ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – СОФИЯ

Факултет Компютърни Системи и Технологии
Катедра Програмиране и Компютърни Технологии

Маг. инж. Мария Савова Евтимова

СЕМАНТИЧНИ АГЕНТИ ЗА ПЕРСОНАЛИЗИРАНО ТЪРСЕНЕ

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: Компютърни системи, комплекси и мрежи
Научен ръководител: Проф. д-р Иван Момчев

СОФИЯ, 2017 г.

Дисертационният труд е обсъден и насочен за защита от Катедрения съвет на катедра „ПКТ“ към Факултет ФКСТ на ТУ-София на редовно заседание, проведено на Протокол №11/26.06.2017г.

Публичната защита на дисертационния труд ще се състои на 19.10.2017г. от 15,00 часа в Конферентната зала на БИЦ на Технически университет – София на открито заседание на научното жури, определено със заповед №ОЖ-261/29.06.2017 г. на Ректора на ТУ-София в състав:

1. Проф. д-р Огнян Наков Наков- председател
2. Проф. д-р Иван Момчилов Момчев- научен секретар
3. Проф. д-р Нина Василевна Синягина
4. Проф. д-р Владимир Димитров Лазаров
5. Проф. д-р Иван Крумов Куртев

Рецензенти:

1. Проф. д-р Нина Василевна Синягина
2. Проф. д-р Владимир Димитров Лазаров

Материалите по защитата са на разположение на интересуващите се в канцеларията на Факултет ФКСТ на ТУ-София, блок №1, кабинет № 1443-А.

Дисертантът е задочен докторант към катедра ПКТ на факултет ФКСТ. Изследванията по дисертационната разработка са направени от автора.

Автор: маг. инж. Мария Евтимова
Заглавие: Семантични агенти за персонализирано търсене
Тираж: 30 броя
Отпечатано в ИПК на Технически университет – София

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

Област на изследване на дисертацията е проблемът с незадоволителното качество на върнатата информация при специфично търсене в дадена област на потребителя. Проблемът за повишаване на качеството на върнатите резултати от системите за търсене е актуален и се разисква непрекъснато, защото обема на данните с информация нараства всекидневно, което затруднява процеса на обработка и връщането на коректни резултати от заявката. Заявките за търсене обикновено са кратки и двусмислени, но персонализацията може да помогне за отстраняване на двусмислените запитвания в интерес на потребителя.

Семантичните технологии намират приложение за реализирането на ново поколение системи за търсене на информация. Технологиите, които използват онтологии са популярни и привличат внимание, защото дават възможност за търсене на информация по смисъл.

Тенденция през последните години е наблюдаването на нарастващ интерес към изучаването на многоагентните системи. Съвременните научни изследвания свързани с многоагентните системи са насочени главно към координиране на поведение на агенти и разпределение и обединение на решения.

Непрекъснато нарастващия брой на уеб сайтове налага необходимостта от персонализиране на върнатите резултати, като създава проблема с обработката на големи данни. Един от основните проблеми за работа с големи данни от семантична гледна точка е извличането чрез търсене и интерпретация на големи данни.

Персонализираното извеждане на резултатите води до по- коректно върната информация на потребителя, т.е. повишава се качеството на върнатата информация на потребителя.

Цел на дисертационния труд, основни задачи и методи за изследване

Целта на дисертацията е да се разработят набор от инструментални програмни средства и модел на система, която да подобрява търсенето дори и в големи масиви данни от неясна и несигурно зададена информация, като използва технологиите за персонализация на заявките, семантични онтологии, интелигентни агенти, и размита и вероятностна логика.

Задачи на настоящия дисертационен труд

1. Да се изследват проблемите на машините за извеждане на логически изводи, при работа с големи масиви от данни

2.Да се предложи нов модел на машина за извеждане на логически изводи, както и онтология, която е подходяща за работа с големи данни, при зададена неясна и непълна информация от потребителя

3.Да се интегрира мобилен агент за извличане на информация в непълна и противоречива информационна среда от интернет източници

4.Да се създаде модел на потребителски профил, за да се повиши качеството на върнатите резултати при търсене на информация от неясно и двусмислено зададена информация и голям обем от данни

5.Да се създаде нов концептуален модел на персонализирана семантична търсеща система, в който да се имплементират предходните задачи.

6.Да се създадат програмни частични реализации на отделни модули, чрез разработване на набор от инструментални програмни средства. Експериментална оценка на възможностите на реализираните средства.

Научна новост

1. Създаден е нов вид хибридна онтология, която е базирана на прецеденти и правила, като е обогатена с елементи на размита и вероятностна логика. Тази онтология предоставя възможност за работа със заявки с размити и неясни данни, които са въведени от потребителя. За практическата реализация на онтологията е избран езика OWL2.
2. Предложен е модел на потребителски профил, който обхваща и процеса на търсене на информация. Ролята на потребителския профил е да персонализира зададената заявка.
3. Предложен е нов хибриден модел на машина за логически изводи, който е подходящ за големи данни и обхваща неяснотата на зададената информация от потребителя. В предложения модел са отчетени техните предимства и недостатъци на познатите машини за извеждане на логически изводи за големи и размити данни.
4. Предложен е концептуален модел на персонализирана търсеща система, който интегрира предходните приноси, като при създаването му е използван и мобилен агент за търсене на информация в интернет. Интегрирането на мобилен агент в предложената мулти- агентна концептуална схема осигурява преимущества при търсенето на информация в интернет. Създадена е структура, както и прототип на динамична персонализирана търсеща система, която връща само един резултат по направената заявка на потребителя (one shot). Тази система подобрява качеството на върнатите резултати при зададена неясна и несигурна информация от потребителя, което е много важно особено в област, като медицината.

Практическа приложимост

Предложения модел за персонализирана търсеща система може да се използва за търсене на информация в областта на туризма(за търсене на хотели), за търсене на филми, за недвижими имоти и дори в медицината. Точността на върнатите резултати при търсене на информация след литературен преглед се определя, като особено важна в областта на медицината. Предложената семантична търсеща система е подходяща за всички потребители, които искат да получат по-качествена информация при въвеждане на симптоми посредством заявка за търсене. Т.е. семантичната търсеща система е насочена към търсене на здравна информация от хора, които искат получат информация за диагноза при съответните оплаквания.

Апробация

Резултатите от дисертационния труд са докладвани изцяло в катедра „ПКТ“ на ФКСТ към ТУ-София. Основния метод е описан в статия на научно списание към ФКСТ „Computer and communications engineering“ със заглавие : „Описание на модел на хибридна онтология за медицинска информация“. Основни анализирани методи са докладвани на International Scientific Conference Computer Science’2015, Albania, която е организирана от ФКСТ и са отпечатани в списание Journal of Communication and Computer 13 (2016). Останалите научни експерименти са публикувани в списание International Journal of Engineering Research and Allied Sciences.

За да се покаже работоспособността на постигнатите резултати се събират 225 заявки от различни източници, като експерти в областта, проучване на потребителите и различни уеб сайтове за здравна информация. Заявките се категоризират в зависимост от симптомите на различните болести, които са взети от сайта за медицинска информация http://bioportal.bioontology.org/ Извършва се диагностика на заявките от системата и от експерт в определената област и се извеждат резултати.

Публикации

Основни постижения и резултати от дисертационния труд са публикувани в 4 научни статии, от които 1 самостоятелна, като 2 са в списания и 1 в сборник от конференция.

Структура и обем на дисертационния труд

Дисертационният труд е в обем от *163* 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 *149* литературни източници, като *144* са на латиница и *2* на кирилица, а останалите са интернет адреси. Работата включва общо *33* фигури, *12* таблици и *12* приложения. Номерата на фигурите, таблиците и на отделните под точки в автореферата съответстват на тези в дисертационния труд.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ГЛАВА 1. Въведение в проблема на настоящия дисертационен труд

1.2.Подход за решаване на поставения проблем

За решаването на дадения проблем предлагам да се интегрират достиженията, които включват, както интелигентни агенти, така и технологиите за персонализирано семантично търсене на информация. В работата се разглеждат и изброяват интелигентните агенти и многоагентни системи от гледна точка на персонализираните семантични търсещи системи. Извършва се анализ и оценка на текущото състояние на семантичните агенти за персонализирано търсене след, което се проучват семантичните агенти и приложените в тях технологии за подобряване на търсенето, което е предпоставка за създаването на нов модел за персонализирана търсеща система.

1.3.Цел и задачи на настоящата дисертация

Целта на дисертацията е в резултат от проведеното изследване на семантичните агенти за персонализирано търсене на информация, да се предложи търсеща система, която да подобрява търсенето на информация дори и в големи данни от неясна и несигурно зададена информация. Това се осъществява, като се създава модел на персонализирана семантична търсеща система, която използва технологиите за интелигентни агенти, както и размита и вероятностна логика.

В труда се поставят следните задачи.

Задачи на настоящия дисертационен труд

1.Да се изследват проблемите на машините за извеждане на логически изводи, като се предвиди възможността за работа с големи данни.

2.Да се предложи нов модел на машина за извеждане на логически изводи, както и онтология, която е подходяща за работа с големи данни, при зададена неясна и непълна информация от потребителя

3.Да се интегрира мобилен агент за извличане на информация в непълна и противоречива информационна среда от интернет източници

4.Да се създаде модел на потребителски профил

5.Да се създаде нов концептуален модел на персонализирана семантична търсеща система, в който да се имплементират предходните задачи.

6.Да се създадат програмни частични реализации на отделни модули, чрез разработване на набор от инструментални програмни средства. Експериментална оценка на възможностите на реализираните средства.

ГЛАВА 2. АНАЛИЗ И ОЦЕНКА НА ТЕКУЩОТО СЪСТОЯНИЕ

2.1.Цели и задачи на настоящата глава

В настоящата глава се извършва анализ и оценка на резултатите от досегашни научни изследвания за семантични агенти за персонализирано търсене. Изследват се възможностите на концептуалните модели на предложените методи, което обосновава необходимостта от създаването на иновативен модел за персонализирано търсене на информация, който решава до известна степен някои проблеми, които възникват при съвременните системи за търсене на информация.

2.2.Оценка на подходи и инструменти за персонализирано търсене на информация

2.2.1.Подходи за персонализирано търсене на информация

Обема от информацията в интернет нараства постоянно. Милионите страници в интернет са неорганизирани и намирането на специфична информация е трудно и изисква отделяне на значително време. За да се подобри качеството на интернет обслужването и да се увеличат посещенията на даден уеб сайт, е важно, за уеб разработчика или уеб дизайнера, да познава поведението на потребителя, да се предвидят страниците, които представляват интерес за потребителя.

2.2.1.1.Уеб персонализация

Технологията за персонализация позволява динамично вмъкване, къстамизиране или предложение на съдържание във всеки формат, който е уместен за съответния индивидуален потребител, въз основа на подразбиращото се поведение на потребителя и предпочитания, и точно определени детайли.

2.2.1.2.1.Стратегии за персонализиране

Според литературата процеса на персонализиране може да се раздели на четири основни фази.

1.Събиране на уеб данни**.**
2.Предварителна обработка на Уеб данните.
3.Анализ на уеб данни.
4.Последен етап е вземане на решение/препоръка.

2.2.2. Оценка на инструментите за персонализирано търсене

В днешно време, онтологиите и агентните технологии се използват широко в семантичните инструменти за търсене на персонализирана информация. Много нови доклади използват размита логика при персонализирано търсене на информация, като включват използването на размити онтологии. Инструментите за уеб търсене представят различни технологии съгласно различни направления при персонализирането. Предложен е преглед на различните методи за персонализиране и примери на наличните инструменти за търсене на информация.

2.4.2.Семантично уеб търсене с онтология

Проблемите, които се наблюдават при стандартната уеб търсачка налагат необходимостта от търсене на нови подходи и техники, които да удовлетворяват потребителя. Това води до създаването на семантична уеб търсачка, която е базирана на RDF. В семантичния уеб, информацията се съхранява в концептуална йерархия, посочена, като разработена онтология с уеб онтологичен език (OWL). Следователно в семантичната мрежа сходството може да се оцени чрез използване на семантиката на понятия в онтология.

Онтологично базирано семантично търсене се определя, като процес на извличане на информация, който използва знанията от домейн онтологията. Домейн онтологията е йерархично структурирана съвкупност от термини, която описва домейн, който може да се използва, като основа за бази знания. Думата, като субект представлява случаи или факти, а свойствата представляват отношения.

Преимущества при създаването на онтология

-дава възможност на машината да използва знания при някои приложения
-дава се възможност множество машини да споделят своите знания
-подпомага разбирането на дадена област на знание

-помага да се постигне консенсус при разбирането на някоя област на знание

След направения литературен преглед на докладите за приложения на онтологии и семантичните технологии при търсене на информация се обособява направление, което представлява значителна част, от докладите за разработване и усъвършенстване на машини за извеждане на логически изводи и съответните необходими онтологии за тях.

Целта на семантичното търсене с онтология е да максимизира прецизност и връщане (recall), F-измерване и средно аритметична прецизност. Стойността на прецизност и връщане е между 0 и 1, като максималната стойност е 1. При използване на общите познания за онтологии за термините от заявката за търсене, прецизността и стойността на връщане ще се увеличи.

2.5.Подходи за създаване на семантични агенти

2.5.1.Анализ на агентните технологии

Семантичните търсещи системи, които са на пазара имат редица недостатъци и не винаги задоволяват потребностите на интернет потребителите. Решение за ефективна значеща търсеща система е семантична търсеща система, която обхваща областите изкуствен интелект и семантичен уеб.

2.5.2. Класификация на информационните агенти

Преимущества на мобилните агенти

1.Намаляване на мрежовото натоварване
2.Преодоляват мрежовата латентност
3.Капсулиране на протоколи
4.Асинхронно и автономно изпълнение
5.Динамично адаптиране
6.Естествена хетерогенност и якост
7.Устойчивост на отказ

2.6.2.Размита и вероятностна логика

Размитата логика е добре познато разширение на дескриптивната логика, което се използва широко в много области. В литературните източници е разгледана взаимовръзката между теорията на вероятностите и размитата логика. В най- простия случай се предполага, че съществува размито множество A във вселената на дискурса U, тогава $a(u)$ е степен на членната функция на u в A. В света на вероятностите, нека X е случайна променлива, която има две стойности A и $A\_$, където $A\_$ означава *not A*. Тогава, $\mu A(u)$ е условна вероятност $Pr(A \mid u)$.

Като по-формална дефиниция, ако $V = \{V1, V2, ..., Vn\}$ *е колекция от избиратели и всеки* Vi представлява вот за класифициране на u като A или $A\_$, тогава $Pr(A \mid u)$ може да се разгледа, като вероятността, че случаен избирател ще класифицира u като A. Чрез тази равностойност на членната функция на размитата логика и условната вероятност, като общ термин се приема фактор на несъвършенство, който покрива несигурността и неяснотата. В литературата съществуват някои подходи, които съчетават вероятностите с размитата логика.

Докато традиционните техники на OWL моделиране не могат да се справят и да изразят моделиране, теорията на размитата логика и теорията на вероятностната логика могат да осигурят разширена логика (извличане на логически изводи) и поддръжка на размити спецификации. Те намират приложение при извличането на информация, където системите за търсене на информация намират трудности при взимането на решение за осигуряването на точна информация.

През последните десет години много изследвания са посветени на добавянето на размитите множества и логика към семантичните уеб технологии. Изследователите ще продължат да градят текущата работа, за да включат размита логика в семантичния уеб и да осигурят иновации необходими да направят създаването и използването на размити онтологии на практика.

2.7.Оценка и перспективи на семантичните агенти

Целта на семантичния уеб не е да замени класическия уеб, но да предостави една структура, която да даде възможност обемистата информация да се обработва автоматично от една машина. В ядрото на Семантичния уеб стоят онтологиите, които са познати още като съществена технология за достигане до семантика, позволяваща представяне на данните в машинно-разбираеми структури. Целта на семантичния уеб е да допринесе за по-добра платформа за представяне на знания на свързаните с тях данни до голяма машинна обработка на цялостен мащаб чрез добавяне на логика, извод и системи с правила в различни приложения и граници. Като обобщение, за целите при посредничеството на семантичната информация, информационния агент трябва да бъде в състояние да се справя с множество, частично припокриващи се, стандартизирани, домейн специфични онтологии.

Основните направления в изследванията базирани на онтологии са търсене базирано на правила, класифициране, групиране, информационно извличане и системи за препоръка. Основни резултати от изследванията са в подобряване на ефективността и качеството на върната информация при търсене в определени аспекти.

2.8.2.Оценка и перспективи за използване на размита онтология

1.Описание на елементите на размита онтология

Изследването на размитите онтологии започва в началото на 2000г. с акцент върху опростените модели, които се използват с цел извличане на информация.

Много изследвания са предложени за разширяване на онтологията до размита онтология. Аспекти от онтологията са представени, като размити, за да могат да се съобразят с нуждите на домейна за представяне на неточна и неясна информация. Това се осъществява със специфичната цел да се подобри търсенето на информация. В по- новите изследвания върху изграждането на размитите онтологии, определението за размита онтология разделя ясната от размитата компонента, които са, обикновени концепции, свойства и връзки от размити концепции, свойства и връзки.

В размитата онтология могат да бъдат намерени субекти, понятия, връзки и аксиоми.

Обикновените онтологии се основават на описателна логика (DL), която не е подходяща за справяне с неточна информация. Възможното решение на този проблем е да се използват размити онтологии. При размитата онтология, субектите могат да установят връзки по размит начин използвайки членна функция. Обикновено се избира интервала [0,1]. В съвременните изследвания, размитата онтология е широко използвана от изследователите.

2.9.2.Анализ на размита дескриптивна логика

Размитата дескриптивна логика може да се използва за представяне и извличане на логически изводи на неясни знания. Това семейство от логически формализми е много разнообразно, всеки член се характеризира от специфичен избор от конструктори, аксиоми, и триъгълни норми, които се използват, за да специфицират семантиката.

DL в класическата им форма не са подходящи за представяне и извличане на логически изводи от неясни и непрецизни данни, които са особено важни в области на знанието, като био-медицината.

В сравнение с класическата дескриптивна логика, размитата дескриптивна логика има допълнителна степен на свобода за избиране, как да интерпретира логическите

конструктори. Стандартния подход, наследен от математическата размита логика, е да се използва непрекъсната триъгълна норма (t-норма), за да се интерпретира връзката. Трите често използвани t-норми са: *Годел, Лукациевич* и *Продукт*, които имат свойството, че всички други непрекъснати норми могат да бъдат представени, като композират копие от тях по определен начин. От избора на t-норма⊗, се определя семантиката на всички други логически конструктори, обобщавайки свойствата на класическите оператори.

В размитите случаи, като допълнение на свързването на индивида със стойността на данните, е възможно да се свърже с размита членна функция.

2.10.Оценка на методите за извличане на логически изводи

Извеждане на логически изводи базирани на правила и на прецеденти са два популярни подхода, които се използват в интелигентните системи. Правилата обикновено представят общи познания, докато прецедентите обхващат знания натрупани при специфични ситуации. Всеки подход има своите предимства и недостатъци, което е доказано, че се допълват в голяма степен. Това доказва съчетаването на логическите изводи базирани на правила и прецеденти в хибриден подход, който надминава недостатъците на всеки от двата подхода по отделно.

2.11.Значение на “големи данни” (Big Data) при персонализирано търсене

Данните се увеличават бързо през последните десетилетия, и големия обем от данни стават все по- популярна тема.

“Големи данни” се характеризират с обем, разнообразие и бързина. Свободно пространство, несигурност и непълни данни са определящи елементи за приложение за големи данни.

Несигурните данни са специален тип реални данни, където всяко поле от данни вече не е детерминирано, но е обект на разпределение на някаква случайна грешка. Това е свързано предимно с домейн специфични приложения с неточни показания на данни и колекции.

При литературен преглед през последните 2-3 години се изразяват ясно тенденциите при работа с големи данни, като използване на машини за логически изводи базирани на прецеденти, проблеми с несигурността и размитостта на данните.

2.13.Заключение

В резултат на направения анализ на съвременното състояние на семантични агенти за персонализирано търсене, могат да се направят следните изводи:

1.Традиционните търсачки не са достатъчно ефективни, за да осигурят точна информация за потребителя, защото не търсят смисъла в потребителската заявка, каквито са семантичните подходи.

2.Направена е оценка на семантичните агенти при персонализирано търсене. Изтъкнати са преимуществата при използване на мобилните агенти в сравнение с другите видове агенти и тяхната приложимост на практика в системи за търсене.

3.В резултат от проведеното проучване при семантичните търсачки се идентифицира, като основен проблем, който влияе на върнатите резултати, алгоритъма на машината за извеждане на логически изводи.

4. Онтологиите присъстват в редици доклади за персонализирано търсене на информация, като ново направление се разглежда размитата онтология.

5.От направения преглед става ясна особената важност на многоагентните технологии и семантичните технологии за успешното реализиране на системата за персонализирано търсене на информация. Като резултат от това се предлага концептуален модел, който съчетава двете технологии.

6.В резултат от изследването се идентифицират няколко основни проблема, които влияят на качеството при търсенето на информация- големи данни, неясна и неточно зададена информация от потребителя и персонализиране на върнатите резултати.

6.1.Големи данни

Свободно пространство, несигурност и непълни данни са определящи елементи за приложение на големи данни. Следователно използването на паралелна изчислителна инфраструктура и съответно поддръжката на език за програмиране и софтуерни модели за ефективно анализиране и търсене на информация в разпределени данни са съществени цели за процесите при големи данни, за да сменят от “количество” към “качество”.

6.2.Неясно и неточно зададена информация от потребителя

6.3.Създаване на потребителски профил за персонализиране на върнатите резултати

Върху този проблем са поставени много доклади, като основни технологии, които се използват са размита логика, вероятностна логика, извеждане на логически изводи базирани на правила и извеждане на логически изводи базирани на прецеденти.

Решенията трябва да се търсят в следните направления:

- усъвършенстване на интелигентните технологии за осъществяване на реално семантично търсене и обработка;

- ефективно използване на машините, търсещи семантичните документи с цел работа с големи данни и с неясна и неточно зададена информация от потребителя

ГЛАВА 3.Моделиране на семантична търсеща система

В настоящата глава се предлага архитектура и концептуален модел на персонализираната търсеща система, при зададена неясна и размита информация, като се базира на последните достижения в областта на семантичното търсене. Тя може да използва интернет страници за събиране на информация за потребителя, чрез мобилни агенти. На основа на семантичен анализ на подадената заявка за търсене тя търси в база с прецеденти необходимата информация, която ще бъде върната от системата на потребителя. Системата връща само един резултат при дадена заявка за търсене. При самото подаване на заявката се отчита неяснотата и несигурността на посочената информация в заявката за търсене, което прави търсачката по-ефективна при намиране на необходимите резултати. Заявката за търсене се съгласува с потребителския профил преди да се търси информация в базата с прецеденти. Базата с прецеденти се представя от онтология, за която е разработена машина за логически изводи, с цел да подобри резултата при дадена заявка, макар заявката да не е достатъчно точна. Машината за логически изводи е създадена да може да работи и с голям обем от данни, което е важно изискване при непрекъснато нарастващия обем от информация в интернет. В системата са приложени нови подходи за търсене на информация, като са взети в предвид обема на информацията в интернет и неяснотата на въведената от потребителя заявка.

3.1.Изисквания към персонализираната търсеща система

Персонализираната търсеща система е предназначена да осигурява качествено търсене на медицинска информация, отговарящо на индивидуалните особености и потребности на хора от всички възрасти, които биха желали да получат информация по даден здравословен проблем в зависимост от техните оплаквания. Тъй като английският е най- широко използваният в момента език и на него могат да се намерят най-много и най-качествени текстови (и семантични) източници във всяка една област, за момента системата ще бъде разработвана и тествана за търсене на английски език. Използваните в нея подходи, методи и алгоритми по принцип биха могли да се реализират за кой да е естествен език, но за това ще са необходими сериозни усилия поради значителните различия между човешките езици, а освен това ефективността на системата зависи от наличните в интернет текстови и семантични източници, което означава, че за момента тя би била най-ефективна на английски език.

Основни изисквания на системата:

- Възможности за извеждане на информация на потребителя по зададени търсещи думи;
- Адаптивност- да се адаптира лесно към потребностите на конкретния потребител;
- Разширяемост- лесно могат да се добавят нови компоненти към системата;
- Да работи ефективно с неясна и неточно зададена информация от потребителя;
- Съвместимост с новите семантични технологии и изследвания ;
- Възможност за работа с големи данни ;
- Повишаване на точност (Precision) на търсенето в интернет според конкретния случай;
- Повишаване на върнатите резултати (Recall) от търсенето според конкретния случай;
- Лесен за използване потребителски интерфейс;
- Платформена независимост;

3.2.Езици и технологии, подходящи за реализация на системата

3.2.1.Избор на език за размита дескриптивна логика

Въпреки, че има няколко езика за определяне на онтология за семантичния уеб, се избира уеб онтологичния език OWL 2, който е обновената версия на OWL 1 и позволява повече експресивност, като се превръща в уеб онтологичен език от W3C. Обосновката за използване свойствата за анотация е, че е възможно да се използва текущия OWL2 редактор за създаване на размити представяния на онтология.

3.2.2. Мотиви за избор на агентна платформа за реализация на системата.

Избора на многоагентна софтуерна платформа е от съществено значение при създаването на многоагентен софтуер.

Jade е много популярна FIPA услужлива агентна платформа. Сред ползите може да се спомене, че има широк набор от инструменти и може да се интегрира с различен софтуер, като Jess (генератор на правила). Мобилността не е ключов елемент от Jade. Поддържа мобилност между контейнерите в една и съща Jade платформа (подобно на идеята за контейнерите при Glasshopper).

Съгласно доклад се определя, като подходяща при избор на мобилна платформа Glasshoper, JADE, AGLETS, като се следва реда на подреждане. Но за Glasshoper трябва да се добави FIPA, за да може да се използва, а AGLETS не поддържа FIPA стандарт.

За нуждите на нашият проект ние избираме платформата JADE, която е с отворен код и е добре документирана.

3.2.3.Мотиви за избор на инструменти за размита онтология

Съществуват различни софтуерни инструменти за създаване и използване на онтологии. Две много важни категории за онтологични инструменти са онтологични редактори и машина за извеждане на логически изводи. Следните представят примери за такива софтуерни инструменти, които са разработени за създаване и управление на размити онтологии.

Един от широко използваните онтологични редактори е Protégé разработен в университета Станфорд. Той има много предимства, като например пълна подкрепа на

новият OWL2 уеб онтологичен език и RDF спецификации от консорциума на световната мрежа.

Protégé плъгина е разработен, за да подпомогне разработването на размита онтология, използвайки представянето на размита онтология. Използвайки този подход, не размитата част на онтологията може да бъде създадена, като се използват свойствата на анотацията. С плъгина Fuzzy OWL2 се изяснява на потребителя, синтаксиса за определяне на всички размити елементи, използвайки свойствата на анотацията.

3.3. Оценка и избор на машина за логически изводи подходяща за големи данни

Сравнителен анализ на изброените машини за логически изводи:

1.CB (машина за логически изводи базирана на последователност)

2.CEL (класификатор за EL)

3.FACT++ (бърза класификация на терминология)

4.HermiT(базиран на хипер табло)

5.Pellet (първата машина, която поддържа OWL DL (SHOIN(D)) поддържа и има разширение за OWL2 (SROIQ(D)))

6.RacerPro (сменя името на ABox и е машина за логически изводи за изразяване на понятия)

От характеристиките представени в дисертацията на представените машини за извеждане на логически изводи и от изследванията представени в дисертацията се избира Pellet, за машина за логически изводи за обикновената онтология, като подходяща за проекта. Pellet поддържа OWL2 профили включително OWL2 EL.

3.4.Основни цели и задачи на разработваната семантична търсеща система

Целта на семантичната търсеща система е да използва нови методи и средства базирани на автономни разпределени агенти за търсене и анализ на информация в Интернет пространството, като се създава нов модел на търсеща система.

Задачи на разработваната семантична търсеща система:

1.Да се създаде семантична търсеща система подходяща за голям обем от данни при подадена от потребителя неясна и несигурна информация

2.Създаване на машина за извеждане на логически изводи.

3.Анализ и оценка на инструментите за създаване на семантичната търсеща система

4.Демонстриране на възможностите за извличане на информация в непълна и противоречива информационна среда чрез съчетаване на софтуерни многоагентни системи.

5.Създаване мобилни агенти при търсенето на информация от интернет.

6.Създаване на концептуален модел на семантичната търсеща система

7.Създаване на програмни реализации на отделни модули.

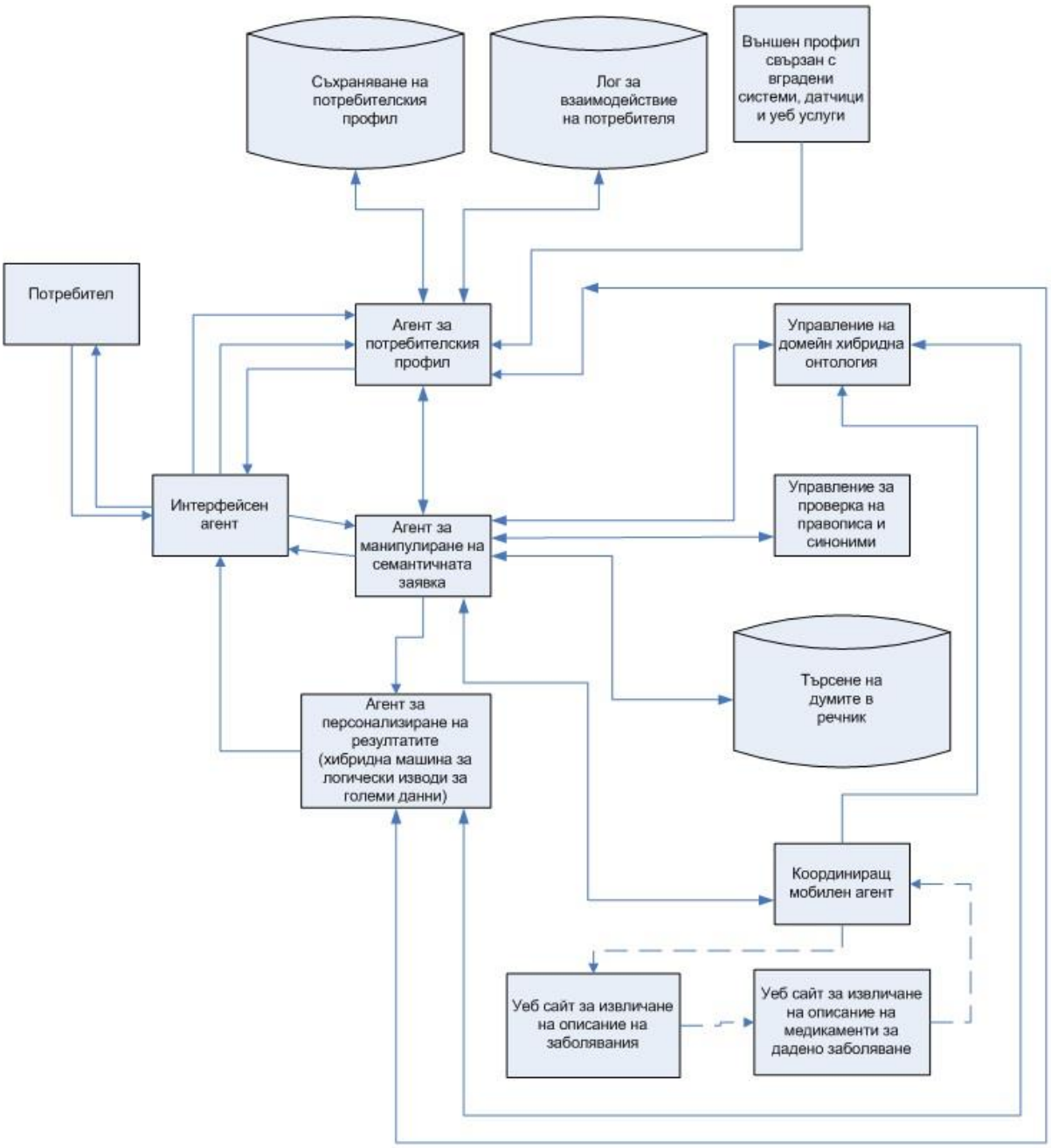
8.Разработване на набор от инструментални програмни средства.

3.4.1.Научно изследователски цели и приложни цели

Целта на трета глава е създаването на персонализирана семантична търсеща система за несигурни и неясни данни, която да има възможност да се приложи и за голям обем от данни.

Предназначението на персонализираната търсеща система е да може да търси от несигурна и неточна подадена заявка, информация в данни от интернет. Области на приложение на семантичната търсеща система са медицина, туризъм и др.

3.5.Концептуална схема на предложената персонализирана търсеща система.



Фиг.12- Концептуална схема за персонализирано търсене на информация при неясно и непълно зададена информация

Предложената концептуална схема има за цел да помогне на потребителите да намерят семантично съответната информация, която отговаря на техните нужди. Архитектурата на концептуалната схема е съставена от три основни компоненти. Първата компонента е за семантично манипулиране на заявката и компонент за персонализиране. Идеята на този концептуален модел е да представи потребителските предпочитания, като анализира потребителската заявка семантично и персонализира намерените резултати. Втората компонента е компонента за управление на размити онтологии, която отговаря за представяне и управление на домейн размити онтологии. Третата компонента е за събиране на данни и компонента за семантична анотация, която отговаря за определянето на доверени източници и анотира информацията, базирана на предварително определена домейн онтология.

Агентите са софтуерни обекти, които имат специфични цели, функционират автономно в определена среда и комуникират с други агенти. Агентно базираното

проектиране се избира поради редицата причини. Агентът може да си комуникира със заобикалящата среда, която подпомага научаването на предпочитанията на потребителя динамично. Като допълнение концептуалната схема изисква много взаимодействия между потребителя и системата, за да предостави на потребителя правилните препоръки и съвети. Това води до много взаимодействия между различните модули. Агентът има ефективна комуникация и механизъм на абстракция, затова се препоръчва използването на агенти.

Предложената концептуална схема има три функционалности. За да се разбере семантично контекстът на въпросите, задавани от потребителя, се използва предефинирана домейн хибридна онтология. Онтологията може да се дефинира, като формално представяне на знания "набор от концепции в рамките на домейн, използвайки общ речник, който да обозначи видовете, свойствата и взаимовръзките на тези понятия". Предварително определената онтология се използва, за да поясни доверените източници за медицинска информация и след това да създаде семантичен потребителски профил. Второ, онтологията помага при анализирането на потребителските предпочитания. В предложения концептуален модел потребителя първо създава профил, като попълва обща информация, медицинска информация и здравно състояние. След това потребителят въвежда въпроси със свободен текст, като използва портален интерфейс. На следващо място потребителската заявка се съчетава с помощта на семантичната логическа машина за изводи и след това се обогатява с помощта на профила на потребителя, така че резултатите се персонализират за нуждите на потребителя. След връщане на персонализирани резултати, поведението на потребителя с резултатите се регистрира, за послужи да се подобрят бъдещите резултати. Агентът, който се учи на предпочитанията прочита тези регистрирани резултати като допълнение към външните източници, като например вградени системи и след това актуализира профила на потребителя.

Предложената концептуална схема се състои от пет агента: Интерфейсен агент, който е отговорен за всички взаимодействия с потребителя, Агент на потребителския профил, който улавя и поддържа потребителските предпочитания, Агент за семантично манипулиране на заявката, който манипулира потребителската заявка, Агент за персонализиране на резултатите, който персонализира търсените резултати и Мобилен агент, който търси информация в интернет страници. Фиг.12 показва концептуалната схема, която е предложена в дисертационния труд.

Функционалността на мобилния агент включва научаване, планиране и търсене на актуална информация в интернет. Събирането на информация е труден процес в зависимост от това в какъв вид информация трябва да бъде събрана, но изследванията се опитват да подобрят сегашните методи или дори да намерят нови. Координиращия мобилен агент приема потребителската заявка от Агента за манипулиране на заявката, намира необходимата категория и не прехвърля заявката на подходящ уеб агент, какъвто трябва да се използва, ако агента не беше мобилен.

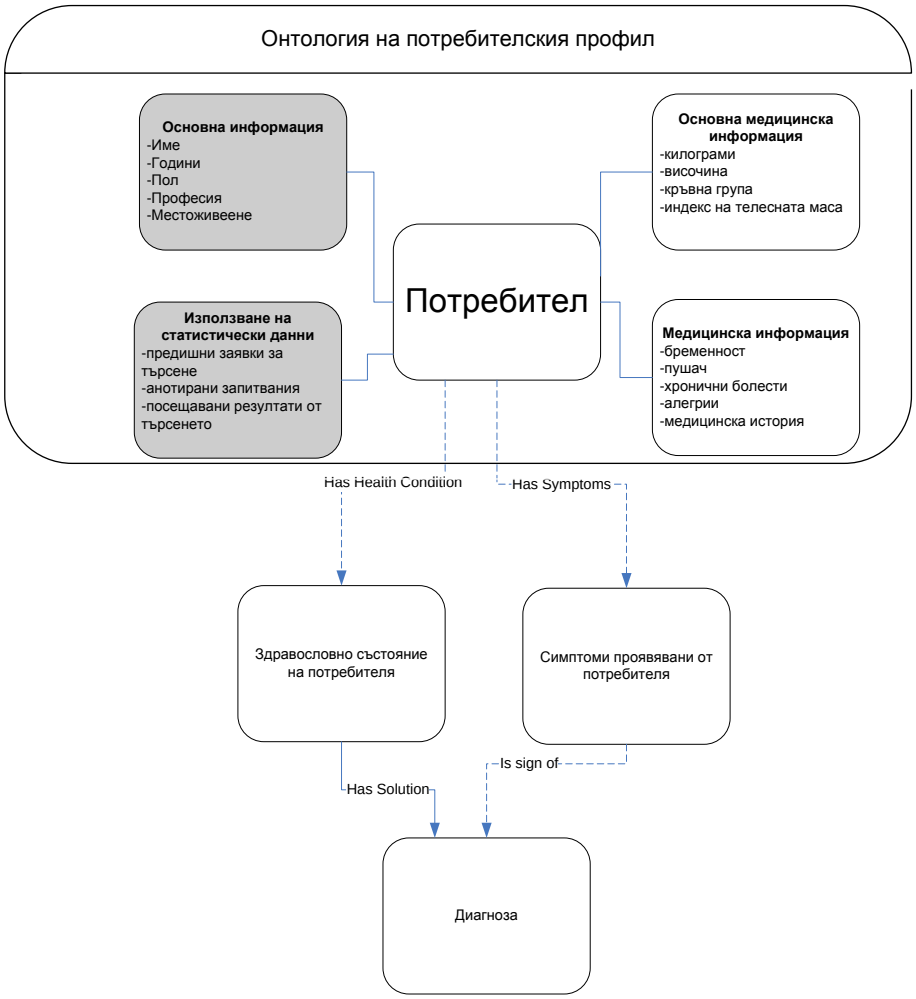
Използването на координиращ мобилен агент вместо обикновен агент спестява създаването на уеб агенти, което прави модела на системата по-опростен.

3.7.Концепция за изграждане на персонализиран потребителски профил

Потребителския профил е необходим, за да се персонализират търсените резултати. Като пример посочвам потребителския профил за търсене на медицинска информация в интернет. Но може да създаде система за търсене и в други научно приложни области. При търсене на медицинска информация в интернет има много неясноти и неточности. Потребителския профил помага по-добре да се разбере контекста на потребителската заявка особено в област в която това се счита за жизнено важно, като медицината.

На фиг.17 е представен профила, като модел на онтология, защото това го прави лесен за интегриране с онтологии в областта на медицината. Първоначално се създава потребителски профил, който съдържа информация за здравословното състояние на

потребителя, като се използва формуляр. След това, този профил се актуализира в резултат от търсената информация на потребителя. На фиг.17 се използва понятието онтология за представяне на профила и за използване за семантично търсене. Показани са детайлите на профила, на онтологията на потребителя на четири части: Първата част съдържа основна информация на потребителя, като име, пол, възраст и др. Втората част съдържа основна информация за здравето на потребителя, като кръвна група, индекс на телесната маса и др. Третата част съдържа медицинско досие – информация за потребителя, например, хронични болести, дали тя е бременна. Последната част съдържа статистически данни за употребата на информация на потребителя, като например неговите по- стари търсения и посетени линкове. Връзката между две понятия е показана, като пунктирана стрелка и се отнася за тройка в условията на RDF.



Фиг.17- Модел на персонализиран потребителски профил за медицинска информация

Модела на фиг.17 е разработен с помощта на експерти в областта на медицината.

3.7.3.1.Запазване на знанията в размитата онтология в предложената система.

Избран е подход за създаване на размитата онтология от обикновена (crisp) онтология.

Обикновената (crisp) онтология доказва своята роля при извличане на логически изводи базирани на прецеденти, размитите онтологии могат да разширят възможностите на обикновените онтологии. Обикновените онтологии не са подходящи за справяне с неясни и неточни знания, което е присъщо на отделните области от реалния свят. Размитата онтология може да се създаде, като се използва един от двата подхода: проектиране на размити бази данни или като разширение на обикновената онтология.

3.8.Описание на процеса на извеждане на логически изводи

Предлагам хибриден механизъм за извеждане на логически изводи, в който се комбинира извеждането на изводи за обикновена онтология и за размита и неясна онтология.

Първа стъпка прилагане на алгоритъм базиран на правила

С навлизане на понятието голям обем от данни се въвежда разпределеното извеждане на логически изводи. Ключова концепция е това да разшири вероятностната база знания в разпределена вероятностна база знания. За да се осъществи това е необходим, набор от правила и данни, които да се разпредели към различните постове. Изборът от първоначалния набор от данни автоматично води до несигурност в резултат на непълна информация.

Тази стъпка може да увеличи коефициента на несъвършенство на информацията, като стъпката се характеризира с несигурност и неяснота. Два проблема могат да бъдат дефинирани от това:

-осъществяване на избор на подгрупа от данни

-определяне на разпределението на данните

Основния проблем със стъпка 1 и 2 е, че съвместното съществуване на някои събития с някои правила, никога няма да доведе до заключение, тъй като те не са свързани.

Например, ако имаме общи познания за болестите на сърцето. Всички болести на сърцето имат болки в гърдите, тогава ние се интересуваме само от прецедентите, които се отнасят за болести на сърцето.

Разпределението на правилата и събитията между постовете трябва да бъде направено по начин, по който изразите, които са свързани трябва, ако е възможно да съществуват съвместно в един пост. Така, че всеки пост е свързан с набор от правила и набор от събития. Нашия свят е образуван от следните субекти за всеки пост:

-концепция основни правила- представя клас от строги и условни ограничения

-концепция прецедент- представя клас от прецеденти

-концепция пост- представя клас от постове

Основните правила имат форма:

Правило: (*b* | *a*[*f*]), означава, че ако **a** държи тогава **b** обикновено държи с коефициента на несъвършенство **f**. Събитията имат форма: Събитие: (*e* | *T* [*f*]), което означава, че събитието е истина с коефициент на несъвършенство **f**, където коефициента представя коефициент на несъвършенство, т.е. коефициент на несигурност и неяснота. Всеки субект може да се асоциира с друг субект, като се посочва, че съществува **корелация**. За да се осъществи това всеки субект може да се характеризира с **ключова стойност**, т.е.

1. ако *a тогава b с коефициент f и ключова стойност k*

2. *a е истина с фактор f и ключова стойност k*

Ролята на ключовата стойност е:

1. Общо описание на правилата- описва правилата, които могат да се използват само когато специфични прецеденти съществуват

2. Описание на прецеденти: Това описва прецеденти, които произлизат от един и същ резултат при едни и същи общи правила т.е.

“ако болест на сърцето има болки в гърдите с фактор 1”, тогава това е истина за всички болести на сърцето, като правилото болки в гърдите е общо.

Това означава, че процедурата за изводи е необходимо да се прилага само веднъж за един прецедент. След като се определи ключовата стойност за всеки субект, корелацията на два субекта се определя както следва:

Ако *e*1 и *e*2 са субекти, след това те са свързани, ако f има една и съща ключова стойност

Така, че първата стъпка в метода е да разпредели прецедентите и правилата между постовете в зависимост от тяхната **ключова стойност.**

Например, ако имаме правило за болестите, които засягат различни части на тялото, тогава ключа тук е честотата на срещане на дадена болест, и следователно болестите могат се разпределят на постове в зависимост от частите на тялото.

Алгоритъма е следния:

Алгоритъм 1– фактор на свързване

1.Определяне на броя на постовете равен на броя на ключовите стойности

2.За всеки субект да се изпълни

3.Свързва субект с пост, който има едно и също число с това на ключовата стойност на субекта

4.За всеки пост се изпълнява намиране на заключение относно прецеденти и правила на тези постове

Ключовата стойност също определя броя на използваните постове: **Броя на използваните постове е равен на броя на използваните ключови стойности.** След определяне на даден пост, всеки пост съдържа броя на прецедентите и правила. Следващата стъпка е процеса на извличане на логически изводи. Този процес се изпълнява независимо на всеки пост, без да се взимат в предвид другите постове. Това позволява обработката на процеса да се извършва паралелно. Проблема е ,че може да възникне конфликт в ситуации, когато две правила имат като резултат едно и също заключение с различни коефициенти на несъвършенство.

Например, ако имаме две правила нервни болести на гръбначния стълб имат болки в гърдите с фактор 0.5 и при болест на сърцето винаги има болка в гърдите, тогава второто правило трябва да се приложи вместо първото.

Като цяло тези конфликти могат да възникнат в ситуации, когато съществуват извънредни класове (като сърдечни болести) и извънредни правила, а оттам извънредните класове се предпочитат за разлика от общия клас.

Алгоритъма за извличане на логически изводи на всеки пост е следния:

Алгоритъм 2- метод за извличане на логически изводи

Начало

Оказване на приоритет 2 на всички извънредни правила

Оказване на приоритет 1 на правила без изключения

Стартиране на двигателя за правила от по-висок към по-нисък приоритет

Отменяне на фактите, когато правилото стартира

Край на алгоритъма

Свиване на фактите служи, като начин да се спре стартирането на правила с нисък приоритет, така че ако правилото което се стартира има приоритет 2, тогава със свиване на факта, който причинява стартиране, правилото с приоритет 1 няма да се стартира.

Втора стъпка прилагане на алгоритъма за класическа онтология и изчисляване на семантичните прилики

Сходството на медицинските понятия може да се провежда без семантика или лексика, или може да бъде направено семантично използвайки онтологията. Избира се втория начин, за измерване на сходство на значенията между понятията. Намерените m прецедента от първия слой влизат в друга оценка на базата на семантични прилики между функциите на случаите (instance). Лексикалната или точната прилика не могат да се използват за сравняване на онтологичните понятия.

Предложена е хибридна мярка въз основа на дължината на пътя и характеристиките на понятието. Първо за дължина на пътеката, се предлага прилика на дълбочината на най-слаб общ прародител (LCA) на две понятия и нивото на близост на понятията до тяхното LCA (Least Common Ancestor) в онтологията, което се основава само на IS-A отношения. С други думи, колкото по-дълбоко е LCA, толкова по-специфично се счита и, по този начин, по-подобни на сравняваните понятия са приети. Колкото по- близки са две понятия до тяхното LCA, те са по-подобни. На второ място, за да се определи количественото сходство

за концептуалните характеристики, трябва да се считат приликите и разликите между понятията.

Използва се предложения алгоритъм, за изчисляване на семантични прилики базирани на пътека и на функции. Използва се следното уравнение

SIM_{Sem} (u, v) =w1sim_{path} (u, v) +w2$sim_{feature}$ (u, v),

където w1,w2(0,1] и са тегла за w1+w2=1;

Чрез тази функция се изчисляват клиничните сходства между две концепции отколкото семантичното разстояние. Клиничното сходство се влияе от клиничната нееднородност на концепциите. Основното правило е, че колкото по-дълбока е концепцията, толкова по-специфична е тя. Например болести е по-обща концепция от болести на сърцето. Защото болести е по-генерално и абстрактно понятие, което означава също и други болести. Накрая се намира решение за най-подобен прецедент и се предлага на потребителя.

Алгоритъма за извеждане на логически изводи е подходящ за работа с голям обем от данни. Той извежда логически изводи освен за неясна, но и за несигурна информация, като включва и вероятността на даден прецедент да бъде извод.

3.9.Заключение

За решаването на задачите определи в дисертационния труд в настоящата глава е разработен архитектурен и концептуален модел на персонализирана търсеща система, която е пригодна за работа с неясни и несигурни зададени данни и има възможност за работа с големи данни. Използваните семантични и агентно- ориентирани технологии гарантират разширяемост, адаптивност, работа в динамична среда и подобряване качеството на търсенето.

1. Търсещата система използва нов хибриден алгоритъм за извеждане на логически изводи от база данни запазена в онтология, която е базирана на прецеденти. Предложения нов алгоритъм за извеждане на логически изводи използва алгоритъм за извеждане на логически базиран на правила, който е подходящ за големи данни и неясна и несигурно зададена информация в комбинация с алгоритъм за извеждане на логически изводи базиран на прецеденти с размита логика. Интегрирането на предложения алгоритъм в системата допринася за подобряване на качеството на върнатите резултати на потребителя при търсене в системата.
2. В системата е интегриран метод за търсена на информация в интернет с мобилен агент, този метод е модифициран и адаптиран за нуждите на системата. Като предлага търсене на информация в интернет по категории. Мобилния агент търси в интернет диагноза спрямо съответната заявка и след това търси метод за лечение за съответната диагноза.
3. С цел осигуряване на научаването на интересите на потребителя бяха разработени:

- Агентно- базиран подход за разработване на онтология на потребителския профил
- Компонентно- базиран подход за разработка на размита онтология базирана на прецеденти

ГЛАВА 4. Частична реализация на системата. Експериментална проверка и оценка на получените резултати

4.1.Цели и задачи на настоящата глава

Основна цел на настоящата глава е описанието на разработения прототип на персонализирана търсеща система за неясна и несигурна информация подходящ за голям обем от данни, за който в предходната глава е представен концептуалния модел и архитектурата. В настоящата глава е направена оценка на качествата на системата на базата на експериментални резултати.

За осъществяване на тези цели се определят следните задачи:

1.Изследване и оценка на възможностите на експерименталните средства.

2.Изследване и оценка на възможностите за реализация на търсещата система на основа системи за разработка на онтология и интелигентни агенти с отворен код.

3.Описание на възможностите на системата и конкретната реализация на интелигентните подходи и методи, които да използва.

4.Описание на възможностите на системата и конкретната реализация на конкретните подходи и методи, които тя използва.

5.Оценка на получените експериментални резултати.

4.2.Преглед и експериментална оценка на използваните софтуерни средства

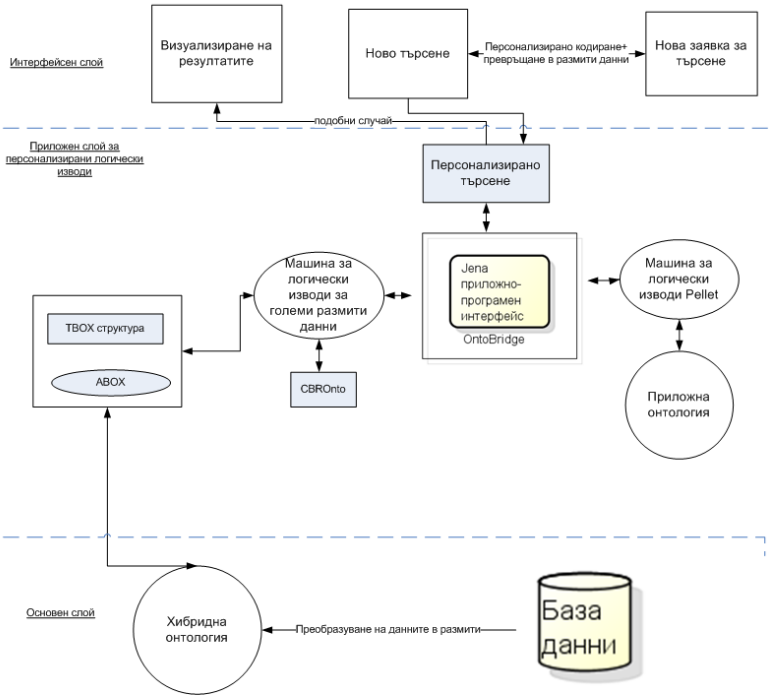
В процеса на реализация на системата се използва библиотеката JColibri при разпознаване на моделирането и извличане на резултати базирано на прецеденти, след което се интегрира в агенти, за да се създаде мултиагентната система. JADE е FIPA съвместима платформа за разработка на многоагентни системи, с възможност предварително да се интегрира в нея програмния интерфейс за разработка на онтологии на Protégé.

4.3.Описание на частичната реализация

4.3.1. Архитектура на персонализираната търсеща система

Създаването на приложението е предложено на фиг.18, като то се разделя на три слоя и всеки слой има специфична задача.

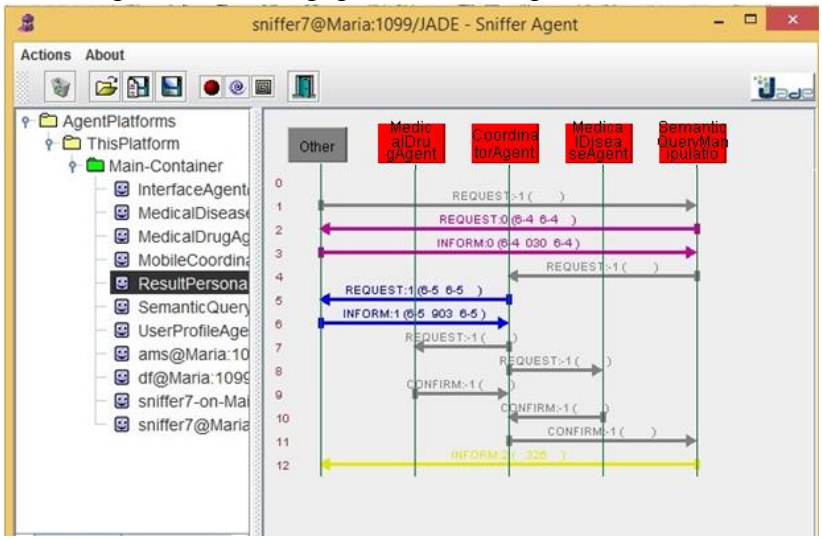
Основния слой подготвя размитата и несигурна онтология. Приложния слой базиран на прецеденти е най-важен за приложението, защото той съдържа извличането на данните от прецедентите. Интерфейсния слой приема заявката от потребителя на приложението и връща най- близкия прецедент. В приложението са въведени само стъпките на представяне и търсене.



Фиг.18- Структура на изпълнение на предложената система

4.3.4.Архитектура и представяне на статичния и мобилния агент при търсене на информация в интернет

Говорещия (sniffer) агент на JADE позволява да се наблюдава в реално време комуникацията между агентите. Тази последователност е описана на фиг.21, като е представено търсенето на информация в интернет със статичен агент.



Фиг.21- Говорещия агент проследява последователността на комуникационния акт между агентите, когато се използва статичен агент

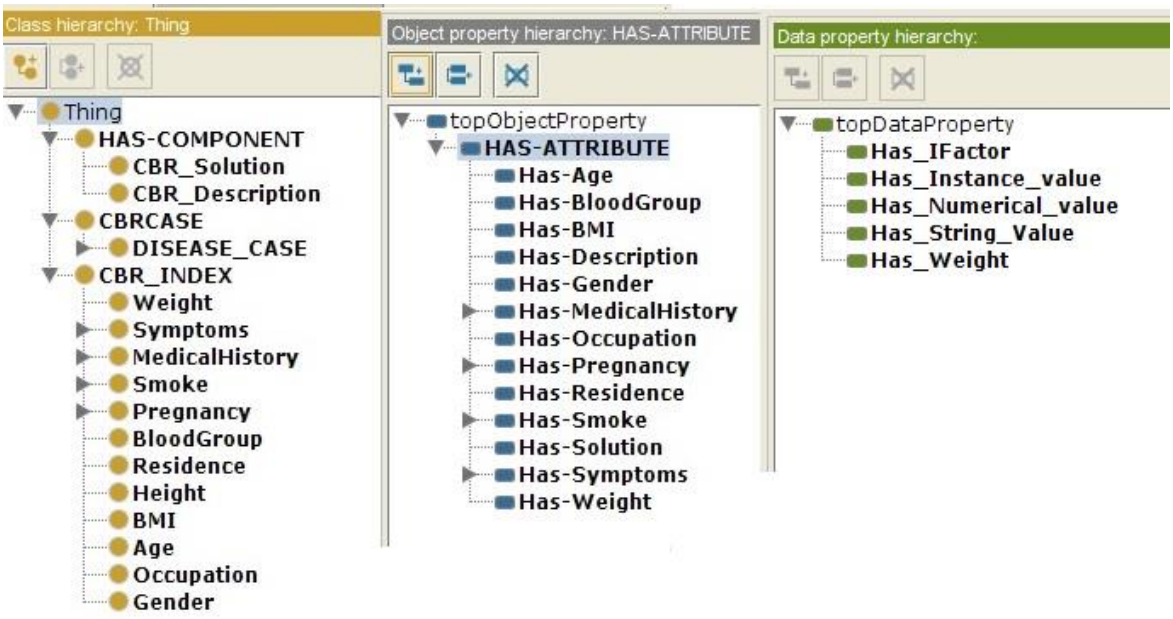
При използването на мобилен агент неговата функционалност замества тази на координиращия агент и уеб агентите, като схемата става по-опростена.

Както се наблюдава на фиг.21, когато Агента за манипулиране на заявката приеме заявка от потребителя, той търси агент, който предлага услугата за координация чрез съобщението за искане и получава отговор, като информира комуникативния акт. След това изпраща искане до координиращ агент, които ще попита за агентите разположени в желаните категории. След като намери, координаторът ще изпрати искания до тези средства, чакащи за техните потвърждения. Когато всички от тях са потвърдили, че ще изпратят съобщение за потвърждение на семантичния агент за манипулиране на заявката, се дава възможност да се научи, когато заявката е завършена.

4.4.Онтологиите в персонализираната търсеща система

4.4.1.Създаване на размита и неясна онтология

За създаване на размита и неясна онтология, която да служи за запазване на прецедентите се създава обикновена (crisp) онтология и след това се създава размитата част на онтологията. На фиг.22 е представена обикновена онтология базирана на прецеденти. Йерархията на онтологията е съвместима с CBROnto. Правилата, както и допълнителните класове не са представени в онтологията. Както е представено на фиг.22 CBR_INDEX включва всички функции на прецедента, CBRCASE включва отделните възможности на прецедента и HAS-COMPONENT двете части на прецедента- предлагане на медикаменти за прецедента и описание на прецедента. В онтологията базирана на прецеденти, прецедентите са представени, като примерни понятия и техните атрибути са представени, като онтологични връзки и свойства. Стойностите, които свързват атрибути могат да се приемат за прецеденти определени в рамките на една и съща домейн онтология.



Фиг.22 - Представя обикновената част на онтологията, която е базирана на прецеденти

Всичките данни за прецедентите и структурата са в онтология базирана на прецеденти.

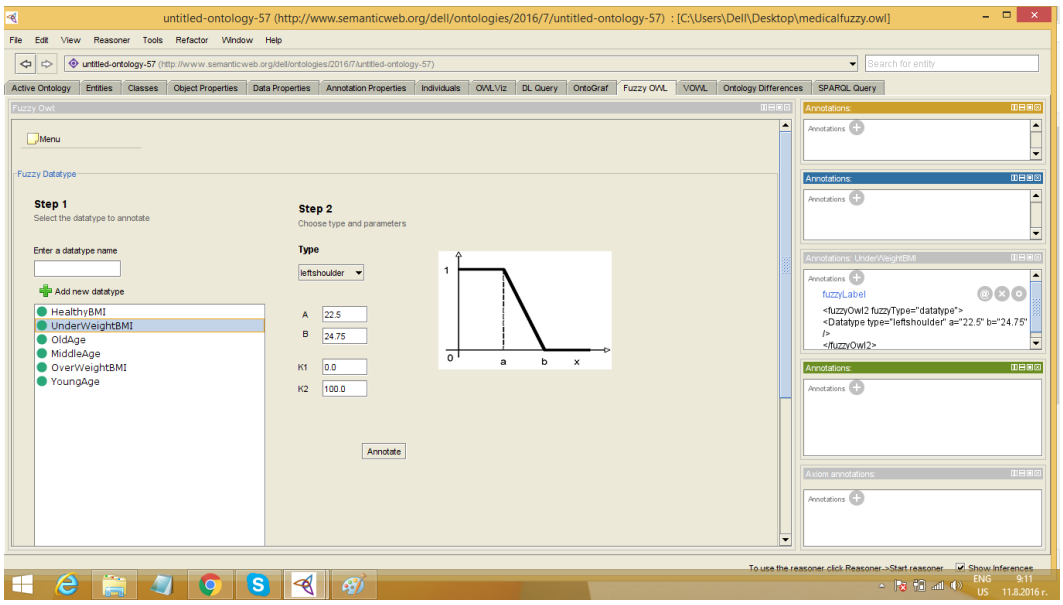
4.4.3. Разширяване на обикновената онтология за създаване на размита онтология

Предложената система използва онтологията за представяне на размита и несигурна информация. В онтологията базирана на прецеденти се добавят следните размити параметри: години, индекс на телесната маса, температура, кръвно налягане, пулс.

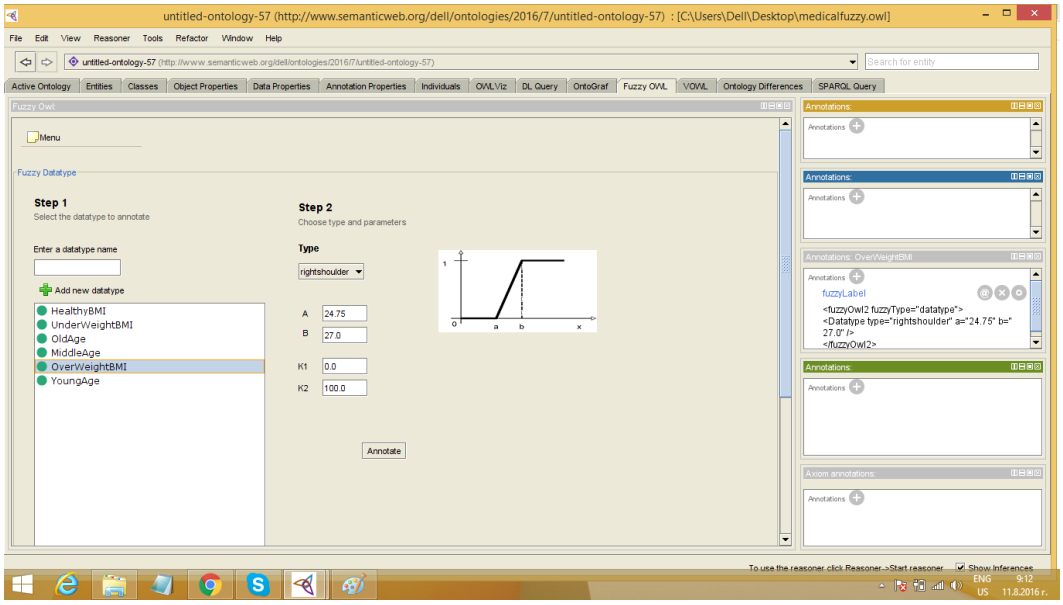
Размити типове данни и размити определени роли (свойства данни). За всяка една от числовите характеристики в базата с прецеденти са определени границите с помощта на литературни източници. Определя се техния обхват, размитите членни функции, техните форми и параметри. За създаване на размитост на тези стойности, се определят две неща- размит тип данни и размита определена роля. Опитът показва, че припокриването на триъгълник- триъгълник и трапец- триъгълник на размитите региони е средно между 25% и 50% от размитата база с множества.

Първо се създават размити типове данни за всеки от неясните термини. За създаване на размитата OWL функция се използва приставката за създаване на размита анотация Fuzzy OWL2 версия 2.21 за Protégé версия 4.3. Тук се разглеждат подробно размити понятия, типове данни и връзки. За размити типове данни, функциите, които се допускат при размито OWL2, се дефинират в интервала [k1, k2] ⊆ Q са d → {лява (k1,k2,a,b) (фиг.25а), дясна (k1,k2,a,b) (фиг.25б), триъгълна(k1,k2,a,b,c) (фиг.25в). Формализирането на всеки елемент в онтологията е показан на фиг.25.

На фиг.25 е представена, като пример, размитата функция на индекса на телесната маса, като за определяне на границите на тази функцията се използват стойности посочени в литературните източници.

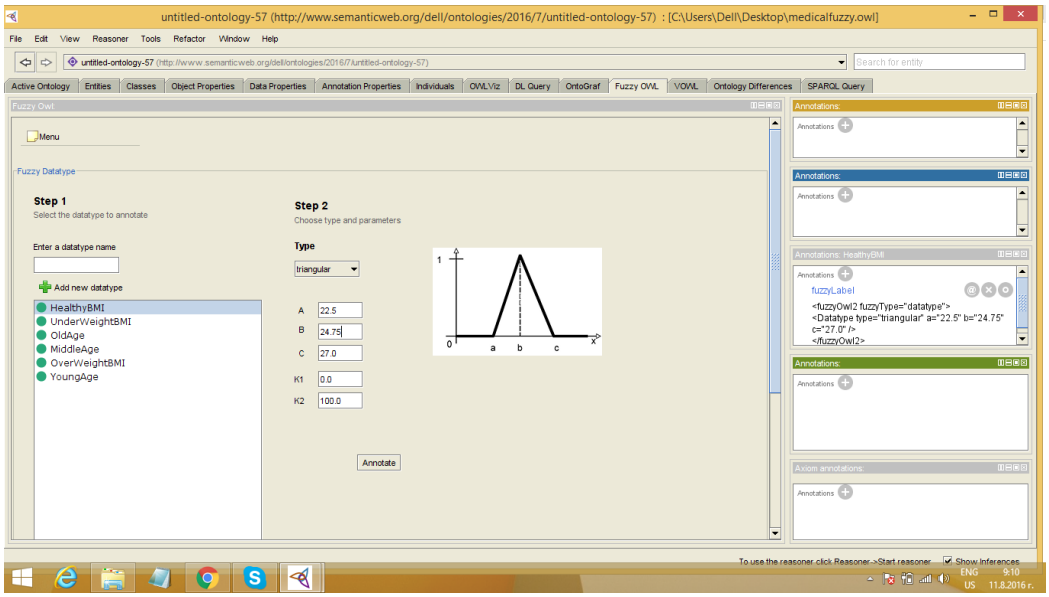


Фиг.25 а)- Представяне на ляво раменна функция на индекса на телесната маса



Фиг.25 б) – Представяне на дясно раменна функция на индекса на телесната маса

В литературата е споменато, че за практически приложения се използва предимно триъгълна функция и по-малко се използва трапецовидната функция, което предопределя избора за функция.



Фиг.25 в)- Представяне на триъгълна функция на индекса на телесната маса

За размита логика на онтологията се избира Заде. Тази анотация на ниво онтология е: <fuzzyOwl2 fuzzyType = “ontology”>

<FuzzyLogic logic = “zadeh”/>

</fuzzyOwl2>

Прецедентите в онтологията се въвеждат от мобилния агент на базата на правила за проектиране определени от W3C правила и се обновяват при работата на системата.

4.4.3. Разширяване на онтологията за работа с неясна информация и с големи данни.

Ролята на онтологията е да запази данните във форма на прецеденти и правила и след това да ги разпредели между постовете в зависимост от техните ключови стойности.

Понятия при добавянето на правила в онтологията:

1.Знания- представят прецедентите и правилата

- Правила- представят всички правила, т.е. субекти с общи познания

- Прецеденти-представят всички субекти от размитата онтология за прецеденти

2.Постове- представляват субектите на всеки отделен пост

3.Условия- това е условната част на правилата

4.Заключение- включва заключителната част на правилата

Също онтологията има следните обектни свойства:

1.if- това е обектно свойство, което свързва субектните правила с условните субекти

2.then- това е обектно свойство, което свързва субектните правила със заключителните субекти

Прилагане към онтологията базирана на прецеденти на свойства на данните:

hasFactor- Представя фактора за несъвършенство, за всяко правило или прецедент

hasPeer- Представя броя на постовете асоцииран с всяко правило или прецедент или с други думи ключова стойност. Като ключова стойност това може да са възможните групи болести.

Стойността на фактора на несъвършенство ползва стойностите на размитата членна функция. Тази информация се допълва към онтологията с помощта на плъгина на Protégé JESS.

4.6.Експериментални резултати

Обсъждане и оценка на експерименталните резултати са направени на някои етапи на работа на персонализираната търсеща система и в заключение е оценена ефективността и като цяло, дискутирани са подходи за подобряване на нейната ефективност. Първоначално са обсъдени подходите за оценка на качеството на върнатите резултати и са предложени метрики подходящи за конкретната задача.

4.6.1.Метрики за оценяване на качеството на върнатите от търсещата машина резултати

Основно при оценяването на качеството на работа на търсеща система се приемат два параметъра за оценяване- прецизност (Precission) и върнати резултати (Recall).

Целта на семантичното търсене с онтология е да максимизира прецизността и връщането, където

$$\textit{Прецизност} = \frac{\text{Брои на намерените съответни документи}}{\text{Брои на намерените документи}}$$

$$\textit{Връщане} = \frac{\text{Брои на намерените документи}}{\text{Брои на съответните документи}}$$

$$\textit{F-измерване} = \frac{\text{Прецизност} * \text{Връщане}}{\text{Прецизност} + \text{Връщане}}$$

За да се направи изследване на параметрите представени по-горе. Се събират 225 заявки от различни източници, като експерти в областта, проучване на потребителите, бази с медицински данни и различни уеб сайтове за здравна информация. Заявките се категоризират в зависимост от симптомите на различните болести. Извършва се диагностика на заявките от системата и от експерт в определената област и се извеждат резултати. Изследването е представено на табл.10.

Таблица 10- Експериментални резултати за определяне на параметрите- прецизност, върнати резултати, F- измерване на предложената система.

Категория симптоми	Общ брой заявки	Диагностицирани от системата	Диагностицирани от експерт	A1	A2	A3	Прецизност	Върнати резултати	F-измерване
Коремни и храносмилателни симптоми	35	34	36	31	1	3	0,96875	0,911764706	0,939393939
Сърдечно съдови симптоми	22	22	22	20	0	2	1	0,909090909	0,952380952
Симптоми на главата и шията	19	19	19	18	0	1	1	0,947368421	0,972972973
Симптоми на имунната система	9	9	8	8	1	0	0,888888889	1	0,941176471
Мускуло- скелетна система	8	10	8	8	0	0	1	1	1
Нервна система	27	26	27	25	0	2	1	0,925925926	0,961538462
Хранене метаболизъм	30	32	30	28	1	1	0,965517241	0,965517241	0,965517241
Репродуктивна система	21	20	20	18	1	2	0,947368421	0,9	0,923076923
Респираторни и гръдни симптоми	32	32	33	31	0	1	1	0,96875	0,984126984
Кожни и обвивни заболявания на съединителната тъкан	13	12	13	12	0	1	1	0,923076923	0,96
Уринарна система	9	9	9	8	0	1	1	0,888888889	0,941176471
Общо	225	225	225						
Средно							0,98	0,94003482	0,96
Легенда:									
A1- Системата и експерта са на едно мнение									
A2- Системата и експерта са на различни мнения, експерта смята, че болестта е в друга категория									
A3-Експерта отсъжда, че е в тази категория, но системата не									

При предложения анализ на системата не са отчетени резултатите, когато и експерта и системата дават различни резултати за дадена болест, поради което не може да се изчисли параметъра отрицателна предсказана стойност.

Като извод от направения анализ се определя, че системата има висока прецизност 98% и върнатите резултати са високи 94%, което показва качеството на системата.

Сравняване на предложената система с други системи представени в литературата е представено на табл.11.

Таблица 11- Сравняване на точността на предложения прототип на система с тази на други изследвания

Вид машина за логически изводи	**Област на приложение**	**Наименование на системата**	**Цел**	**Точност (%)**	**Прецизност**	**Върнати резултати**
Хибридна машина за логически изводи за голям обем от размити и неясни данни	Медицина, туризъм и други	Предложената система	Извеждане на логически изводи от несигурни и неясни данни	95.5	98	94
Машина за логически изводи за голям обем от данни базирана на прецеденти	Медицина	Improving biomedical signal search results in big data case-based reasoning environments	Представяне на монте карло техника на сближаване за последовател-но съвпадение	-	100	0.99
Машина за логически изводи за размити данни базирани на правила	Търсене на футболно видео	Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video	Размита машина за логически изводи базирана на правила	-	76.12	81.78
Машина за логически изводи за размити данни базирани на прецеденти	Система за диагностика на диабет	A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis	Размита машина за логически изводи базирана на прецеденти	97.67	100	96.43

Хибридна машина за логически изводи за размити данни	Медицинска диагностика	A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification				
			Средна точност на прогнозиране на рак на гърдата от базата с данни за рак на гърдата Уисконсин	98.4	-	-
			Проблеми с черния дроб от база с данни за проблеми с черния дроб	81.6	-	-
Традиционна машина за логически изводи	Персонализирана система за подобряване на живота (wellness)	Multimodal hybrid reasoning methodology for personalized wellbeing services	Хибридна методология, която интегрира изводи базирани на правила, изводи базирани на и изводи прецеденти в линейна комбинация, която позволява персонализирането и препоръки	94	94	97

4.6.2.Оценка на използването на мобилни агенти при търсене на информация в интернет

За да се направи обосновка на използването на мобилен агент при търсенето на информация е направен кратък сравнителен анализ с резултати, който е представен на табл.12. Тези резултати са получени чрез провеждане на търсенето на данни в интернет с мобилен и със статичен агент. За тестването на системата се използва компютър със следната системна конфигурация: Intel Core i5-5200 CPU @2.20GHz, 4GB RAM. Подробно описание на получените резултати са представени в следващите параграфи.

Таблица 12- Сравнителни резултати

	Използване при търсене стационарни агенти	**Използване при търсене мобилен агент**
Време за отговор	80524ms	75123ms

Общото време необходимо за получаване на отговор от страна на архитектурата със статични агенти ще бъде максималното време необходимо за един уеб агент да събере всички данни от уеб страницата, плюс времето за комуникация между агентите. В сравнение

с това, общото време, необходимо за архитектурата с мобилен агент ще бъде сумата от времената, необходими за събирането на данните от всяка уеб страница. Времето необходимо за комуникация е намалено при архитектурата с мобилни агенти, което се дължи на по-малкия брои агенти, които участват.

Въз основа на тези резултати се определя, като по-предпочитана архитектурата с мобилен агент, поради необходимост от управление на ресурсите. Друго предимство, което може да се добави е сигурността на канала за комуникация по който комуникират агентите. Резултатите от теста могат да се променят при стартиране на приложението на различна системна конфигурация с друга скорост на интернет връзката.

4.7.Заключение

В настоящата глава е разработен прототип на персонализирана търсеща система, базирана на семантични технологии, използваща Уеб текстови и семантични документи, интелигентни и многоагентни технологии с цел събиране, съхраняване и използване на знания относно интересите на потребителя за подобряване качеството на търсене на информация на потребителя. Системата е тествана с множество заявки събрани от интернет, чрез проучване на потребители или експертни мнения в дадена област. Постигнати са следните резултати:

1. На основа извършеното структурно и функционално изследване на библиотеката за разпределени приложения ALADIN. Лексическата база Wordnet, използвана в множество проекти на семантичния интернет с цел определяне на нейната пригодност в конкретната задача се достига до следните заключения:

- ALADIN (jcolibri) е пригоден за използване на системи базирани на прецеденти. Aladin има възможност за интегриране на онтологии и предоставя възможност за търсене на информация в онтологии. ALADIN може да комуникира с JADE платформата.
- Основен недостатък на ALADIN, за предложената система за търсене е невъзможността на агентите да мигрират. Тоест, Aladin не поддържа функция мобилност на агентите. Затова се налага в системата да се използват и JADE агенти.

2. Намиране на подходящи Уеб онтологии, позволяващи смисловото структуриране на термините от Уеб документите в общия случай не е гарантирано. Почти всички онтологии в интернет предлагат структурирани знания само на английски и при това тези знания често или са прекалено общи, или тясно специализирани, но не точно в необходимата област и се оказват недостатъчни за пълно изграждане на базата данни за търсене. За това се налага използването на мобилен агент за търсене на информация в интернет, който да се интегрира в онтологията базирана на прецеденти и правила.
3. В резултат на извършените изследвания с разработения прототип на персонализирана търсеща система се установи, че:

- При търсене на неясна и несигурна информация системата предлага висока прецизност на върнатия резултат.
- Задаване на потребителския профил е важно за определянето на върнатия резултат.
- Големината на обработваните прецеденти влияе на времето на търсене на информация в системата.
- Използването на мобилен агент за търсене на информация в системата намалява времето на търсене на информация в интернет.

5.Бъдещото усъвършенстване на персонализираната търсеща система е предимно при:

- Проучване на възможността за интегриране на готови данни с прецеденти към системата. Както и анализ на получените резултати при работа с различно голям обем от данни, която предоставя системата.
- Добавяне на повече уеб сайтове за събиране на информация и намиране на начин за динамично модифициране на списъка и след това да се използва по-обща техника за събиране на информация, която може да се приложи за всички уеб сайтове в списъка

- Проучване и анализиране на системата за работа в други научно-приложни области, като туризъм, търсене на имоти и други.

ЗАКЛЮЧЕНИЕ

Извършеният анализ в дисертацията показва, че семантичните технологии са особено полезни за подпомагане работата на експерти в много научни области, където са натрупани и систематизирани голям обем знания, за по-нататъшната обработка и ефективно използване. Онтологиите и наскоро утвърдилият се, като стандарт език за разработка на онтологии OWL, се използва все по-широко за представяне и обработка на знания в области, като био информатика, медицина, туризъм, електронното обучение и много други. В последните година са генерирани множество доклади в научно- приложната област медицина и използването на медицинските данни. Семантичните технологии се използват успешно с интегриране на размитите множества, както и теорията на вероятностите, като дават добри резултати и нови възможности за изследвания. Семантичните технологии са основополагаща технология за много системи за търсене на информация и в много изследвания свързани със семантичните уеб услуги.

Настоящата работа въвежда нови идеи, подходи и модели за представяне на субективните знания чрез специфично разширяване на дескриптивната логика с елементи на размитата и вероятностна логика. Създаден е агентно– базиран модел и прототип на семантична търсеща система, използваща онтология за извличане на информация, която съответства на потребителската заявка, като използва и съхранява знанията за предпочитанията на потребителя. Персонализираното търсене е един от основните подходи за подобряване на качеството на системите за търсене на информация. Използваната в конкретната разработка, многоагентна технология за изграждане на персонализираната семантична система подобрява работата в динамичната и бързо променяща се Уеб среда и предлага разширяемост и възможности за използване на последните достижения в областта на семантичните технологии и многоагентните системи.

На основа постигнатите резултати бъдещите изследвания могат да бъдат насочени в няколко основни направления- разработване на оптимизиращи алгоритми за предложената хибридна машина за логически изводи и експериментално изследване на тяхната ефективност чрез програмната им реализация в машина за изводи за голям обем от данни с отворен код FACT++. Изследване на възможностите за подобряване ефективността на генерирането на потребителския профил. Разработка на онтология с елементи на български имена и средства за извличане на знания от български текст.

ПРИНОСИ

Научни и научно-приложни приноси:

1.Създадена е хибридна онтология, която е базирана на прецеденти и правила, като е обогатена с елементи на размита и вероятностна логика. Тази онтология предоставя възможност за работа със заявки с размити и неясни данни, които са въведени от потребителя. За практическата реализация на онтологията е избран езикът OWL2.

2.Предложен е модел на потребителски профил, който обхваща и процеса на търсене на информация. Ролята на потребителския профил е да персонализира и уточни зададената заявка.

3.Предложен е нов хибриден модел на машина за логически изводи, който е подходящ при големи данни и неяснота на зададената информация от потребителя. В предложения модел са отчетени техните предимствата и недостатъците на познатите машини за извеждане на логически изводи за големи и размити данни.

4.Предложена е концептуална схема на персонализирана търсеща система, която интегрира предходните приноси, като при създаването му е използван и мобилен агент за търсене на информация в интернет.

5.Създадена е структура, както и прототип на динамична персонализирана търсеща система, която връща само един резултат по направената заявка на потребителя (one shot). Тази система подобрява качеството на върнатите резултати при зададена неясна и несигурна информация от потребителя, което е много важно.

Приложни приноси:

1.За да се покаже реализирането на постигнатите приноси е извършена реализация на отделни компоненти от цялостната система и анализ на експерименталните резултати. Извършено е сравнение в много приложни области като в областта на медицината се счита, че качеството на върнатата информация при търсене е особено важност.

2.Получени са резултати от апробиране на системата в конкретно приложена област за която са достъпни голямо количество от данни в интернет.

СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИОННИЯ ТРУД

1.Evtimova M., “Comparative analysis of multi-agent platforms for creation of semantic mobile agents for personalize searching” ,НСНТК, 2014, България, Созопол, pp.225-229
2.Evtimova M., Momtchev I., Personalize Web Searching Strategies Classification and Comparison, International Scientific Conference Computer Science’2015, Albania, pp.107-112
3.Evtimova M., Momtchev I., A reasoner for uncertain and vague big data, International Journal of Engineering Research and Allied Sciences, ISSN:2455-9660, Vol.01, Issue 08, 2016, India, p.12-14
4. Evitmova M., Momtchev I., “A model of hybrid ontology for medical information”, Journal: Computer and communications engineering, ISSN 1314 – 2291, Bulgaria , in printing

Summary

The thesis of the doctorate is focused on exploring new methods and tools for semantic agents for personalized search. Because of this, were analyzed and evaluated different platforms for the realization of mobile software agents. And the possibility of extracting information in an incomplete and contradictory information environment was investigated through the combination of software multiagent systems.

Furthermore, an investigation of the personalization of the searched query and was created a model for personalized user profile. The problems of the reasoners were analyzed in conditions of incomplete and contradictory information and a big volume of data and then was proposed a new hybrid algorithm for reasoning.

The types of ontologies were analyzed and a new model of ontology was proposed that can be used with the proposed reasoning algorithm. An investigation of conceptual models was performed in the field of semantics, ontologies and also in multiagent systems, so that to create a new conceptual model for personalized semantic system for searching that is suitable for fuzzy and uncertain information defined from the user and has a possibility to work with big data. Software programs were created for the separated modules. In addition, a set of instrumental programming tools was developed that was suitable for the proposed system. Finally, an experimental evaluation was performed for the possibilities of the realized modules.