РЕЦЕНЗИЯ
върху дисертационния труд за придобиване
на образователна и научна степен „доктор”
в професионално направление 4.5. Математика,
научна специалност 01.01.01 „Математическа логика”

Автор на дисертационния труд: ас. Стефан Владимиров Герджиков
от катедра „Математическа логика и приложенията ѝ”
към Факултета по математика и информатика на СУ „Свети Климент Охридски”.
Тема на дисертационния труд: „Ефективен алгоритъм за приближено търсене в регулярни множества”.
Член на научното жури: акад. проф. дмн Веселин Стоянов Дренски от ИМИ към БАН.

Дисертационният труд е в област, която е гранична за няколко математически дисциплини – математическа логика (теория на алгоритмите и теория на изчислимостта), математическа лингвистика и математическа информатика (теория на формалните езици), алгебра (теория на полугрупите и линейна алгебра), комбинаторика (теория на графите и теория на производящите функции), информатика (част от резултатите са представени във вид на работещи процедури, а някои от тях са реализирани програмно). Той е написан на английски и е изложен на xvi + 181 „стандартни” TEX-страници. Състои се от увод, 8 глави, апендикс, заключение и списък на използваната литература.

1. Актуалност на разработвания проблем. Много често при предаването на информация полученото съобщение се различава от изпратеното поради наличието на смущения (или „шум”) по канала. Това явление се наблюдава в много различни области и се проявява по най-различни начини. Става въпрос не само за грешки при обичайното предаване на съобщения (по радиото или по телефона, с телеграми, SMSи или от една част на компютъра до друга). Проблемът възниква и в много други случаи, например в естествените езици. Появяването на грешки е обичайно при набирането на писмен текст, а при обработката на стари текстове се налага отчитането на развитието и изменението на езика в течение на времето. Възниква естествената и важна задача за разработване на ефективни методи за възстановяване на първоначалното съобщение. Един математически подход е разглеждането на съобщението като дума от език, използващ азбука от краен брой символи. След това се въвежда подходящо „разстояние” между крайните редици от букви. Задачата се свежда до намирането на тези думи от езика, които са „най-близко” до полученото съобщение, а след това, след допълнителни разглеждания, се решава коя от тези думи е оригиналното съобщение. При това се прави естественото ограничение, че „силата” на шума трябва да е относително „малка”, за да не отдалечава твърде много полученото съобщение от изпратеното. В теория на кодирането, където думите са с еднаква дължина, се въвежда разстоянието на Хеминг, което отчита броя на сгрешените символи при предаването на съобщението. В представената дисертация се разглеждат езици от думи с различна дължина. Поради тази причина се разглежда разстоянието на Левенщтейн, което е естествено обобщение на разстоянието на Хеминг. Предполага се, че при предаването на информацията се допускат три типа грешки – (1) заместване на някоя от буквите в оригиналното съобщение с друга буква; (2) изтриване на някоя от буквите в съобщението; (3) вмъкване на нови букви. Разстоянието на Левенщтейн между две думи отчита минималния брой на измененията, необходими за превръщането на

едната дума в другата. В дисертацията се разглежда и обобщеното разстояние на Левенщтейн, при което се допускат и други изменения (например разместване на букви или добавяне на група от символи), като съответните изменения имат „цена" или „тегло". Примери за такива грешки могат да се намерят и в дисертацията:
(1) на стр. 35, ред 3 отдолу нагоре: „th**A**n" в текста „shorter candidates first and than longer" трябва да бъде „th**E**n" („e" е сменено с „a");
(2) на стр. 4, ред 3 отдолу нагоре и на други места в текста: „set of operation" трябва да бъде „set of operations" (изпуснато окончанието „s");
(3) на стр. 56, ред 1 на доказателството на лема 5.3.1: в „the the" има излишно „the";
(4) на страница 161, ред 3 от параграф 8.3 „Bulgarina" трябва да бъде „Bulgarian" (смяна на местата на две букви).

2. Степен на познаване на състоянието на проблема и творческа интерпретация на литературата. Дисертантът познава много добре състоянието на проблема. Дисертационният труд съдържа списък от 67 литературни източника. За актуалността на списъка говори фактът, че съществена част от публикациите – 23 – са след 2000 г. „Образователната" страна на степента „доктор**"** изисква кандидатът да докаже, че е навлязъл в областта. В случая Стефан Герджиков не само показва, че е навлязъл в трудна съвременна област, но и е изложил състоянието на проблемите и своите резултати по такъв начин, че да ги направи достъпни за начинаещия читател. Например, авторът разяснява всяка по-сложна стъпка първо на идейно ниво, а след това дава формалните разсъждения. Трудът може успешно да служи за навлизане в тематиката на магистри и докторанти (и дори на бакалаври), което несъмнено е негово достойнство. Това, че е написан на английски език, допълнително разширява кръга от потенциалните читатели.

3. Научни и научноприложни приноси. Съществуват множество алгоритми и техни програмни реализации, работещи в различни ситуации и почиващи на различни идеи, които решават задачата при зададен език по дадена дума да се намерят най-близките до нея думи от езика. Основният проблем е, че при увеличаването на броя на думите в езика и на дължината на съобщенията, нарастват времето за намирането на най-близките думи и обемът на компютърната памет, необходима за решаването на тази задача. Ето защо, задачата за намиране на нови по-бързи алгоритми, използващи по-малко памет, е изключително актуална. Едно от основните постижения на представената дисертация е създаването на такива алгоритми. Работи се с регулярни езици над крайни азбуки. Това са езици, които се разпознават от детерминирани крайни автомати без допълнителна памет. На интуитивно ниво, автоматът има краен брой състояния, едно от които е начално, а няколко са крайни. Когато му се зададе дума, автоматът започва да чете последователно буквите, като на всяка стъпка сменя състоянието си (или спира, защото не може да изпълни операцията). Думите, които могат да се прочетат от автомата така, че на последната стъпка той се намира в крайно състояние, образуват регулярен език. Например, всички крайни езици са регулярни. В дисертацията се поставя естественото допълнително условие, че броят на грешките не надхвърля $q|V|$, където q е фиксирано число в интервала (0,1), а $|V|$ е дължината на полученото съобщение V. При фиксирани регулярен език $\mathcal{L}$ и параметър q, създаденият алгоритъм получава на входа дума V (полученото съобщение), а задава на изхода всички думи от $\mathcal{L}$, от които V може да се получи с не-повече от $q|V|$ грешки. Алгоритъмът почива на няколко основни идеи, които сами по себе си не са нови (на Михов и Шулц, на Наваро и Баеса-Ятес, допълнени с идеи и на други автори), но комбинирането им, заедно с тяхната конкретна реализация, съвсем

не е тривиален проблем. Предаването на съобщението се разглежда като редица от алгебрични операции, наречена подравняване (редица от предаване на букви без грешки, предаване с грешки, добавяне или изтриване на символи, със съответна цена на грешката). Тъй като някои от основните стъпки за решаването на задачата са от експоненциална сложност, дисертантът използва принципа *Divide et impera* (*Разделяй и владей*). Той разделя полученото съобщение на две поддуми и обработва поотделно двете половинки (като втората половинка се чете отзад напред). Този процес на деление продължава и по-нататък, което подобрява чувствително изчислителната сложност. Това налага да се работи не само с целите думи в езика $\mathcal{L}$, но и с техните поддуми (които също образуват регулярен език, със съответен детерминиран автомат, който може да се построи ефективно на базата на автомата, съответен на $\mathcal{L}$). След намирането на всички поддуми U от $\mathcal{L}$, които са достатъчно близки до поддумите на полученото съобщение V, се оставят само истинските думи U. При това, на всяка стъпка дисертантът се стреми да оптимизира пресмятанията, което води до допълнително подобряване на ефективността и икономия на памет. Например, при пресмятането на разстоянието между две думи дисертантът разглежда само тези поддуми, разстоянието между които е достатъчно малко; кандидатите за корекции се генерират само по веднъж, а не по няколко пъти и т.н. Като се използват разнообразни техники, се дава оценка на изчислителната сложност. В случая на крайни езици са направени допълнителни упростявания на предложените алгоритми.

Първите две глави на дисертацията са уводни. Трета глава е много полезна за по-нататъшното изложение. В нея се разглежда пример, който на наивно ниво обяснява как работи основният алгоритъм. В четвърта глава се дава теоретичната обосновка за по-нанататъшните разглеждания – апаратът за подравнявания от думи и списъци от разстояния. Пета глава е посветена на алгоритъма за приближено търсене в регулярно множество от думи спрямо обичайното разстояние на Левенщейн. Глава шеста пренася идеите от предишната глава за случая на обобщено разстояние на Левенщейн. Основният проблем е, че методът за деление на думите на две части не работи веднага. Налага се да се добави допълнителна операция при разделянето, което води и до съответни изменения в разглежданията. Седма глава предлага вероятностен подход, който теоретично аргументира ефективността на алгоритъма. Докато в предишните глави се отчита растоянието между началния и изходния текст, в глава осма се подхожда по принципно нов начин, който представлява нов оригинален подход за дефиниране на близост между думи. Като приложение, тръгвайки от оригинала (написан на съответния за своето време език) се взема най-вероятното съвременното тълкуване. При това част от оригиналния текст се използва за „самообучение" в духа на методите, разработени в предишните глави. Направена е програмна реализация на метода, която е тествана на текстове на български и английски, с различна дъжина и от различни периоди. Резултатите показват предимствата на разработката в сравнение с други известни методи, както и широкия обхват на нейните възможности за приложение.

4. Преценка на публикациите по дисертационния труд: Бройката и качеството на публикациите удовлетворяват изискванията, предявявани във ФМИ на СУ. По темата на дисертацията са публикувани 3 статии. От тях едната статия е самостоятелна в „Доклади на БАН" (с импакт-фактор 0.211 за 2012 г., когато е публикувана статията). Тъй като аз представях тази статия, искам да отбележа, че тя получи много ласкава оценка от рецензента, който е утвърден специалист в областта. Другите две статии са съвместни с

научния ръководител Стоян Михов и други трима съавтори (съответно с Петър Митанкин и Клаус Шулц и с Владислав Ненчев) и са публикувани в трудовете на авторитетни международни конференции в Италия и САЩ. Всички публикации са минали рецензиране преди да излязат. Подробен вариант на едната от двете съвместни статии е качен в Интернет в базата от препринти arXiv, която се следи от цялата математическа колегия по света. Дисертантът много акуратно е отбелязал своите приноси в съвместните публикации. Нямам данни за цитиране на резултатите от дисертацията. Освен това, резултатите от дисертацията са докладвани на международни форуми в Италия, САЩ и Великобритания и на научни сесии на ФМИ на СУ.

5. Мнения, препоръки и бележки: (1) Обикновено най-важните факти в една публикация по математика са формулирани като теореми. Всички твърдения в дисертационния труд са кръстени „Лема", „Твърдение" или „Следствие" и нито едно не е обявено за „Теорема". Мисля, че трудът съдържа достатъчно много сериозни факти, които биха могли да бъдат наречени „Теорема". (2) В параграф 8.3, при сравняването на съществуващите методи с тези, разработени от дисертанта и неговите съавтори, се коментира точността на получените резултати, но не се казва нищо за времето и компютърната памет, необходими за пресмятанията. Такива данни, поне за методите от дисертацията, биха били полезни за допълнителната оценка на ефективността на разработката. (3) Има известен разнобой в цитирането на литературата. Някои от източниците са дадени без страници или без издателство. Това се отнася и за две от статиите по дисертацията. Обичайна практика в англоезичната литература е за статиите на езици, които не използват латиница, да се указва езикът, на който са написани. В дисертацията това не е направено. (4) Авторът би трябвало да обърне по-голямо внимание на езика на автореферата. Някои от изразите са буквален превод от английски и звучат лошо на български (например „Левенщайн разстояние" би трябвало да бъде „разстояние на Левенштейн"). Мисля, че горните недостатъци са лесно отстраними и не развалят общото положително впечатление от дисертацията.

6. Авторефератът и справката за приносите са написани достатъчно подробно и дават ясна и адекватна представа за съдържанието и основните резултати на дисертацията.

Заключение: Представеният дисертационен труд е в актуална област. Той е на високо образователно и научно ниво и удовлетворява всички изисквания, поставени пред един дисертационен труд в областта на математиката и нейните приложения. Убедено препоръчвам на почитаемото Научно жури да присъди на ас. Стефан Владимиров Герджиков образователната и научна степен „доктор" в професионално направление 4.5. Математика, научна специалност „Математическа логика".

София, 3 януари 2014 г. Рецензент:

(акад. д.м.н. В. Дренски)