z důkazu samotného!) lze domyslet, že student má pravděpodobně na mysli rekurzivní backtracking. Práce obsahuje důkaz, že automat lze determinizovat. Chybí ale jakékoli zdůvodnění, proč nakonec používá jen nedeterministickou verzi. S ohledem na to, že výsledná časová složitost vyhledávacího algoritmu je exponenciální, by takové zdůvodnění bylo na místě.

K navrženým algoritmům mám ještě jednu připomínku: dojde k nahlášení jednoho výskytu stromu vícekrát, s více vzdálenostmi najednou. Algoritmus neobsahuje způsob, jak najít a nahlásit daný výskyt s pouze nejmenší vzdáleností. Algoritmus tedy neodpovídá studentově definici vzdálenosti, která minimalitu obsahuje. Škoda, že se student neinspiroval v konstrukci vyhledávacích automatů pro řetězce a editační vzdálenost, které jsou probírané v oborovém předmětu a které se s tímto jevem vyrovnávají.

Za nedostatečné považuji i testování. Testování správnosti podle textu spočívá ve spuštění referenční implementace oproti jednomu vstupu (podle přílohy se zdá, že to tak naštěstí nebylo).

Oceňuji, že student použil snadno pochopitelné definice stromů i že mnoho definic a vět doplňoval vysvětlením vlastními slovy.

Z hlediska typografického i jazykového práce obsahuje chyby, není jich však mnoho a s přihlédnutím ke skutečnosti, že student nepsal ve svém rodném jazyce, je práce z tohoto pohledu na dobré úrovni.

Nenašel jsem porušení citační etiky, využití zdrojů má však své rezervy. Kromě již zmíněné nedostatečné rešerše jsou často citované zdroje nerelevantní i v kontextu, kdy je na ně odkazováno. Příklady: odkaz na barovou notaci v úvodu vede na nepublikovanou přednášku; je uvedeno, že myšlenky v práci vychází z amerického patentu, zároveň nikde v práci není rozebráno, zda je takové použití patentu zákonné; odkaz na publikaci o zásobníkových automatech v místě, kde se mluví o automatech řetězcových (příklad 2.1).

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
3. Nepísemná část, přílohy	*80 (B)*

Popis kritéria:
Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce
– opakovatelnost experimentů

Komentář:
Nepísemnou částí je implementace navržených algoritmů. Není jasné, proč se student rozhodl pro jazyk Java (nikde není zdůvodnění). Implementace zřejmě odpovídá navrženému algoritmu a na mé nahodilé testování dávala výstupy odpovídající algoritmu. Výhrady mám k programátorskému stylu, především globálním proměnným, dále k velkému množství zakomentovaného kódu a přitom malému množství vysvětlujících komentářů, stejně tak k ladícím výpisům.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*

4. Hodnocení výsledků, jejich využitelnost	*50 (E)*

Popis kritéria:
Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Komentář:

Za přínos práce považuji nápad na vzdálenost dvou stromů: block type matching v kombinaci se Selkowovou vzdáleností. Z důvodu malé pečlivosti v ostatních částech práce: rešerše a algoritmy, by ostatní části práce byly použitelné po značném úsilí. Skutečnost, že navržený algoritmus neodpovídá vzdálenosti a že je z důvodu ne úplně dobré práce s nedeterminismem exponenciální, sráží přínosy této práce.

Hodnotící kritérium:	*Způsob hodnocení – nehodnotí se*

5. Otázky k obhajobě

Popis kritéria:
Uveďte případné dotazy, které by měl student zodpovědět při obhajobě ZP před komisí (body oddělte odrážkami).

Otázky:

1. Jakým způsobem byly použity výsledky J. Oomena z amerického patentu ve Vaší práci?
2. Jakým způsobem/algoritmem pracujete při vyhledávání s nedeterministickým zásobníkovým automatem? Proč jste se rozhodl nepoužít konstrukci deterministického zásobníkového automatu a následné vyhledávání nad ním?

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*

6. Celkové hodnocení	*65 (D)*

Popis kritéria:
Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.

Text hodnocení:

Práce přináší zajímavou myšlenku na výpočet vzdálenosti dvou stromů použitelnou v praktickém vyhledávání. Dojem z práce kazí nedotažený algoritmus na konstrukci vyhledávacího zásobníkového automatu (který v důsledku toho nefunguje) i nedostatečné vyrovnání se s výzvami, které nedeterminismus přináší; stejně tak i řada formálních chyb. V práci prakticky chybí přehled "state-of-the-art". Implementace navrženého algoritmu je průměrná.

Podpis oponenta práce: