FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Student: Lukáš Renc
Oponent práce: Ing. Jan Trávníček
Název práce: Automata Approach to XML Data Indexing: Implementation and Experimental Evaluation
Obor: Teoretická informatika

Datum vytvoření: 27. 5. 2018

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:
Práce je čistě implementační a i přes nutnost nastudovat různé možnosti implementace automatů (které efektivně tvoří indexovací struktury ze zadání), nemyslím si, že by zadání mělo být hodnoceno jako náročnější.

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:
Rešerše možných implementací automatů by mohla být podrobnější, jen na naší faultě vznikl článek zabývající se tímto tématem: Holub, J. (2007). Finite automata implementations considering CPU cache. Acta Polytechnica, 47(6). Zároveň bych si představoval více implementací automatů. Na druhou stranu kladně hodnotím více implementací konstrukčních algoritmů.

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:
Práce je, i přes nějaké nepřirozeně prázdné oblasti v textu, dostatečné délky a všechny kapitoly mají v práci své opodstatnění; navíc je text psaný v angličtině.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Věcná a logická úroveň práce	*70 (C)*

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:
V kapitole 1 Definice v sekci 1.1.6 je deterministický konečný automat definován na základě konečného automatu, kdy je definován pouze nedeterministický konečný automat.
V sekci 1.2 je popsáno XML jako struktura s uzly, kde tyto uzly jsou pojmenované, ale zároveň mají přiřazené pořadí odpovídající preorder průchodu. Toto zní spíše jako implementační rozhodnutí reprezentace uzlů XML.
Dále pak v sekci 1.2.1 se hovoří o XPath dotazech, které jsou nadefinované až v sekci 1.3.
V sekcích 2.1., 2.2. a 2.3. jsou popsány implementované konečné automaty, které mají přijímat podčást jazyka XPath. Tyto podčásti jazyka XPath jsou podle textu práce generované pomocí bezkontextových gramatiky. Tento fakt ale v práci nikde není diskutován.
V sekci 2.1.1. je chybně přeložená časová složitost jako "complexity time" místo "time complexity" navíc je asymptotická složitost O označena jako minimální časová složitost.
V algoritmu 3 je Q1 i Q2 množina hodnot tedy Q obsahuje typově hodnoty; proměnná d je typově množina, protože reprezentuje d-subset; i když je d později do množiny Q přidávaná, troufám si tvrdit, že definice a použití množiny Q by mělo byt přesnější.
Vzhledem k tématu práce - indexování - čekal bych sekci 3.4. mnohem obsáhlejší (hlavní důvod pro hodnocení splnění zadání).
V kapitole měření je pod označením M1, M2, M3 a M4 odkazováno na algoritmy původní, nový semideterministický, nový nedeterministický a referenční. Později ale podle tabulek měření i grafů vypadá označení M1, M2, M3 a M4 spíše jako semideterministický, nedeterministický, původní, a referenční.
V měření jsou pod jedním popisem agregovány až čtyři grafy, které ovšem nejsou patřičeně popsány a napojeny na zbytek textu.
V závěru měření je těžké posoudit, na základě jakého měření byla potvrzena složitost vyhledávání O(|query| + |occ|), když všechny dotazy Q1 až Q9 měly efektivně stejnou velikost a buď téměř stejný nebo zanedbatelný absolutní rozdíl v počtu výsledků dotazů.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*70 (C)*

Popis kritéria:
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 26/2017, článek 3.

Komentář:
Angličtina je v práci v některých místech stylově méně odborná. Volba některých slov nebo slovních spojení (crutial part, obviously, tricky, ...) není příliž vhodná pro odborný text.
Stejně tak gramaticky není text bez chyb (have/has), chybějící členy, ...). Dále text obsahuje i nějaké strojově odhalitelné překlepy (paralllel, ...).
Z typografického pohledu je vidět, že byla práce psána v kratší době a na opravu mnoha detailů nezbyl čas. Tabulky a grafy v kapitole měření přetékají. V některých místech i text.
Další typografické chyby v kapitole 1. sekce a podsekce často neobsahují jediný jiný text než jen definici. Příklad 1.2.1 je bez textu jen čistě figura. Chybí tečky na konci vět v sekci 1.2.1 v seznamu. Príklady 1.2.2., 1.2.3., 1.2.4. také obsahují téměř čistě jen figury.
V kapitole 2, příklady 2.0.1. a 2.0.2 obsahují prázdnou položku seznamu. Algoritmus 1 má položky seznamu příkladů uvozené "ht!". Uvození příkladu 2.1.2. je na posledním řádku stránky.
Figury 2.8 a 2.9 jsou jedna na výšku a jedna na šířku, i když by obě mohly být jak na výšku tak na šířku. Figura 2.10. přetéká do hřbetu vazby.
Figura 4.1. nevypadá jako figura. Odkazy v algoritmu 4 nejsou definované.
Grafy v kapitole měření mají na ose x vynesený krok algoritmu i pozdějšího vyhledávání jakoby spojitě. Navíc bez specifikované jednotky. Některé tabulky jsou vysázené větším, některé menším písmem.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*95 (A)*

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:
Za nejpodstatnější část práce považuji vlastní rešerši a nakonec i volbu způsobu implementace indexu. Toto je ve zdrojích pokryto třemi články, což se mi zdá málo. Jinak ale považuji práci se zdroji v pořádku.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*60 (D)*

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:
Text práce navrhuje nový algoritmus konstrukce již publikovaného automatu indexujícího výsledky podmnožiny dotazů XPath. Měření konstrukčních algoritmů bych uvítal podrobnější než omezené na 10 XML souborů a 9 XPath dotazů a 4 algoritmy.

Podle popisu implementace by měla třída Automaton být rozšířena (extended) vlastními datovými třídami jednotlivých indexů (TSPA, TSPSA a TPA). V implementaci je takto vytvořena jen třída TPA. Ostatní třídy tohoto jména neexistují, i když text na ně odkazuje později jako dTSPA, dTSPSA a dTPA (ty ale nedědí od třídy Automaton).
Konstrukce indexovacích automatů v třídě AutomatonFactory je přinejmenším matoucí. Termíny z textu práce jako semideterministic se v implementaci vůbec neobjevují, a je tedy podstatně složitější se v implementaci zorientovat. Factory metody getdTSPSA a getdTSPA mají stejnou implementaci. Neměly by se lišit?
Factory metody konstrukce indexovacích automatů automaticky serializují výsledný indexovací automat, případně se ho snaží deserializovat ze souboru daného jménem indexovaného XML. (Co když bude tento soubor jiný index než autor předpokládá?)
Dokumentace kódu je pouze občasná, případně komentáře připomínají dookomentovat kód nebo domyslet řešení.

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:
Na práci bude dle mého názoru nutné ještě navázat, především implementaci bude potřeba zrevidovat, ale i tak považuji výsledky práce za zajímavé a užitečné.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*
9. Otázky k obhajobě	

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

Otázky:
2 otázky viz hodnocení výsledků.
Jakým způsobem bylo provedeno měření paměti?

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*65 (D)*

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:
Na textu práce je bohužel vidět, že nezbýval čas na jeho dopracování. Mnoho typografických i věcných chyb. Implementace by bohužel také potřebovala nějaké revize.

Nový algoritmus konstrukce indexovacího algoritmu je na druhou stranu pozitivní.

Celkově práci považuji za zajímavou. Kvalita textu i implementace snižují mé hodnocení až na 65 bodů, tedy D.

Podpis oponenta práce: