ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – СОФИЯ
Факултет Електроника и Автоматика
Катедра Компютърни Системи и Технологии

маг. инж. Милена Цветанова Ангелова

ИНТЕЛИГЕНТНИ ПОДХОДИ ЗА ПРОФИЛИРАНЕ, АНАЛИЗ И ЛОКАЛИЗИРАНЕ НА ЕКСПЕРТИЗА ОТ ПУБЛИЧНО ДОСТЪПНИ ИЗТОЧНИЦИ

А В Т О Р Е Ф Е Р А Т

на дисертация за придобиване на образователна и научна степен
"ДОКТОР"

Област: 5. Технически науки
(шифър и наименование)

Професионално направление: 5.3 Комуникационна и компютърна техника
(шифър и наименование на направлението в посочената област)

Научна специалност: 02.21.05 Системи с изкуствен интелект
(наименование на научната специалност)

Научен ръководител: Проф. д-р Веселка Боева и д-р Елена Ципоркова

СОФИЯ, 2020 г.

Дисертационният труд е обсъден и насочен за защита от Катедрения съвет на катедра „Компютърни системи и технологии“ към Факултет Електроника и Автоматика на ТУ-София на редовно заседание, проведено на 16.07.2020 г..

Публичната защита на дисертационния труд ще се състои на 18.11.2020 г. от 13,00 часа в зала 4425 в IV корпус на Технически университет – София, филиал Пловдив на открито заседание на научното жури, определено със заповед № ОЖ-5.3-22 / 21.07.2020 г. на Ректора на ТУ-София в състав:

1. проф. д-р инж. Михайл Петров– председател
2. проф. д-р Веселка Боева – научен секретар
3. проф. д-р Георги Тотков
4. проф. д-р Кольо Онков
5. доц. д-р Елена Сомова

Рецензенти:

1. доц. д-р инж. Диляна Будакова
2. проф. д.т.н. инж. Тодор Стоилов

Материалите по защитата са на разположение на интересуващите се в канцеларията на Факултет Електроника и Автоматика на ТУ-София, Филиал Пловдив кабинет № 4242.

Дисертантът е редовен/задочен докторант към катедра „Компютърни системи и технологии“ на факултет Електроника и Автоматика. Изследванията по дисертационната разработка са направени от автора.

Автор: маг. инж. Милена Ангелова
Заглавие: Интелигентни подходи за профилиране, анализ и локализиране на експертиза от публично достъпни източници
Тираж: 30 броя
Отпечатано в Пловдив Принт Дизайн ЕООД

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

Разработването на алгоритми и техники, които да профилират, анализират и локализират експертиза се явява актуален проблем, тъй като организациите в различни области имат нужда от системи, които да им помагат в ефективното управление на техния наличен човешки потенциал. Необходимостта от такива софтуерни системи допълнително се засилва от големия обем информация, която се събира и генерира в организациите в наши дни. Цялата тази информация може да бъде полезна за съответната организация чрез подходящ анализ и органицация, които да улеснят бързото намиране на хора с търсената комптентност и умения. На първо място е проблемът с групирането и представянето на наличната информация, както и оценяването на нивото на експертните познания. Като последното е времеотнемащ процес, който не е съвсем автоматизиран в повечето организации. Наличието на новопостъпващи данни периодично довежда до нуждата от системи, които да могат да адаптират по-гъвкаво получените данни.

Цел на дисертационния труд, основни задачи и методи за изследване

Цел: Да се разработят и реализират нови подходи за профилиране, анализ и локализиране на експертиза като се използват публично достъпни източници. За постигането на тази цел са поставени следните научно-изследователски задачи:

1. Улесняване на търсенето на експертиза в дадена област чрез йерархична организация на експертите по тематични направления на базата на налична онлайн-достъпна информация относно тяхната компетентност.
2. Разработване на техники за количествена оценка на нивата на експертност и използване на тези оценки за по-прецизно представяне и локализиране на експертиза.
3. Разработване на техники за организиране на експертизата по тематични направления и адаптиране на съществуваща организация при постъпване на нова информация.
4. Оценяване и демонстриране на разработените техники за представяне, локализиране и анализ на експертиза върху тестови набори от данни извлечени от различни онлайн-достъпни източници.

Научна новост

Разработен е нов алгоритъм за йерархично организиране на експерти от дадена експертна област в тематични направления. Предложена е и нова техника за по-прецизно описание и представяне на експертните познания при използване на количествена оценка на нивото на експертиза на експертите от дадена предметна област. Разработен е

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

протип на препоръчваща система в сферата на академичните среди. Предложени са и два еволюционни клъстериращи алгоритъма, които гъвкаво адаптират съществуващо клъстериращо решение чрез новопостъпили данни.

Практическа приложимост

Разработените алгоритми и техники могат да се използват в различни предметни области за организиране на наличната експертиза и улесняване на търсенето и обновяването. Това следва да покаже тяхната гъвкавост и възможност да се интегрират и с други техники. Биха били подходящи както за големи масиви от данни така и за по-малки. Като приложение в организациите, техниките биха могли да се използват за групиране на експертите по тематични области и след това да се приложи търсене.

Апробация

Основните резултати от проведените изследвания в дисертацята са докладвани и обсъдени на нучни форуми, национални и международни конференции: In 8th IEEE International Conference on Intelligent Systems, 2016, Sofia, Bulgaria; 9th International Conference on Agents and Artificial Intelligence 2017, Porto, Portugal; 10th International Conference on Agents and Artificial Intelligence, 2018, Funchal, Madeira, Portugal; The 22nd International Conference on Electronic Publishing-Revised Selected Papers, 2018, Canada, Toronto; 30th Annual Workshop of the Swedish Artificial Intelligence Society, SAIS 2017.

Публикации

Резултатите от дисертационният труд са отразени в 11 публикации, от които: 2 глави в книги, 1 в научно списание, която е самостоятелна, 8 на международни конференции и три абстракта.

Структура и обем на дисертационният труд

Дисертационният труд е в обем от **157** страници, като включва увод, **4** глави за решаване на формулираните основни задачи, списък на основните приноси, списък на публикациите по дисертацията и използвана литература. Цитирани са общо **150** литературни източници, като всички са на латиница, част от тях са интернет адреси. Работата включва общо **13** фигури и **18** таблици. Номерата на фигурите и таблиците в автореферата съответстват на тези в дисертационния труд.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ГЛАВА 1. ОТ ИЗВЛИЧАНЕ НА ДОКУМЕНТИ ДО ИЗВЛИЧАНЕ НА ЕКСПЕРТИЗА

1.1 Извличане на документи (Document Retrieval)

Извличането на документи се определя като извличане на списък, които е свързван с въведена потребителска заявка. Основно ако трябва да опишем как работят тези системи, трябва да споменем, че те имат две основни задачи, които следва да бъдат изпълнени: (1) търсене на документи, които да съответстват на зададената заявка от потребителя и (2) поставяне на тегла на получените резултати, с цел да резултатите да могат да се подредят във възходящ ред на подобие. Под документи трябва да се подразбира всякакъв неструктуриран текст като вестници, статии и уеб страници, или структуриран като имейл съобщения. Потребителските заявки може да варират от няколко ключови думи до цели изречения, с цел да се опише по-точно търсената информация. Днес, Интернет търсачките са се превърнали в класически инструменти за търсене и извличане на информация под формата на документи. Водещите уеб търсачки като Google или Yahoo, осигурят ефективен достъп до милиарди уеб страници. Нещо повече, ако се върнем назад във времето, търсенето в библиотеките е било на физическо ниво, докато сега в днешно време съществуват новоразработени системи, които работят по традиционният начин - чрез търсене на информация по съвпадение или такива, които използват интелигентно търсене базирани на системи за извличане на знания.

1.2 Извличане на експертиза (Expertise Retrieval)

Някои от най-ценните знания в една организация са тези на нейните служители. Необходимо е предприятията да комбинират цифрова информация със знанията и опита на служителите си. Организациите може да имат много ценни експерти, които да са разпръснати географски. Независимо от това, че повечето организации позволяват на техните служители да работят от различни части на света, това не би попречило на тяхното сътрудничество и доставяне на нужният ресурс независимо, къде се намират. Най-ефективният начин за обмен на знания е контактът - човек с човек. Все пак намирането на подходящия човек, с който да се свържем е нещо, което с информационните технологии може да се постигне. Разглеждането на проблема с определянето на експертни познания в една организация довежда до създаването на клас от търсещи машини, известни като търсене на експерти.

Ранните подходи на този тип експертно търсене са използвали база данни, съдържаща описание на уменията на хората в дадена организация. Изясняването и създаването на такава информация за всеки отделен човек в организацията се явява скъпо струваща задача и разбира се времеотнемаща. Статичният характер на базите често ги прави непълни и по-трудни за работа. Нещо повече, заявките за търсене на експерти обикновено

са фини и специфични, но описанията на експертизите са тенденциозно насочени към по-общ вид; следователно стандартният начин на представяне на експертните познания няма да е подходящ за системите за търсене на експертиза.
За справяне с тези недостатъци са предложени редица системи, насочени към автоматично откриване на актуална експертна информация от вторични източници. Обикновено това се извършва в специфични тематични области, например в контекста на разработката на софтуерно инженерство.

ГЛАВА 2. МЕТОДИ И ТЕХНИКИ ЗА ОБРАБОТКА НА ЕСТЕСТВЕН ЕЗИК И АНАЛИЗ НА ДАННИ

В тази глава са разгледани някои техники за обработка на естествен език и направен преглед на клъстериращите техники като повече внимание е отделено на тези използвани в настоящия дисертационен труд. Техниките за обработка на естествен език предлагат един друг прочит в представянето на експертните познания. Те дават възможност да се извлекат група от думи най-често срещани от някакъв вид документ и след това да се приложи техника за обобщаването на думите от една и съща тема. Въз основа на думите, групирани заедно, темите могат да бъдат изведени от групите думи. Въз основа на създадените клъстери от думи, следващата стъпка да се групират хората по създадените клъстери от думи, така те ще са групирани по сходимост в определени области. Друг подход е чрез използване на клъстериращи методи, които да групират експертите описани чрез своите експертни профили в различни тематични области спрямо тяхната сходимост.

2.1 Тематично моделиране на естествен език

Тематичното моделиране е техника на машинното обучение, която автоматично анализира текстовите данни за да определи клъстер от думи върху множество от документи. Известно е още като обучение без учител, тъй като не се изисква предварително определен списък от теми или данни за обучение, които преди са били класифицирани от хора. Моделирането на теми е бърз и ефикасен начин за обобщаване и индексиране на големи обеми от текст, като се използват много малко думи. Това означава, че има много голяма вероятност думите свързани с една и съща тема да се появят в един и същ документ, отколкото в други. Когато работим с тематично моделиране не е сигурно, че всяка една дума от списък с думи за дадена тема ще са свързани една с друга. Съществуват различни алгоритми, в чиято основа стои тематично моделиране. Ще разгледаме част от тях в настоящата секция:

- Латентен Семантичен анализ (Latent Semantic Analysis, LSA) е техника в обработката на естествен език за анализ на връзките между множество от документи и термините, които съдържат. Като резултат се генерира списък от понятия свързани с документите и термините. LSA предполага, че думите, които са близки по значение, ще се появят в подобен по смисъл текст.
- Вероятностен латентен семантичен анализ (Probabilistic Latent Semantic Analysis, pLSA) е техника използвана за моделиране на информация в рамките на вероятността рамка. Нарича се латентен, защото темите се третират като скрити или скрити променливи.
- Латентно разпределение на Дирихле (Latent Dirichlet allocation, LDA) осигурява рамка за анализ на голям обем от данни, при които наблюденията се събират в групи. LDA е най-често срещаният тип тематичен модел, тъй като той разширява pLSA и адресира проблемите, които възникват при него.

2.2 Клъстериращи методи

Клъстерирането е важна техника за откриване на относително плътни подрегиони или подпространства на многоизмерно разпределение на данни. Клъстерирането се използва при извличане на информация за много различни цели, като разширяване на заявки, групиране на експерти, индексиране на полученият резултат и визуализация на резултатите от търсенето. В областта на извличането на информация, клъстерирането свързва с подобряване на ефективността и ефикасността на процеса на извличане на информация.

Съществуват различни методи за клъстериране и всеки един от тях си има своите предимства и недостатъци. От съществено значение е, кой алгоритъм ще бъде избран спрямо избраните данни, така, че клъстериращото решение да бъде оптимално. Най-често използваните клъстериращи техники са йерархични клъстериращи техники, мрежовидни, k-means, k-medoids, които са използвани в дисертационният труд.

2.3 Клъстер-валидиращи техники

Терминът валидиране на клъстери се използва при оценяване на резултатите на даден клъстериращ алгоритъм. Разделени са в следните категории:

- **Клъстер-валидиращи техники базирани на вътрешен критерии**, които при процесът на валидиране използват само с наличните данни, за да оценят до колко е добро клъстерирането. Може да се използват и като инструмент за търсене на оптималният брой на клъстерите. Техника използвана в настоящият дисертационнен труд е Silhouette Index.

- **Клъстер-валидиращи техники базирани на външен критерии**, които сравняват резултатите от клъстериращия анализ с външно известен резултат, като например предоставени етикети на класовете. При тези критерии се измерва степента на съвпадение на етикетите на клъстерите и външно представените. Тъй като предварително съществува информация за етикетите на клъстерите, този подход, може да се използва и за намирането на правилният алгоритъм за клъстериране на конкретен набор от данни. Техники използвани в дисертационният труд: F-Measure, Jaccard Index.
- **Относителни клъстер-валидиращи техники**, които оценяват структурата на клъстерирането чрез промяна на различни стойности на параметрите за един и същ алгоритъм (например: промяна на броя на клъстерите k). Обикновено се използва за определяне на оптималния брой клъстери.

ГЛАВА 3. ПРЕДЛОЖЕНИ НОВИ ТЕХНИКИ ЗА ПРОФИЛИРАНЕ И ИЗВЛИЧАНЕ НА ЕКСПЕРТИЗА

3.1 Извличане на група от тематични експерти с Формален Концептуален Анализ

Формалният концептуален анализ е математически формализъм, позволяващ да се изведе концептуална решетка от формален контекст, съставен от множество от обекти O, множество от атрибути A и двоична връзка, дефинирана върху декартовото произведение $O \times A$. Контекстът е описан като таблица, редовете съответстват на обекти, а колоните - на атрибути или свойства, а знакът кръст в клетката на таблицата означава, че „обектът притежава свойство“. Концептуалната решетка е съставена от формални понятия или просто концепции, организирани в йерархия чрез частично подреждане. Интуитивно понятието е двойка (X, Y), където $X \subseteq O$, $Y \subseteq A$ и X е максималният набор от обекти, споделящи целия набор от атрибути в Y и обратно. Множеството X се нарича обхват, а множеството Y - цел на концепцията (X, Y). Съотношението между концепциите се дефинира, както следва: $(X_1, Y_1) \prec (X_2, Y_2) \Leftrightarrow X_1 \subseteq X_2 \, (Y_1 \subseteq Y_2)$. Разчитайки на това съотношение на релацията $(X_1, Y_1) \prec (X_2, Y_2)$, множеството от всички понятия, извлечени от контекста, е организиран в рамките на цялостна решетка, което означава, че за всеки набор от понятия има най-малката супер концепция (superconcept) и най-голямата под концепция (subconcept), наречен решетка от концепции.

Изграждане на експертните профили Определяме профила на експерта като списък от ключови думи, извлечени от свободно достъпна информация за самите експерти. Ключовите думи описват неговата/нейната експертиза. Предполагаме, че n на брой различни експертни профили са създадени и всеки експерт i $(i = 1, 2, ..., n)$ е представен като списък от ключови думи p_i.

Семантично подобие (Semantic similarity) Семантичното подобие се определя като, търсене на подобие между областта на компетентност на даден експерт и търсената заявка или търсене на подобие между два експерта, тоест доколко те могат да имат близка експертиза. Семантичното подобие се представя като:

$$S_{ij} = \sum_{l=1}^{p_i} \sum_{m=1}^{p_j} s(k_{il}, k_{jm}), \tag{3.1}$$

където $s(k_{il}, k_{jm})$ е семантичното подбие между ключовите думи k_{il} и k_{jm}. Ако експертните профили са описани не само от списък с ключови думи, но е добавена и информация за количествена оценка на придобитото знание. В такъв случай, наличната информация позволява да се разшири формулата за търсене на семантично подобие в следният вид:

$$S_{ij} = \sum_{l=1}^{p_i} \sum_{m=1}^{p_j} W_{lm} s(k_{il}, k_{jm}), \tag{3.2}$$

където W_{lm} е количествената оценка отнасяща се към семантичното подобие $s(k_{il}, k_{jm})$ между ключовите думи k_{il} и k_{jm}.

Групиране на експерти Областта на знание може да се представи чрез k на брой основни категории $C_1, C_2, ..., C_k$. Нека да отбележим b_{ij} броят на ключовите думи от профила на експерт i, който принадлежи на категорията C_j. Така всеки експерт i $(i = 1, 2, ..., n)$ може да се представи чрез вектор $e_i = (e_{i1}, e_{e2}, ..., e_{ik})$, където $e_{ij} = \frac{b_{ij}}{p_i} (j = 1, 2, ..., k)$ и p_i е общият бройна ключовите думи на експертният профил. Така всеки един експерт i е представен като вектор с дължина k, като векторът съдържа изчислени стойности, които определят степента на познание на една или повече от различните k категории.

3.2 Подход базиран на количествена оценка на нивото на експертиза

Подходът базиран на количествена оценка на нивото на експертиза, е иновативен подход, които разширява предходният (Формален Концептуален анализ) като добавя допълнителна функционалност. Този метод разширява експертните профили, като добавя допълнителна количествена оценка на всяка експертна област на компетентност на всеки експерт. Като целта е да се представи профила на експерта чрез описание на темите, по които той/тя е експерт, плюс оценка на нивото на знания или опит, които той/тя има по различните теми. Важен въпрос, който стои на преден план е по какъв начин да се търси оценка на сходството на експертните знания между експертите.

Оценяване на експертиза. При този подход, използваме тегла, за да получим достъп до относителните нива на знания или опит, които даден индивид има по определени

теми, за които той/тя е показал, че има експертиза. Да предположим, че се използва метод за количествена оценка, подходящ за съответната област и като резултат за всяка ключова дума k_{ij} от експертният профил i ($i = 1, ..., n$) е свързана с тегло w_{ij}, изразяващо относителното ниво на експертиза което притежава въпросният експерт за ключова дума k_{ij}, тоест $\sum_{j=1}^{p_i} w_{ij} = 1$ и $w_{ij} \in (0, 1]$ за $i = 1, ..., n$. По този начин всеки експертен профил се описва чрез темите, в които е експерт плюс нивата на знания или опит, които има в различните теми.

Подобие на експертиза. Нека s да бъде мярка за сходство, която да е подходяща за оценка на семантичната свързаност между всяка двойка ключови думи, използвани за описване на експертните профили в разглежданата област. Тогава подобието на експертизата S_{ij} на двата експертни профила i и j, може да се дефинира чрез използване на средното претеглено семантично сходство между съответните ключови думи:

$$S_{ij} = \sum_{l=1}^{p_i} \sum_{m=1}^{p_j} W_{lm} \cdot s\left(k_{il}, k_{jm}\right), \tag{3.3}$$

където $W_{lm} = w_{il} \cdot w_{jm}$ е тежестта, свързана със семантичното сходство $s\left(k_{il}, k_{il}\right)$ между ключови думи k_{il} и k_{jm}, и $W_{lm} \in (0, 1]$ за $l = 1, ..., p_i$ и $m = 1, ..., p_j$. Лесно може да се изведе и че $\sum_{l=1}^{p_i} \sum_{m=1}^{p_j} W_{lm} = 1$.

Търсене на експерти. Потребителят може да търси експерти, например по име или по зададена една или няколко области на знание, системата би върнала списък с подходящи експерти близки до търсената заявка. За да може да се изгради такава група от експерти, следва да се намери подобието между заявката, например примерен експерт i и списък от експерти, които имат близка експертиза до неговата. Експертният профил j ще бъде включен към списъкът от подобни експерти на експерт i, ако следното неравенство $S_{ij} \leq T$ е валидно, където $T \in (0, 1)$ е предварително определен праг. Идентифицираните експерти могат да бъдат подредени по отношение на сходството им спрямо тяхната експертиза с примерният експерт i.

3.3 Система за препоръчване на експерти базирана на DiVA данни

В текущата работа е представена система, която дава препоръка при търсене на ръководител за дипломна теза спрямо неговата експертиза, която е изградена върху данни, извлечени от институционалната система на хранилище DiVA (Digital Scientific Archive) на шведският университет Blekinge Institute of Technology (BTH). DiVA е платформа

за публикации и архивиране на изследователски публикации и есета за студенти, използвани от 46 публично финансирани университети и власти в Швеция и останалите Скандинавски държави (www.diva-portal.org).

Профилиране на експертиза. Както беше дискутирано в 3.1 експертните профили се представят като двойка съдържаща лична информация за експерта и информация описваща неговата сфера на компетентност с ключови думи. Наличието на лична информация спомага за унифицирането на експертните профили в резултатното множество. Като допълнение към последното, изчислява се сходство между двойките експерти за да може да се определи дали те да бъдат обединени в един общ експертен профил. По този начин, се сформира множество от уникални експерти.

Подобие на експертиза. Използва се модифицирана мярка на Wu и Palmer, за да се изчисли сходството между два онтологични елемента. В разглеждания контекст всички профили на експерти се описват от списък с термини, специфични за домейна, където той/тя е експерт. Да приемем, че всеки профил на експерт i е описан от списък с ключови думи P_j. В нашите експерименти използваме измерената мярка Wu и Palmer, която се определя както следва: $s = \frac{\left(2N.e^{\frac{-\lambda L}{D}}\right)}{N1+N2}$, където L е най-краткото разстояние между две дадени понятия, D е дълбочината на онтологията, N е разстоянието от най-малко разпространения прародител до корена, $N1$ и $N2$ са съответно, разстоянията от двете разглеждани понятия до корена, и λ е коефициент, който е 0, когато понятията са в една и съща йерархия и 1, когато са съседни понятия. Изменената мярка на Wu и Palmer (отбелязана с s тук) намира най-краткия път между две концепции в онтологичното дърво и дълбочината на цялата онтология. Тогава приликата на експертизата S_{ij} между два експертни профила i и j $(i \neq j)$, могат да бъдат определени чрез средно аритметичното от семантични сходства между съответните ключови думи, тоест

$$S_{ij} = \frac{1}{p_i.p_j} \sum_{l=1}^{p_i} \sum_{m=1}^{p_j} s\left(k_{il}, k_{jm}\right), \tag{3.4}$$

където $s\left(k_{ij}, k_{jm}\right)$ е семантичното подобие между ключовите думи k_{ij} и k_{jm}.

Търсене на експерт. Задачата за търсене на експерт(и) може да се разглежда като задача за попълване на списък, тоест потребителят трябва да предостави малък брой примерни експерти, които в миналото са били използвани за работа по подобни проблеми, а системата трябва да върне експерти със сходна област на компетентност. В нашия контекст потребителят може да бъде студент, директор на учебна програма или друг административен персонал, а върнатите експерти се препоръчват като дипломни ръководители.

Описание на онтологията. Определихме нашата онтология като формализация на понятията и отношенията между тях. Тематичните концепциите са описани от класове и релациите между тях. Моделът в това проучване е създаден с помощта на Национални тематични категории в DiVA, което е стандарт за шведското изброяване на изследователски теми 2017 и класификационни термини от извлечени данни. Националните тематични категории в DiVA са разделени на три нива. А именно основните категории (най-високо ниво) са селскостопанските науки, инженерството и технологиите, хуманитарните науки, медицинските и здравните науки, естествените науки и социалните науки. Тези основни категории са разположени на първото ниво на онтологията. Останалите две нива са описани от под-предметни области, които принадлежат към всяка от основните категории. Второто ниво има 26 предметни категории, а третото ниво се състои съответно от 257 категории. От извлечените данни сме взели предметните термини (конкретни ключови думи на домейни) и сме ги добавили в структурата на онтологичния модел като класове. По този начин създаденото дърво на онтологията има дълбочина четири нива.

3.4 Експериментално тестване и резултати

В тази секция ще разгледаме ново предложените подходи представени в настоящата глава сред които са: подход базиран на Формален Концептуален Анализ и метод базиран на количествена оценка на експертиза, които са валидирани с данни от хранилището PubMed.

3.4.1 Експериментално оценяване на подхода базиран на Формален Концептуален Анализ

Първоначално множество от 9212 български автори е извлечено от хранилището PubMed. След като е разрешен въпросът с повтаряемостта на профилите на авторите, множеството е редуцирано до 242 различни изследователи. След което всеки един от авторите е представен като вектор от MeSH термини, като тези термини се използват за да опишат каква експертиза има всеки един от тях.
MeSH терминте са групирани в 16 основни категории. Приложен е формалният концептуален анализ върху извлечените данни от PubMed, като са използвани и тези 16 основни категории за да се опишат експертните профили и да се изгради след това и концептуалната решетка, която генерира и финалното групиране на експертите спрямо тяхната област(и) на компетентност. Изчислените степени на принадлежност са трансформирани до бинарни стойности, използвайки предварително заден праг. В разглеждания контекст прагът е определен чрез намиране на средната стойност от всички изчислени степени на принадлежност. Стойност 1 е дадена само на тези автори, чиито степени на принадлежност на техните ключови думи е по-висок или равен на 0.45.

Всички останали степени на принадлежност се заменят с бинарна стойност 0. Списъкът с клъстерираните автори по категории може да се види на Таблица 3.1. Концептуалната решетка съдържа 23 не празни концепции, където всяка една концепция представя подмножество от автори, които принадлежат на някакъв брой предметни категории. По този начин, българските експертни профили са разделени в 23 непресичащи се експертни области, спрямо основните MeSH категории. Таблици 3.1 и 3.2 показват броят на авторите разделени в главните MeSH категории както и тези които са принадлежащи на повече от една категория. Авторите принадлежащи на повече от една категория, следва да имат експертиза в повече от една експертна област. Резултатите показват, че 205 изследователи са били разпределени в 12 различни непресичащи се концепции, 5 от авторите принадлежат на празна концепция (това не е добавено в Таблица 3.2) и други 32 изследователя, чиято експертиза е в повече от една категория. Последните са разделени в 10 концепции съдържащи няколко категории, виж Таблица 3.2. Очевидно полученото групиране от експерти добре отразява разпределението на експертните знания в разглежданата област по отношение на основните категории. Освен това, той улеснява идентифицирането на хората с необходимата компетентност.

Таблица 3.1: Броят на авторите разпределени в основните MeSH категории

Етикет	Име на категорията	Брой автори
A	Anatomy	2
B	Organisms	2
C	Diseases	24
D	Chemicals and Drugs	16
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment	39
G	Phenomena and Processes	51
H	Disciplines and Occupations	7
I	Antropology, Education, Socialogy and Social Phenomena	5
J	Technology, Industry, Arguculture	1
L	Information Science	4
M	Named Groups	10
N	Health Care	44

Таблица 3.2: Броят на авторите в обединените MeSH категории

Обединени категории	Брой автори
$\{D, N\}, \{E, N\}$	3
$\{H, I\}$	2
$\{F, G\}, \{G, N\}$	5
$\{E, G\}, \{D, H\}, \{D, G\}$	1
$\{J, N\}$	7
$\{H, N\}$	4

3.4.2 Експериментално оценяване на подхода базиран на количествена оценка на експертиза

В настоящият експеримент е свалено ново множество от PubMed което съдържа 4343 български автори. След разрешаването на проблема с дублирането на експертни профили, които е допустимо да се отнасят за един и същ човек, е редуцирано до 3753 различни изследователи. Всеки един от авторите (изследователи) е представен чрез два компонента: списък от различни MeSH термини, които описват тяхната експертиза и вектор от тегла, изразяващ каква е степента на експертното познание. Теглото за всеки термин е изчислено като се вземе предвид отношението между броят на повторение на даденият термин спрямо общият брой термини, описващ експертният профил. Списък на примерни 10 експертни профила са дадени в Таблица 3.3, а в Таблица 3.4 списъка с изчислените тегла. Валидирането на резултатите е извършено чрез изчисляване на resemblance r и containment c, като те са използвани за да сравнят две резултатни решения приложени върху данни от PubMed, представени като експертни профили. Примерни резултати може да се видят на Фигури 3.1 и 3.2. Фигура 3.1 представя резултатите от валидиращите метрики r и c, които са изчислени върху списък от фиксиран брой (50) експерта, предоставени за всеки един от 10-те експерта представени в Таблица 3.3. От резултатите може да се забележи, че експертите под номера 1, 2, 3, и 10 имат стойности $r = 1$ и $c = 1$, което е основание, че експертизите им са еднакви и добре разпределени в различни MeSH области. Докато при останалите 4, 5, 6, 7 и 8, резултатите са сравнително по-ниски, можем да приемем, че по-голяма част от множеството от експерти има прекалено различни области на знание. За да се подобрят резултатите е приложен и тегловният метод върху данните, виж Фигура 3.2.

Таблица 3.3: Експертни профили описани чрез MeSH термини

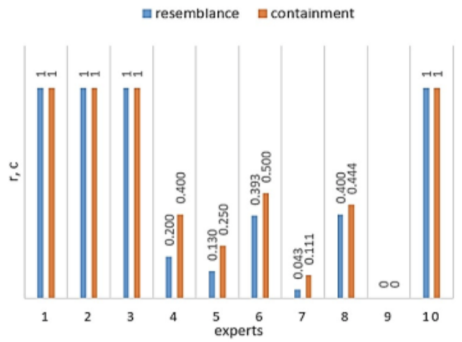
Експерт	MeSH термини
1	Kidney Transplantation; Liver Transplantation
2	Health Behavior
3	Drinking; Health Behavior; Health Knowledge, Attitudes, Practice; Program Evaluation
4	Models, Biological; Temperature; Models, Neurological; Water
5	Computer Simulation; Models, Molecular; Protons; Thermodynamics; Molecular Conformation
6	Vibration; Models, Molecular; Infrared Rays; Hydrogen Bonding
7	Monte Carlo Method; Models, Theoretical; Phase Transition; Thermodynamics
8	Photosynthesis; Quantum Theory
9	Health Behavior; Desicion Support Techniques; ...(more than 20 MeSH terms
10	Polymorphism, Genetic

Таблица 3.4: Получените тегла за всеки един MeSH термин от даден експертен профил

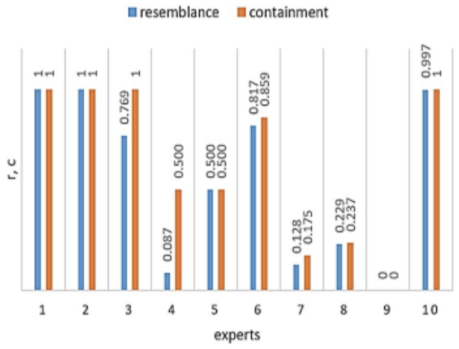
Експерти	Тегла на MeSH терминте
1	0.5; 0.5
2	1
3	0.25; 0.25; 0.25; 0.25
4	0.166; 0.333; 0.166; 0.333
5	0.285; 0.285; 0.142; 0.142; 0.142
6	0.5; 0.166; 0.166; 0.166
7	0.428; 0.285; 0.142; 0.142
8	0.75; 0.25
9	0.22; ...; 0.045; ...; 0.068; ...; 0.25
10	1

3.4.3 Валидиране на системата за препоръчване на експерти базирана на DiVA данни

Два основни сценария са валидарани при този подход. В първият сценарий потребителят трябва да избере примерен научен ръководител, а системата да върне списък с подобни експерти. В Таблица 3.5 са представени и получените оценки от подобието на експертизите на участващите в експеримента ръководители. Те варират между 0,046 и 0,77. Всеки един експерт в резултатната таблица е описан чрез ключови думи. Както може да се забележи, всички най-високо класирани ръководители притежават експертен опит, който се припокрива в различна степен с компетенцията на дадения примерен



Фигура 3.1: Стойности от валидиращи метрики r и c изчислени върху резултатите от извличането на експертизата



Фигура 3.2: Стойности от валидиращи метрики r и c изчислени върху резултатите от извличането на експертизата

експерт.

Във вторият сценарий, множеството от ръководители е групирано в групи от експерти с подобен опит чрез прилагане на алгоритъм за разделяне k-means. Избраната оптимална стойност за k е 16, а средната стойност от SI за $k = 16$ е 0,38. В този сценарий, за да изберат подходящите индивиди за посочената тема на дипломна теза, потребителят може да ограничи своите съображения само до тези ръководители, които са в клъстера, който е идентичен с (или поне най-подобен на) дадената тема. Експертите в избрания клъстер могат да бъдат класирани по отношение на сходството на техните експертни профили с определената тема. Например, ако се нуждаем от ръководител с опит в база данни и паралелни изчисления, ще използваме клъстер 0 (виж Таблица 3.6), в която всички експерти имат компетентност или в едната, или в двете области.

Таблица 3.5: Десетте най-високо оценени тематични ръководители на дипломанти, препоръчани от системата от първия сценарий на експеримента

Експерт	Ключови думи	Семантично подобие
1	user experience, usability	0.770
2	privacy, security, cloud computing	0.763
3	cloud computing, security metrics, security threats, ...	0.760
4	procedural city generation, perlin noise, ...	0.760
5	machine learning, parallel computing, ...	0.760
6	mobile, power, consumption, android, native, ...	0.754
7	mongodb, couchdb, python, pymongo, ...	0.754
8	compression, sms, arithmetic, lambda, huffman, ...	0.751
9	non-functional search-based software testing, ...	0.75
10	digital multimedia broadcasting, mpeg-2 standard, mpeg-4 standard, ...	0.75

Таблица 3.6: Групиране на 10-те най-високо класирани ръководители, препоръчани от системата в първия експериментален сценарий (тези, изброени в Таблица 3.5).

Експерти	Клъстер	Описание на клъстерите
3,4,7,8	0	usability; tessellation; android; security threats; main-memory database; database; distributed databases; parallel computing; security; data mining и други
2, 5, 6, 10	3	usability; data mining; performance monitoring; systematic review; video streaming; parallel computing; mpeg-2 standard; mpeg-2 standard; nosql database; machine learning; cloud computing и други.
1, 9	15	usability; quality of experience; urban design; systematic literature review и други.

3.5 Научно-приложни приноси

В проведените експерименти, Формалният Концептуален Анализ беше изпълнен върху множество от данни на учени от България в сферата на Медицината и Здравеопазването. Получените резултати показаха, че разработеният алгоритъм е подходяща клъстерираща техника за организиране на експертизата на експерти от дадена тематична общност в йерархична класификация от различни тематични направления. Показали сме също експериментално, че генерираната йерархична класификация се доближава до традиционното йерархично клъстериране, но го превъзхожда по отношение липса на допълнителни входни данни или прекалено много обхождания върху данните до достигането на финалното клъстериращо решение.
Предложена е техника за количествена оценка и представяне на експертизата на експерти по зададени тематични направления. Експериментално, техниката е тествана върху данни които са публично достъпни. В отговор на потребителска заявка съдържаща примерен експерт или група от експерти се генерира подреден в низходящ ред списък от експерти с близки експертни познания. Чрез този подход ние предлагаме възможност да се оцени нивото на познание на експерти в зададени тематични направления.
Предложен е прототип на препоръчваща система в сферата на академичните среди. Прототипната система генерира в отговор на потребителска заявка, списък от имената на подходящи ръководители за дипломната теза описана в заявката, които са подредени в низходящ ред по отношение на тяхната близост до тази теза на база на оценка на техния предишен опит. В експерименталната част са използвани реални данни на шведски университет. Предложената система е базирана върху онлайн информация, която се обновява периодично. Предложен е също семантичен модел, който организира в йерархична структура наличните тематични категории в университетската система.

АДАПТИРАЩИ ТЕХНИКИ ЗА ИНТЕГРИРАНЕ НА НОВОПОСТЪПВАЩИ ДАННИ

4.1 Еволюционно двустранно клъстериране (Bipartite Evolutionary Clustering)

Алгоритъмът е в състояние да анализира връзките между две клъстерни решения C и C' и въз основа на откритите модели той третира съществуващите клъстери (C) по различни начини. По този начин някои клъстери ще бъдат актуализирани чрез сливане с тези от новоизградените клъстериране C', докато други ще бъдат трансформирани чрез разделяне на техните елементи между няколко нови клъстера. Нашият еволюционен клъстериращ алгоритъм се основава на алгоритъма PivotBiCluster. Да допуснем, че всеки клъстер от клъсериращи решения C и C' е представен от списък от специфични за

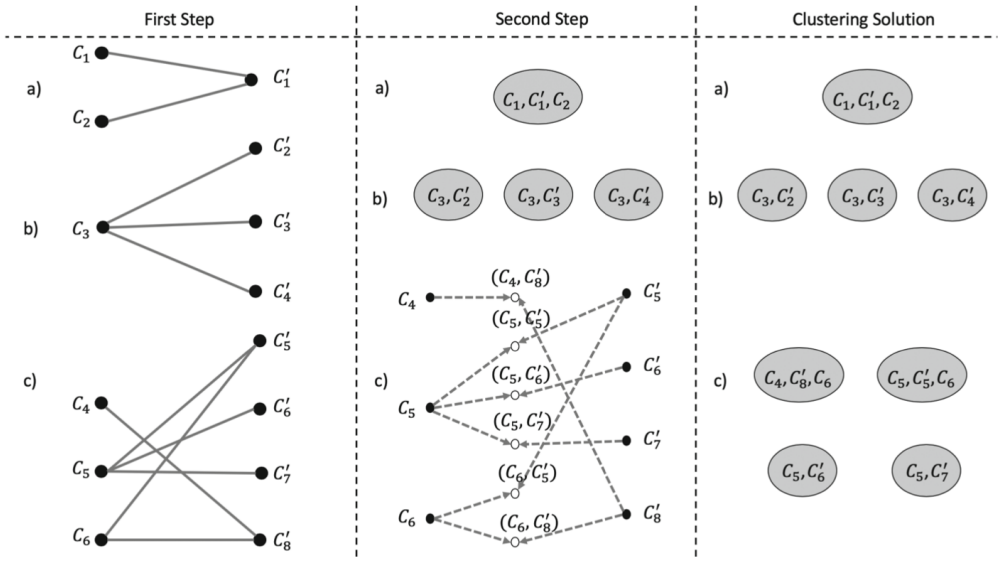
областта теми, които описват неговата експертна област. Входният граф $G = (C, C', E)$, където C и C' са множествата от левите и десните възли, а E е подмножество на $C \times C'$, което представя корелациите между възлите на двете множества. Подробно обяснение на предложения Merge-Split PivotBiCluster е дадено в Алгоритъм 1.

Algorithm 1 Merge-Split PivotBiCluster

function MERGE-SPLIT PBC($G = (C, C', E)$)
  for all nodes $c \in C \cup C'$ **do**
    if c is an unreachable node **then**
      Turn c into a singleton and remove it from G
    end if
  end for
  while $C \neq \emptyset$ **do**
    Choose c_1 uniformly at random from C
    if c_1 takes part in a bi-clique connecting it with several nodes from C' **then**
      Split c_1 among the corresponding nodes from C'
    else
      From a new cluster by merging c_1 with its neighbors from C'
      The neighbors of c_1 is denoted by $N(c_1)$.
      for all nodes $c_2 \in C \setminus \{c_1\}$ **do**
        Consider the sets: $R_1 = N(c_1) \setminus N(c_2), R_2 = N(c_2) \setminus N(c_1)$ and $R_{1,2} = N(c_1) \cup N(c_2)$
        Calculate probability $p = min\{|R_{1,2}| \setminus |R_2|, 1\}$
        if $|R_{1,2}| \geq |R_1|$ **then**
          with probability p append c_2 to the above cluster
        end if
      end for
    end if
    Remove all clustered nodes from G
  end while
  return all connected components (bi-cliques) as clusters of $C \cup C'$
end function

4.2 Разделяш-сливаш еволюционен клъстериращ алгоритъм (Split-Merge Evolutionary Clustering)

Алгоритъмът представлява рамка за разделяне-сливане, която може да се използва за адаптиране на съществуващото клъстерно решение към новопостъпили данни. Нашата рамка също така моделира две клъстериращи решения (съществуващото и новопостроеното) като двустранен граф, който се разлага на свързани компоненти (двустранни клики) (виж Фигура 4.1, (a), (b) и (c)). Всеки компонент се анализира допълнително и ако е необходимо, той се разлага на подкомпоненти (виж Фигура 4.1, (c) и (d). След това подкомпонентите се вземат предвид при производството на крайното клъстерно решение. Например, ако съществуващ клъстер е пренаселен (Фигура 4.1, буква (b)), т.е. той пресича два или повече клъстера в новото клъстериране, той се разделя между тях. Ако няколко съществуващи клъстера се пресичат с един и същ нов клъстер, т.е. те са припокриващи се (Фигура 4.1, буква a)), и се сливат с този клъстер.



Фигура 4.1: Алгоритъм Split-Merge Clustering

Algorithm 2 Split-Merge Evolutionary Clustering Algorithm

function SPLIT-MERGE($G = (C, C', E)$)
  for all nodes $c \in C \cup C'$ **do** (*First step*)
    if c is an unreachable node **then**
      Turn c into a singleton and remove it from G
    end if
  end for
  for all nodes $c \in C \cup C'$ **do** (*Second step*)
    if c_1 is the only node from C that takes part in a bi-clique connecting it with one or several nodes from C' **then**
      Split c_1 among the corresponding nodes from C'
    end if
  end for
  for all nodes $c \in C \cup C'$ **do**
    if c'_1 is the only node from C' that takes part in a bi-clique connecting it with one or several nodes from C **then**
      Merge c'_1 with the corresponding nodes from C
    end if
  end for
  for all nodes $c \in C$ **do** (*Third step*)
    Split c_1 among its adjacent nodes from C' and form new temporary clusters
  end for
  for all nodes $c' \in C'$ **do**
    Merge c'_1 with its adjacent nodes from the built set of temporary clusters
    Remove the clustered nodes from G
  end for
  return all connected components (bi-cliques) as clusters of $X \cup X'$
end function

4.3 Експериментално тестване и резултати

4.3.1 Експериментално тестване на Evolutionary Bipartite Clustering

Експерименталното тестване на предложеният алгоритъм е разделено на два различни сценария. Първият сценарий е насочен към областта на извличане на експертизи, докато вторият показва, че алгоритъмът би могъл да се използва и в друга област - здравеопазването. Разработеният клъстериращ алгоритъм е с общо предназначение и би могъл да се прилага в различни приложни области за еволюционно групиране на данни. Това е валидно също и за другият разработен и дискутиран в тази глава еволюционен алгоритъм.

Таблица 4.1: Експеримент 1: Средни стойности за F-measure и SI генерирани върху клъстериращото решение на 10-те тестови двойки

Metrics	PivotBiCluster	Split Merge Clustering
F-measure	0.618	0.582
SI	-0.145	-0.129

Генерираните клъстериращи решения са оценени с SI като средните им стойности са както следва: -0.158 (PivotBiCluster) и 0.058 (Split-Merge Evolutionary Clustering). Очевидно, предложеният алгоритъм Split-Merge Evolutionary Clustering превъзхожда PivotBiCluster върху тези данни. Вярваме, че това се дължи на факта, че той се приспособява по-добре към данните, като може не само да обедини онези клъстери, които са подклъстерирани, но и да раздели тези, които са пренаселени.

Таблица 4.2: Експеримент 2: Средни стойности за F-measure и SI генерирани върху клъстериращото решение на 10-те тестови двойки

Metrics	PivotBiCluster	Split Merge Clustering
F-measure	0.321	0.331
SI	0.137	0.157

Във вторият сценарии получените средни резултати за SI са съответно -0.013 (PivotBiCluster) и -0.170 (Split-Merge Evolutionary Clustering). Очевидно, PivotBiCluster превъзхожда алгоритъма Split-Merge Clustering по отношение на тези критерии за оценка. Резултатите, получени чрез F-measure, също показват по-добрата производителност на PivotBiCluster (0,71 срещу 0,46 за Split-Merge Evolutionary Clustering) върху това

множество от данни. Интересно е обаче да се отбележи, че броят на клъстерите в клъстериращите решения, генерирани от PivotBiCluster, варира от 1 до 5, докато при Split-Merge Evolutionary Clustering, индивидите се групират в 5 или 6 клъстера. Забележете, че еталонно клъстериране (benchmark) има 6 клъстера. Горните резултати ни мотивираха да използваме и Jaccard Index за допълнително сравняване на двата изследвани алгоритъма. А именно, ние приложихме Jaccard Index, за да намерим сходството между генерираните клъстериращи решения и еталонно решение. Съответните стойности са 0.081 (PivotBiCluster) и 0.291 (Split-Merge Evolutionary Clustering), алгоритъма Split-Merge Clustering генерира по-висок среден резултат от колкото резултата генериран за PivotBiCluster.

4.3.2 Експериментално тестване на Split-Merge Evolutionary Clustering

Експерименталното тестване на предложеният алгоритъм е разделено на два различни сценария. Първият сценарий е насочен към областта на извличане на експертизи, докато вторият показва, че алгоритъмът би могъл да се използва и в друга област - здравеопазването. И в двата сценария алгоритъма е сравнен с друг алгоритъм подобен на него, а именно PivotBiCluster.

Използвано е множество с мощност 3753 експертни профили. Всеки експертен профил е представен като вектор от ключови думи, описващи неговата сфера на компетентност. Изследователите от това множество са случайно разделени на две множества. Едното множество съдържа 2407 експерта групирани в 122 клъстера чрез използване на алгоритъма k-means, а другото съдържа 1346 експерта групирани в 112 клъстера, отново чрез прилагане на k-means. Броят на клъстерите е определен чрез клъстериране на всяко едно множество чрез прилагане на k-means за различни k, а получените решения оценени чрез SI. След това двата клъстериращи алгоритъма са изпълнени два пъти за да интегрират тези две клъстериращи решения в едно. Полученото клъстериращо решение от алгоритъма PivotBiCluster има 95 клъстера, докато предложеният Split-Merge Evolutionary Clustering има генерирано решение с 104 клъстера. Генерираните клъстериращи решения са оценени с SI като средните им стойности са както следва: -0.158 (PivotBiCluster) и 0.058 (Split-Merge Evolutionary Clustering). Очевидно, предложеният алгоритъм Split-Merge Evolutionary Clustering превъзхожда PivotBiCluster върху тези данни.

Използвано е и еталонното решение от 102 различни експертни профила за да се генерират 10 тестови множества от двойки. Всяко тестова двойка разделя изследователите случайно на две множества. Едно множество (съдържащо 70 експерти) от всяка двойка представя наличния набор от експерти, а другото (32 експерти) е набор от новоизвлечени експерти. Проучихме два различни експериментални сценария. В първия сценарий

експертите във всеки тестов набор са групирани в клъстери от експерти с подобна експертиза въз основа на информацията от сесията на научна конференцията, т.е. всеки набор е разделен на 8 клъстера. Във втория сценарий за всеки набор от данни оптималният брой клъстери се определя чрез групиране на множеството, като се прилага k-means за различни k и оценяване на получените решения чрез SI. Получените резултати от двата експеримента са дадени на Фигури 4.3 и 4.4. PivotBiCluster превъзхожда алгоритъма Split-Merge Evolutionary Clustering само в един случай. А именно, той е генерирал по-висока стойност за F-measure от алгоритъма Split-Merge Evolutionary Clustering в първият експеримент. Във вторият опит (виж Таблица 4.4) оценките за SI са не само по-високи в сравнение с тези, генерирани в първият опит, но са и положителни. Използването на оптималния брой клъстери значително подобрява качеството на генерираните клъстериращите решения с отнасяне към компактността и свойствата на разделяне на клъстерите.

Таблица 4.3: Експеримент 1: Средни стойности за F-measure и SI генерирани върху клъстериращото решение на 10-те тестови двойки

Metrics	PivotBiCluster	Split Merge Clustering
F-measure	0.618	0.582
SI	-0.145	-0.129

Таблица 4.4: Експеримент 2: Средни стойности за F-measure и SI генерирани върху клъстериращото решение на 10-те тестови двойки

Metrics	PivotBiCluster	Split Merge Clustering
F-measure	0.321	0.331
SI	0.137	0.157

При вторият сценарий е използвано тестово множество което съдържа 400 различни профила и от него се генерират 10 тестови двойки чрез случайно разделяне на индивидите на две множества. Едното множество (280 пациенти) във всяка двойка представящо наличният набор от индивидуални профили, а другото (120 индивида) е набор от новосъбрани профили на пациенти. Профилите на пациентите от всяко множество са групирани в 6 групи спрямо тяхното кръвно налягане. Получените клъстери са представени от техните центроиди. Генерираните клъстериращи решения отново са оценени с SI и F-measure. Получените средни резултати за SI са съответно -0.013 (PivotBiCluster) и -0.170 (Split-Merge Evolutionary Clustering). Очевидно, PivotBiCluster превъзхожда алгоритъма Split-Merge Clustering по отношение на тези критерии за оценка. Резултатите,

получени чрез F-measure, също показват по-добрата производителност на PivotBiCluster (0,71 срещу 0,46 за Split-Merge Evolutionary Clustering) върху това множество от данни. Броят на клъстерите в клъстериращите решения, генерирани от PivotBiCluster, варира от 1 до 5, докато при Split-Merge Evolutionary Clustering, индивидите се групират в 5 или 6 клъстера. Приложена е и метриката Jaccard Index, за да се намери сходството между генерираните клъстериращи решения и еталонно решение. Съответните стойности са 0.081 (PivotBiCluster) и 0.291 (Split-Merge Evolutionary Clustering), алгоритъма Split-Merge Clustering генерира по-висок среден резултат от колкото резултата генериран за PivotBiCluster.

Предложеният алгоритъм е допълнително тестван с нови набори от данни в различни експериментални сценарии като е сравнено представянето му с друг подобен алгоритъм Dynamic split-and-merge. Отново са изследвани два научни сценария.

При първият сценарии са използвани три алгоритъма PivotBiCluster, Dynamic-split-and-merge и Split-Merge Evolutionary Clustering. Изпълнени са върху две различни множества от данни, като са създадени от тях 10 тестови двойки, генерирани на случаен принцип.Всяко едно от тези 10 тестови двойки съдържа едно множество с 70% от данните и още едно с 30% от данните.По-малкото множество са данните, които са нови и следва да бъдат адаптирани към вече съществуващите (по-голямото множество от 70%).

Във вторият сценарий ние изследваме до колко трите алгоритъма са чувствителни спрямо размера на новопостъпващите данни. Използвали сме другите две множества anthropometric и yeast. Отново са създадени 10 тестови двойки за всяко едно от двете множества и всеки един от алгоритмите е изпълнен върху тях 4-ри пъти. Десетте двойки от множества са разделени в различни съотношения 50/50, 60/40, 70/30 и 80/20, като първата стойност отговаря на процентното съдържание което се съдържа в множеството представящо текущите съществуващи данни, а втората стойност представя новопостъпващите данни. Резултатите от изпълнението на първият сценарий са показни в Таблици 4.5 и 4.6.

Таблица 4.5: Експеримент 1: Средните валидиращи стойности върху изпълнението на трите алгоритъма върху 10-те тестови множества на cover-type

Метрики	PivotBiCluster	Split-Merge clustering	Dynamic split-and-merge
SI	0.194	0.034	0.196
F-measure	0.903	0.759	0.376
Jaccard Index	0.231	0.021	0.161

Таблица 4.6: Експеримент 1: Средните валидиращи стойности върху изпълнението на трите алгоритъма върху 10-те тестови множества на wine quality

Метрики	PivotBiCluster	Split-Merge clustering	Dynamic split-and-merge
SI	-0.111	-0.129	0.143
F-measure	0.676	0.461	0.311
Jaccard Index	0.269	0.143	0.137

Резултатите получени при изпълнение на вторият експеримент са показани в Таблици 4.7, 4.8, 4.9, 4.10, 4.11 и 4.12.

Таблица 4.7: Експеримент 2: Средните валидиращи стойности за F-measure генерирани върху 4 x 10 тестови двойки на множеството antropometric

Алгоритми	50/50	60/40	70/30	80/20
PivotBiCluster	0.677	0.624	0.544	0.676
Split-Merge clustering	0.546	0.504	0.519	0.481
Dynamic split-and-merge	0.374	0.389	0.442	0.482

Таблица 4.8: Експеримент 2: Средните валидиращи стойности за F-measure генерирани върху 4 x 10 тестови двойки на множеството yeast

Алгоритми	50/50	60/40	70/30	80/20
PivotBiCluster	0.700	0.710	0.821	0.858
Split-Merge clustering	0.576	0.522	0.496	0.489
Dynamic split-and-merge	0.419	0.423	0.426	0.410

Таблица 4.9: Експеримент 2: Средните валидиращи стойности за SI генерирани върху 4 x 10 тестови двойки на множеството antropometric

Алгоритми	50/50	60/40	70/30	80/20
PivotBiCluster	-0.344	-0.327	-0.231	-0.178
Split-Merge clustering	-0.212	-0.189	-0.108	-0.096
Dynamic split-and-merge	0.2	0.238	0.188	0.170

4.4 Научно-приложни приноси

Разработеният нов алгоритъм за еволюционно двустранно клъстериране (Evolutionary Bipartite Clustering) успява да се справи със задачата за обновяване на клъстер решението при пристигане на нови данни. Предложеният алгоритъм е сравнен с два други

Таблица 4.10: Експеримент 2: Средните валидиращи стойности за SI генерирани върху 4 x 10 тестови двойки на множеството yeast

Алгоритми	50/50	60/40	70/30	80/20
PivotBiCluster	0.142	0.068	0.044	0.076
Split-Merge clustering	-0.061	-0.061	-0.048	-0.036
Dynamic split-and-merge	0.164	0.157	0.158	0.150

Таблица 4.11: Експеримент 2: Средните валидиращи стойности за Jaccard Index генерирани върху 4 x 10 тестови двойки на множеството antropometric

Алгоритми	50/50	60/40	70/30	80/20
PivotBiCluster	0.021	0.015	0.068	0.058
Split-Merge clustering	-0.077	0.164	0.107	0.074
Dynamic split-and-merge	0.156	0.199	0.119	0.094

Таблица 4.12: Експеримент 2: Средните валидиращи стойности за Jaccard Index генерирани върху 4 x 10 тестови двойки на множеството yeast

Алгоритми	50/50	60/40	70/30	80/20
PivotBiCluster	0.014	0.022	0.034	0.020
Split-Merge clustering	0.086	0.089	0.090	0.086
Dynamic split-and-merge	0.099	0.136	0.105	0.118

алгоритъма върху различни тестови набори от данни извлечени от онлайн-достъпни хранилища и демонстрира по-добро представяне. Предложеният еволюционен алгоритъм превъзхожда традиционните клъстериращи алгоритми, поради факта, че не е необходимо да се подават и настройват допълнителни входни параметри като например, броя клъстери. Последното е доказано, че влияе на бързината на клъстериращия алгоритъм, тъй като не се губи време за намиране на оптималния брой клъстери. Доказано е също, че разработената техника, може да се използва в областта на извличане на експертиза.

Разработен е също и алгоритъм разделяш-сливаш еволюционен клъстериращ алгоритъм (Split-Merge Evolutionary Clustering). Предложеният алгоритъм е оценен и сравнен с два други клъстериращи алгоритъма в различни експериментални сценарии. Получените експериментални резултати показват, че предложеният алгоритъм превъзхожда другите два в много от проведените експерименти или има доста близко представяне

като успява да генерира решения близки до еталонните клъстер решения. Разработеният алгоритъм успява да обработва периодичното постъпване на нови данни и да адаптира съществуващото клъстер решение в съответствие. Последното демонстрира, че предложеният алгоритъм може успешно да се справя с concept drift феномена.

НАУЧНО-ПРИЛОЖНИ И ПРИЛОЖНИ ПРИНОСИ

1. Предложен е алгоритъм за йерархично организиране на експерти от дадена област в тематични направления. Алгоритъмът е показан да предоставя по-добро решение от традиционните клъстериращи алгоритми. Представянето на алгоритъма е оценено върху тестови набори данни изтеглени от PubMed хранилището.

2. Предложена е техника за количествена оценка на нивото на експертиза на експерти в дадена област. Предложената техника предлага по-прецизно описание и представяне на експертните познания. Техниката е оценена и демонстрирана върху тестови набори от данни извлечени от PubMed хранилището. Показано, е че техниката може да се използва и за извличане на подреден списък от експерти с подобна експертиза спрямо дадена заявка.

3. Предложен е прототип на препоръчваща система в сферата на академичните среди. Прототипът е оценен върху тестови данни относно защитени магистърски дипломни работи на Шведски университет представени в онлайн-достъпна система. Изграден е концептуален модел на областта на базата на извлечена информация от системата. Реализираната система е подходяща и може лесно да бъде адаптиране за използване в други области.

4. Разработени са два еволюционни клъстериращи алгоритми, които се справят с concept drift феномена като гъвкаво адаптират съществуващо клъстер решение когато постъпят нови данни. Представянето на всеки алгоритъм от предложените е сравнено с това на други два подобни алгоритъма. За целта са използвани тестови набори от данни от няколко различни онлайн източника. Получените резултати са съпоставими или превъзхождат тези генерирани от използваните еталони алгоритми. Предложените алгоритми са с общо предназначение, но е показано, че успешно могат да бъдат използвани и при извличането на експертиза.

СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИОННИЯ ТРУД

Глави в редактирани книги

[1] V. Boeva, M. Angelova, V. Manasa Devagiri, and E. Tsiporkova. Bipartite split-merge evolutionary clustering. In *Agents and Artificial Intelligence*, pages 204–223. Springer, 2019.

[2] M. Angelova, V. M. Devagiri, V. Boeva, P. Linde, and N. Lavesson. An expertise recommender system based on data from institutional repository (diva). In *Connecting the Knowledge Commons – from projects to sustainable infrastructure*, pages 135–149, 2019.

Статии в научни списания

[1] M. Angelova. Clustering technologies for analysis of large datasets. *Fundamental Science and Applications*, 23:113–118, 2017. ISSN 1310-8271.

Статии от участия в научни конференции

[1] V. Boeva, M. Angelova, V. Manasa Devagiri, and E. Tsiporkova. A split-merge framework for evolutionary clustering. *31th SAIS 2019, Umeå, Sweden, June 18-19*, 2019.

[2] V. Boeva, M. Angelova, and E. Tsiporkova. A split-merge evolutionary clustering algorithm. In *Proc. of 11th ICAART, Volume 2, Prague, Czech Republic, February 19-21*, pages 337–346, 2019.

[3] M. Angelova, V. M. Devagiri, V. Boeva, P. Linde, and N. Lavesson. An Expertise Recommender System Based on Data from an Institutional Repository (DiVA). In *ELPUB 2018*, June 2018.

[4] V. Boeva, M. Angelova, N. Lavesson, O. Rosander, and E. Tsiporkova. Evolutionary clustering techniques for expertise mining scenarios. In *Proc. of the 10th ICAART, January 16-18*, pages 523–530, 2018.

[5] V. Boeva, M. Angelova, and E. Tsiporkova. Data-driven techniques for expert finding. In *9th ICAART 2017*, pages 535–542, 2017.

[6] V. Boeva, M. Angelova, and E. Tsiporkova. Identifying subject experts through clustering analysis. In *Benelearn 2017*, pages 95–97, 2017.

[7] M. Angelova, V. Boeva, and E. Tsiporkova. Advanced data-driven techniques for mining expertise. In *30th SAIS, May 15-16, 2017, Karlskrona, Sweden*, pages 137–005, 2017.

[8] V. Boeva, M. Angelova, L. Boneva, and E. Tsiporkova. Identifying a group of subject experts using formal concept analysis. In *8th IEEE, IS'16, Sofia, Bulgaria, September 4-6*, IS IEEE, pages 464–469, 2016.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Абстракти

[1] V. Boeva, M. Angelova, V. Manasa Devagiri, and E. Tsiporkova. Patient profiling using evolutionary clustering. womENcourage 2019 Rome, Italy, 16-18 September 2019, December 2019.

[2] V. Boeva, M. Angelova, and E. Tsiporkova. A bipartite-graph based approach for split-merge evolutionary clustering. ECDA 2019 Bayureth, Germany, March 18-20, March 18-20 2019.

[3] V. Boeva, M. Angelova, E. Tsiporkova, and N. Lavesson. Evolutionary clustering techniques. WiML2017, December 2017.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

MILENA ANGELOVA, M.Sc

INTELLIGENT APPROACHES FOR PROFILING, ANALYSIS AND LOCALIZATION OF EXPERTISE FROM PUBLIC ONLINE SOURCES

ABSTRACT OF Ph.D. THESIS

Many organizations in different fields have a need for systems that can help them in managing effectively their human potential. The need for such software systems that can deal with the big volume of information that is collected and generated in the organizations, further increases nowadays. The available information can be useful for many organizations by being able of analysing and extracting valuable knowledge about their employees. For example, this can facilitate the process of finding the right person with the required skills and knowledge for a given task. However, organizing human potential by analyzing, assessing, and grouping the available information about the skills and knowledge of the employees is a challenging problem for many companies and organizations. This is a time-consuming process that is still not automated in many organizations. Moreover, human expertise is dynamic and data is periodically increasing, which leads to seeking new techniques for efficient adapting of the existing expertise organization when new data becomes available.

The main aim of this thesis is to develop and implement new approaches for profiling, analyzing, and localization of expertise using publicly available sources. In order to achieve the above aim four research questions (tasks) have been set. Initially, we are interested in developing a novel approach for facilitating the finding of expertise by organizing the experts in a given field in a hierarchical clustering structure based on available online information about their competence. The second question is related to the development of techniques for quantitative assessment of expertise levels and the use of these assessments for more refined profiling and localization of expertise. The third task is proposing new clustering techniques that are capable of organizing the expertise in subject areas and adapting the existing organization when new data arrives. Finally, the developed techniques for profiling, analyzing, and localization of expertise need to be evaluated and demonstrated on test datasets extracted from different online accessible sources.

One of the results of this thesis is the development of a new FCA (Formal Concept Analysis)-based approach for hierarchical clustering of a group of experts with respect to given subject areas. Furthermore, it has been proposed techniques that optimize expert knowledge representation by assessing the level of competence. Namely, a weighting method has been used to evaluate the levels of expertise of a given expert with respect to considered domain-specific topics. It has further been shown how this can be applied for refining the expert finding process. In addition, two evolutionary clustering methods have been developed that are

capable of dealing with the concept drift phenomenon by adapting the existing clustering solution when new data become available. The last is a prototype system for ranking potential supervision experience and expertise retrieval approaches.

The proposed system is based on online information and uses a built ontology model that conceptualises the thesis domain supported by the university. The proposed method based on formal concept analysis demonstrate that the proposed approach is a robust clustering technique that is suitable to deal with sparse data. The second proposed approach - the weighting method is considered as expert identification technique that return similar experts ones provided by the user. Both proposed evolutionary clustering techniques are supposed to be more robust to concept drift scenarios by providing the flexibility to update the existing clustering solution by considering the clusters derived from a new portion of data. The thesis is elaborated at the Technical University of Sofia, branch Plovdiv, Department of Computer Systems and Technologies. The thesis is written in Bulgarian and contains introduction, four chapters, scientific contributions, publications, bibliography of 150 titles, a volume of 152 pages, 13 figures, and 18 tables. The results are published in 11 scientific papers.