Р Е Ц Е Н З И Я

относно кандидатурата на доц. дн Стоян Милков Михов за участие в конкурс за заемане на академичната длъжност „професор“ по професионално направление 4.6. Информатика и компютърни науки, специалност 01.01.12. Информатика

от проф. дмн Галя Ангелова, ИИКТ-БАН

Конкурсът е обявен в „Държавен вестник” бр. 45 (от 28 май 2021 г.) за нуждите на ИИКТ-БАН, секция „Изкуствен интелект и езикови технологии”. Единствен кандидат за конкурса е доц. дн Стоян Милков Михов. Съгласно регламента на *Правилника за заемане на академични длъжности в ИИКТ-БАН* относно изпълнение на минималните изискавания за заемане на длъжността, кандидатите за академичната длъжност „професор” трябва да имат поне 50 точки по показател А, 100 точки по показател В, 260 точки по показател Г, 140 точки по показател Д и 150 точки по показател Е. Доц. Михов представя попълнена справка за НАЦИД, която съдържа 50 точки по показател А, 100 точки по показател В (от монография, отпечатана в Cambridge University Press през 2019 г.), 470 точки по показател Г, 1704 точки по показател Д от цитирания и 318 точки по показател Е. Доц. Михов има над 30 години трудов стаж по специалността - като научен работник и преподавател по информатика във висши учебни заведения, както и диплома на ВАК от 2000 г. за образователната и научна степен „доктор” по научната специалност 01.01.12 Информатика и диплома от ИИКТ-БАН за присъждане на научната степен „доктор на науките“ през 2020 г. С това формалните изисквания на Правилника са не само изпълнени, но и значително надхвърлени особено при показател Д за цитирания.

Кратки биографични данни за кандидата

Доц. Михов завършва магистратура във ФМИ на СУ „Кл. Охридски” през 1993 г., като защитава дипломна работа на тема „Унификация на корегулярни множества“ с ръководител доц. Анатолий Буда. През 2000 г. защитава в ИИКТ-БАН кандидатска (по сегашната терминология – докторска) дисертация на тема „Минимални ациклични автомати: конструкции, алгоритми, приложения“ с ръководител проф. дмн Димитър Скордев. Става доцент по информатика в ИИКТ-БАН през 2006 г. в секция „Лингвистично моделиране“ (сега „Изкуствен интелект и езикови технологии”). През 2020 г. защитава в ИИКТ-БАН дисертация на тема „Крайни автомати, трансдюсери и бимашини: алгоритмични конструкции и приложения“ за получаване на научната степен „доктор на науките“. През повечето години на професионална активност съвместява научната работа в ИИКТ-БАН, свързана с участие в различни задачи и ръководство на множество проекти, с редовна преподавателска дейност в СУ „Св. Кл. Охридски” и индустриални изследвания в български фирми. В последните години научните му интереси към задълбочени теоретични изследвания и създаване на практически приложения позволиха в ИИКТ да се създаде устойчива група от млади учени за изследвания по теория на крайните автомати и разработка на системи за приближено търсене, както и синтез и разпознаване на реч.

Общо описание на представените материали за процедурата

За конкурса са представени 17 научни статии (16 от тях индексирани в Scopus и/или Web of Science, както и една в архив) и един патент, регистриран в САЩ. Всички материали са в съавторство и характеризират съвместната работа на кандидата с млади учени, сътрудници от СУ „Св. Кл. Охридски", чуждестранни колеги (предимно от университета „Лудвиг Максимилиан" в Мюнхен) и индустриални партньори. Съавторстовото в публикациите не намалява значението на постиженията на доц. Михов, а по-скоро подчертава важността на неговата позиция като ценèн и търсен сътрудник, партньор и ръководител.

Списъкът с цитиранията съдържа 213 цитата към шест от най-популярните статии на кандидата:

- за „Fast string correction with Levenshtein automata" са представени 93 цитирания,
- за „Fast approximate search in large dictionaries" са представени 46 цитирания,
- за „Incremental construction of minimal acyclic finite-state automata" са представени 30 цитирания,
- за „Orthographic errors in Web pages: Toward cleaner Web corpora" са представени 23 цитирания,
- за „Lexical postcorrection of OCR-results: The web as a dynamic secondary dictionary?" са представени 12 цитирания, и
- за „Adaptive Text Correction with Web-Crawled Domain-Dependent Dictionaries" са представени 9 цитирания.

Всички тези статии са посветени на ефективна обработка на низове (думи) и съдържащи ги речници с прилагане на теорията на крайните автомати.

Под ръководството на доц. Михов има двама защитили докторанти – Стефан Герджиков (защитил през 2014 г. с дисертация на тема „Ефективни алгоритми за приближено търсене в регулярни множества", който вече е доцент във ФМИ на СУ „Св. Кл. Охридски") и Петър Митанкин (защитил през 2010 г. с дисертация на тема „Универсални автомати за ефективно определяне на близост между символни низове", в момента старши сътрудник в „Онтотекст"). Понастоящем доц. Михов е ръководител на Георги Шопов, редовен докторант на ИИКТ.

В справката, подготвена за НАЦИД, при показател Е са споменати участието на доц. Михов в проекта АКОМИН (2012-2016 г.), ръководство на българския екип в проекта ИМПАКТ (Improving Access to Text), както и значителни привлечени средстава по проекти с фирми H-TECH ЕООД и СТАТСОФТ ЕООД. Дейността на кандидата в тези проекти показва неговия стремеж към систематично натрупване на ресурси и програмен код за автоматична обработка на текст, реч и аудио:

- в проекта АКОМИН беше създаден фонетичен корпус на българския език, с използване на изградената (по предложение на доц. Михов) модерна звукозаписна лаборатория;
- в проекта ИМПАКТ беше създаден исторически речник на българския език през края XIX век, езиков корпус на българския език през края на XIX век и специализирана OCR система за разпознаване на стара кирилица (съвместно с фирмата ABBYY създала FineReader).
- в проектите с фирма H-TECH е създаден индустриален софтуер за приближено

търсене на подобни аудиозаписи, който се използва за идентификация на излъчвания на телевизионни реклами в реално време;

- На фирмата СТАТСОФТ ЕООД беше предоставен български фонетичен корпус, съдържащ записи на реч анотирани на фонетично ниво. Корпусът съдържа 21891 записа направени от 140 диктора върху 319 фонетично богати изречения на български език.

Научни резултати и приноси в материалите за конкурса

Тематиката и постиженията, описани на представените 17 статии и един патент, могат да бъдат групирани в пет направления:

- **Теоретични** (статия 2 – предлага алтернативен принцип за построяване на бимашини, наречен принцип за акумулиране на изравнители. Предложената конструкция има пространствена сложност, близка до оптималната, и може да се прилага за широк клас рационални функции; статия 3 – въвежда методология за конструиране на вероятностни модели с подпоследователни преобразуватели с преходи при неуспех, предлага ефективни алгоритми за композиция и канонизиране, и показва полезни приложения на алгоритмите например при разпознаване на реч).
- **Корекция на текст с използване на крайни автомати** (статии 1, 4, 9, 10, 11, 12, 16 и 17 – предлага се метод с оптимална ефективност за презаписване на текст чрез конструиране на специфичен подпоследователен преобразовател за линейно време; представен е алгоритъм за преобразуване на даден контекстен речник за презаписване като детерминиран подпоследователен преобразовател, при който се използват преходи при неуспех за намаляване на размера /f-преобразовател/, като е предложен и алгоритъм за композиране на f-преобразователи; разработени са методи за ефективно откриване на правописни грешки от различен тип в Интернет и маркирането им с използване на бази от езикови знания, автоматично извлечени от мрежата; предложен е ефективен метод за подбор на кандидати за лексикална корекция на сгрешени думи в текст чрез използване на универсален Левенщайн автомат с усъвършенствано основно разстояние; показан е подход за автоматично изчисляване на профила на грешки, които се нуждаят от правописна корекция, и адаптивно подбиране на кандидати за корекция; показан е начин за извличане от Интернет на речници и езикови модели, които могат да се използват за коригиране на входен текст; описан е нов подход за извличане на вариации на правописа от списък с примери и алгоритъм за корекция, който предлага и ранкира кандидатите за корекция на определена грешна дума; представен е и нов общ и езиковонезависим подход за корекция на текстове, разработен на основата на функционалните автомати в проекта CULTURA);
- **Приближено търсене с използване на крайни автомати** (статии 13, 14 и 15 – представени са ефективни методи и алгоритми за търсене на близост на входен шаблон в голяма база данни; предложена е процедура за приближено търсене в електронен речник, организирано като подтърсения, които винаги започват с точно частично съвпадение където подниз от входния шаблон е подравнен с подниз от речниковата дума и след това частичното съвпадение се разширява постепенно до по-големи поднизове; представена е и системата WallBreaker за търсене на близост,

която спечели състезание за ефективно търсене на низове през 2013 г.);

- **Обработка на реч** (статии 5, 6, 7 и 8 – представени са последователни версии на система за разпознаване на непрекъсната българска реч от първата версия през 2009 г., с подобряване на производителността чрез частично компилиране на решетката на думата като детерминиран краен автомат през 2016, дизайн и съдържание на речевата база данни BulPhonC; изграждане на речев корпус BG-PARLAMA от записите на пленарните сесии на българския парламент през 2019).
- **Метод за автоматичен анализ на взаимодействие между инфлуенсъри в социални мрежи** – патент 18.

Тематиката на представените статии характеризира широката област на изследвания на кандидата, както и разнообразието от разглеждани задачи и приложения, както и множеството сътрудници в различни проекти и разработки.

Цялостен поглед върху постиженията на кандидата

Доц. Стоян Михов е световно-известен специалист по използване на крайните автомати в компютърната лингвистика (т.е. като инструментариум при автоматична обработка на език и реч) и приближено търсене (както в низове, така и в други ресурси като напр. аудиозаписи). Често приложните му резултати се основават върху оригинални приноси към теорията на крайните автомати. Доц. Михов има над 60 научни труда, повечето от които са индексирани в Scopus и/или WoS. В представената автобиография той посочва 426 цитирания на неговите трудове в Scopus, но в Google scholar те са 1148 като голяма част от тях не са автоцитирания.

Още през 2000 г. в докторската си дисертация доц. Михов предложи алгоритъм за директно построяване на минимален ацикличен краен автомат по речник от думи, зададени в лексикографски нареден списък, и публикува този резултат в списание Computational Linguistics като статия, която се цитира и до днес. Дългогодишната му работа с различни видове автомати, непрекъснатият стремеж към съчетаване на теоретичните обосновки с практически приложения му позволиха да предложи цялостно изложение на теорията на крайните автомати от абстрактна алгебрична гледна точка, което е построено с оглед въвеждане и изследване на изчислително-ефективни конструкции. Тези резултати са представени в монографията, публикувана в Cambridge University Press през 2019 г.

На база на тази монография през 2020 г. доц. Михов защити дисертация за присъждане на научната степен „доктор на науките“ със следните теоретични приноси: разработен метод за тестване за ограничена вариация за краен преобразувател, който се интегрира в конструкцията за секвенциализация; предложен нов алгоритъм с полиномиална сложност за канонизация на подпоследователен преобразувател; предложена нова ефективна конструкция за построяване на бимашина от краен преобразувател; разработена е конструкция за директно композиране на бимашини с доказателство за коректност. Като приложен резултат на дисертацията е представен езикът за програмиране C(M) и софтуерна библиотека от 45 програми на C(M) за построяване на автоматни конструкции и приложения в реални изчислителни задачи.

Доц. Михов е ръководител на колектива, създал програмата Wallbreaker, спечелила световно състезание за приближено търсене през 2013 г. (https://www2.informatik.hu-

berlin.de/~leser/searchjoincompetition2013/Results.html).

През последните години под негово ръководство с много труд, търпение и превъзмогване на затруднения с финансирането, беше създаден прототип на диктофон за българска реч (система speech-to-text). Разработката започна над корпус от парламентарна реч като първо учебно множество данни и продължава с обучение на системата над записи на медицинска реч, които се изготвят по Националната научна програма еЗдраве. През последните няколко месеца Съюзът на слепите инвестира в разработка за синтез на българска реч от текст (система text-to speech), която да бъде осъществена от групата в ИИКТ. Благодарение на постоянството и усилията на доц. Михов, в момента ИИКТ може да планира създаване на гласови интерфейси на български език за различни системи.

Лични впечатления

Познавам кандидата от много години, откакто той постъпи в ИИКТ като редовен докторант. Освен таланта, голямо впечатление правят неговата самостоятелност и всеотдайност. Със собствени усилия той успя да привлече млади хора и да изгради група от учени, разработващи по оригинален начин теорията на крайните автомати и нейните приложения за обработка на естествен език. Неговата активна и непрестанна преподавателска дейност привлича студенти, които се интересуват от задълбочени научни изследвания. Впечатлена съм и от търпението, с което доц. Михов планира и изпълнява договори за индустриални проекти.

Заключение

Считам, че доц. Стоян Михов е рядък пример на талантлив математик, който е истински заинтересован от създаване на реални системи и е готов да работи като професионален програмист за тяхното изграждане. Представените за конкурса материали доказват наличието на задълбочени знания, водеща роля при формулиране на амбициозни изследователски цели, способност за работа в екип, както и постоянство, прецизност и стремеж към достигане на световното ниво. Количеството и качеството на неговите статии от по-ранен етап, които са представени за конкурса, и техните цитати показват, че той е разпознат от международната научна общност като водещ изследовател с оригинални идеи от около 15 години. Проектите, финансирани от фирма HTech, доказват капацитета му да създава индустриален софтуер за прилагане на генерираните идеи. **Подкрепям убедено избора на доц. дн Стоян Михов за професор в секция „Изкуствен интелект и езикови технологии" на ИИКТ-БАН и предлагам на уважаемите членове на Научното жури единодушно да гласуват в подкрепа на такова решение.**

27 септември 2021 г.
София

Член на Научното жури за процедурата:

проф. дмн Галя Ангелова