Софийски университет
"Св. Климент Охридски"

Факултет по математика
и информатика

Петър Ангелов Петров

Интелигентни системи в биоинформатиката: *намиране на съответствия между* и *обединяване на* анатомични онтологии

Автореферат
на дисертация

за присъждане на образователна и научна степен
"Доктор" по научна специалност 01.01.12 Информатика

Научен ръководител:
доц. д-р Антоний Попов

София, 2012 г.

Дисертационният труд е обсъден и насочен за
публична защита на 21.11.2012 г. на заседание
на катедра "Информационни технологии" към
Факултета по математика и информатика (ФМИ) на
Софийския университет (СУ) "Св. Климент Охридски".

Публичната защита на дисертационния труд
ще се състои на г.
от часа
в
на открито заседание. Материалите по защитата
са на разположение в библиотеката на ФМИ на СУ
(София, бул. Джеймс Баучър №5).

Пълният обем на дисертацията е 182 страници. Дисертацията се състои от 7 глави (сред които увод и заключение), 60 фигури, списък на публикациите, списък на цитираната литература и приложения в електронен вид. Цитираната литература включва 98 заглавия – всички на английски език. Списъкът от публикации на докторанта по същността на дисертацията включва 4 заглавия.

Забележка: Всички препратки от този автореферат към съкращения (Атп), термини (Втп) и цитирана литература ([т] и [тп]) използват съответната номерация от основния текст на дисертацията (тук т и п означават цифри).

Съдържание

Обща характеристика на дисертацията

Важност на проблема и мотивация

Проблемите по ***намиране на съответствия между онтологии*** (B11) и ***обединяване на онтологии*** (B12) са ключови при изследването на онтологии по принцип [3]. Често се използва още и обобщеното понятие ***интегриране на онтологии*** (B33).

Важността на интегрирането на онтологии произхожда от факта, че самите онтологиите обикновено биват проектирани и разработвани от различни страни (научни групи и институти, софтуерни фирми, други организации). Това води до появата на множество от хетерогенни онтологии, всяка от които моделира подобни или дори еднакви предметни области. По множество причини (най-вече икономически), трудно може да се постигне съгласуваност между страните за използването на една обща онтология, обединяваща знанията, съдържащи се в отделните хетерогенни онтологии. Това прави обмяната на знания и информация между тези отделни страни и между техните софтуерни системи трудна или дори практически невъзможна. Интегрирането на онтологии цели създаването на среда за свободна обмяна на знания и информация между различните страни, спомети по-горе, на базата на обща онтология, получена след интегрирането на отделните хетерогенни онтологии [2].

В настоящия труд са използвани анатомични онтологии, като целта е интегрирането на тези онтологии. Условно са дефинирани два основни етапа за интегрирането на тези онтологии: 1) установяване или намиране на съответствия помежду им; 2) тяхното обединяване или сливане.

Мотивацията за разработването на дисертационния труд може да се разгледа в контекста на следните три предизвикателства.

Често в биологията експерименталните данни, получени за даден (напр. моделен) организъм могат да се окажат по-общи и приложими и за други организми. Настоящото състояние на знанията и информацията от различните видово-специфични анатомични онтологии не позволява извършването на интелигентно търсене в структурирани данни, простиращи се отвъд контекста на на конкретен организъм.

Индивидуалните анатомични онтологии са полезни за извличане на данни от различни помежду си бази от данни, посветени на конкретни организми. Извършването обаче на интегриращи заявки, които да се допитват до множество хетерогенни анатомични бази от данни, все още не е лесна задача. Това е така, понеже всяка отделна анатомична база от данни използва своя собствена онтология, а различните онтологии са проектирани и разработени, базирайки се на различни стратегии, принципи и цели. Все още съществува явен недостиг на междуонтологични връзки между отделните анатомични онтологии и почти пълна липса на връзки от анатомията на даден организъм към други биологични области на изследване на този организъм като неговите генотип и фенотип [84].

Понастоящем се наблюдава и липса на надеждни механизми за допитване до анатомични данни за човека (*homo sapiens*) от една страна и до подобни данни за различни моделни и немоделни организми, от друга страна поради големите различия в техните терминологии [84]. Това също води до сериозни трудности при

опитите за пренасянето на знанията за тези моделни и немоделни организми към въпросите на човешкото здраве и медицината.

Цели и задачи на дисертацията

Първата основна цел на настоящата работа е разработването на метод и алгоритъм, който да служи за *намиране на съответствия между* и за *обединяване на* анатомични онтологии.

Втората основна цел е реализацията на този метод под формата на интелигентна софтуерна програма, която да се използва от специалисти по биология и анатомия и която да полуавтоматизира процеса по обединяване на две и повече анатомични онтологии на базата на експертни знания, налични в различни външни източници на знания (ВИЗ – А02).

За постигането на така поставените цели се налага да бъдат решени следните по-конкретни задачи:

1. дефиниране на формална постановка на задачата;

2. задаване във формален вид на използваните (i) входни анатомични онтологии – O_1 и O_2; (ii) външни източници на знания – UMLS [8,9], FMA [10,11], WordNet [12,13,14];

3. създаване на взаимно допълващи се аналитични модели, описващи процесите по намиране на съответствия между анатомични онтологии и тяхното обединяване;

4. разработване на алгоритъм на базата на тези модели, който изпълнява следните функции:

4.1. създаване на синонимни анатомични речници (В04) за входните анатомични онтологии;

4.2. намиране на съответствия между две дадени/входни анатомични онтологии посредством установяване на семантични релации между понятията, дефинирани в тях;

4.3. обединяване на две дадени анатомични онтологии и генериране на една обща изходна супер-онтология, която включва в себе си знанията и информацията, налични в двете входни онтологии;

5. проектиране и реализиране на интегрирана софтуерна програма, реализираща разработения алгоритъм, и включваща следните модули:

5.1. модул за комуникация, служещ за комуникация (обмен на данни) с наличните външни източници на знания (представени под формата на релационни бази от данни);

5.2. модул за визуализиране, служещ за ясно, графично представяне на конкретни части от двете входни онтологии, както и на семантичните връзки, установени между тях в хода на изпълнение на алгоритъма;

5.3. модул за ръчно редактиране и коригиране (В26), позволяващ на потребителя да приема/отхвърля конкретните предложения/предсказания на програмата за наличието на (потенциални, евентуални) семантични връзки между понятията от двете входни онтологии;

6. осигуряване на съвместимост на софтуерната програма с различни общоприети стандарти за декларативно представяне на онтологии.

Обем и структура на дисертацията

Пълният обем на дисертацията е 182 страници, като тя се състои от 7 глави, 60 фигури, списък на цитираната литература и приложения в електронен вид. Използваната литература включва 98 заглавия, всички на английски език. Списъкът с публикации по същността на дисертацията съдържа 4 заглавия.

Глава I въвежда накратко в проблематиката и задачите по намиране на съответствия между и за обединяване на анатомични онтологии. В тази глава се изяснява важността на проблема. Определят се целите и задачите на дисертацията.

Глава II представлява подробен обзор на проблемната област, на известните методи, подходи, алгоритми и софтуерни програми за решаване на задачи, подобни на задачата от нашата дисертация. В тази глава се дефинира понятието онтология, разглежда се приложението на онтологиите в естествените науки (като биология и медицина), разглежда се въпросът за интегриране на онтологии, моделиращи подобни или идентични предметни области и в частност въпросът за интегриране на анатомични онтологии. Части от текста на тази глава могат да бъдат намерени в [1A].

Глава III представя използваните от нас модели от теория на графите, а именно – насочени ациклични графи (НАГ – A06), за представяне на онтологиите. В тази глава се формализира поставения проблем с цел по-ефективното му решаване посредством алгоритмични процедури. Части от текста на тази глава могат да бъдат намерени в [1A].

Глава IV представлява пълно и формално описание на процедурите, използвани за решаване на проблема. Алгоритмичните процедури DM (direct matching) и SMP (source matching predictions), описани в рамките на тази глава представляват до известна степен вече известни подходи за откриване на съответствия между и за обединяване на онтологии. В дисертационния труд те са описани в контекста на нашата формалната постановка на задачата с цел преминаване към процедурата, която се изпълнява след тях и която е наречена CMP (child matching predictions). Така към DM и SMP ние добавяме процедурата CMP, която е оригинална процедура, създадена в рамките на дисертационния труд, работеща върху НАГ (A06) и търсеща в тях няколко предварително фиксирани типа шаблони на свързаност (B21). На базата на тези шаблони се откриват и предсказват нови междуонтологични връзки, които не могат да бъдат установени посредством DM или SMP. И трите процедури (DM, SMP, CMP) се базират на вероятностна схема за оценяване на откриваните в хода на изпълнението им междуонтологични връзки, която също е представена в тази глава. Текстът на тази глава е формиран от [1A], [2A], [3A].

Глава V представлява пълно описание на софтуерната програма AnatOM, разработена в хода на нашата работа и реализираща алгоритмичните процедури DM, SMP, CMP. Нейното име е своеобразно съкращение от ***Anat****omical* ***O****ntologies* ***M****erger*. Представена е цялостната архитектура на софтуерната програма и в детайли са описани основните модули от нея. Програмата не използва никакви външни библиотеки и затова модулите в нея са също оригинални от техническа (реализационна) гледна точка. Програмата AnatOM представлява едно цялостно, завършено, решение за полуавтоматично интегриране (откриване на съответствия

и обединяване) на анатомични онтологии. Текстът на тази глава е формиран на базата на [4A].

Глава VI представя анализ на проведените експерименти и на получените резултати.

Глава VII представя виждането ни за приносите (научни и приложни) от нашата работа, както и идеите ни за бъдещото развитие на работата.

Дисертацията завършва със списък на цитираните литературни източници. Пълният програмен код на програмата AnatOM, както и съдържанието на базите от данни, използвани от програмата, могат да бъдат намерени в приложенията към дисертацията, които са предоставени от автора в електронен вид.

Кратко съдържание на дисертацията

Глава I. Увод

Основният предмет на разглеждане в настоящата работа са задачите по намиране на съответствия между (B11) и обединяване на (B12) анатомични онтологии.

Проблемът за намиране на съответствия между дадени анатомични онтологии и за тяхното обединяване се състои в приемането на няколко анатомични онтологии като входни данни, установяването на наличните релации между тях и генерирането на една обща, обединяваща, анатомична онтология като изход, която включва в себе си знанията от входните онтологии. Онтологиите, използвани в дисертационния труд, описват анатомиите на различни категории организми (напр. мишка (B15), жаба (B41), риба Данио (B16) и други).

В идеалния случай, намирането на съответствия и обединяването трябва да бъдат извършени по един адекватен начин в анатомичен, биологичен и еволюционен аспект. Това означава, че дадена анатомична част (орган, тъкан, клетка и др.) от един организъм (например мишка – B15), трябва да бъде съпоставена на някаква анатомична част от друг организъм (например риба Данио – B16), само ако двете части са анатомично подобни или ако едната част е произлязла от другата в хода на еволюцията.

В някои случаи това съпоставяне изглежда просто и дори тривиално. Например, очевидно е, че следните анатомични понятия са съответни: *мозък (мишка) = мозък (риба Данио) = мозък (жаба)*.

Трудностите обаче идват оттам, че в мнозинството от случаите това терминологично съответствие далеч не е толкова очевидно. Това може да се види от следните примери.

капиляр (мишка) = микроскопичен кръвоносен съд (риба Данио)

ухо (мишка) = слухов апарат (жаба)

миелоиден левкоцит (риба Данио) = миелоидна клетка (жаба)

Можем да разглеждаме това като ***проблем #1***, а именно – намирането на съответствия между и обединяването на анатомичните понятия от отделните входни онтологии.

Още трудности идват оттам, че след като е намерено съответствие между дадени анатомични понятия, като например

капиляр (мишка) = микроскопичен кръвоносен съд (риба Данио),

възниква въпросът как следва да бъдат съпоставени техните понятия-деца и понятия-родители, дефинирани във входните видово-специфични онтологии.

Можем да разглеждаме това като ***проблем #2*** – намирането на съответствия между и обединяването на релациите от двете входни онтологии и оттам и на самите онтологии като цяло.

Глава II. Анализ на състоянието на проблема

Онтология – определения, компоненти, типове онтологии, езици за описание на онтологии

Вероятно най-популярното определение за понятието онтология в смисъла на информатиката е това, дадено от Том Грубер в [1], което гласи, че онтологията е "*спецификация на концептуализация*". Тази дефиниция е подобна, но не е идентична с първоначалното значение на понятието онтология, известно от философията[1].

Дефиницията за онтология (отново в смисъла на информатиката), дадено в Уикипедия (свободната Интернет енциклопедия) е по-подробна от краткото определение на Грубер, което приведохме по-горе и гласи, че "*всяка онтология представлява формално представяне на знания като множество от понятия/термини, служещи за назоваване на обектите, съществуващи в дадена предметна област и множество от релации между представените понятия; при това онтологията може да се използва за извеждането на нови знания относно обектите от предметната област и за описание (в общ смисъл) на предметната област*".

Компонентите[2], съставляващи една онтология са: класове, релации, атрибути, индивидуални термове, функционални термове, ограничения, правила, аксиоми, събития. Не всички компоненти са задължителни. Най-важни сред тези компоненти са класовете и релациите.

Онтологиите могат да бъдат класифицирани на базата на различни признаци, които те притежават. Според своята ***цел*** онтологиите се разделят на ***приложни онтологии*** (application ontologies) и ***референтни онтологии*** (reference ontologies). Според своята ***специфичност*** онтологиите се делят на три групи: ***общи онтологии*** или онтологии от високо ниво (generic, upper-level, top-level ontologies), ***същински (междинни) онтологии*** (core ontologies), ***предметни онтологии*** (domain ontologies). Според своята ***изразителност*** онтологиите обикновено се делят на ***леки*** (lightweight) и ***тежки*** (heavyweight).

Онтологиите и използването им като модели за представяне и извод на знания произлизат от някои по-ранни и по-неформални модели с подобно на тяхното предназначение като например семантичните мрежи [52,53,46,54] и фреймовите езици [44,45].

Съществуват множество езици за описание на онтологии, сред които вероятно най-важните към днешна дата са RDF[3]/RDFS[4] [49], OWL[5] [46,51], OBO[6] [6].

---

1 http://en.wikipedia.org/wiki/Ontology
2 http://en.wikipedia.org/wiki/Ontology_components
3 http://www.w3.org/RDF/
4 http://www.w3.org/TR/rdf-schema/
5 http://www.w3.org/TR/owl-features/
6 http://www.geneontology.org/GO.format.obo-1_2.shtml

Нашата работа има допирни точки главно с езика OBO, тъй като това е езикът, чрез който са зададени повечето публично достъпни анатомични онтологии.

Приложение на онтологиите в естествените науки

Онтологиите като модели за представяне (B50) и извод (B51, B34) на знания намират широко приложение в естествените науки като биология, биомедицина, медицина, анатомия, генетика, протеомика и др. Тези онтологии наричаме биоонтологии, а съответните изследователски проекти, в които те се използват като средство за моделиране и извод на знания – биоонтологични проекти. В тази част от дисертацията е направен подробен преглед на някои от най-значимите биоонтологични проекти и системи: **Gene Ontology (GO)**[1] [57], **GALEN**[2] [65,66,67,68], **UMLS**[3] [8,9], **FMA**[4] [10,11], **OBO** и **OBOFoundry**[5] [5,6,81].

Интегриране на онтологии – намиране на съответствия между онтологии и обединяване на онтологии

Силното развитие и популяризиране на онтологиите като модели за представяне на знания през последните десетина години е повлияно от идеята за създаването на ***семантична уеб мрежа***[6] (Semantic Web – B64).

Семантичната уеб мрежа представлява глобална идея и движение за по-нататъшното развитие на ***световната уеб мрежа*** (B65), стремящо се към обогатяването на уеб мрежата със семантична информация, което да направи възможно обработването на информацията, съдържаща се в нея от автоматизирани системи и агенти, превръщайки я по този начин от световна уеб мрежа в глобална семантична уеб мрежа. В рамките на семантичната уеб мрежа се предполага, че данните ще бъдат анотирани посредством онтологии. Затова, идеята за семантична уеб мрежа е пряко свързана с въпросите за интегриране на различни по произход, но подобни по предметна област, онтологии [3].

Един от проблемите пред реализацията на семантичната уеб мрежа в оригиналния ѝ вид е, че няма механизъм, по който да може да се наложи на различните индивиди, научни и бизнес организации използването на едно общоприето, стандартно множество от онтологии [3]. Не може да се очаква, че тези различни индивиди и организации ще се съгласят някога помежду си относно използването на една обща терминология или на едно общо множество от стандартни, общовалидни онтологии [88], които да описват предметните области – обект на човешкото познание. Следователно, трябва да се търсят други решения, а не просто налагането от когото и да било на подобно стандартно множество от онтологии. Алтернативата, за да се осигури взаимодействието между приложения и системи, използващи хетерогенни онтологии, е различията между тези онтологии да бъдат изгладени или преодолени след тяхното създаване.

---

1 http://www.geneontology.org/
2 http://www.opengalen.org/, http://www.openclinical.org/prj_galen.html
3 http://www.nlm.nih.gov/research/umls/
4 http://sig.biostr.washington.edu/projects/fm/
5 http://obofoundry.org/
6 http://en.wikipedia.org/wiki/Semantic_Web

Целта при интегрирането на онтологии е да стане възможна тяхната повторната употреба (reuse) и споделянето (sharing) на знанията, анотирани чрез тях, между хетерогенни софтуерни програми и инструменти [3].

В това се състои важността на проблема по интегрирането на хетерогенни онтологии, притежаващи сходни предметни област [22,23]. От своя страна, проблемът по интегриране на анатомичните онтологии на различни *биологични категории от организми* (видове, родове, семейства и др.) се явява именно частен случай на такъв проблем.

Терминологията, свързана с процеса по интегриране на онтологии, която използваме тук, сме възприели от [3]. Така ние използваме термина ***интегриране на онтологии*** (ontology mediation, ontology integration – B33) като сборно понятие, което обозначава решаването на всякакви задачи, свързани с: (i) намиране на съответствия между онтологии (ontology mapping – B63); (ii) съпоставяне на онтологии (ontology alignment, ontology matching – B62); (iii) обединяване на онтологии (ontology merging – B48).

При ***намирането на съответствия между онтологии***, установените съответствия между онтологиите се съхраняват отделно от онтологиите, те не са стават част от дадените онтологии. Те могат да бъдат използвани за комуникация посредством определени заявки с множество от хетерогенни източници на знания посредством някакъв общ интерфейс, а също и за трансформиране на знания и информация между различни представяния и формати. Автоматичното или полуавтоматично установяване на съответствия между онтологии се обозначава като ***съпоставяне на онтологии***. При ***обединяването на онтологии*** се създава нова онтология, която включва в себе си знанията от дадените онтологии. Основните трудности при обединяването на онтологии се състоят в това да се осигури, че както всички *съответствия*, така и всички *различия* между онтологиите са отразени в новата онтология.

В нашата работа приемаме, че разликата между термините *намиране на съответствия между онтологии* и *съпоставяне на онтологии* е доста малка и използваме първия от двата термина, като смислово имаме предвид значението и на двата.

В основния текст на дисертацията са представени някои известни методи, алгоритми, приложения и инструменти с общо предназначение за намиране на съответствия между и за обединяване на онтологии като **MAFRA**[1] [89]**, RDFT** [91], **PROMPT** [94,95], **Anchor-PROMPT** [96], **QOM** [97,98], **OntoMerge**[2] [55,78]. Казваме, че те са с общо предназначение, понеже те могат да бъдат използвани за интегриране на онтологии с произволни предметни области.

Намиране на съответствия между и обединяване на анатомични онтологии – проектът Uberon

Що се отнася до намирането на съответствия между и обединяването на *анатомични онтологии*, най-мащабният проект е Uberon[3]. Основна цел на проекта Uberon е обединяването на отделните съществуващи анатомични онтологии (на

[1] http://mafra-toolkit.sourceforge.net/
[2] http://cs-www.cs.yale.edu/homes/dvm/daml/ontology-translation.html
[3] http://uberon.org/

различни биологични категории организми от животинското царство) в една или в няколко общи организмово-неутрални анатомични онтологии.

Проектът стартира през 2008-2009 година [56]. Една от целите му е да се запълни празнината, формирана от липсата на единна организмово-неутрална онтология, описваща анатомиите на множество животински видове. Тази липса е основната пречка при превеждането на изследванията, извършени върху моделни организми, към области като тези на човешкото здраве и медицината. Друга цел на проекта е да се запълни празнината между референтната онтология CARO[1] (онтология от високо, абстрактно ниво) и различните конкретни, съществуващи вече или разработвани в бъдеще, видово-специфични анатомични онтологии. В рамките на Uberon се изследват множество публично достъпни онтологии, някои от които са чисто анатомични, а други съдържат в себе си в неявен вид вградени анатомични онтологии.

В началото на 2012 година авторите на Uberon анонсират първата завършена версия на онтологията Uberon [84]. Освен вътрешни понятия и релации, Uberon съдържа също така и множество изходящи връзки към различните съществуващи видово-специфични анатомични онтологии, с които онтологията Uberon е интегрирана към настоящия момент.

Глава III. Формализация на проблема

В тази глава формализираме входните данни и използваните външни източници на знания (ВИЗ – А02), използвани в настоящата работа. При това изхождаме от нашето възприемане на онтологиите, с които работим, като насочени ациклични графи (НАГ – А06).

Представени са три модела: модел #1 – модел на входните онтологии, представени като НАГ; модел #2 – модел на входните онтологии след намирането на съответствията между тях; модел #3 – модел на изходната т.нар. супер-онтология.

За двете входни онтологии възприемаме следните представяния.

$$O_1: G_1 = DAG_1 = (V_1, E_1); F_1 : E_1 \rightarrow C = \{ c_1, c_2, \ldots, c_n \}$$

$$O_2: G_2 = DAG_2 = (V_2, E_2); F_2 : E_2 \rightarrow C = \{ c_1, c_2, \ldots, c_n \}$$

Тук O_1 и O_2 са двете входни онтологии. Всяка от тях се състои от НАГ, означен с G_k или DAG_k, и функция на оцветяване на ребрата на този граф F_k. Върховете на тези графи представляват понятията от двете дадени анатомични онтологии, а ребрата им – релациите между понятията. $C = \{ c1, c2, \ldots, c_n \}$ е множеството от цветовете, използвани при оцветяването. Всеки цвят обозначава един от няколко възможни видове вътрешноонтологични релации. Тези релации са релации на включване (В24) от определен тип.

Най-често използвани в анатомичните онтологии са релацията ***is_a*** (генерализация/специализация) и ***part_of*** (агрегация/принадлежност). Понякога се използват и други релации, но за целите на нашата работа приемаме, че разполагаме само с тези два вида релации, т.е. приемаме, че $n=2, c_1=is\_a, c_2=part\_of$.

В означенията, въведени току-що, V_1 е множеството от понятия (термини) от анатомичната онтология на един даден организъм, а V_2 е множеството от

[1] http://www.bioontology.org/wiki/index.php/CARO:Main_Page

понятия (термини) от анатомичната онтология на друг (също даден) организъм. Въвеждаме още следните означения.

$$V_1 = \{v_{11}, v_{12}, ..., v_{1n_1}\},\ |V_1| = n_1;\ V_2 = \{v_{21}, v_{22}, ..., v_{2n_2}\},\ |V_2| = n_2$$

Релациите в тези два дадени графа наричаме вътрешноонтологични релации – те са винаги от типа ***родител-дете***, като този тип от своя страна се разделя на няколко други вида релации, най-важни и най-широко използвани сред които са ***is_a*** и ***part_of***. Вътрешноонтологичните релации във входните онтологии са асиметрични, т.е. при тях е от значение кой е родителя и кое е детето. Различните видове вътрешноонтологични релации моделираме като различни цветове, в които са оцветени съответстващите им в двата графа дъги/ребра.

Следват няколко важни забележки, които целят да изяснят някои основни аспекти от настоящата работа.

Първо, навсякъде в текста на дисертацията под ***релация от типа родител-дете*** се разбира произволна асиметрична релация. Тъй като ***is_a*** и ***part_of*** са асиметрични, за нас те и двете са релации от типа родител-дете.

Второ, анатомичните онтологии, които използваме тук, понякога съдържат не само информация за възрастния организъм, но и описват различни фази от развитието на организма – било то пренатално или постнатално. Затова в някои от тези онтологии се срещат също и понятия и релации, свързани именно с процесите и фазите на развитие (development). Примери за такива са например релациите: ***develops_from***, ***start_stage***, ***end_stage***, ***preceded_by***. Ние не разглеждаме тези релации в нашата работа, понеже се интересуваме само от анатомиите на възрастните организми.

Трето, важно е да се отбележи, че като цяло в настоящата работа не се занимаваме с вътрешноонтологични релации, различни от ***is_a*** и ***part_of***. Постъпваме така, понеже много от съществуващите анатомични онтологии въобще не съдържат други видове релации освен ***is_a*** и ***part_of***, а за онези онтологии, при които такива допълнителни релации съществуват, липсват унифицирани смислови значения на тези релации, които да се простират отвъд границите на конкретната онтология. В теоретичен план обаче, алгоритмичният метод за намиране на съответствия между анатомични онтологии и процедурата за обединяване на такива онтологии, които са предложени в настоящата работа, могат да се прилагат и при разглеждане на други видове асиметрични вътрешноонтологични релации, а не само на ***is_a*** и ***part_of***.

На четвърто място, следва да подчертаем, че навсякъде в текста на дисертацията, когато става дума за онтологии и в частност за анатомични онтологии, сме използвали понятията "понятие" и "термин" като синоними. Като синоними използваме също така и понятията "ребро" и "дъга" от теорията на графите, въпреки че според някои автори понятието "дъга" (на английски arc) следва да се използва само за ориентирани т.е. за насочени графи, а понятието "ребро" (на английски edge) – само за неориентирани т.е. за ненасочени графи.

Наличните ВИЗ и информацията, съдържаща се в тях, представяме формално като съвкупност от *множеството от термините* и *множеството от релациите*, дефинирани в тези ВИЗ.

● *Множество от понятия/термини*

$$M_s = \{t_{s1}, t_{s2}, ..., t_{sm_s}\}$$

Тук $t_{sk} = (id_{sk}; name_{sk})$ е термин, id_{sk} е идентификатор на термина t_{sk} (низ), $name_{sk}$ е име на термина t_{sk} (също низ), m_s е броят термини във ВИЗ T_s.

● *Множество от релации на включване*

$$R'_{T_s} = R^{is\_a}_{T_s} \subseteq M_s \times M_s ;\ R''_{T_s} = R^{part\_of}_{T_s} \subseteq M_s \times M_s$$

Това са релациите, които всеки ВИЗ дефинира над множествата от своите термини. Тук става дума за релациите ***is_a*** и ***part_of***, но по начина, по който те се дефинират от съответния ВИЗ, означен с $\mathbf{T_s}$ (s=1,2,3). Използваните ВИЗ обикновено дефинират и други релации (освен ***is_a*** и ***part_of***), но ние отново се ограничаваме само до разглеждането на тези две релации.

На базата на въведените тук означения и формални дефиниции на входните анатомични онтологии и на използваните ВИЗ, в следващата глава предлагаме формална формулировка на задачата (по намиране на съответствия между тези онтологии и по обединяването им) и представяме описание на предложения алгоритмичен метод за решаването ѝ.

Въпреки, че в работата си сме се занимавали с анатомични онтологии, предложените тук метод и алгоритъм са *достатъчно общи* и биха могли по естествен начин да бъдат приложени и към друга предметна област, различна от тази на анатомията. Условието за това е да са налице достатъчно близки по тематика (до тази друга предметна област) външни източници на знания (ВИЗ – А02). Всъщност единият от източниците на знания, използвани от нас (WordNet), би могъл да се прилага почти към всяка предметна област, тъй като той е източник на знания с общо предназначение.

Глава IV. Откриване на междуонтологични връзки за установяване на съответствия между анатомични онтологии. Обединяване на анатомични онтологии

Постановка на задачата

Нашата задача се състои в откриването на семантични връзки между две дадени анатомични онтологии O_1 и O_2 – например между онтология #1 (на мишка – В15) и онтология #2 (на риба Данио – В16) – чрез използването на наличните ВИЗ (А02) и на ***is_a*** и ***part_of*** релациите, които тези ВИЗ дефинират между своите понятия/термини. В нашия случай ВИЗ са три на брой – T_1, T_2, T_3 ***(UMLS, FMA, WordNet)***. Целта при това е да се генерира или предскаже множество от надеждни биологично/анатомично обосновани семантични релации/връзки между термините на входните онтологии Тези семантични релации трябва да бъдат от следните типове: $R_1 = R_{syn}$ (синоними), $R_2 = R_{hyper}$ (хиперними), $R_3 = R_{hypo}$ (хипоними), $R_4 = R_{holo}$ (холоними), $R_5 = R_{mero}$ (меронйми).

Така нашата цел е да се установят релации от изброените тук типове, такива че $R_k \subseteq (V_1 \times V_2) \cup (V_2 \times V_1)$, за $k=1,2,3,4,5$ и такива че тези релации са анатомично обосновани. От най-голям интерес е установяването на синонимните релации, тъй като те най-директно ни позволяват да стигнем до намирането на

съответствия (B11) между двете входни онтологии O_1 и O_2 и в крайна сметка до тяхното обединяване (B12) в една обща (изходна) онтология O_{super}, която наричаме супер-онтология.

Алгоритмично решение

Етап 1 – Генериране на речници.

Това е първият подготвителен етап на алгоритъма. При него от анатомичните онтологии O_k се построяват техните синонимни речници Th_k (B04) ***(за k=1,2)***. Речниците (B04) Th_k са таблични структури, подобни на хеш-таблици (B05). За идентификаторите ***id*** на всевъзможните понятия $t \in V_k$ таблицата Th_k поддържа списък $Th_k[t.id]$, съдържащ както основното (първичното) име, така и алтернативните (вторичните) имена, ако такива съществуват, на понятието $t \in V_k$. Списъците $Th_i[t.id]$ (за $i=1,2$) съдържат низове, които задават имената на понятието ***t,*** чиито идентификатор е ***id***.

Етап 2 – Съпоставяне на двете входни онтологии към наличните външни източници на знания.

Това е вторият подготвителен етап на алгоритъма. Тук всяка от двете входни онтологии O_1 и O_2 бива съпоставена (B01) към всеки от наличните ВИЗ. Както вече споменахме, в нашия конкретен случай това са $T_1=T_{UMLS}$, $T_2=T_{FMA}$, $T_3=T_{WordNet}$, но техният брой не е от съществено значение. На практика на този етап не самите онтологии, а речниците Th_1 и Th_2, генерирани от тях, биват съпоставени (B01) към наличните ВИЗ. Съпоставянето става по начина, описан по-долу.

Процедура 2.1: За всеки идентификатор k на термин $t \in V_1 \rightarrow$ извличаме списъка $L = Th_1[k]$ от построения вече речник Th_1 на O_1.

Процедура 2.2: За всяко име на термин $s \in L \rightarrow$ извличаме от ВИЗ T_1 всички различни идентификатори (идентификатори, дефинирани от T_1), които съответстват на името s, т.е. извличаме:

$$RS_1 = \{(t^I.id) \mid t^I \in T_1 \ \& \ t^I.name = s\}$$

Стъпка 2.2.1: За всеки идентификатор ***id*** from $RS_1 \rightarrow$ извличаме от ВИЗ T_1 множеството:

$$RS_2 = \{(t^{II}.id, t^{II}.name) \mid t^{II} \in T_1 \ \& \ t^{II}.id = t^I.id\}.$$

След като тази стъпка е изпълнена, резултатът е, че синонимите на s (дефинирани от ВИЗ T_1) са вече известни. Наричаме ги ***T_1–синоними на s***. Формално погледнато, това е множеството:

$$RS_2^* = \{t^{II}.id \mid t^{II} \in T_1 \ \& \ t^{II}.id = t^I.id\}$$, състоящо се от първите компоненти на наредените двойки, съдържащи се в RS_2.

Стъпка 2.2.2: За всеки идентификатор ***id*** от $RS_1 \rightarrow$ извличаме от ВИЗ T_1 множеството:

$$RS_3 = \{(t^{III}.id) \mid t^{III} \in T_1 \ u \ ((t^{III}, t^I) \in R_{T_1}^{is\_a} \text{ или } (t^{III}, t^I) \in R_{T_1}^{part\_of})\}$$

След като тази стъпка е изпълнена, резултатът е, че следните две множества са вече известни:

• Меронимите на s, дефинирани от T_1 и наричани още ***T_1–мерономи на s***; формално погледнато, това е множеството:

$$RS_{3,1} = \{t^{III}.id \mid t^{III} \in T_1 \ \& \ (t^{III}, t^{I}) \in R_{T_1}^{part\_of}\}.$$

• Хипонимите на ***s***, дефинирани от T_1 и наричани още ***T_1–хипоними на s***; формално погледнато, това е множеството:

$$RS_{3,2} = \{t^{III}.id \mid t^{III} \in T_1 \ \& \ (t^{III}, t^{I}) \in R_{T_1}^{is\_a}\}.$$

Тук следва да се отбележи, че са изпълнени следните съотношения:

$$RS_{3,1} \cup RS_{3,2} = RS_3 \ \ \& \ \ RS_{3,1} \cap RS_{3,2} = \emptyset.$$

Стъпка 2.2.3: За всеки идентификатор ***id*** от $RS_1 \rightarrow$ извличаме от ВИЗ T_1 множеството

$$RS_4 = \{(t^{IV}.id) \mid t^{IV} \in T_1 \ \textit{и} \ ((t^{I}, t^{IV}) \in R_{T_1}^{is\_a} \ \textit{или} \ (t^{I}, t^{IV}) \in R_{T_1}^{part\_of})\}.$$

След изпълнението на тази стъпка, резултатът е, че следните множества са вече известни:

• Холонимите на ***s***, дефинирани от T_1 и наричани още ***T_1–холоними на s***; формално погледнато, това е множеството:

$$RS_{4,1} = \{t^{IV}.id \mid t^{IV} \in T_1 \ \& \ (t^{I}, t^{IV}) \in R_{T_1}^{part\_of}\}.$$

• Хипернимите на ***s***, дефинирани от T_1 и наричани още ***T_1–хиперними на s***; формално погледнато това е множеството:

$$RS_{4,2} = \{t^{IV}.id \mid t^{IV} \in T_1 \ \& \ (t^{I}, t^{IV}) \in R_{T_1}^{is\_a}\}.$$

Тук следва да се отбележи, че са изпълнени следните съотношения

$$RS_{4,1} \cup RS_{4,2} = RS_4 \ \ \& \ \ RS_{4,1} \cap RS_{4,2} = \emptyset.$$

Етап 3 – Откриване на междуонтологични синонимни релации и междуонтологични релации от типа родител-дете (is_a/part_of)

В рамките на този етап се прилагат три отделни алгоритмични процедури, които означаваме с ***DM (direct matching)***, ***SMP (source matching predictions)*** и ***CMP (child matching predictions).*** След прилагането на всяка една от тези алгоритмични процедури двата входни НАГ са свързани с множество междуонтологични връзки (DM, SMP, CMP връзки). Полученият така граф ***G=(V,E)*** наричаме *текущ резултантен граф*. Всяка от процедурите, които описваме тук, добавя нови междуонтологични връзки към текущия резултантен граф ***G***, променяйки го по този начин.

Процедура 3.1: В хода на тази процедура (**DM**) се откриват директни (текстови, синтактични) съвпадения и на тяхна база се генерират предсказания за междуонтологични синоними. Между понятията в двете онтологии се търси текстово съвпадение на имената им. Процедурата е практически тривиална – итерира се по всички термини $t_1 \in V_1$ и $t_2 \in V_2$ и се проверява дали е изпълнено $t_1.name = t_2.name$. При откриване на такова съвпадение t_1 и t_2 се маркират като синоними и се отбелязва, че предсказанието произхожда от директно съвпадение (от ***DM***). Тези предсказания наричаме още ***DM–предсказания***.

Процедура 3.2: В хода на тази процедура биват генерирани още предсказания. Те се отнасят и за синонимни междуонтологични връзки, и за междуонтологични връзки родител-дете (***is_a/part_of***). В този момент двете входни

онтологии вече са съпоставени към наличните ***ВИЗ***. На базата на това се прилагат прости логически правила от едно предварително конструирано множество. Прилагането на тези правила генерира предсказания, които наричаме още ***SMP–предсказания***. Тук изброяваме логическите правила.

Правило (A) Ако два термина $t_M \in O_1$ и $t_Z \in O_2$ са били открити като синоними на един и същи термин $t \in T_k$, то тогава t_M и t_Z биват маркирани като предсказани (посредством SMP) междуонтологични синоними един на друг.

Правило (B) Ако терминът $t_j \in O_j$ е бил открит като синоним на термина $t \in T_s$ ***(s=1,2,3)*** и ако терминът $t_{3-j} \in O_{3-j}$ е бил открит като (***is_a/part_of***) дете/родител на t, то t_j бива маркиран като предсказан междуонтологичен (***is_a/part_of***) родител/дете на t_{3-j} (тук ***j=1 или 2*** и съответно ***3-j=2 или 1***).

Чрез прилагането на горните правила се установява едно (отнапред неизвестно) множество от междуонтологични връзки/релации (синонимни и родител-дете) между върховете на графите DAG_1 and DAG_2 (т.е. между понятията от O_1 и O_2). Доказателствата/предпоставките за тяхното съществуване се съдържат в информацията, съхранявана в наличните ВИЗ. За установените (чрез ***SMP***) релации от типа родител-дете информацията за това дали те са ***is_a*** или ***part_of*** връзки бива запазена.

Процедура 3.3: Тук привеждаме описанието на т.нар. *процедура по намиране на съвпадения на базата на децата* (child matching procedure – ***CMP***). Тя води до установяването на още междуонтологични връзки (предсказания), които наричаме ***CMP-предсказания***.

Паралелно с описанието на ***CMP*** привеждаме някои дефиниции, които задават оценки на всяка една връзка (била тя вътрешноонтологична или пък междуонтологична) от текущия резултантен граф G във вида, който той има преди началото на изпълнението на тази процедура. Дефинициите са дадени, следвайки един йерархичен принцип, движейки се от по-просто към по-сложно, надграждайки една над друга. Те ни позволяват да достигнем до едно число, наречено ***финална (обобщена) CMP оценка*** на ***финалната (обобщената) CMP връзка***, която бива установена между произволни два върха $\mathbf{t_1} \in V_1$ и $\mathbf{t_2} \in V_2$.

CMP процедурата се опитва да намери нови връзки от типовете R_1, R_2, R_3, R_4 и R_5 между термини $\mathbf{t_1} \in \mathbf{V_1}$ и $\mathbf{t_2} \in \mathbf{V_2}$. Тя разглежда ***шаблони на свързаност***, в които участват $\mathbf{t_1} \in \mathbf{V_1}$ (родител 1) и $\mathbf{t_2} \in \mathbf{V_2}$ (родител 2), както и децата на $\mathbf{t_1}$ и $\mathbf{t_2}$ от двете входни онтологии. Тези шаблони се търсят в графа G, получен след прилагането на DM и SMP процедурите. Три различни вида шаблони на свързаност се разглеждат от процедурата CMP:

(1) $\mathbf{t_1} \in \mathbf{V_1} \leftarrow \mathbf{t_{ch1}} \in \mathbf{V_1} \leftrightarrow \mathbf{t_{ch2}} \in \mathbf{V_2} \rightarrow \mathbf{t_2} \in \mathbf{V_2}$ (наричаме това **U-pattern**);

(2) $\mathbf{t_1} \in \mathbf{V_1} \leftarrow \mathbf{t_{ch2}} \in \mathbf{V_2} \leftrightarrow \mathbf{t_{ch1}} \in \mathbf{V_1} \rightarrow \mathbf{t_2} \in \mathbf{V_2}$ (наричаме това **X-pattern**);

(3) $\mathbf{t_1} \in \mathbf{V_1} \leftarrow \mathbf{t_{ch1}} \in \mathbf{V_1} \rightarrow \mathbf{t_2} \in \mathbf{V_2}$ или $\mathbf{t_1} \in \mathbf{V_1} \leftarrow \mathbf{t_{ch2}} \in \mathbf{V_2} \rightarrow \mathbf{t_2} \in \mathbf{V_2}$ (наричаме това **V-pattern**).

В тази нотация стрелките → и ← обозначават множества от връзки от типа родител-дете, които или произхождат от DM/SMP, или са част от входните онтологии (IO връзки). Стрелките винаги са насочени от детето към родителя; това са асиметрични връзки. Стрелките ↔, от друга страна, обозначават множества от синонимни връзки, произхождащи от DM или SMP; те са симетрични връзки.

Всяко срещане на такъв шаблон между $\mathbf{t_1}$ и $\mathbf{t_2}$ (между двата върха-родители) наричаме ***инстанция на съответния шаблон***. Всички асиметрични връзки в

рамките на даден шаблон обозначават или ***is_a***, или ***part_of*** връзки, т.е. тук не позволяваме смесването на тези два вида асиметрични връзки в рамките на една инстанция на даден шаблон.

На базата на тези шаблони на свързаност в рамките на настоящата процедура въвеждаме нови междуонтологични връзки (CMP връзки) между $\mathbf{t_1}$ и $\mathbf{t_2}$. Наричаме ги индивидуални ***CMP връзки***. За да им присвоим оценки, въвеждаме понятията ***оценка на множество от не-CMP връзки*** и ***оценка на инстанция на шаблон (или оценка на индивидуална CMP връзка)***. Накрая оценките на всички индивидуални CMP връзки между $\mathbf{t_1}$ и $\mathbf{t_2}$ се агрегират посредством т.нар. ***агрегираща функция***. По-долу привеждаме най-важните дефиниции, свързани с CMP процедурата.

Дефиниция 1 (Conj): Функцията ***Conj*** приема N аргумента от [0, 1] и връща резултат в [0, 1]. Дефинираме я рекурентно както следва:

5.1. $Conj\ (A_1, A_2) = A_1 \bullet A_2$;

5.2. $Conj\ (A_1, A_2, \ldots, A_N) = Conj\ (\ Conj\ (A_1, A_2, \ldots, A_{N-1}), A_N\ )$, за $N \geq 3$. □

Дефиниция 2 (Disj): Функцията ***Disj*** приема N аргумента от [0, 1] и връща резултат в [0, 1]. Дефинираме я рекурентно както следва:

6.1. $Disj\ (A_1, A_2) = A_1 + A_2 - A_1 \bullet A_2$;

6.2. $Disj\ (A_1, A_2, \ldots, A_N) = Disj\ (\ Disj\ (A_1, A_2, \ldots, A_{N-1}), A_N\ )$, за $N \geq 3$. □

Дефиниция 3 (оценка на не-CMP връзка): Оценка на произволна връзка, която не произлиза от CMP, дефинираме както следва:

$$score(s_{ij}) = \begin{cases} I, & \text{ако } s_{ij} \text{ е IO връзка} \\ D, & \text{ако } s_{ij} \text{ е DM връзка} \\ f(T), & \text{ако } s_{ij} \text{ е SMP връзка, която произлиза от ВИЗ } T \end{cases}$$

В тази формула: IO означава вътрешноонтологична връзка; DM означава връзка, която произлиза от DM процедурата; SMP означава връзка, която произлиза от SMP процедурата; s_{ij} е произволна връзка от видовете IO, DM или SMP; I и D са фиксирани константи в [0, 1] (най-често единици); $f(T)$ е също константа, зависеща от конкретния ВИЗ T. □

Дефиниция 4 (оценка на множество от не-CMP връзки): Оценка на множество от не-CMP връзки дефинираме така: $score(\bar{S}_i) = Disj_{j=1}^{m_i}(score(s_{ij}))$. Тук ***Disj*** е функцията от дефиниция 2, s_{ij} са връзки (IO/DM/SMP връзки), а ***Disj*** се прилага върху всички връзки, участващи в множеството $\bar{S}_i$. □

Дефиниция 5 (оценка на инстанция на шаблон, т.е. на индивидуална CMP връзка): Оценка на индивидуална CMP връзка e (оценка на конкретната инстанция на шаблона, породил тази връзка) дефинираме по следния начин: $score(e) = p.Conj_{i=1}^{n}(score(\bar{S}_i))$. Тук числото $p \in (0, 1)$ е фиксирана константа (CMP наказателна константа), ***Conj*** е функцията от дефиниция 1, а функцията ***Conj*** прилагаме върху всички множества от връзки, участващи в конкретната инстанция на шаблона, от която произлиза индивидуалната CMP връзка e. □

Дефиниция 6: Нека t_1=*Par1*$\in V_1$ и t_2=*Par2*$\in V_2$ да означават два термина от двете входни онтологии. Нека ***G*** означава графа, получен от DAG_1 и DAG_2 след като всички **DM връзки** и **SMP връзки** са вече установени (от процедури 3.1 и 3.2). Нека също така:

(6.1) $u = \{u_1, u_2, \ldots, u_{N_u}\}$ да е множеството от всички конкретни срещания на ***U-шаблони,*** в които t_1 и t_2 участват като родители ($N_u \geq 1$);

(6.2) $x = \{x_1, x_2, \ldots, x_{N_u}\}$ да е множеството от всички конкретни срещания на ***X-шаблони,*** в които t_1 и t_2 участват като родители ($N_x \geq 1$);

(6.3) $w = \{w_1, w_2, \ldots, w_{N_w}\}$ да е множеството от всички конкретни срещания на ***V-шаблони,*** в които t_1 и t_2 участват като родители ($N_w \geq 1$);

(6.4) $N_u + N_x + N_w > 0$;

(6.5) $PIS(t_1, t_2) = u \cup x \cup w$, като тук с ***PIS*** (A10) е означено множеството от конкретните срещания на шаблони, в които t_1 и t_2 участват като родители;

(6.6) $|PIS(t_1, t_2)| > 0$.

Тогава при изпълнение на CMP процедурата прекарваме ***обобщена (агрегирана, финална) CMP връзка*** $e_{CMP}(t_1, t_2)$ между t_1 и t_2 и дефинираме нейната оценка както следва:

$$score_{CMP}(t_1, t_2) = \underset{\forall p \in PIS(t_1, t_2)}{MAX} (score(p)).$$

Тук p означава конкретна инстанция/срещане на шаблон, т.е. функцията максимум се взема по всички срещания на шаблони, в които t_1 и t_2 участват като родители. □

Функцията максимум играе ролята на една ***агрегираща функция***, т.е. максимумът се явява частен случай на такава функция. При практическа реализация на ***CMP*** процедурата бихме могли да използваме за агрегиращи функции и някои алтернативи на функцията максимум.

В хода на представеното тук описание на ***CMP*** процедурата беше показано как индивидуални синонимни ***CMP*** връзки могат да бъдат построени между произволни два върха $t_1 \in V_1$ и $t_2 \in V_2$, стига да е изпълнено условието **(6.4)** или еквивалентното на него **(6.6)**. После беше дефинирана оценка за всяка една от тези индивидуални ***CMP*** връзки. Накрая беше показано как множество от индивидуални ***CMP*** връзки между два върха $t_1 \in V_1$ и $t_2 \in V_2$ могат да бъдат

агрегирани в една единствена т.нар. обобщена или финална ***CMP*** връзка $e_{CMP}(t_1, t_2)$ между t_1 и t_2. Именно тези финални връзки заедно с техните числови оценки $score_{CMP}(t_1, t_2)$ формират резултата от процедурата ***CMP***.

Обединяване на две онтологии

Тук представяме някои дефиниции и доказваме (в основния текст на дисертацията) две твърдения, позволяващи трансформирането на ***текущия резултантен граф G*** до един специален, обобщен граф ***G**** и генерирането на изходна т.нар. супер-онтология (при определени условия) от получения по този начин граф ***G****.

Дефиниция 8: Нека RZ е релация между върховете на G (т.е. $RZ \subseteq V \times V$), дефинирана по следния начин: $(v_1, v_2) \in RZ$, ако между върховете v_1 и v_2 съществува поне една синонимна връзка в графа G. Нека също така приемем, че $(v, v) \in RZ$ за всяко $v \in V$. □

Ясно е, че така дефинираната релация RZ е симетрична, т.е. ако е налице $(v_1, v_2) \in RZ$, то и $(v_2, v_1) \in RZ$. Това е така, понеже синонимните връзки в графа G по своява същност са двупосочни връзки (т.е., което е равнозначно на първото, са ненасочени). По дефиниция приемаме релацията RZ за рефлексивна.

Дефиниция 9: Нека RZ^* означава транзитивното затваряне на релацията RZ. □

Лесно се вижда, че RZ^* е релация на еквивалентност, дефинирана върху множеството от върховете на графа G.

Дефиниция 10: Дефинираме графа $G^* = (V^*, E^*)$ по следния начин: нека $V^* = \{C_1^*, C_2^*, ..., C_n^*\}$ е множеството от класовете на еквивалентност (във V), породени от RZ^*; нека също така произволна дъга $e = (C_k^*, C_l^*)$ с цвят $c \in \{is\_a, part\_of\}$ принадлежи на E^* тогава и само тогава, когато $\exists u \in C_k^*$ и $\exists v \in C_l^*$, такива че $u, v \in V$ и освен това между u и v съществува ребро $(u, v) \in E$ със същия цвят c. □

Дефиниция 11: Даден цикъл в графа G наричаме допустим, ако всички ребра, участващи в него, представляват синонимни връзки, т.е. ако той не съдържа нито едно ребро от типовете ***is_a*** и ***part_of***. Даден цикъл в графа G наричаме недопустим, ако той не е допустим, т.е. ако той съдържа поне едно ребро от типа ***is_a*** или ***part_of***. □

Твърдение 1: Ако графът $\boldsymbol{G}$ съдържа само ***допустими цикли***, то графът $\boldsymbol{G}^*$ е ***ацикличен***. □

Твърдение 2: Ако графът $\boldsymbol{G}^*$ е ***ацикличен*** то графът $\boldsymbol{G}$ съдържа само ***допустими цикли***. □

Тези две твърдения задават едно необходимо и достатъчно условие графът $\boldsymbol{G}^*$ да бъде ацикличен, а това е равносилно на условието от него да може да бъде генерирана изходна супер-онтология. Двете твърдения са доказани в основния текст на дисертацията.

Глава V. AnatOM – софтуерно решение за намиране на съответствия между и за обединяване на анатомични онтологии

В тази глава е описана софтуерната програма AnatOM, чието наименование е съкращение от ***Anat****omical* ***O****ntologies* ***M****erger*. Програмата AnatOM реализира представените в предходните две глави теоретични модели и алгоритмични процедури.

Програмата AnatOM е разработена като част от тази работа с една конкретна практическа цел, а именно – да полуавтоматизира намирането на съответствия между анатомични онтологии и обединяването на тези онтологии. За целта трите алгоритмични процедури са интегрирани в програмата AnatOM, за да се получи едно завършено, цялостно софтуерно приложение, решаващо споменатите задачи. Програмата следва да се използва от специалисти по анатомия, които да обединяват с нейна помощ анатомични онтологии от различни категории организми. Специалистите по анатомия следва да редактират ръчно (to manually curate) автоматично генерираните от програмата предсказания и накрая да обединяват входните видово-специфични онтологии в изходна организмово-неутрална супер-онтология.

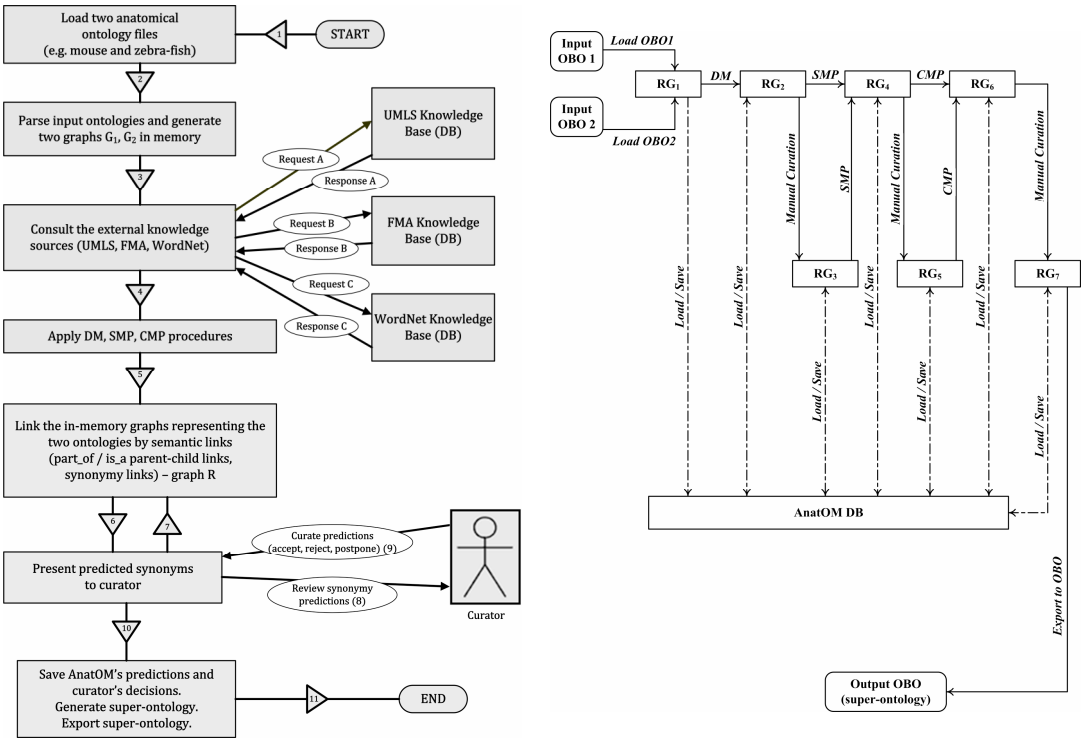
Програмата AnatOM представлява едно типично самостоятелно десктоп-базирано приложение (В45) с графичен потребителски интерфейс (ГПИ–А17), реализирано на платформата Microsoft.NET[1]. Езикът за програмиране, използван за имплементация на програмата е C# – най-важният и най-популярният език от платформата Microsoft.NET.

Входните онтологии се подават на програмата AnatOM под формата на OBO файлове (файлове с декларативни знания, описани чрез езика OBO[2]). Наличните три ВИЗ (UMLS, WordNet, FMA) са представени във вид на релационни бази от данни, управлявани от MySQL СУРБД (А11) [18]. Освен тях програмата AnatOM използва една допълнителна релационна MySQL база от данни, в която тя записва различни междинни резултати от изчисленията, извършвани в хода на изпълнението ѝ.

---

[1] http://www.microsoft.com/net

[2] http://www.geneontology.org/GO.format.obo-1_2.shtml



Фиг. 1: Основни етапи в хода на изпълнението на програмата AnatOM

На фиг. 1 са представени две диаграми, илюстриращи основните етапи и преходи в хода на изпълнение на програмата AnatOM. На втората диаграма с RG_k сме означили ***текущия резултантен граф*** и различните му състояния в хода на изпълнение на програмата. При подадена команда от менютата на програмата тя прочита и зарежда файловете, задаващи двете входни онтологии. Това се извършва от ***модула OBOParser.NET***. След това специален конструиращ обект (B47) конвертира обектите, получени от това прочитане и ги трансформира в Graph/Node/Edge обекти, т.е. в обекти дефинирани от ***модула Graph.NET***. Отново от менютата на програмата се стартират отделни нишки (threads), които изпълняват трите алгоритмични процедури – DM, SMP, CMP. Нишките съдържат основната логика на програмата и формират т.нар. ***логически модул*** на програмата. Тези нишки работят върху Graph обектите G_1 и G_2 на двете входни онтологии, като в хода на работата си добавят нови ребра към тези Graph обекти. Добавените ребра са такива, че единият им край е в G_1, а другият – в G_2. Това са междуонтологичните връзки, описани в изложението по-горе. DM и CMP процедурите нямат нужда от комуникация с трите релационни бази от данни, представящи трите налични ВИЗ (UMLS, FMA, WordNet). SMP е онази от трите процедури, която комуникира с наличните ВИЗ. За тази комуникация SMP процедурата използва ***модула DataAccess***, осигуряващ връзката с използваните релационни бази от данни. Логиката по изпълнение на DM, SMP, CMP е реализирана в класове, дефиниращи три нишки (B44). Всяка от трите нишки работи върху т.нар. ***текущ резултантен граф***. Преработеният Graph обект се подава на ***визуализиращия модул*** за представянето му в графичен вид. Накрая

експортиращият модул на програмата дава възможност за експортиране на получения като резултат граф под формата на супер-онтология в OBO формат. Следва описание на модулите, от които се състои програмата AnatOM.

Модул OBOParser.NET

Модулът OBOParser.NET представлява специализирана библиотека, разработена с цел прочитане на OBO файлове. Действието му се базира на събития. Този модул дефинира структури, описващи основните елементи от синтаксиса на OBO [4] – строфи, термини, идентификатори на термини, имена на термини, синоними на термини, релации между термини. Тези структури се предоставят на клиентите на модула след прочитане от него на съответния OBO файл. Това е реализация на типичен парсър с рекурсивно спускане (A13). Модулът OBOParser.NET се извиква при зареждане от програмата AnatOM на входните анатомични онтологии; това действие се инициира от главното меню File на програмата AnatOM.

Модул Graph.NET

Модулът Graph.NET е библиотека с общо предназначение за работа с графи. Тя дефинира класовете Node, Edge, Graph и позволява също така съхраняването на произволни обекти като характеристики (properties) към всеки Node обект и към всеки Edge обект от графа.

На своите клиенти, модулът предоставя методи за: добавяне на върхове, добавяне и премахване на ребра, извличане на всички ребра на графа, извличане на всички ребра между два дадени върха от графа, извличане на всички родители/деца на даден връх, извличане на началния и крайния връх на дадено ребро, преброяване на всички върхове/ребра на графа, прилагане на различни полезни алгоритми, работещи върху графи като например топологично сортиране, алгоритъм на Tarjan [10], алгоритъм на Johnson [9].

Библиотеката може лесно да бъде разширена и обогатена и с други полезни алгоритми за работа с графи. Тя е достатъчно обща и може лесно да се интегрира с други програми, а не само с AnatOM.

Модул DataAccess

Модулът DataAccess е стандартен модул, позволяващ на програмата AnatOM да изпълнява различните видове SQL заявки (select, insert, update, delete), а също и SQL процедури и SQL функции, инсталирани като част от четирите релационни бази от данни, споменати по-горе.

Той съдържа класове за комуникация с четирите релационни бази от данни, използвани от програмата AnatOM – umls, fma, wordnet и anatom. Първите три от тях представляват трите налични ВИЗ, с които програмата AnatOM комуникира. Четвъртата база от данни съдържа: (i) нашата т.нар. KNOWLEDGE таблица, която съдържа множество важни, получени чрез ръчно въвеждане от специалист по анатомия, организмово-неутрални твърдения от анатомично естество (към този момент твърденията са валидни основно за категорията на гръбначните животни); (ii) всички междинни данни от изчисленията, извършвани от програмата AnatOM.

Графичен потребителски интерфейс и логически модул

Логическият модул поддържа текущия резултантен граф и реализира трите алгоритмични процедури – DM, SMP, CMP. Графичният потребителски интерфейс (ГПИ – A17) служи за контрол над тяхното изпълнение и за представяне на резултатите от тях в табличен и в графичен вид. Трите алгоритмични процедури се изпълняват от програмата AnatOM в отделни нишки (threads – B44). Във всеки момент от работата на програмата потребителят има достъп до текущия резултантен граф и може да го разглежда и редактира. Трите нишки, съставляващи логическия модул се стартират от главното меню Action на ГПИ на програмата.

Визуализиращ модул – GraphVisualizer.NET

Модулът GraphVisualizer.NET служи за визуализиране на графите, с които работи програмата AnatOM. Той реализира подхода, базиращ се на уравновесяването на физични сили [11], поради неговата относителна простота и интуитивност. При този подход графът се моделира като физична система, чиито върхове се разглеждат като материални точки, притежаващи определени електрически заряди и чиито ребра се приемат за пружини, свързващи съответните материални точки. Алгоритъмът е итеративен и действа дотогава, докато така дефинираната физична система достигне до своя еквилибриум – състоянието, в което системата притежава най-малка (практически нулева) кинетична енергия.

Визуализиращият модул се стартира, когато потребителят избере в ГПИ една конкретна междуонтологична връзка, предсказана от AnatOM. Когато това се случи, двата върха, свързани с тази връзка, се визуализират от модула GraphVisualizer.NET заедно с всички ребра, прилежащи към тях.

Експортиращ модул

Експортиращият модул действа върху текущия резултантен граф и при определено условие генерира като изход супер-онтология под формата на OBO файл. Необходимото и достатъчно условие (НДУ – B69), за да бъде възможно генерирането от програмата на изходна, обединяваща, т.нар. супер-онтология, е това текущият резултантен граф да не съдържа в себе си недопустими цикли (това НДУ е доказано в предходната глава в основния текст на дисертацията). Програмата AnatOM предоставя две възможности за проследяване на наличните в текущия резултантен граф цикли – *броене на циклите* и *експортиране на циклите* в текстов файл.

Експортиращият модул се стартира от главното меню Result на ГПИ на програмата. Изпълнението му е успешно, когато споменатото НДУ е изпълнено. В противен случай е нужна намесата на специалист по анатомия, който да извърши съответна работа с програмата и ръчно редактиране на някои от предсказанията, генерирани от нея, докато условието за липса на недопустими цикли бъде удовлетворено.

Глава VI. Анализ на проведените експерименти. Резултати. Дискусия

В тази предпоследна глава анализираме получените резултати при намирането на съответствия между анатомичните онтологии на три категории от организми (мишка /B15/, риба Данио /B16/, жаба /B41/) и при обединяването им (по двойки). Анализът се базира на оценки и резултати, получени от ръчно редактиране/оценяване, извършено от експерт по анатомия. Тези експертни оценки са представени с числата 1, 2, 3 и имат следния смисъл:

• **1 – напълно точно предсказание**: двата термина (от двете входни онтологии) наистина са свързани с релацията, предсказана от програмата AnatOM; типът на релацията (***is_a***, ***part_of***, ***synonymy)*** също е предсказан правилно;

• **2 – частично точно предсказание**: двата термина (от двете входни онтологии) наистина са свързани с релация, но тя не е точно релацията, предсказана от програмата AnatOM, например – програмата AnatOM е предсказала релация ***synonymy***, а реално релацията, съществуваща между двата термина, е ***is_a*** или ***part_of*** (или обратно).

• **3 – напълно неточно предсказание**: предсказаната релация от програмата е невярна и няма анатомичен смисъл; не съществува и друга, близка до предсказаната, релация между двата термина (от двете входни онтологии), с която евентуално да я е сбъркала програмата AnatOM.

Тези три вида оценки са използвани в анализа на получените от програмата AnatOM резултати.

Анализ на DM процедурата

DM процедурата генерира само предсказания за синонимни релации. Тя се базира само на текстови съвпадения между термините от двете входни онтологии. При ръчно оценяване на предсказаните от DM процедурата релации се установява много висок процент на напълно точните предсказания. Този процент е повече от 95% и при трите двойки сравнявани организми. Това е очаквано висок резултат, имайки предвид самото естество на DM процедурата.

Анализ на SMP процедурата

SMP процедурата използва силно наличните ВИЗ (A02) и генерира предсказания както за релации от типа родител-дете (***is_a***/***part_of***), така и за синонимни релации (***synonymy***). Процентът на напълно точните съвпадения е много висок и е над 86% и за двата типа предсказания, разгледани и за трите двойки организми. Това също е очаквано поради самото естество на SMP процедурата.

Анализ на CMP процедурата

CMP процедурата води до генерирането само на синонимни релации между термини от входните онтологии. Тя не използва пряко никой от наличните ВИЗ (A02), а само косвено (понеже входните данни за CMP процедурата са изходните данни от DM и SMP процедурите, а SMP процедурата използва наличните външни източници на знания).

Резултатите от работата на CMP процедурата е систематизирана в таблицата от фиг. 2. В първата колона на таблицата е представена двойката от организми, явяваща се обект на действие на процедурата. Във втората колона е показан произхода на предсказаните релации. В следващите четири колони са представени общият брой на предсказанията, както и бройките предсказания, разбити според получената оценка от ръчното оценяване, извършено от специалиста по анатомия.

Двойка	**Произход**	**Общ брой**	**Оценка 1**	**Оценка 2**	**Оценка 3**
Mus-Danio	CMP Only	693	21 (3.03%)	517 (74.60%)	155 (22.37%)
Mus-Danio	CMP + Other	109	104 (95.41%)	4 (3.67%)	1 (0.92%)
Mus-Danio	CMP Any	802	125 (15.59%)	521 (64.96%)	156 (19.45%)
Mus-Xenopus	CMP Only	595	26 (4.37%)	503 (84.54%)	66 (11.09%)
Mus-Xenopus	CMP + Other	125	120 (96.00%)	4 (3.20%)	1 (0.80%)
Mus-Xenopus	CMP Any	720	146 (20.28%)	507 (70.42%)	67 (9.30%)
Danio-Xenopus	CMP Only	566	23 (4.06%)	427 (75.44%)	116 (20.50%)
Danio-Xenopus	CMP + Other	146	137 (93.84%)	7 (4.79%)	2 (1.37%)
Danio-Xenopus	CMP Any	712	160 (22.47%)	434 (60.96%)	118 (16.57%)

Фиг. 2: Анализ на резултатите от CMP процедурата

За произхода на предсказанията (за втората колона от таблицата) са използвани следните означения: (i) "CMP Only" означава, че тези релации са предсказани единствено от процедурата CMP; те не са предсказани преди това нито от DM, нито от SMP процедурата; (ii) "CMP + Other" означава, че тези релации са предсказани както от процедурата CMP, така и от някоя от другите процедури – DM или SMP; (iii) "CMP Any" означава, че тези релации са предсказани от CMP процедурата, като тук се включват както релации, непотвърдени от никоя от другите процедури, така и релации, потвърдени от поне една от тях; оттук следва, че бройките на реда 3.n+1 (от таблицата от фиг. 2) са суми от бройките на редовете 3.n и 3.n-1 (за n=1,2,3).

При "CMP + Other" процентите на напълно точните предсказания са много високи, което е очаквано – те са около 93% и дори по-високи. Оттук следва извода, че предсказанията направени от CMP процедурата са с най-голям шанс да са верни, когато са потвърдени от поне една от другите две процедури.

При редовете, маркирани с "CMP Any", процентите на напълно верните предсказания (тези с оценка 1) са задоволително високи – около 15%–22%. В този случай огромният брой от CMP предсказанията са частично точни (имат оценка 2), което се дължи най-вече на факта, че използваните две входни анатомични онтологии притежават различни степени на *дълбочина* и на *грануларност*. Добър показател се забелязва що се отнася до напълно неточните предсказания (оценка 3). Процентно този показател не е висок – достига най-много до около 19%.

При редовете, маркирани с "CMP Only", се вижда, че има известен брой напълно верни предсказания (оценка 1) – най-често около 3%-4%, което е положителен резултат. Това означава, че CMP процедурата, която е чисто алгоритмична и не използва допитване до ВИЗ (A02), успява да намери релации, които са пропуснати и от DM, и от SMP процедурата. Процентът на напълно неточните предсказания (оценка 3) отново е задоволително нисък – не надминава

22%. Вижда се, че и тук огромният брой от CMP предсказанията са частично точни (оценка 2), което, както вече споменахме, се дължи на факта, че използваните входни анатомични онтологии притежават различни помежду си степени на *дълбочина* и на *гранулярност*.

Гранулярността обозначава размера (в смислово отношение) на стъпката между дадено понятие-родител и дадено негово понятие-дете в дадена онтология. Нека имаме примерна онтология A, която декларира, че

finger (пръст) *part_of* forelimb (преден крайник) (1)

и друга примерна онтология B, която декларира, че

finger (пръст) *part_of* hand (ръка) (2)
hand (ръка) *part_of* forelimb (преден крайник) (3)

Вижда се, че това твърдение, което онтологията A декларира само в рамките на единствена стъпка – (1), онтологията B формира в рамките на две стъпки – (2) и (3). Затова казваме, че онтология A е "по-груба" от B, а онтология B е "по-фина" от A или още, че гранулярността при онтология B е по-фина (fine grained), а гранулярността при A е по-груба (coarse grained).

Дълбочината (или още височината) обозначава разликата в нивата между най-детайлните термини (или листата) и най-общите термини (или корените) в дадена онтология. Най-общите термини в дадена онтология са тези, които не притежават нито един родител, а най-детайлните термини са тези, които не притежават нито едно дете. Понятието дълбочина е добре известно от теорията на графите. Там то най-често се използва за дървета, а тук е използвано за насочени ациклични графи (НАГ – A06).

Анализ на процеса на сливане и на генерираните от него изходни онтологии (супер-онтологии)

Тук са представени някои факти и наблюдения, отнасящи се до обединяването на анатомичните онтологии на мишката (B15) и на рибата Данио (B16) при работа върху тях с програмата AnatOM и по-точно – след прилагане върху тях на DM и SMP процедурите.

Анатомичната онтология на мишката се отнася само до *анатомията на възрастния организъм* – тя не съдържа термини, описващи развитието на организма в периода преди или след раждането. При анатомичната онтология на рибата Данио имаме *както анатомични термини, отнасящи се до възрастния организъм, така и термини, отнасящи се до различни фази от развитието на организма*. Термините от онтологията на мишката започват с низа "MA", а термините от онтологията на рибата започват или с низа "ZFA" (когато се отнасят до анатомични понятия), или с низа "ZFS" (когато се отнасят до понятия, обозначаващи фази от развитието на рибата). В двете онтологии имаме общо 2982 MA понятия, 2712 ZFA понятия и 46 ZFS понятия.

Броят на *оригиналните понятия* (от входните онтологии), от които произхожда дадено *обобщено понятие* (от супер-онтологията) наричаме степен на това обобщено понятие. Ясно е, че ZFS понятията нямат съответни в онтологията на мишката. Затова при сливане на двете онтологии те водят до обобщени понятия от степен 1. Ние се интересуваме предимно от броят на обобщените понятия от степен поне 2. Оказва се, че този брой е 255, т.е. около 10% от понятията

от анатомичните онтологии на мишката и на рибата Данио имат съответни в другия организъм. Подробен анализ разкрива, че супер-онтологията съдържа общо 5470 понятия, от които 1 понятие е със степен 5, 12 понятия са със степен 3, 242 понятия са със степен 2 и останалите 5215 понятия са със степен 1.

Тези резултати са приведени тук най-вече с илюстративна цел, тъй като те се получават само след прилагането на DM и SMP процедурите и понеже при тях е използвана само автоматичната работа на програмата AnatOM. При включване на CMP процедурата е нужна по-сериозна работа от страна на експерт по анатомия, който ръчно да редактира предсказаните от програмата връзки преди да бъде генерирана изходна онтология.

Като цяло, най-важен тук е факта, че при анатомичните онтологии на мишката и на рибата Данио около 10% от двете онтологии могат да бъдат "наложени" или "припокрити" една върху друга (B11) посредством синонимни релации, което води до генерирането в супер-онтологията на обобщени понятия със степен 2 или по-висока.

Дискусия относно някои възникнали в хода на работата проблеми

В нашата работа е експериментирано най-много с анатомичните онтологии на мишката (B15) и на рибата Данио (B16), но немалък брой експерименти са проведени също така и с анатомичната онтология на жабата (B41), търсейки връзките ѝ с току-що споменатите две.

От биологична гледна точка, трябва да обърнем внимание на някои проблеми, за да обясним по-пълно резултатите, получени при *намирането на съответствия между* и при *обединяването на* анатомични онтологии на различни организми.

Първо, трябва да се отбележи, че понякога *входните анатомични онтологии не са хомогенни по отношение на своя предмет*. Някои (риба Данио) включват фази от развитието на организма, а други (мишка) са фокусирани само върху анатомията на възрастния организъм. От този факт възникват някои проблеми, тъй като е почти невъзможно да се намерят съотвествия между ембрионални анатомични понятия (от едната онтология) и понятия, отнасящи се до възрастния организъм (от другата онтология). Пример за такъв проблем е този с понятието "coelom" от онтологията на рибата Данио и релацията му с понятието "pericardium" от онтологията на мишката. От анатомична гледна точка тези две понятия могат да бъдат свързани с връзката ***pericardium is_a coelom***, но и с връзката ***pericardium part_of coelom***. С други думи, тук имаме нееднозначност, отнасяща се до вида на връзката, съществуваща между тези две понятия.

Второ, трябва да отбележим друг проблем, произтичащ от *различните описания на потенциално идентични структури в латинския и в английския*. Във връзка с това се сблъскваме с доста проблеми при избора на адекватна релация между някои грубо (loosely) формулирани термини като напр. "cardiac muscle tissue" и някои стриктно (strictly) формулирани термини като напр. "myocardium". Често двата термина се използват като синоними, но стриктно погледнато те би следвало да са свързани по-скоро с друга релация (например с ***part_of***). Този избор на релация обикновено зависи от смисъла, който са придали на съответните понятия авторите на двете онтологи, а и от интерпретацията на експерта, който ги анализира. В идеалния случай двете би следвало да съвпадат. Този проблем се

корени в *различната семантична натовареност на някои термини*. Особено трудни за обработка са термини, които са семантично претоварени (overloaded) или термини, които са семантично недостатъчно натоварени (underloaded). Примери за претоварени понятия са "cardiac muscle tissue", а също и такива понятия като "bowel" и "gut". Недостатъчно натоварени понятия има доста, например: "portion of organism substance", "portion of tissue", "Xenopus anatomical entity", "mouse anatomical entity", "acellular anatomical structure", "anatomical set", "multi-tissue structure", "anatomical space", "surface structure".

Трето, съществуват известни *проблеми, свързани с дефинициите на някои термини в рамките на самите входни онтологии*. Добър пример за това е твърдението ***bulbus arteriosus part_of heart*** от онтологията на рибата Данио. В действителност дефиницията на понятието "bulbus arteriosus" при рибата Данио води до извода, че тази структура не е част от сърцето, а е по-скоро част от артериалната система и като такава, притежава гладка, а не сърдечна мускулатура. Все пак при обединяване на кои да е две онтологии, винаги когато подобни твърдения не са предизвиквали наличието на цикли в текущия резултантен граф (представляващ скелета на изходната супер-онтология), ние сме се стремили да ги запазим в оригиналния им вид, за да запазим входните онтологии максимално автентични и за да избегнем излишното коригиране на знанията, зададени чрез тях.

Накрая, трябва да спомена за проблема, отнасящ се до *различните релации, съществуващи между фиксирани (едни и същи) термини в зависимост от анатомичния контекст*. Анатомичната онтология на мишката дефинира твърдението ***maxilla part_of upper jaw***. От своя страна, програмата AnatOM открива, че е налице и релацията ***maxilla synonym upper jaw*** на базата на знанията, налични в WordNet. Тези две твърдения водят до някои цикли в текущия резултантен граф при обединяването на анатомичните онтологии на мишката и на рибата Данио. Проблемът тук се корени във факта, че първото твърдение е валидно за всички бозайници, а второто твърдение е валидно за всички други челюстни организми. Това разкрива и някои ограничения при използването от AnatOM на проектираната от нас т.нар. KNOWLEDGE таблица. При евентуална бъдеща работа вероятно добре би било KNOWLEDGE таблицата да се специализира, т.е. да се превърне в множество от KNOWLEDGE таблици, всяка от които се отнася до определен анатомичен контекст, т.е. до определена група от организми – бозайници, риби, земноводни, влечуги и други.

Глава VII. Перспективи за развитие на работата. Заключение

Съществуват множество различни насоки за бъдещо развитие на настоящия дисертационнен труд. Тук изброяваме най-важните от тях.

1. Би могло да се работи по подобряване на *чувствителността* и *специфичността* на процедурата CMP. Тук трудността идва от различните степени на детайлност и гранулярност на входните анатомични онтологии. Постигането на макар и малко подобрение в тази насока, би довело до сериозно подобрие по отношение на получените резултати от процедурата (що се отнася до тяхната *адекватност от анатомична гледна точка*).

2. Би могло да се работи по търсене на алтернативни, подобрени схеми (били те вероятностни като настоящите или не) за оценяване на автоматично

генерираните междуонтологични връзки. Естествено е да се приеме, че една схема за оценяване е по-добра от друга, ако автоматично генерираните от нея оценки са по-близки до реалните оценки, които би дал експерт по анатомия.

3. Би могло да се работи над разработване на алтернативна или надграждаща процедура на настоящата процедура CMP. Една такава процедура би могла да разглежда някои други, по-различни от настоящите (и евентуално по-сложни от тях), шаблони на свързаност в текущия резултантен граф, получен след изпълнението на DM и SMP. Креативното в една такава задача се състои в това да се предвиди точно какви шаблони на свързаност да бъдат разглеждани от алгоритъма, така че новополучената процедура да бъде по-добра (в семантично отношение) от настоящата процедура CMP.

4. Биха могли да се извършат по-сериозни тестове по обединяване на три и повече анатомични онтологии. Понастоящем са извършени сериозни тестове с три двойки организми, които бяха споменати по-горе в текста. Що се отнася до обединяването на три, четири и повече анатомични онтологии в изходна супер-онтология, са извършени само някои базисни тестове. Затова в тази посока има още доста поле за работа както по извършване на допълнителни тестове и последващо ръчно редактиране и оценяване, така и по подобряване на съществуващите вече алгоритмични процедури (най-вече CMP).

5. Би могло да се работи по разработването на процедура за автоматично елиминиране на наличните цикли в текущия резултантен граф.

6. Би могло да се работи над подобряване на алгоритъма за визуализиране на графи, използван в програмата AnatOM. Понастоящем при него в някои случаи възниква известен проблем, свързан с твърде "близкото" визуализиране на някои дъги, поради това че ъгълът между тях е твърде малък, което е нежелателно. Затова добре би било след завършване на работата на алгоритъма да се изпълни някаква "коригираща" процедура, която да увеличи ъглите между онези дъги, за които те са твърде малки.

7. Би могло да се работи също така над малка модификация на програмата AnatOM с цел поддръжката от нея и на други езици за описание на онтологии (като например езиците OWL и/или RDFS), а не само на езика OBO. Понастоящем обаче всички публично достъпни анатомични онтологии са зададени именно чрез езика OBO, който на практика представлява единствения практически общоприет езиков стандарт за описание на биологични и биомедицински онтологии.

8. Би могло да се работи и над реализацията на някои по-конкретни и по-практически ориентирани задачи, свързани с обединяване на повече анатомични онтологии в мащабна обща онтология, интегрираща анатомичните знания от едно доста по-широко множество от организми. Тази обща онтология може да служи като модел за интелигентно организмово-независимо текстово търсене или текст майнинг в научната литература, публикувана в електронен вид или в уеб пространството в по-общ план.

Авторска справка за приносите в дисертационния труд

Научни приноси

1) Направен е подробен обзор на предметната област (глава II). Описани са множество съществуващи системи с общо предназначение за намиране на съответствия между онтологии (B11, B63) и за обединяване на онтологии (B12, B48). Прякото сравняване между тези системи и системата AnatOM, разработена като част от настоящата работа, е практически невъзможно поради две причини: (i) повечето от известните системи не поддържат езика OBO, който е основният стандарт в областта на биомедицинските онтологии и на анатомичните онтологии в частност; (ii) липсват ясни критерии, позволяващи прякото сравняване на резултатите от програмата AnatOM с тези от съществуващите системи, както и начини за автоматизиране на такова сравняване, понеже става дума за семантично сравняване; единственият възможен подход е неавтоматичен и е свързан с намесата на специалист по анатомия.

2) В обща теоретична рамка са формализирани трите алгоритмични процедури – DM, SMP, CMP (глава IV), всяка от които допълва предишната в процеса на откриване на междуонтологични семантични връзки между две дадени анатомични онтологии. Тук ще поясним произхода на трите процедури. ***Процедурата DM*** е широко известна. Тя се използва почти винаги, когато става дума за интегриране на онтологии (B33) и представлява базата за намиране на междуонтологични съответствия. ***Някои идеи за процедурата SMP*** сме черпили от научната литература (напр. от [70] и [85]), но в настоящата работа тези идеи са допълнени, формализирани и пренесени точно към анатомичните онтологии и към използваните в нашата работа ВИЗ (A02). ***Процедурата CMP е изцяло оригинална***, разработена в хода на настоящата работа. Тази процедура позволява откриването на смислени от анатомична гледна точка междуонтологични връзки, които се пропускат както от DM, така и от SMP процедурата.

3) Доказано е едно НДУ (в глава IV), чрез използването на което е възможно да се достигне от модел#2 до модел#3 (двата модела, въведени в глава III). Това НДУ указва кога е възможно генерирането на *валидна, изходна, обединяваща, организмово-неутрална анатомична онтология* (наричана още супер-онтология) от две *дадени, входни, видово-специфични анатомични онтологии*. Доказателството на това НДУ задава явна процедура как това генериране може да бъде извършено.

4) Въпреки че работим само с анатомични онтологии, предложените от нас метод и алгоритъм са *достатъчно общи* и биха могли по естествен начин да бъдат пренесени и приложени и към друга предметна област, различна от тази на анатомията. Единственото условие за това е да са налице достатъчно близки по тематика до тази нова предметна област ВИЗ (A02). Тук следва да отбележим, че източникът на знания WordNet на практика може да се прилага почти към всяка предметна област, понеже той е ИЗ (A01) с общо предназначение, явяваващ се лексична база от данни на английския език като цяло.

Приложни приноси

Разработена е програмата AnatOM – цялостно, завършено решение за полуавтоматично интегриране (B33) на анатомични онтологии.

В хода на разработването на AnatOM са реализирани множество модули, които са ценни от практическа гледна точка, дори и разгледани сами по себе си (извън контекста на програмата AnatOM), като например: модул с общо предназначение за работа с графи и мултиграфи; модул за визуализиране на графи; модул за прочитане/парсване на OBO файлове, описващи онтологии; модул за експортиране на онтологии в OBO формат от съответните им графи.

Програмно са реализирани формализираните предварително в теоретичен вид процедури DM и SMP. Реализирана е и оригиналната процедура CMP, предложена от нас в текста на настоящата дисертация. Тези три реализации формират логическия модул на програмата AnatOM.

Реализирани са и няколко класически алгоритъма за работа с графи като алгоритъма на Tarjan [33] и алгоритъма на Johnson [32], както и не толкова известния алгоритъм за визуализиране на графи, базиращ се на уравновесяване на физични сили [24].

Публикации

Основните резултати от дисертацията са отразени под формата на статии в следните рецензирани издания.

[1A] Peter Petrov, Milko Krachunov, Elena Todorovska, Dimitar Vassilev, "An intelligent system approach for integrating anatomical ontologies", Biotechnology and Biotechnological Equipment 26(4):3173-3181, 2012.

Части от текста на тази статия могат да бъдат намерени в глави II, III, IV на дисертацията.

[2A] Peter Petrov, Milko Krachunov, Ognyan Kulev, Maria Nisheva, Dimitar Vassilev, "Predicting and Scoring Links in Anatomical Ontology Mapping", To appear in: Proceedings of BIOMATH 2012, vol. 3, 2012.

Части от тази статия са включени в глава IV (и по-точно в параграф 4.4.) на текста на дисертацията.

[3A] Peter Petrov, Milko Krachunov, Ernest A.A van Ophuizen, Dimitar Vassilev, "An algorithmic approach to inferring cross-ontology links while mapping anatomical ontologies", To appear in: Serdica Journal of Computing, ISSN 1312-6555, vol. 6, 2012.

Текстът на тази статия формира голяма част (параграфи 4.1., 4.2., 4.3.) от глава IV на дисертацията.

[4A] Peter Petrov, Nikolay Natchev, Dimitar Vassilev, Milko Krachounov, Maria Nisheva, Ognyan Kulev, "AnatOM – An intelligent software program for semi-automatic mapping and merging of anatomy ontologies", To apperar in: Proceedings of the 6th International Conference on Information Systems & Grid Technologies (ISGT, Sofia, 1-3 June 2012), Sofia, St. Kliment Ohridski University Press, 2012.

Текстът на тази статия е свързан предимно с глава V на дисертацията.

Цитирания

Доколкото ни е известно, към този момент изброените публикации по същността на настоящата дисертация нямат цитирания от други автори, което вероятно се дължи в някаква степен на обстоятелството, че тези публикации станаха факт едва в рамките на последните 12-15 месеца преди защитата на дисертационния труд.

Благодарности

За достигането ми от февруари 2007 г., когато бях зачислен като редовен докторант във ФМИ, до ноември 2012 г. и до този финален етап, предшестващ защитата на настоящата дисертация, дължа благодарности на доста хора.

Най-големи благодарности дължа на моя неформален научен ръководител доц. д-р Димитър Василев, на моя много добър приятел и бивш колега Димитър Франгов, на покойния вече холандски учен проф. Джак Люнисен и на моето семейство – на сина ми Методи Петров и на майка му Екатерина Николова (най-близките ми хора), на нейните родители и на моите родители.

Благодаря още на моя формален научен ръководител – покойния вече доц. д-р Антоний Попов, на доц. д-р Мария Нишева, на колегите докторанти Милко Крачунов и Ернест ван Опхюйцен, на д-р Николай Начев, на доц. д-р Елена Тодоровска, на доц. д-р Маргарита Камбурова и на доц. д-р Нели Димитрова.