УНИВЕРСИТЕТ ЗА НАЦИОНАЛНО И СВЕТОВНО СТОПАНСТВО – СОФИЯ

Катедра „Информационни технологии и комуникации”

Явор Николов Табов

ПОДХОДИ ЗА СТРУКТУРИРАНЕ НА ДАННИ В УЕБ БАЗИРАНИ ИНФОРМАЦИОННИ СИСТЕМИ

А В Т О Р Е Ф Е Р А Т

на дисертационен труд за присъждане
на образователна и научна степен „доктор“

Научен ръководител:
доц. д-р Пламен Милев

София, 2022

Дисертационният труд се състои от увод, три глави, приноси, заключение, библиография и приложения. Той е в общ обем от 217 страници (202 страници без приложенията). Използваните литературни източници са 108 броя, от тях 28 на български език, 61 на английски език и 19 интернет адреса. В изложението присъстват 77 фигури и 21 таблици.

Дисертационният труд е обсъден в катедра „Информационни технологии и комуникации" при Университет за национално и световно стопанство – София и е насочен за защита пред научно жури.

Авторът на дисертационния труд е бил редовен докторант към катедра "Информационни технологии и комуникации" при Университет за национално и световно стопанство – София.

Защитата на дисертационния труд ще се състои на 10.06.2022 г. от 13:30 часа в зала 2032А на Университет за национално и световно стопанство – София на заседание на научното жури.

Материалите по защитата са на разположение на интересуващите се в Университет за национално и световно стопанство – София.

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

1. Актуалност на темата

През последните години данните в публичното пространство значително се увеличават и представляват все по-голям интерес за организациите. Въпреки че данните се събират в голям обем и с висока честота, не начинът им на съхранение, а възможностите за аналитичната им обработка се превръщат в основно предизвикателство. Организациите разчитат на обработката на колкото се може повече данни за вземане на адекватни управленски решения. В резултат от развитието на уеб концепцията, голяма част от публичните данни са налични в уеб базирани информационни системи с различни предназначения. Поради сериозния обем на този вид данни и естествената им неструктурирана форма, съществено предизвикателство представлява привеждането им във форма, която позволява аналитичната им обработка. Неструктурираните данни представляват предизвикателство за специалистите в областта на информационните технологии, защото често изискват повече от предварително предвиденото време за привеждане в структурирана форма и последваща подготовка за аналитична обработка. Неструктурираните данни в уеб базираните информационни системи са с много голям обем. Това предполага приложението на специфични подходи при работа с този вид данни. В контекста на необходимостта от аналитична обработка на данните в уеб базираните информационни системи и предвид естествената неструктурирана форма на този вид данни, се откроява необходимостта от представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

2. Обект, предмет и цел на изследването

Обект са уеб базираните информационни системи;

Предмет е представянето на текстовите данни от уеб базираните информационни системи в структурирана форма;

Целта на дисертационния труд е: като се изследват спецификите на данните в уеб базираните информационни системи, да се предложи концептуален подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

3. Хипотеза и задачи на изследването

Представената цел на дисертационния труд дава основание за формулиране на следната **хипотеза** на научното изследване: на основата на анализ на съвременните измерения на структурирането на данни в уеб базирани информационни системи, може да се разработи концептуален подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

Основните задачи на разработката в контекста на темата, целта и хипотезата са:

1. Изследване на особеностите на данните в уеб базираните информационни системи;
2. Разглеждане на възможностите за извличане на данни от уеб базирани информационни системи;
3. Проучване и сравнителен анализ на съвременни софтуерни решения за представяне на текстови данни от уеб базирани информационни системи в структурирана форма;
4. Разработване на концептуален подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма;
5. Проектиране на система от методи в рамките на модел на софтуерно решение за представяне на текстови данни от уеб базирани информационни системи в структурирана форма;
6. Създаване на експериментален прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма;
7. Провеждане на експеримент с помощта на разработения прототип върху множество от уеб базирани информационни системи;
8. Анализ на резултатите от експерименталното приложение на подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

Резултатите от разработването на изследователската проблематика могат да намерят конкретно практическо приложение в изграждането на софтуерни решения за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

II. СТРУКТУРА И СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Увод

Глава Първа. Съвременни измерения на структурирането на данни в уеб базирани информационни системи

1.1.Особености на данните в уеб базираните информационни системи

1.2.Възможности за извличане на данни от уеб базирани информационни системи

1.3.Софтуерни решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма

1.4.Изводи

Глава Втора. Представяне на текстови данни от уеб базирани информационни системи в структурирана форма

2.1.Разработване на концептуален подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма

2.2.Описание на системата от методи в състава на подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма

2.3.Проектиране на модел на софтуерно решение за представяне на текстови данни от уеб базирани информационни системи в структурирана форма

2.4.Изводи

Глава Трета. Анализ на приложимостта на предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма

3.1.Създаване на експериментален прототип по предложения подход

3.2.Практическо приложение на експерименталния прототип

3.3.Оценка на резултатите от практическото приложение на експерименталния прототип

3.4.Изводи

Приноси

Заключение

Библиография

Приложения

III. КРАТКО ИЗЛОЖЕНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Глава Първа. Съвременни измерения на структурирането на данни в уеб базирани информационни системи

Необходимостта от представяне на текстови данни от уеб базирани информационни системи в структурирана форма се открива най-вече в последващите възможности за анализ на представените по такъв начин текстови данни. В контекста на повишаване на нивото на управленските решения, които се вземат в организациите, от съществено значение е наличието на данни във вид, който позволява тяхната последваща аналитична обработка.

Данните съществуват в различна форма навсякъде около нас, но тяхното съществуване не винаги е обвързано с конкретна цел. Много автори акцентират, че данните не са информация, а представляват сурови факти или наблюдения за физически процеси или бизнес трансакции[1]. Данните са обективна мярка за стойности на атрибути (характеристики) на обекти от реалния свят. Обичайно се представят под формата на символи – букви, цифри, специални знаци[2]. Преди широкото използване на изчислителната техника, данните за различни дейности, като преброяване на населението, научни експерименти, извадкови изследвания и др., са записвани на хартия. Събирането на данни е ставало при предварително дефиниране на въпросите, на които се търсят отговори, като резултатните данни са се представяли във вид на таблици с подредени редове и колони, така че да са податливи на традиционните статистически методи за анализ. В средата на XX век част от данните за научни изследвания се съхраняват в електронен вид, което помага да се облекчи този процес на събиране и обработка на данни. С началото на уеб концепцията и нейното последвало бързо развитие става все по–осъществимо да се генерират, събират, съхраняват и анализират данни по електронен път. Този процес на технологичен прогрес генерира неизбежни проблеми, свързани с обработването на все по–големи и разнообразни масиви от данни и преобразуването им в полезна информация[3]. Данните, които са налични в уеб пространството, могат да бъдат класифицирани като структурирани, неструктурирани и полуструктурирани. Традиционно структурираните данни се съхраняват електронно

[1] Struhl, S., Practical Text Analytics: Interpreting Text and Unstructured Data for Business Intelligence, Kogan Page, 2015, ISBN 978-0-7494-7401-0

[2] Мурджева, А. и колектив, Информатика, Издателски комплекс - УНСС, София, 2018, ISBN 978-619-232-123-9

[3] Holmes, D., Big Data: A Very Short Introduction, Oxford University Press, 2017, ISBN 978-0-19-877957-5

под формата на електронни таблици или бази от данни с таблици, съставени от редове и колони, като всеки ред представлява запис и всяка колона описва точно определено поле или характеристика. Съхраняваните в табличен вид структурирани данни са сравнително лесни за управление и подлежат на аналитични обработка. Неструктурираните данни, от друга страна, не са податливи лесно на категоризация и обработка. Те представляват данни, които са генерирани от машини или са създадени от човек. Някои автори ги класифицират в два типа[4]:

- Нетекстови тип под формата на мултимедия – изображения, видео файлове, аудио файлове и др.;
- Текстови тип под формата на имейл съобщения, незабавни съобщения, бележки, текстообработващи документи, презентации и др.

С развитието и разширяването на уеб концепцията се оказва, че данните в рамките на софтуерните решения, които са част от тази концепция, остават недостъпни за аналитична обработка в естествената им форма, защото им липсва необходимата структура за прилагане на съществуващите аналитични техники. Въпреки това, чрез идентифициране на ключови характеристики, данните, които изглеждат на пръв поглед неструктурирани, може да не са напълно без структура[5]. Електронните писма например съдържат структурирани метаданни в заглавието, както и действителното неструктурирано съобщение в текста. Метаданните представляват структурирана информация, която описва, обяснява, локализира или по някакъв начин улеснява операциите на извличане, използване или управление на даден информационен ресурс. Метаданни често се дефинират като данни за данни или информация за информация[6]. На тази база електронните писма могат да бъдат класифицирани и като полуструктурирани данни. Етикети с метаданни, които по същество са описателни, могат да се използват за добавяне на структура към неструктурирани данни. Добавянето на етикет с думи към изображение в рамките на уеб базирана информационна система прави същото това изображение разпознаваемо и по-лесно за търсене. Полуструктурирани данни могат да се открият и в социалните мрежи, които използват хаштагове (описателни етикети), така че съобщенията по дадена тема, които по същество са неструктурирани данни, могат да бъдат идентифицирани[7]. Основните дейности в процеса на обработка на

[4] Kanimozhi, K. and Venkatesan, M., Unstructured Data Analysis-A Survey. International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 3, 2015
[5] Holmes, D., Big Data: A Very Short Introduction, Oxford University Press, 2017, ISBN 978-0-19-877957-5
[6] https://www.lter.uaf.edu/metadata_files/UnderstandingMetadata.pdf
[7] Holmes, D., Big Data: A Very Short Introduction, Oxford University Press, 2017, ISBN 978-0-19-877957-5

неструктурираните данни в рамките на уеб концепцията могат да бъдат дефинирани в следната последователност[8]:

- Събиране на неструктурирани данни (създаване, измерване, запис и сравнение);
- Преобразуване на събраните неструктурирани данни в удобен за обработка вид (кодиране, класифициране, верификация и трансформация);
- Манипулиране на преобразуваните данни (подреждане, изчисление, обобщаване и сравнение);
- Съхранение на подредените и обобщени данни и извличане на съответната информация от тях;
- Извеждане и предаване на извлечената информация в подходящ вид.

В контекста на особеностите на данните много автори разглеждат и изследват от различни гледни точки понятието информация. По своята същност информацията представлява обработени данни, предоставени на заинтересована страна в полезна за нея форма с цел решаване на някакъв информационен проблем[9]. Информацията може да се определи и като данни, получили значение чрез релационна връзка.

Значението обикновено носи полезност в някаква конкретна сфера на приложение. В областта на приложение на изчислителната техника една релационна база от данни може да генерира информация от съхранените в нея данни[10]. Базата от данни е добре организирана колекция от данни, които са свързани по смислен начин и които могат да бъдат достъпни в различни логически последователности. Системите за управление на бази от данни са системи, в които съхранението на данни и тяхната интерпретацията под формата на информация са от първостепенно значение. Необходимостта от дългосрочно съхранение на данните и достъпът до данните от голям брой потребители обикновено характеризира системите за управление на бази от данни[11].

Наличието на голям обем данни не винаги води до тяхната интерпретация под формата на полезна информация. Това е значителна проблематика в контекста на наличието на неструктурирани данни във все по-големи обеми. Познанието къде

[8] Мурджева, А. и колектив, Информатика, Издателски комплекс - УНСС, София, 2018, ISBN 978-619-232-123-9
[9] Пак там
[10] https://digitalcommons.kennesaw.edu/facpubs/1370/
[11] Sumathi, S., Esakkirajan S., Fundamentals of Relational Database Management Systems, Springer, 2007, ISSN 1860-949X

да се търси, както и как да се приложи анализ, става критично по отношение на данните, които не са структурирани[12].

За да може да се определи необходимостта и значимостта от приложението на подходи за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, е необходимо да се изследва развитието на уеб концепцията и по-специално особеностите на данните в уеб базираните информационни системи. Условно развитието на уеб концепцията може да се раздели на три етапа[13]:

- Уеб 1.0;
- Уеб 2.0;
- Уеб 3.0.

Първият етап от развитието на уеб концепцията – Уеб 1.0, се отнася за достъпването на информация и влизането в Интернет[14]. Характерен пример е разработването и поддържането на представителни уеб страници през 90-те години, чието съдържание представлява информация за компании и техните продукти[15]. За концепцията Уеб 1.0 може да се каже, че започва да се прилага от 1996-та година и е актуална до 2004-та година. При нея се говори за уеб страници и хиперлинкове. Съдържанието на уеб страниците е статично (не се променя) и е предвидено основно за четене. За тази концепция са характерни частните представителни уеб сайтове и уеб сайтове за съобщения.

Вторият етап от развитието на уеб концепцията – Уеб 2.0, се отнася за свързването на хора, които поставят „аз“ в потребителския интерфейс и „ние“ в мрежа на социалното участие. Периодът, който обхваща развитието и приложението на концепцията Уеб 2.0, започва да се прилага от 2004-та година. При тази концепция вече се говори за социален уеб. Съдържанието на уеб страниците е динамично (може да се променя) и е предвидено не само за четене, но и за редакция. Появяват се уеб блогове и портали за общности. Аналог на представителните уеб сайтове стават профили от социални мрежи. Концепцията Уеб 2.0 представлява от една страна платформа, върху която са изградени иновативни технологии, а от друга страна пространство, в рамките на което потребителите се третират като обекти с най–голямо значение. Концепцията Уеб 2.0

[12] Struhl, S., Practical Text Analytics: Interpreting Text and Unstructured Data for Business Intelligence, Kogan Page, 2015, ISBN 978-0-7494-7401-0

[13] Върбанов, Р., Подход към използването на Enterprise 2.0 от Компаниите, Бизнес управление, бр.1, 2013, ISSN 0861-6604

[14] Naik, U., Shivalingaiah, D., Comparative Study of Web 1.0, Web 2.0 and Web 3.0, International CALIBER-2008

[15] Славова, М., Дигитална трансформация на бизнеса, Икономически и социални алтернативи, брой 3, 2016

се състои от нови технологии, на база на които са изградени различни популярни социални мрежи. Примери за такива социални мрежи са Facebook, LinkedIn и др. Във всички социални мрежи потребителите са също толкова важни, колкото съдържанието, което създават и споделят с останалите потребители[16]. Голям брой софтуерни решения и услуги в Интернет осъществяват социални взаимодействия между потребителите чрез богат обмен на мултимедийна информация. Примери за такива софтуерни решения и услуги представляват блогове, форуми, социално маркиране, peer to peer мрежи с инструменти с отворен код, фото и видео споделяне, онлайн бизнес мрежи и др.[17]. Съществената новост в концепцията Уеб 2.0 е, че потребителите в Интернет от консуматори на съдържание се превръщат в създатели на съдържание. Поради тази причина в рамките на концепцията Уеб 2.0 са реализирани множество технологични помощни средства с цел максимално увеличаване на потенциала за създаване на съдържание. Демократичният характер на концепцията Уеб 2.0 е илюстриран от създаването на голям брой социални кръгове, групирани по общи интереси в различни направления, които могат да обменят съдържание от всякакъв вид. Едни от най–популярните Уеб 2.0 инструменти са социалните мрежи Facebook, Google+, LinkedIn, платформата Twitter и др.[18]. Според изследване от 2015 г., над 84% от интернет потребителите имат регистрация в някоя от тези социални мрежи, а средното време, прекарвано в тях е между 2 и 4 часа на ден[19]. По отношение на концепцията Уеб 2.0 някои автори обръщат голямо внимание на развитието на корпоративните интранет портали. Интранет порталите представляват инструменти, които помагат на дадена компания в управлението на своите данни, приложения и информация, чрез персонализирани потребителски изгледи. Може да се каже, че моделът на използване на концепцията Уеб 2.0 е съпроводен със създаване на потребителско съдържание, предназначено за ползване от цялата интернет общност[20]. Възможността на потребителите да създават интернет съдържание променя изцяло облика на Уеб. Потребителите могат да създават, публикуват, коментират и споделят съдържание. От технологична гледна точка такива системи за публикуване на съдържание в Интернет могат да се третират като уеб базирани информационни системи за управление на съдържание. В контекста на популярните социални мрежи този тип системи се свързва с

[16] Cormode, G., Krishnamurthy, B., Key differences between Web 1.0 and Web 2.0, First Monday, Volume 13 Number 6 - 2 June 2008, ISSN 1396-0466

[17] Върбанов, Р., Подход към използването на Enterprise 2.0 от Компаниите, Бизнес управление, бр.1, 2013, ISSN 0861-6604

[18] Ванков, Н., Онлайн ресурси и инструментариум за дигитален PR, 2014

[19] Върбанов, Р., Потенциалът на социалните медии в бизнеса на компаниите, Бизнес управление, бр. 3, 2015, ISSN 0861-6604

[20] Георгиев, В., Насоки за развитие на корпоративните комуникации чрез приложение на средства от социалните мрежи, Икономически и социални алтернативи, брой 3, 2014

понятието големи данни[21]. Под големи данни се разбира нарастването на количества и типове данни, които стават неудобни за управление чрез традиционните информационни средства. Големите данни се характеризират с необходимостта от прилагането на съвкупност от средства, инструменти и процедури, които позволяват на компаниите да създават и управляват данни, достигащи до размери от порядъка на петабайти и повече[22]. Важно е да се отбележи, че големите данни се свързват с един основен проблем, който не се изразява в големия обем, а във факта, че голяма част от тях не съответства на традиционния структуриран формат на базите от данни – неструктурирани данни (видео записи, текстови документи, машинен код и др.)[23].

Третият етап от развитието на уеб концепцията – Уеб 3.0, поставя акцент върху представянето на значения, свързване на знания и адаптирането им да работят по начини, които правят опита на потребителите в Интернет по-подходящ, полезен и приятен[24]. Концепцията Уеб 3.0 датира след 2016-та година. С нея се свързват понятия като семантичен уеб и изкуствен интелект. Говори се за изпълними приложения в уеб. Характерното за тази концепция е, че при нея потребители създават приложения, предоставящи възможност за публикуване на съдържание от други потребители.

Като обобщение на развитието на уеб концепцията, могат да бъдат направени следните изводи:

- Уеб 1.0 се счита за уеб, който представя информация;
- Уеб 2.0 надгражда Уеб 1.0, като се фокусира върху създаването на съдържанието от страна на потребителите;
- Уеб 3.0 е насочен към свързаните набори от данни.

На база на изложените характеристики на концепцията Уеб 2.0 и особеностите на етапите на развитие на Интернет, може да се обобщи следният извод: разнообразието от възможностите, които предоставят уеб базираните информационни системи, е предпоставка за наличието на много данни с разнообразно тематично съдържание и липса на централизиран контрол по отношение на тяхната форма на представяне. Ако в рамките на една уеб базирана информационна система данните обичайно са представени по определен начин, то в

[21] Милев, П., Тенденции в развитието на архитектурните решения при системите за уеб публикации, Бизнес управление, кн. 2, 2019, ISSN 0861-6604
[22] Стефанова, К., Кончев, К., Аналитични приложения за обработка на големи обеми данни в реално време, Икономически и социални алтернативи, брой 3, 2012
[23] Йорданова, С., Стефанова, К., Предизвикателствата на големите данни, Научни трудове на УНСС, том 1/2019, ИК – УНСС, София
[24] Naik, U., Shivalingaiah, D., Comparative Study of Web 1.0, Web 2.0 and Web 3.0, International CALIBER-2008

рамките на група от множество уеб базирани информационни системи представянето и категоризирането на един и същ тип данни обикновено е различно за всяка уеб базирана информационна система и цялото множество от данни може да се приеме като множество от неструктурирани данни.

В днешно време голяма част от данните, които представляват интерес за анализ от страна на организациите, са в неструктурирана форма в рамките на различни уеб базирани информационни системи. В този контекст извличането на данни от уеб базирани информационни системи е ключова дейност, която е от изключителна важност за организациите. Според някои проучвания в предметната област повече от 80% от полезните бизнес данни са в неструктурирана форма[25]. Изследвания по тематиката показват, че използването на големи количества данни е характерно основно за големите компании, като тенденцията е броят на тези компании да нараства[26]. Уеб пространството е ярък пример за хранилище на огромни набори от данни.

Според редица изследвания в предметната област неструктурираните данни постоянно увеличават обема си, като това се дължи на данните, които ежедневно се създават в съвременните световноизвестни платформи с богато потребителско съдържание, като YouTube, Facebook, Twitter, LinkedIn и др.[27]. Представените системи за управление на съдържание по същество представляват уеб базирани информационни системи. Уеб базираните информационни системи могат да бъдат описани под формата на различни видове, от гледна точка на своето предназначение и услугите, които предоставят:

- Уеб сайтове;
- Интернет магазини;
- Социални мрежи;
- Блогове и форуми;
- Системи за управление на съдържание;
- Системи за споделяне на съдържание;
- Системи за публикуване на съдържание и др.

[25] Das, T., Kumar, P., BIG Data Analytics: A Framework for Unstructured Data Analysis, International Journal of Engineering and Technology (IJET), Vol 5 No 1 Feb-Mar 2013, India, ISSN : 0975-4024

[26] Kirilov, R., Approaches to Research the Costs and Benefits of Implementation of Information Technologies, Proceedings of the 9th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE – 2019), UNWE - Sofia, 2020, pp. 216 - 218, ISSN 2367-7635

[27] Eberendu, A., Unstructured Data: an overview of the data of Big Data, International Journal of Computer Trends and Technology (IJCTT), Volume 38 Number 1 - August 2016, ISSN: 2231-2803; Khan, N., Yaqoob, I., Hashem, I., Inayat, Z., Ali, W., Alam, M., Shiraz, M., Gani, A., Big Data: Survey, Technologies, Opportunities, and Challenges, The Scientific World Journal, 2014

В настоящото изследване ще се придържаме към разбирането, че всички системи, които предоставят потребителски изгледи, описани в текстов вид чрез езика за маркиране на хипертекст HTML, представляват уеб базирани информационни системи. В контекста на това, че всички системи функционират в някаква среда[28], уеб базираните информационни системи функционират в рамките на уеб концепцията. Уеб базираните информационни системи са изградени от множество от уеб страници, които се достъпват с помощта на уеб браузър посредством протокола HTTP и с помощта на комуникация от вида заявка – отговор[29]. По своята същност уеб базираните информационни системи притежават обичайните за всяка система характеристики[30]. Уеб базираните информационни системи съществуват в рамките на уеб концепцията и в този смисъл Интернет представлява тяхната заобикаляща среда. От технологична гледна точка уеб базираните информационни системи са софтуерни решения, изградени с помощта на уеб базирани технологии. Удовлетвореността на потребителите в Интернет от този вид софтуерни решения се превръща в един от най-важните показатели за успеха им от началото на дигиталната ера. За разлика от традиционните информационни системи, които са специално създадени за професионални потребители в организациите, уеб базираните информационни системи са предназначени за широката аудитория в Интернет с различни компютърни умения. Това включва и по-прякото потребление и взаимодействие с данните в уеб базираните информационни системи, в сравнение с традиционните информационни системи[31]. Характерна особеност за уеб базираните информационни системи е тяхната многослойна логическа архитектура, съставена обичайно от три слоя[32]:

- Потребителски интерфейс;
- Бизнес логика;
- Данни.

Потребителският интерфейс предоставя достъп до съдържанието и функционалността на уеб базираната информационна система в удобна за

[28] Gupta, H., Management Information System, INTERNATIONAL BOOK HOUSE PVT. LTD., 2011, ISBN 978-93-81335-05-5. Gupta, H., Management Information System, INTERNATIONAL BOOK HOUSE PVT. LTD., 2011, ISBN 978-93-81335-05-5

[29] https://www.sciencedirect.com/science/article/pii/B9780128040188000073

[30] Bagad, V., Management Information Systems, Technical Publications, 2008, ISBN 9788184313673

[31] Cheung, C., Lee, M., The Structure of Web-Based Information Systems Satisfaction: An Application of Confirmatory Factor Analysis, Springer, Berlin, Heidelberg, 2008, ISBN 978-3-540-68257-8

[32] https://www.researchgate.net/publication/4056556_A_three-tier_system_architecture_design_and_development_for_hurricane_occurrence_simulation

потребителя форма. В рамките на този слой обичайно се показват различни отчети и информация за управление[33].

Бизнес логиката включва набор от софтуерни модули, които извършват преобразуването на съответните данни на уеб базираната информационна система в полезна информация съгласно правилата на предметната област.

Данните репрезентират съответното съдържание на уеб базираната информационна система.

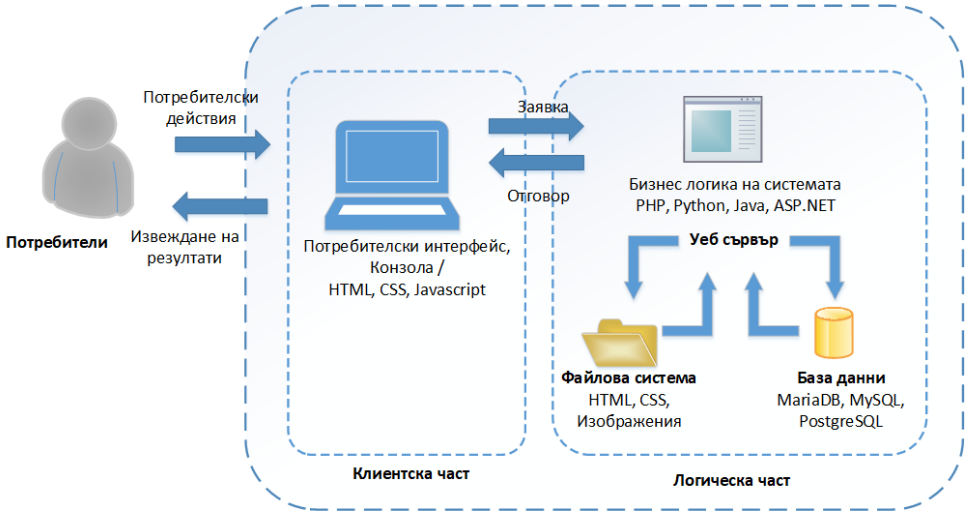
Според други източници в предметната област, уеб базираните информационни системи представляват системи, базирани на уеб технологии, които споделят различни видове Интернет протоколи и играят важна роля за събиране, съхраняване и предаване на данни от различни източници в рамките на уеб концепцията[34]. Този вид информационни системи споделят висока степен на формиране на общност и генериране на съдържание на ниво потребител[35]. Бизнес организациите използват уеб базирани информационни системи за подобряване на нивото на управленските процеси в организациите. Развитието на много бизнес сфери е пряко свързано в приложението на уеб базирани информационни системи, включително за целите на управлението на комуникация с други организации. В този смисъл уеб базираните информационни системи и съответните технологии за тяхното изграждане играят значима роля в развитието на бизнеса[36].

С навлизането на концепцията Уеб 2.0 все по-популярни и разпространени стават уеб базираните информационни системи, чрез които се формират онлайн общности от потребителите им. Класически пример за такъв вид системи са социалните мрежи. Уеб базираните информационни системи намират приложение и в различни организации от публичния сектори, като образователни институции, медийни компании, държавни институции и др.

Погледнато от технологичен аспект, уеб базираната информационна система може да се разглежда като софтуерно решение, което се използва за публикуване и поддържане на данни под формата на хипертекст. Според някои автори уеб базираната информационна система е комбинация от едно или повече софтуерни решения[37]. Изграждането на уеб базирани информационни системи обикновено е специализирано, за да приспособи начините, по които уеб браузърите взаимодействат с различни уеб базирани технологии. Тази променливост в уеб

[33] Kirilov, R., The challenge – systems with open business logic, Proceedings of the 5th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE – 2015), UNWE - Sofia, 2016, pp. 608 - 611, ISSN 2367-7635
[34] https://www.ukessays.com/essays/information-systems/web-based-information-systems.php
[35] https://udspace.udel.edu/handle/19716/28502
[36] https://www.ukessays.com/essays/information-systems/web-based-information-systems.php
[37] https://udspace.udel.edu/handle/19716/28502

браузърите, заедно с вероятността от използването им от потенциални потребители, обикновено се взема предвид при проектирането на съответните уеб базирани информационни системи[38]. Логическата архитектура на уеб базирана информационна система е представена на фиг. 1[39].



Фиг. 1 Логическа архитектура на уеб базирана информационна система

Логическата архитектурата на уеб базираната информационна система е съставена от две части – клиентска част и логическа част. Клиентската част включва уеб базирани технологии, които се използват за визуализация на потребителски интерфейси. Логическата част се състои от няколко основни компонента:

- Файлова система за съхранение на изпълними файлове, изображения, документи и др.;
- Система за управление на база от данни;
- Уеб сървър;
- Бизнес логика на системата.

В съвременното дигитално общество социалният живот се влияе значително от данните, които се откриват в различните видове уеб базираните информационни системи[40]:

- Онлайн общности – блогове, форуми, социални мрежи и др.

[38] https://www.sciencedirect.com/science/article/pii/B9780128040188000073
[39] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2638623/
[40] https://www.ukessays.com/essays/information-systems/web-based-information-systems.php

- Софтуерни решения в системата на образованието;
- Онлайн медии;
- Публични услуги в рамките на електронното управление и др.

Онлайн общността е виртуална общност, която съществува в резултат от наличието на Интернет и развитието на уеб концепцията. Онлайн общността обикновено съществува под формата на уеб базирана информационна система, където всеки може да публикува съдържание. Онлайн общността се използва от различни социални и професионални групи, взаимодействащи помежду си чрез уеб базирани технологии.

Уеб базираните информационни системи в сферата на образованието се използват от повечето образователни институции за комуникация с ученици и студенти. В рамките на тези системи могат да бъдат намерени учебни материали, графици на обучение, както и всякаква друга полезна информация чрез уеб базирана информационна система. Също така, следва да се отбележи, че с помощта на уеб базираните технологии съвременните изследователи имат достъп до множество проучвания и данни, налични в Интернет.

Предвид основното приложение на уеб базираните информационни системи за създаване, съхранение, манипулиране и разпространение на информация, онлайн медиите са обичаен пример за уеб базирани информационни системи. Чрез този вид системи може да се получават актуални новини от цял свят под формата на разнообразни уеб публикации от страна на съответните медии.

Развитието на уеб базираните информационни системи помага и на публичния сектор да подобрява онлайн услугите, които са предназначени за бизнеса и гражданите. Публичните услуги в рамките на електронното управление се предоставят най-вече с помощта на съответните уеб базирани информационни системи на електронното правителство.

Уеб базираните информационни системи в световен мащаб съхраняват данни по различни теми и направления. Представянето на тези данни е разнообразно и докато в рамките на една уеб базирана информационна система данните могат да имат някаква собствена за системата организация и структура, то в рамките на множество от уеб базирани информационни системи, всичките им данни взети заедно нямат единна структура и организация, поради това, че произлизат от различни информационни системи, които се различават по направление, технологии, организация и начин на представяне на данните. По тази причина, от гледна точка на множеството от уеб базирани информационни системи, може да се приеме, че данните му са неструктурирани. На същата база може да се каже, че всички уеб базирани информационни системи, достъпни в Интернет, могат да се

определят като най-голямото множество от системи с данни в неструктурирана форма.

Необходимостта от представянето на текстови данни от уеб базирани информационни системи в структурирана форма, с цел организирането им във формат, удачен за подлагането им на аналитична обработка, е от съществено значение за организациите в контекста на необходимостта за подпомагане на вземането на управленски решения в съответните организации. Интегрирането на данни, съхранени в структурирани и неструктурирани формати, може да добави допълнителна стойност към една организация. Аналитичната обработка на неструктурирани данни, в обединение с традиционните структурирани данни, осигурява по-голямо корпоративно прозрение за вземащите управленски решения[41]. Според някои автори, в съвременното дигитално общество, информацията се счита за най-важният ресурс на всяка организация[42]. По тази причина, за да се подобри цялостната дейност на организацията, е изключително важно процесите на вземане на управленски решения да се основават на навременна и точна информация. За да усъвършенстват своята дейност, организациите са принудени да събират, интерпретират и използват данните така, че да обезпечават осъществяването на оптимално управление[43]. Стремежът за реализация на инструменти за подпомагане на вземането на управленски решения води до появата на различни технологии и средства за бизнес анализ, познати под наименованието бизнес интелигентност. Системите, предназначени да подпомагат вземането на управленски решения чрез приложението на различни аналитични средства, се свързват с понятието бизнес интелигентни системи. Няма конкретна дефиниция за бизнес интелигентна система. Различни автори предлагат различни определения, в зависимост от избрания фокус на дефиниране – цели и предназначение, градивни компоненти, използвани технологии и др. Според някои автори бизнес интелигентните системи извършват събиране и съхранение на данни, както и управление на знания с помощта на аналитични средства, с цел да се представи сложна и важна за бизнеса информация, необходима за планиране и взимане на управленски решения[44]. Според други автори бизнес интелигентната система представлява архитектура и съвкупност от интегрирани приложения за обслужване на оперативната дейност и подпомагане процесите на взимане на управленски решения. Независимо от конкретните

[41] Orobor, I., Integration and Analysis of Unstructured Data for Decision Making: Text Analytics Approach. International Journal of Open Information Technologies, ISSN 2307-8162, 2016
[42] Стефанова, К., Фактори и насоки за проектиране и изграждане на бизнес интелигентни системи, Годишник на УНСС, 2008, ISSN 1312-5486
[43] Пак там
[44] https://digitalcommons.kennesaw.edu/facpubs/1370/

дефиниции, бизнес интелигентните системи се свързват с два важни въпроса, а именно[45]:

- Интегриране на данните (обединяване на данни от различни източници и с различни формати);
- Осигуряване на техники за анализ и визуализация на информацията.

Настоящото изследване обхваща проблематиката на първия въпрос, а именно зареждането на бизнес интелигентните системи с данни от различни източници и в различни формати или казано по друг начин представянето на текстови данни от уеб базирани информационни системи в структурирана форма, съобразена с изискванията на дадена бизнес интелигентна система. Трябва да се отбележи, че бизнес интелигентните системи следва да включват всички източници на данни, използвани за бизнес решения, които могат да бъдат[46]:

- Операционни системи;
- Ключови електронни таблици;
- Ведомствени данни, включително и уеб страници.

Източниците на данни за една система могат да бъдат вътрешни (в рамките на организацията) и външни. В настоящия дисертационен труд всички източници от Интернет, който се разглежда като едно огромно множество от източници на неструктурирани данни, които не принадлежат на дадена организация, се възприемат като външни източници от гледна точка на организацията. В този смисъл, за дадена организация, всички източници на данни от уеб базирани информационни системи, които не са част от информационната инфраструктура на съответната организация, могат да се възприемат като външни източници. Един от факторите за безпроблемното функциониране на бизнес интелигентните системи е качеството на данните, което е фактор по принцип за управлението на всяка информационна система. От техническа гледна точка качеството на данните се определя основно от възможностите за приемане, съхранение, защита, предоставяне и извличане на данните[47].

В контекста на значимостта на данните от външни източници за бизнес интелигентите системи при осъществяване на аналитичност на висше управленско

[45] Стефанова, К., Кабакчиева, Д., Бизнес интелигентни системи, ИЗДАТЕЛСКИ КОМПЛЕКС - УНСС, София, 2015
[46] Stefanova, K., Kabakchieva, D., Management Aspects of the Business Intelligent Systems Development, Economic Alternatives, issue 1, 2012
[47] Кисимов, В., Управление на качеството на бизнес данните в ИС, Икономически алтернативи, брой 6, 2008

ниво[48] и предвид тясната обвързаност на структурирането с извличането и предоставянето на данни от уеб базирани информационни системи, може да се обобщи, че проблематиката за представянето на текстови данни от уеб базирани информационни системи в структурирана форма е от изключителна важност за бизнес интелигентните системи. В този смисъл следва да се допусне, че при нарастването на текстовите данни в уеб базираните информационни системи, предизвикателствата в контекста на тази проблематика ще стават все повече. В литературата са формулирани различни подходи в контекста на възможностите за представяне на текстови данни в структурирана форма, а именно[49]:

- Обработка на естествения език;
- Разпознаване на реч;
- Машинен превод;
- Генериране на текст;
- Откриване на знания в текст.

В настоящия дисертационен труд ще се обърне внимание на два от подходите, а именно обработка на естествения език и откриване на знания в текст. Според някои автори това са двата основни подхода в контекста на възможностите за представяне на текстови данни в структурирана форма, които се използват в практиката[50].

Обработката на естествения език обхваща различни техники, които се базират на езикови особености, като даден текст се анализира синтактично с помощта на приложение на правила от формална граматика, а резултатните данни се интерпретират семантично, за да се извлече информация от тях. Обработката на естествения език може да се прилага задълбочено, като се анализира всяка част от всяко изречение и се прави опит за интерпретация или да се прилага повърхностно чрез ограничен семантичен анализ на само някои пасажи или фрази в изречения. Обработката на естествения език позволява да се прилагат техники за разграничаване на определени думи или части от думи. Този подход се използва от голяма част от съвременните информационни системи, предназначени за автоматичен превод на текст.

При подхода за откриване на знания в текст се използват основно аналитични техники, свързани с извличането на информация и машинното

[48] Стефанова, К., Кабакчиева, Д., Бизнес интелигентни системи, ИЗДАТЕЛСКИ КОМПЛЕКС - УНСС, София, 2015
[49] Kao, A. and Poteet, S., Text Mining and Natural Language Processing – Introduction for the Special Issue, SIGKDD Explorations, Volume 7, Issue 1, 2005
[50] Das, T., Kumar, P., BIG Data Analytics: A Framework for Unstructured Data Analysis, International Journal of Engineering and Technology (IJET), Vol 5 No 1 Feb-Mar 2013, India, ISSN : 0975-4024

самообучение. Целта на този подход е да извлича модели за обработка на голям брой текстови документи. Този подход обичайно се използва от съвременните търсещи машини. В подхода за откриване на знания в текст се включват различни дейности, а именно:

- Автоматична класификация на текст съгласно определен набор от категории;
- Групиране на текстове по общи признаци;
- Автоматично обобщаване;
- Извличане на теми от текстове и анализ на тематични тенденции в текстови потоци.

Подходът за откриване на знания в текст се използва в практиката от бизнес интелигентните системи с различно предназначение[51].

Логическата архитектурата на една бизнес интелигентна система може да се разглежда като съвкупност от две среди, а именно:

- Среда на данните;
- Аналитична среда.

Средата на данните включва съответните източници на данни, които могат да бъдат вътрешни и външни от гледна точка на организацията, ETL процес и склад от данни[52]. От своя страна ETL процесът включва извличане на данните от източниците, трансформиране на извлечените данни в подходяща форма и зареждането им в склада от данни на съответна бизнес интелигентна система. В рамките на логическата архитектура на бизнес интелигента система, която включва данни от уеб базирани информационни системи, извличането на данни от уеб базираните информационни системи се явява като част от ETL процеса. В този контекст извличането на данни от този вид системи преминава през един междинен слой, наречен в конкретния случай уеб услуга. Въпросната уеб услуга включва извличане на данни от уеб базирани информационни системи, някаква форма на временно съхранение на тази данни, както и възможност за предоставянето им по зададен критерий, посредством наложения като стандарт протокол за комуникация HTTP, в някой от популярните съвременни формати за обмен на данни HTML, XML, JSON и др. Има и вариант, при който ETL процесът се осъществява с директно извличане на данните от уеб базираните информационни системи без преминаването през междинен слой. Независимо от това дали една бизнес интелигентна система ползва или не ползва междинен слой за извличането на данни

[51] Stefanova, K., Yordanova, S., Knowledge Discovery from Unstructured Data using Sentiment Analysis, Economical and Social Alternatives, Issue 1, 2017, ISSN 1314–6556

[52] Стефанова, К., Кабакчиева, Д., Бизнес интелигентни системи, ИЗДАТЕЛСКИ КОМПЛЕКС - УНСС, София, 2015

от уеб базирани информационни системи, процесът на извличане на данни от този вид системи може да се реализира по модела на търсещите машини.

Проблемът с анализа на текст придобива нарастващо внимание през последните години поради големия обем текстови данни, които се създават в разнообразието от уеб базирани информационни системи. Предвид обичайната неструктурирана форма на данните в този вид системи съществува необходимост от приложение на подходи за представяне на текстовите данни от уеб базираните информационни системи в структурирана форма. Представянето на текстовите данни от уеб базирани информационни системи в структурирана форма предполага предшестващо извличане на текстовите данни от съответните системи в тяхната обичайна неструктурирана форма. Извличането на текстови данни от уеб базирани информационни системи е област, в която през последните години се наблюдава бързо развитие[53], поради огромния напредък в хардуерните и софтуерните технологии. Технологичният напредък е съпътстван със създаването и приложението на различни типове данни. Това важи особено за текстовите данни, които в резултат от развитието на уеб базираните информационни системи нараснаха с много сериозни темпове. Развитието на уеб концепцията като цяло е технологичен фактор, който насърчава създаването на голямо количество текстово съдържание от различни потребители във форма, която е сложна за съхранение и обработка. Нарастващите количества текстови данни в уеб базираните информационни системи създават необходимост от приложение на съответните подходи за представяне на текстовите данни от този вид системи в структурирана форма. Докато структурираните данни обикновено се управляват с помощта на система за управление на бази от данни, то текстовите данни в неструктурирана форма обикновено се управляват чрез търсеща машина поради липсата на структурираност[54]. Търсещата машина позволява на потребителя да намери полезна информация в колекцията от неструктурирани текстови данни. Подобряването на начина на работа на търсещата машина е съществена изследователска проблематика в областта на откриването на информация[55], тъй като много свързани с търсещите машини въпроси, като групиране на текст, категоризиране на текст, обобщаване на текст и препоръчване на текст, също са обект на изследване[56]. С бързото развитие

[53] Aggarwal, C., Zhai, C., Mining Text Data, Springer, 2012, ISBN 978-1-4614-3222-7
[54] Пак там
[55] Baeza-Yates, R., Ribeiro-Neto, B., Modern Information Retrieval - the concepts and technology behind search, Second edition, Pearson Education Ltd., 2011, ISBN: 978-0321416919; Jones, K., Willett, P., Readings in information retrieval, Morgan Kaufmann Publishers Inc, 1997, ISBN:1-55860-454-5
[56] Grossman A., Frieder, O., Information Retrieval: Algorithms and Heuristics, Springer, 2004, ISBN: 978-8181289179; Manning, C., Raghavan, P., Schutze, H., Introduction to Information Retrieval, Cambridge University Press, 2008, ISBN: 978-0521865715

на уеб концепцията, търсещите машини се превръщат в неразделна част от ежедневния живот на интернет потребителите. Търсещите машини, които работят със структурирани данни, могат да предоставят на потребителите по-добри резултати спрямо търсещите машини, които работят с неструктурирани данни[57].

След навлизането на концепцията Уеб 2.0 Интернет може да се разглежда като хранилище на обединена информация, което се поддържа от набор от съгласувани протоколи и формати за данни, сред които:

- Протокола за управление на обмена на информация (Transmission Control Protocol, TCP);
- Системата за имена на домейните (Domain Name System, DNS);
- Протокола за пренос на хипертекст (Hypertext Transfer Protocol, HTTP);
- Езика за маркиране на хипертекст (Hypertext Markup Language, HTML);
- Протокола за изключване на уеб роботи и др.

Наличието на тези протоколи и формати прави възможно извличането на текстови данни от уеб базираните информационни системи. С напредъка на интернет технологиите се разработват и развиват агрегатори на съдържание. Агрегаторите на съдържание представляват системи за събиране и обединяване на данни от уеб базирани информационни системи. В този смисъл, към агрегаторите на съдържание, спадат и търсещите машини. За усъвършенстване на възможностите на агрегаторите на съдържание по отношение на извличането на данни от уеб базираните информационни системи има два основни подхода[58]:

- Агрегаторите на съдържание могат да възприемат модел на извличане на данни от уеб базирани информационни системи, при който те активно да претърсват уеб базираните информационни системи в Интернет за нови или актуализирани данни;
- Агрегаторите на съдържание могат да използват създадена конвенция и набор от протоколи, които позволяват на уеб базираните информационни системи, в ролята им на доставчици на интернет съдържание, да направят лесно достъпни своите данни, които представляват интерес за агрегаторите на съдържание.

По своята същност данните в уеб базираните информационни системи са част от документи, написани на HTML език. Основното предназначение на този език е да предложи структурирано описание на текстови данни, независимо от

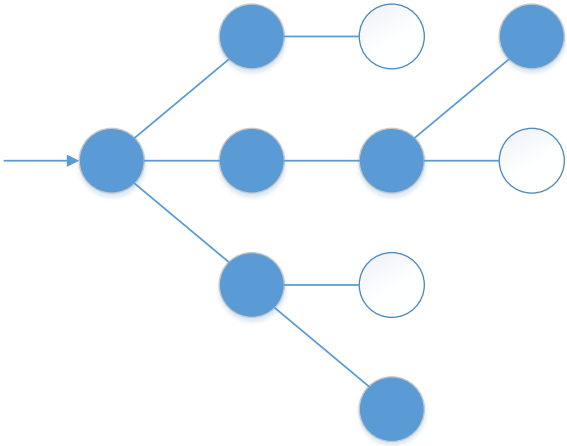
[57] Aggarwal, C., Zhai, C., Mining Text Data, Springer, 2012, ISBN 978-1-4614-3222-7
[58] http://i.stanford.edu/~olston/publications/crawling_survey.pdf

начина, по който те се визуализират в рамките на потребителския интерфейс на съответната уеб базирана информационна система[59]. Подобно на електронните съобщения, отделните уеб страници, които изграждат дадена уеб базирана информационна система, могат да се разглеждат като неструктурирани данни с добавена посредством HTML езика структура. Неструктурираните източници на данни са много големи по обем. Някои автори изследват предизвикателства при работа с тях[60]:

- Извличане и получаване на полезна от гледна точка на ползвателя информация;
- Преобразуване на данните в знание;
- Анализ с цел откриване на модели и тенденции;
- Съхраняване, което да осигурява бърз и ефективен достъп;
- Съставяне на полезни бизнес интелигентни доклади и др.

В уеб базираните информационни системи се съдържа голям обем от данни, които се отнасят до различни теми. Тези теми могат да представляват интерес както за бизнес организациите, така и за организациите от публичния сектор. За разлика от традиционните хранилища на данни, в Интернет и по-специално в рамките на концепцията Уеб няма централно организирано хранилище на съдържанието. Обичаен начин за извличане на текстови данни от уеб базираните информационни системи представлява приложението на уеб робот[61]. Уеб роботът представлява технологична реализация под формата на софтуерно решение на концепцията за автоматично изтегляне на съдържанието на уеб базирани информационни системи. Уеб роботът много често се нарича и уеб паяк. Начинът на работа на този вид софтуерни решения е познат в научната литература като уеб обхождане. Уеб обхождането представлява концепция за събиране на данни от уеб базирани информационни системи. Тази концепция може да бъде описана като задача за търсене в граф, тъй като уеб базираните информационни системи в рамките на уеб концепцията могат да се разглеждат като голям граф, където върховете на графа представляват отделните уеб страници, които изграждат уеб базираните информационни системи, а ребрата на графа репрезентират хипервръзки към други уеб страници. Обхождането на уеб страници от уеб робот е графично представено на фиг. 2.

[59] White, B., HTML and the Art of Authoring for the World Wide Web, Springer, 1996, ISBN 978-0792396918
[60] Orobor, I., Integration and Analysis of Unstructured Data for Decision Making: Text Analytics Approach.International Journal of Open Information Technologies, ISSN 2307-8162, 2016
[61] Ahuja, M., Singh Bal, J. and Varnica, Web Crawler: Extracting the Web Data. International Journal of Computer Trends and Technology (IJCTT), volume 13, 2014



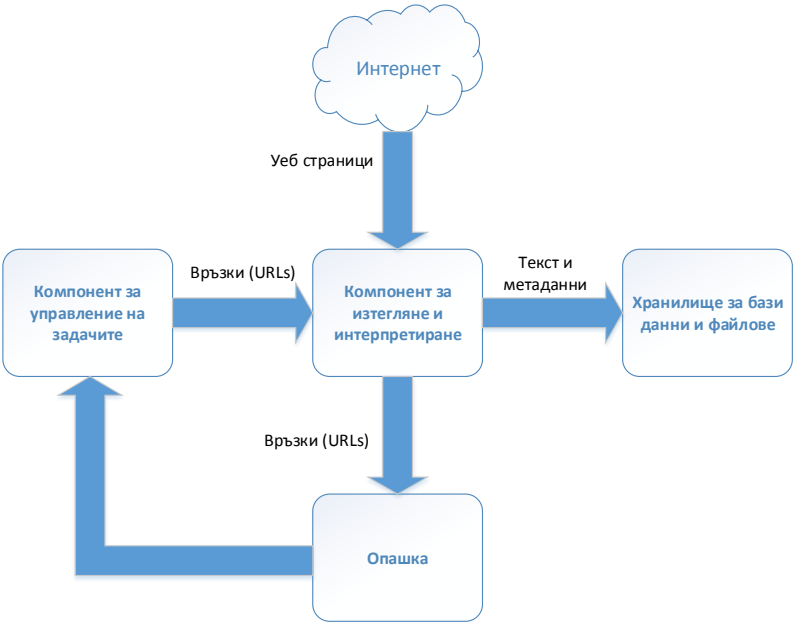
Фиг. 2 Обхождане на уеб страници от уеб робот

Фигурата илюстрира преминаването на уеб робот през произволни уеб страници, които са означени с тъмни и бели кръгове. Кръговете са свързани помежду си и формират граф с разклонения. С тъмните кръгове са означени уеб страници, които съдържат в себе си връзки (под формата на хиперлинкове) към други необходени страници. С белите кръгове са представени уеб страници, които или не притежават никакви връзки или връзките, които притежават са вече обходени. По тази причина белите кръгове са крайната точка на обхождане на дадено разклонение. Обхождането на уеб робот би могло да продължи докато се премине през всички възможни разклонения до прочитането и индексирането на всички страници в рамките на даден обхват, който може да бъде уеб сайт, домейн, компютърна мрежа и др.

Начинът на работа на уеб робот е представен на фиг. 3, където са изобразени основните компоненти, които изпълняват процеса на работа на уеб робот. Самият процес протича в следните стъпки:

- Проверява се страница за изтегляне по даден уеб адрес, като уеб адресите се съхраняват в опашка от адреси;
- Компонента за управление на задачите проверява дали за дадената страница е разрешено да бъде изтеглена – проверява се конфигурационен файл за изключване на роботи;
- Компонента за изтегляне и интерпретиране изтегля цялата страница;

- Извличат се всички връзки от страницата (адреси към уеб страници) и се добавят към опашката от адреси, за да бъдат изтеглени на следващ етап;
- Извличат се всички думи и се записват в база данни, като създадените записи се асоциират с тази страница;
- Преминава се към следващия уеб адрес от опашката и за него се изпълнява целият процес отначало.



Фиг. 3 Начин на работа на уеб робот

Уеб роботите могат да имат различно приложение, но обичайно целта им е да извличат данни от уеб базирани информационни системи. Уеб роботите изпращат заявки за достъп до данни в уеб базирани информационни системи, местонахожденията на които са предварително зададени. Съществуват различни видове уеб роботи в контекста на съответния начин, по който уеб базираните информационни системи се обхождат. Най-популярните уеб роботи са следните[62]:

- Избирателни уеб роботи;

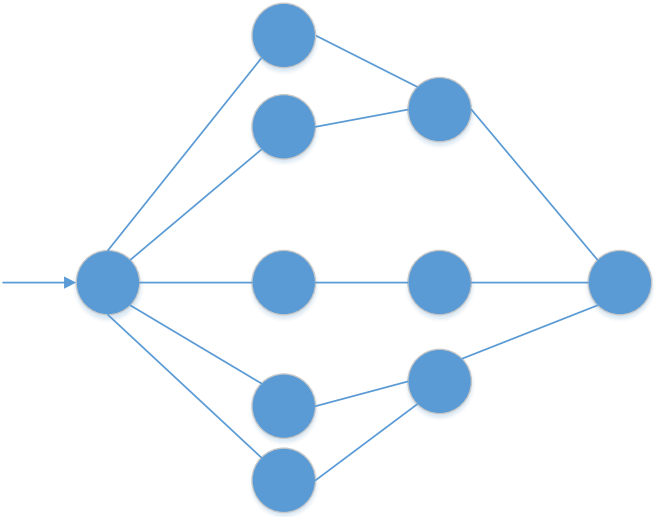
[62] Ahuja, M., Singh Bal, J. and Varnica, Web Crawler: Extracting the Web Data. International Journal of Computer Trends and Technology (IJCTT), volume 13, 2014

- Инкрементални уеб роботи;
- Фокусирани уеб роботи, известни още като целеви уеб роботи;
- Уеб роботи, ориентирани към формуляри;
- Уеб роботи, ориентирани към скрити страници;
- Паралелни уеб роботи;
- Разпределени уеб роботи.

Има случаи, в които не е необходимо уеб роботът да обхваща всички уеб базирани информационни системи. Понякога е необходимо да се фокусира върху конкретни теми, които представляват интерес само за конкретна потребителска група. В този контекст подходящ пример е такъв уеб робот, който извлича данни от специализирани уеб базирани информационни системи за конкретна тематика, като в същото време съответният уеб робот е програмиран да пропуска уеб базирани информационни системи със съдържание, което е специализирано в други тематики. Фокусираният уеб робот може да бъде дефиниран като уеб робот, който извлича данни от уеб базирани информационни системи, свързани само с определени теми. Броят на уеб базираните информационни системи в конкретен домейн може да е достатъчно малък, така че фокусираният уеб робот да може извлече данните от тях в относително кратък период от време[63]. Фокусираният уеб робот представлява софтуерно решение, което извлича данни от уеб базирани информационни системи, отговарящи на конкретни критерии, като внимателно извършва приоритизиране на границите за обхождане. Граница на обхождане представлява връзката в уеб страница, която може да бъде избрана от фокусирания уеб робот в рамките на съответното обхождане. Някои фокусирани уеб роботи могат да бъдат предварително програмирани да обхождат само уеб страници от домейни с конкретно разширение, а други – уеб страници с висок рейтинг на база на някакви предварително дефинирани критерии.

На фиг. 4 е представена схема за обхождане на уеб базирана информационна система от фокусиран уеб робот, като с тъмни кръгове са изобразени уеб страниците на уеб базираната информационна система, които са свързани по между си с хиперлинкове и формират граф. Критерий на обхождане за фокусирания уеб робот е да се преминава само през уеб страниците, които са част от системата и не се отнасят за страници извън нея, които се явяват уеб страници на външни уеб базирани информационни системи. По тази причина обхождането се осъществява в ограничен обхват, започва от начална точка, преминава през възможните разклонения и приключва в крайна точка.

[63] Sampat, J., Jain, A. and Mistry, D., Focused Web Crawler and its Approaches. International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 – 5161



Фиг. 4 Обхождане на уеб страници от фокусиран уеб робот

Предвид спецификата на фокусираните уеб роботи, могат да бъдат дефинирани следните предимства на фокусираните уеб роботи спрямо традиционните уеб роботи за обхождане на уеб страници:

- Извличане на данни от уеб базирани информационни системи на база на домейни, приложения и заявки;
- По-голяма уместност на обхожданите уеб страници поради предварително зададените критерии за обхождане;
- Висока производителност на обхождане на уеб страниците.

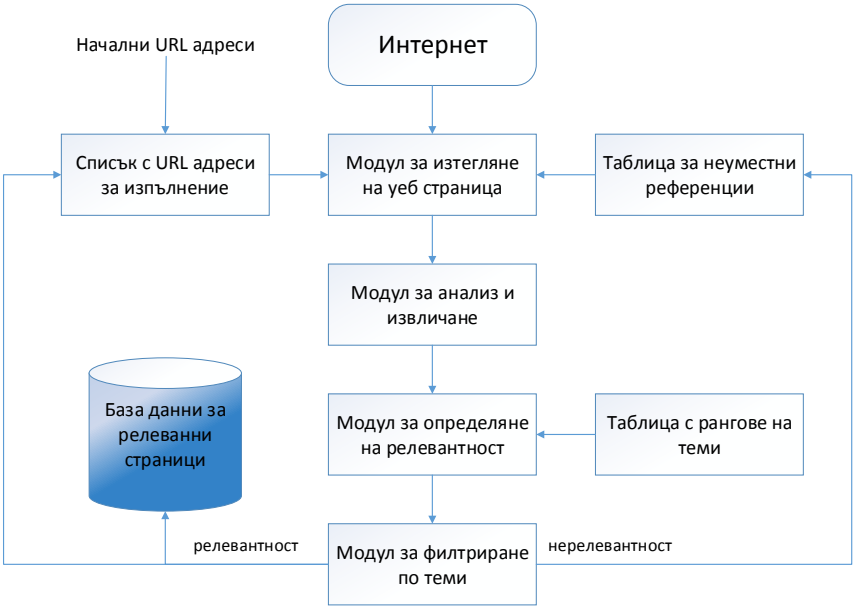
Фокусираните уеб роботи се характеризират и с някои технологични ограничения спрямо традиционните уеб роботи за обхождане на уеб страници:

- Необходимост от повече хардуерни и софтуерни ресурси;
- Зависимост на производителността на фокусирания уеб робот от количеството хипервръзки в рамките на конкретния домейн.

За целите на настоящото изследване интерес представлява фокусираният уеб робот, защото той функционира по определени, предварително зададени критерии, а начинът му на обхождане на уеб базирани информационни системи е избирателен. В този контекст, за извличането на данни от уеб базирани информационни системи, избирателният подход за обхождане на уеб страници е подходящ, защото при него могат да се систематизират определени групи от уеб страници и за всяка група биха могли да се приложат подходи за представяне на

текстовите данни от съответните уеб базирани информационни системи в структурирана форма.

На фиг. 5 са представени компонентите на фокусирания уеб робот[64].



Фиг. 5 Компоненти на фокусиран уеб робот

В опашката от интернет адреси се съдържат извлечените уеб адреси, които се предвиждат за обхождане. Тези адреси се зареждат в модул за изтегляне на уеб страници. След като се извърши това зареждане, модул за анализ и извличане разпознава определени фрагменти, като текстови области и хипервръзки от изтеглените страници. Модулът за определяне на релевантността изчислява съответствието на страницата по отношение на темата и присвоява резултата към същите уеб адреси, извлечени от страницата. Модул за филтриране по теми определя дали съдържанието на анализираните страници е свързано с темата или не. Ако страницата е тематично свързана, извлечените от нея уеб адреси се добавят към списък с URL адреси за изпълнение. В противен случай се добавят към таблица, съхраняваща всички адреси, които се отнасят до уеб страници, които не отговарят на критериите на темата. В контекста на представените компоненти и тяхното

[64] Pranav, A., Chauhan, S., Efficient Focused Web Crawling Approach for Search Engine, International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 5, May 2015, pg.545 – 551, ISSN 2320–088X

взаимодействие, от технологична гледна точка съществуват няколко разновидности на фокусиран уеб робот[65]:

- Фокусиран уеб робот с приоритет;
- Фокусиран уеб робот, базиран на структурата;
- Фокусиран уеб робот, основан на контекст;
- Фокусиран уеб робот, основан на обучение.

При фокусирания уеб робот с приоритет, уеб страницата, която съответства на URL адреса, се изтегля от интернет и се изчислява относителният резултат на уеб страницата за изтегляне с дума за фокус. При тази технология уеб адресът от страницата се съхранява в приоритетната опашка вместо в опашката по подразбиране. По този начин всеки път уеб роботът взима от опашката уеб адрес, който е с най-високия приоритет.

При фокусирания уеб робот, базиран на структурата, основно се разчита на оценка на релевантността на уеб страницата с темата. В технологичен аспект съществуват няколко възможности. Първата възможност е да се изтеглят само онези уеб страници, чийто резултат е висок. Оценката на уеб страниците се изчислява на база на приложимостта на родителските уеб страници на определената връзка. Втората възможност е да се приложи комбиниране на фокусирано обхождане на базата на съдържание и връзки. В този случай, от една страна се използва информацията за съдържанието на уеб страницата, за да определи дали страницата е подходяща за темата, а от друга страна се оценява референтната информация за уеб страницата[66].

При фокусирания уеб робот, основан на контекст, се анализа контекстът на уеб страницата. Този подход има концептуални предимства спрямо подходите на предходните две разновидности на фокусиран уеб робот, защото при тях обхождането се възприема като „черна кутия“, в резултат на което потребителският контекст не се взема предвид.

При фокусирания уеб робот, основан на обучение, се извършва обучение на уеб робота в няколко аспект, а именно:

- Съответствие на URL адреса с темата;
- Съответствие на текста на уеб страницата с темата;
- Съответствие на родителската страница с темата;
- Съответствие на заобикалящите страници с темата.

[65] Pranav, A., Chauhan, S., Efficient Focused Web Crawling Approach for Search Engine, International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 5, May 2015, pg.545 – 551, ISSN 2320–088X
[66] Пак там

Уеб роботите имат пряко отношение към логическата архитектура на търсещите машини в ролята им на софтуерни решения за извличане на данни от уеб базирани информационни системи. Логическата архитектура представлява фундаменталната организация на една информационна система, нейните компоненти, отношенията между тях и обкръжаващата среда, както и принципите на управление и развитие[67]. Логическата архитектурата се използва за описание на дадена информационна система на определено ниво на абстракция[68]. Описанието на логическата архитектура на информационна система се състои в дефиниране на нейните базови компоненти, връзките помежду им и обкръжаващата ги среда, както и в посочване на целите и задачите, които всеки един компонент реализира. Описанията на логическата архитектура се представят в графичен, словесен или смесен вид и се наричат модели на логическата архитектура[69]. В научната литература са дефинирани различни видове логически архитектури:

- Хардуерна архитектура;
- Софтуерна архитектура;
- Корпоративна архитектура;
- Архитектура на производствена система и др.

Софтуерната архитектура основно се състои от софтуерни компоненти, интерфейсите, предоставени от тези компоненти и отношенията между тях. Логическата софтуерна архитектура на софтуерно решение, което по своята същност представлява търсеща машина, следва да се проектира така, че от една страна софтуерното решение да работи с възможно най-подходящият набор от данни от уеб базирани информационни системи, а от друга страна софтуерното решение да обработва заявките за търсене в извлечените данни от страна на потребителите възможно най-бързо. При работа с текстови данни, търсещите машини използват специализирани структури от текстови данни, които са оптимизирани за бърз достъп до съответните данни. За получаване на резултати, търсещите машини обработват текста и съхраняват текстови статистики, които помагат за подобряване на уместността на резултатите. Търсещите машини поддържат две основни функционалности, а именно:

- Индексиране на извлечени от уеб базирани информационни системи данни;

[67] Мурджева, А. и колектив, Информатика, Издателски комплекс - УНСС, София, 2018, ISBN 978-619-232-123-9
[68] Croft, W., Metzler, D., Strohman, T., Search Engines Information Retrieval in Practice, Pearson Education, Inc., 2011, ISBN 9780133001594
[69] Мурджева, А. и колектив, Информатика, Издателски комплекс - УНСС, София, 2018, ISBN 978-619-232-123-9

- Обработване на заявки за търсене в извлечените данни.

Основните градивни елементи за реализиране на индексиране на извлечени от уеб базирани информационни системи данни представляват компоненти за:

- Извличане на текст;
- Преобразуване на текст;
- Създаване на индекс.

Задачата на компонента за извличане на текст е да идентифицира и направи достъпни документите, които ще бъдат търсени. Под документите може да се разбира всякакви електронни материали, съхраняващи текстови данни – електронни писма, уеб страници, статии и др. От една страна компонентът за извличане на текст изпълнява уеб обхождане и извличане на документи в рамките на предвидения за това обхват – от обикновен настолен компютър до областта на целия Интернет. От друга страна този компонент създава хранилище за документи, където се съхраняват всички извлечени материали. Извлечените документи се подават към компонента за преобразуване на текст. Компонентът за преобразуване на текст трансформира документите в индексни характеристики. Индексните характеристики представляват части от документа, които се съхраняват в индекси и се използват при търсене. Най–елементарната единица за такава характеристика е дума от текст. Индексите могат да съхраняват фрази, имена на хора, дати и др. Колекция, съставена от индекси, се нарича индексен речник. Компонентът за създаване на индекс приема резултатите от компонента за преобразуване на текст и създава индекси или структури от данни, които позволяват търсене в тези индекси.

Основните градивни елементи на функционалността за обработване на заявки за търсене, представляват компоненти за:

- Взаимодействие с потребителя;
- Класиране;
- Оценяване.

Компонентът за взаимодействие с потребителя осигурява потребителския интерфейс между потребителя и търсещата машина. Една задача на този компонент е приемането на потребителската заявка и превръщането ѝ в индексни характеристики. Друга задача е да използва подредения списък с документи, върнат от търсещата машина и да го организира в резултати, които да се показват на потребителя. Компонентът за класиране е ядрото на търсачката. Използва трансформираната заявка от компонента за взаимодействие с потребителя и генерира подреден списък. Елементите в списъка са класирани по определен критерий. Задачата на компонента за оценка е да регистрира и анализира поведението на потребителите с помощта на регистрационни данни. Резултатите от

анализа се използват за настройка и подобряване на класирането на резултатите от търсещата машина.

Извършените проучвания, свързани с проблематиката на настоящото изследване, показват, че съществуват разработени софтуерни решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма. Проучването на характеристиките на някои от най-популярните софтуерни решения в предметната област е от съществено значение от гледна точка на дефинирания в дисертационния труд изследователски проблем. След като се разгледат съответните характеристики на съществуващите софтуерни решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма, може да се определи дали изследователският проблем на дисертационния труд може да бъде решен с някое от наличните софтуерни решения или следва да се разработва собствен подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма и софтуерен инструментариум, базиран на разработения подход.

Софтуерните решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма прилагат различни подходи за постигане на целите си. От технологична гледна точка съответните подходи имат много общи черти. Проучването на характеристиките на най-популярните софтуерни решения в тази област може да се използва като основание за обобщение на характеристиките на този вид софтуерни решения като цяло. Предварително очакване на проучването е този вид софтуерни решения да имат минималистичен потребителски интерфейс за конфигуриране на извличането на данни от съответните уеб базирани информационни системи. Целта на използването на специализиран инструментариум е да бъдат извлечени от уеб базираните информационни системи само някои части от съответните уеб страници, които представляват интерес от гледна точка на данните, които съдържат.

Според някои от проучванията в предметната област, най-популярните софтуерни решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма са[70]:

- Web Scrapper;
- Import.io;
- Octoparse;
- Scrappy.

[70] https://www.dev.to/justinhellderson/best-web-scraping-tools-in-2020-3maa/

Избраните за проучването софтуерни решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма са представени в таблица 1.

Таблица 1 Избрани за проучването софтуерни решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма

Софтуерно решение	**Уеб адрес**
Import.io	https://www.import.io/
Octoparse	https://www.octoparse.com/
Scrappy	https://www.scrapy.org/
Web Scrapper	https://webscraper.io/

Web Scraper извлича данни от уеб базирани информационни системи с множество нива на навигация. Той може да навигира в дадена уеб базирана информационна система на всички нива[71]. Това софтуерно решение дава възможност да се приспособи извличането на данни от уеб базирани информационни системи към различни структури на съответната уеб страница, като се използват конфигурации. След като процесът на представяне на текстови данни от уеб базирани информационни системи в структурирана форма приключи, данните могат да бъдат експортирани в различни файлови формати. Функционалността на това софтуерно решение позволява избирането на различни елементи от въпросния потребителски интерфейс като изображения, параграфи от текст, таблици, числови стойности и др. Всички тези елементи се запаметяват в шаблон, на базата на който Web Scrapper извлича данни от всички уеб базирани интерфейси с подобна организация на елементите.

Import.io извлича данни от уеб базирани информационни системи, както и от документи и различни формати на данни. Той заснема екранни снимки по време на извличане на данни от уеб базирани информационни системи, изтегля файлове, изображения и изходния HTML сорс код за допълнителни обработки на по-късен етап. Чрез уеб базиран потребителски интерфейс инструментът предоставя възможност за зареждане на уеб страници от зададени уеб адреси. Задават се настройки за елементите с данни, които следва да бъдат извлечени и по този начин се формира шаблон, по който да се извличат данни от подобни уеб страници. Има възможност и за задаване на график за изпълнение на процесите на извличане на

[71] Web Scrapper, https://webscraper.io/

данни от уеб базирани информационни системи[72]. Import.io предоставя функционалности за зареждане на уеб страници от достъпни през Интернет уеб базирани информационни системи. Тези функционалности включват преминаване през различните уеб страници стъпка по стъпка, тяхното визуализиране и възможност да се избират различни елементи от всяка страница. По този начин се предоставя възможност да се създават шаблони за извличане на данните от избраните елементи. Тези шаблони се използват при обхождане на уеб страниците за представянето на текстовото съдържание от тези страници в структурирана форма.

Octoparse е софтуерно решение, което извлича данни от уеб базирани информационни системи в три стъпки[73], а именно:

- Идентифициране на уеб страницата, от която ще бъдат извлечени данни;
- Избор на целевите данни за извличане;
- Стартиране на процеса на извличане на данните.

Разширеният конфигурационен режим на този инструмент дава възможност за изпълнение на задачи по график, включително циклично изпълнение на съответните задачи за извличане на данни от уеб базирани информационни системи.

Scrapy е софтуерно решение с отворен код за извличане на данни от уеб базирани информационни системи[74]. Логическата архитектура на това софтуерно решение включва потока от данни, който съдържа първоначални заявки, планиране на заявки, извличане на данни, обработка на отговори и т.н. Софтуерното решение включва в себе си програмна библиотека за разработване на други софтуерни решения за извличане на данни от уеб базирани информационни системи. Всеки компонент от логическата архитектура е представен в библиотеката програмно. Това позволява да се програмира изцяло ново софтуерно решение, което да следва модела на взаимодействие на компонентите по логическата архитектура, като има възможност да се зададат програмно специфични настройки, съгласно поставените изисквания и цели за извличане на данни от уеб базирани информационни системи.

Проучените четири софтуерни решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма представляват интерес за изследователите в областта на извличане на данни от уеб базирани информационни системи. Софтуерните решения в предметната област имат своите различия, но притежават сходна бизнес логика. Представянето на текстови данни от уеб базирани информационни системи

[72] Import.io, https://www.import.io/
[73] Octoparse, https://www.octoparse.com/
[74] Scrappy, https://www.scrapy.org/

в структурирана форма е многостранна задача, но изясняването на бизнес логиката на този вид софтуерни решения е от съществено значение. Извършеното проучване показва, че някои от установените общи характеристики на софтуерните решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма са:

- Възможност за конфигуриране на отделните елементи на уеб страници с помощта на графичен, най-често уеб базиран, потребителски интерфейс;
- Възможност за експортиране на текстови данни в популярни формати, като json, csv, xls и др.;
- Възможност за симулиране на реално поведение на потребителя;
- Възможност за експортиране на данни, които не са текстови данни, от уеб страници, като снимки и видеоклипове.

Извършеното проучване на характеристиките на софтуерните решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма не показва наличието на цялостни решения по поставената в дисертационния труд изследователска проблематика. В този смисъл, за целите на настоящото изследване, следва да се предложи собствен подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

Глава Втора. Представяне на текстови данни от уеб базирани информационни системи в структурирана форма

За да бъде възможно да се проектира и изгради подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, е необходимо да бъдат направени няколко предварителни допускания, които имат характер на ограничителни условия по отношение на самия подход.

Първо: Целият процес по анализ на текстови данни от уеб базирани информационни системи може да бъде разделен на три условни части: извличане на текстовите данни от уеб базираните информационни системи; представяне на вече извлечените текстови данни в структурирана форма и аналитична обработка на представените в структурирана форма текстови данни. Тези три части са представени графично на фиг. 6.

Фиг. 6 Извличане, представяне в структурирана форма и аналитична обработка на текстови данни от уеб базирани информационни системи

Второ: Подходът за представяне на текстови данни от уеб базирани информационни системи в структурирана форма цели установяване на съответни методи и технологичен инструментариум само по отношение на представянето на текстовите данни в структурирана форма. Това означава, че приложението на подхода предполага допускането, че текстовите данните от уеб базираните информационни системи са вече извлечени и се намират във временно хранилище.

Трето: Подходът не предполага включването на методи, свързани с аналитична обработка на вече представените в структурирана форма текстови данни. Аналитичната обработка е обект на допълнителни разработки извън обхвата на настоящия дисертационен труд. По своята същност аналитичната обработка

може да включва извършването на контекстен анализ, анализ на конкретни фрагменти от текста и др.

Тези три допускания следва да се възприемат като ограничителни условия, по отношение на обхвата на разработвания подход. Неговата основна цел е да се постигне правилно представяне на вече извлечените текстови данни в структурирана форма, което да завърши със записването на съответните данни в хранилище за данни, а това да позволи последващ техен анализ.

Подходът за представяне на текстови данни от уеб базирани информационни системи в структурирана форма включва система от методи, които имат специфичен характер:

- Метод за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма;
- Метод за разпознаване на фрагменти в структури от текстови данни;
- Метод за създаване на множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма;
- Метод за изчистване на излишни фрагменти в структури от текстови данни;
- Метод за съхранение на конфигурации за представяне на текстови данни от уеб базирани информационни системи в структурирана форма;
- Метод за визуализиране на представените в структурирана форма текстови данни.

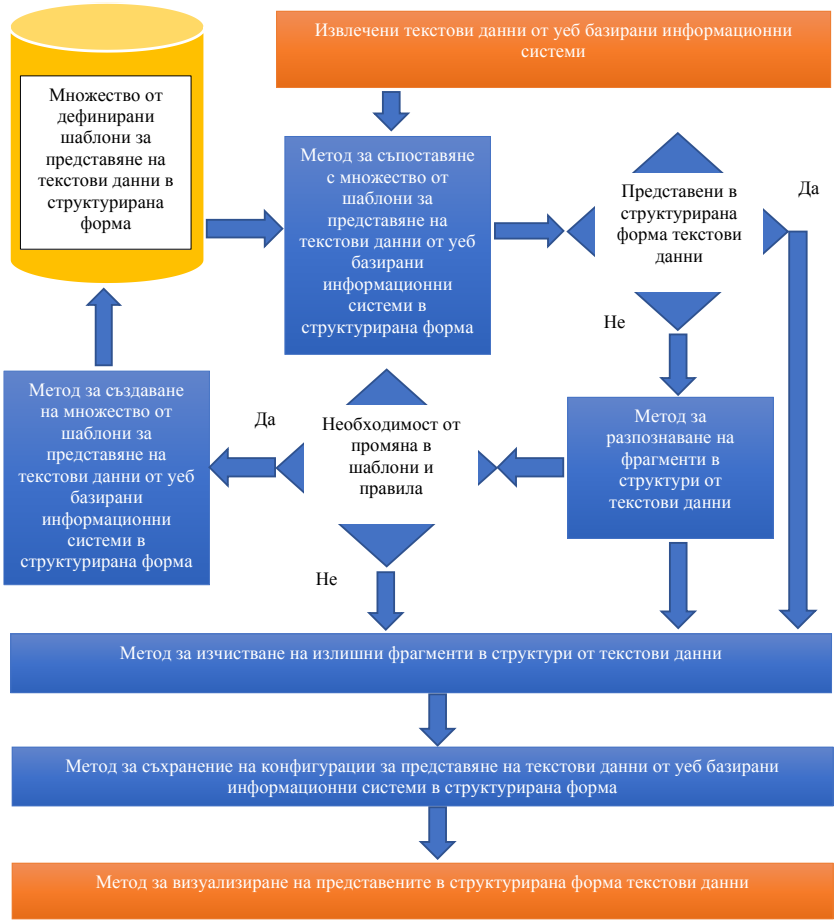
На следващ етап предложената система от методи следва да бъде прецизно специфицирана. Това описание трябва да изпълнява функциите на проектна документация, която да позволи софтуерната реализация на съответните методи. От друга страна, за да е възможно по-ясното описание на характеристиките на всеки от методите е добре да се използва общ подход в специфицирането, а именно спецификационна матрица на метод, която включва:

- Наименование на метода;
- Поредност на метода в системата от методи;
- Задължителност на изпълнението на метода;
- Цел на метода;
- Задачи на метода;
- Графично описание на метода;

- Резултати от изпълнението на метода;

Поради спецификите на проектиране на всеки един от заложените в подхода методи е възможно да настъпят известни изменения или параметризиране в настройките за описание и съответната спецификационна матрица на метод.

Проектирането на системата от методи налага и изисква както детайлното описание и специфициране на всеки от методите в състава на подхода, така и описание на тяхната взаимовръзка. Поради тези причини, а и с оглед целта на дисертационния труд, се предлага следната последователност на изпълнение на предложените методи (фиг. 7).



Фиг. 7 Взаимовръзка между изпълняваните методи

Резултат от изпълнението на първия метод е открито съвпадение с предварително дефиниран шаблон.

Резултатът от изпълнението на втория метод представлява множество от разпознати фрагменти в структури от текстови данни, което включва описание на съответните елементи и тяхното съдържание под формата на хипертекст.

Резултатът от изпълнението на третия метод представлява множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

Резултатът от изпълнението на четвъртия метод представлява създаването на алгоритъм за разпознаване на блокове от текстове в рамките на уеб страница, които не представляват интерес за извличане (изчистени данни).

Резултатът от изпълнението на петия метод представлява хранилище с конфигурации за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

Резултатът от изпълнението на шестия метод представлява визуализирани представени в структурирана форма текстови данни, които могат да бъдат експортирани към външни системи за анализ.

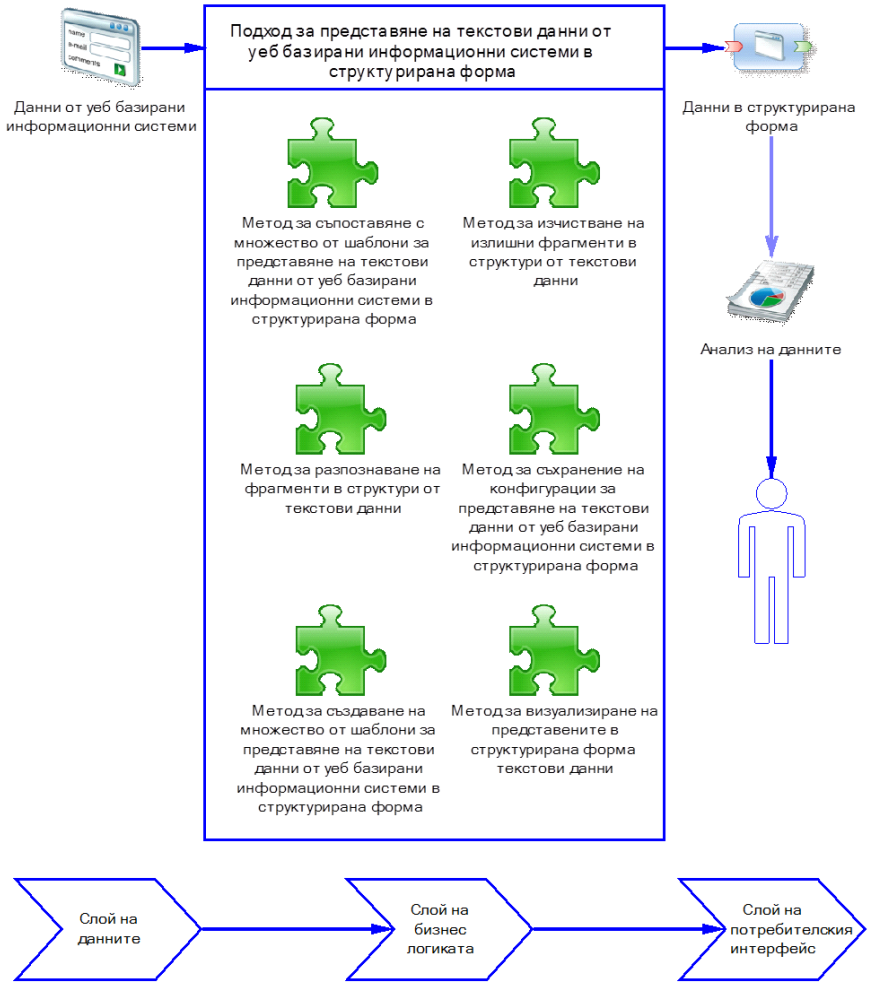
Проектирането на всеки от методите в състава на подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма дава основание да се създаде модел на софтуерно решение. Този модел включва взаимоотношенията между системата от методи и обкръжаващата среда, от където се осигуряват входните текстови данни от уеб базирани информационни системи. В настоящия дисертационен труд се предлага модел на софтуерно решение, който е базиран на предложения подход и системата от методи. Моделът на това софтуерно решение се проектира съобразно спецификите на класическата трислойна архитектура на уеб базираните информационни системи, която се състои от слой на данните, слой на бизнес логиката и слой на потребителския интерфейс (фиг. 8).

Проектирането на системата от методи предшества създаването на експериментален прототип, който представлява технологична реализация на предложения проект на модел на софтуерно решение за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

От архитектурна гледна точка проектирането на отделните методи в състава на подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма може да се раздели на два основни етапа:

- Проектиране на програмна част, която включва програмните класове и библиотеки за последващата реализацията на методите и потребителския интерфейс;

- Проектиране на база от данни, която съхранява всички необходими данни, които програмната част използва (данни за шаблони, данни за фрагменти, данни за конфигурации).



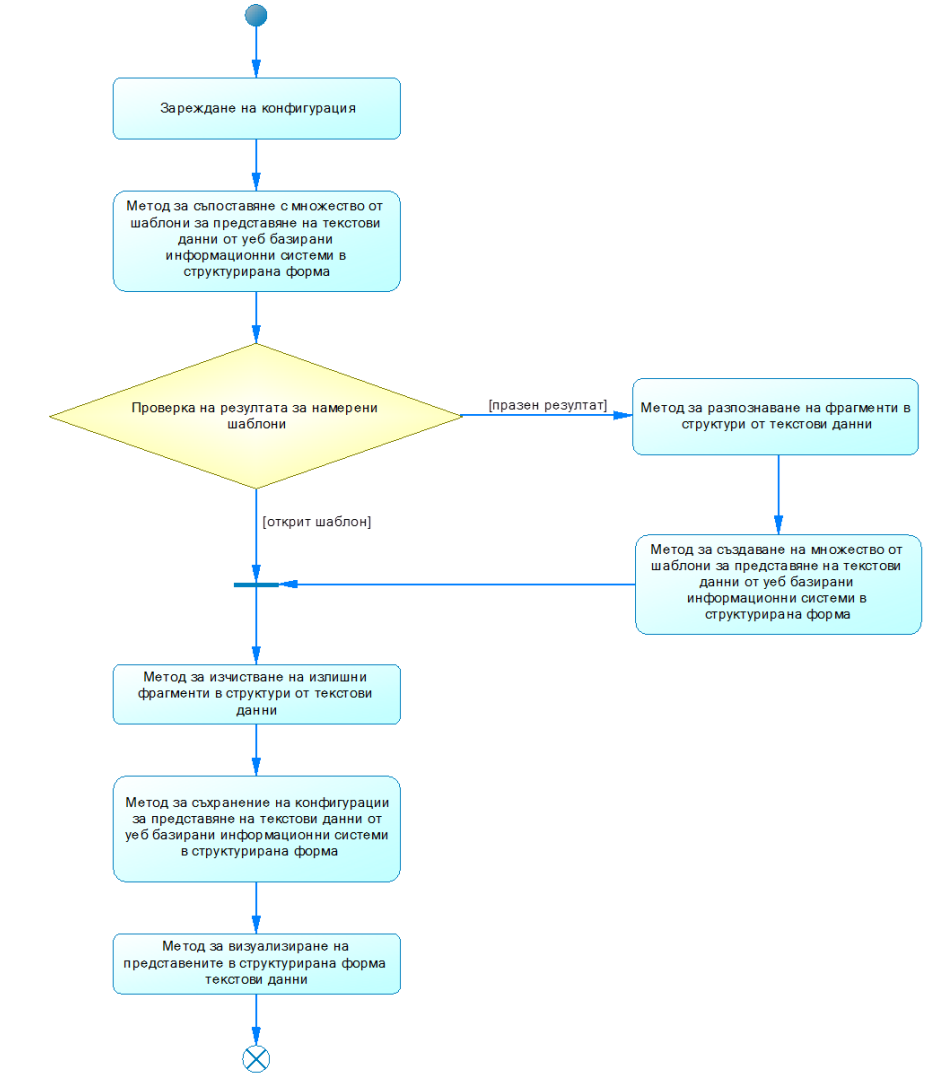
Фиг. 8 Модел на софтуерно решение за представяне на текстови данни от уеб базирани информационни системи в структурирана форма

Приложението на шаблони и фрагменти върху текстови данни от уеб базирани информационни системи води до разпознаване и извличане на определени части от данните и групирането на тези части по определен начин. Разпознатите

части и техния начин на групиране представляват структурираната форма на представяне на текстовите данни. На практика шаблоните и фрагментите определят структурираната форма на представяне на текстови данни. Самите шаблони и фрагменти описват както критериите, по които отделните части от текстовите данни се разпознават, така и начина на групиране на тези части, като е възможно наличието на повече от една група. Всички групи имат отделно описание в даден шаблон или фрагмент. За целите на настоящото изследване тези групи се наричат елементи за групиране като общо понятие. Всеки такъв елемент се характеризира с тип. Типът е наименование на елемента за групиране, с което той се превръща в разпознаваем. Елементи за групиране, които имат дефиниран тип, се наричат накратко елементи от съответния тип. Така например, елемент за групиране, който описва извличането на части от текстовите данни и организирането им в списък, може да бъде дефиниран като елемент от тип списък.

На фиг. 9 е изобразена поточна (технологична) диаграма на изпълнение на отделните методи в състава на подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, като изпълнението на системата от методи започва с първоначална конфигурация, която задължително включва уеб адрес на уеб базирана информационна система, върху който ще се приложат отделните методи. След зареждане на съответната конфигурация се извличат текстовите данни от съответния адрес и се пристъпва към изпълнение на метода за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма. Зареждат се наличните шаблони, които се прилагат върху извлечените текстови данни. При откриване на съответстващ шаблон, се пристъпва към метода за изчистване на излишни фрагменти в структури от текстови данни. В случай че няма нито един намерен шаблон, се пристъпва към метода за разпознаване на фрагменти в структури от текстови данни. Зареждат се наличните правила за разпознаване на фрагменти в структури от текстови данни. Резултатът от този метод представлява динамично генериран шаблон или част от шаблон. След изпълнението на метода за разпознаване на фрагменти в структури от текстови данни, ако има налични резултати, следва да им се зададе конфигурация, която представлява одобряване или отхвърляне на всеки един от тях. След това се пристъпва към изпълнението на метода за създаване на множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма. В рамките на този метод се създава крайният шаблон, който да се приложи в процеса на представяне на извлечените текстови данни в структурирана форма. Проектирането на метода за изчистване на излишни фрагменти в структури от текстови данни включва визуален инструмент за задаване на специфики, които по определен начин описват

фрагменти от разпознатите структури, които трябва да бъдат премахнати от крайния резултат.



Фиг. 9 Поточна диаграма на изпълнение на отделните методи в състава на подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма

Потвърждаването на окончателните настройки води до преминаване към метода за съхранение на конфигурации за представяне на текстови данни от уеб базирани информационни системи в структурирана форма. Изпълнението на този метод представлява запис на избраните конфигурации от предишните методи. След това се пристъпва към изпълнението на метода за визуализиране на представените в структурирана форма текстови данни, при който се показва практическият резултат от приложението на подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

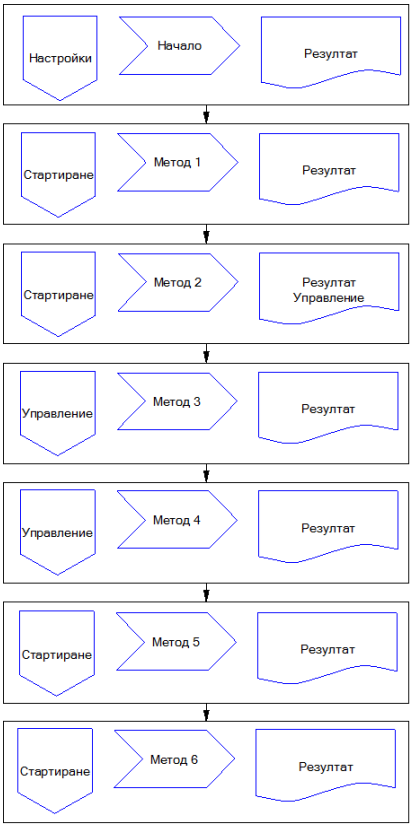
Глава Трета. Анализ на приложимостта на предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма

Създаденият за целите на настоящото изследване експериментален прототип представлява технологичната реализация на представената под формата на подход концепция за представяне на текстови данни от уеб базирани информационни системи в структурирана форма. За създаването на експерименталния прототип следва да бъдат изпълнени следните задачи:

- Дефиниране на жизнен цикъл на приложение на експерименталния прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма;
- Избор на програмни технологии за реализацията на експерименталния прототип;
- Съставяне на логическа архитектура, описваща градивните компоненти на експерименталния прототип и техните взаимодействия;
- Изготвяне на модел за описание и съхранение на представените във втора глава шаблони и фрагменти;
- Залагане на набор от готови шаблони и фрагменти, необходими за работата на експерименталния прототип при изпълнение на метода за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма и метода за разпознаване на фрагменти в структури от текстови данни от представената във втора глава система от методи;
- Дефиниране на резултатите от изпълнението на експерименталния прототип.

Съгласно първата задача, по отношение на създаването на експерименталния прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, е необходимо предварително да бъде дефиниран жизнен цикъл на експеримента, на базата на който ще се реализира експерименталният прототип.

На фиг. 10 са представени етапите от жизнения цикъл на експеримента, като всеки етап е представен като съвкупност от 3 подетапа, а именно начална инициализация, изпълнение на основна дейност в рамките на етапа и край на етапа.



Фиг. 10 Жизнен цикъл на провеждане на експеримента

Изпълнението на подетапите е последователно отляво надясно, а изпълнението на етапите е последователно отгоре надолу. Първият етап отбелязва началото на експеримента. Задават се първоначални настройки, които се запазват и се извеждат като резултат, който илюстрира въпросните настройки. Преминава се на втория етап, където в инициализацията се стартира изпълнението на метода за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма и се връща резултат. След това се преминава на третия етап, където се инициира стартиране на изпълнението на метода за разпознаване на фрагменти в структури от текстови данни и се връща резултат, който може да се управлява. На следващия четвърти етап се извършва управление на настройките и изпълнение на метода за създаване на множество от шаблони за представяне на текстови данни от уеб базирани

информационни системи в структурирана форма, като се връща резултат. Преминава се на следващия пети етап. В него се извършва управление, изпълнява се методът за изчистване на излишни фрагменти в структури от текстови данни и отново се получава резултат. Преминава се на следващия шести етап. В него се стартира приложението на метода за съхранение на конфигурации за представяне на текстови данни от уеб базирани информационни системи в структурирана форма и се връща съответен резултат. Накрая се преминава към последния етап, при който се стартира приложението на метода за визуализиране на представените в структурирана форма текстови данни и се връща резултат. С това целият цикъл на експеримента приключва до започването на нов експеримент, който стартира отново с първия етап от жизнения цикъл.

Съгласно втората задача, по отношение на създаването на експерименталния прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, следва да се състави схема на програмната реализация на експерименталния прототип, която включва:

- Избор на програмни технологии;
- Проектиране на градивните компоненти и техните взаимодействия;
- Съставяне на схема на приложение в реална среда.

На фиг. 11 е представена схема на експерименталния прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, която включва основните компоненти на прототипа и взаимодействията между тях.

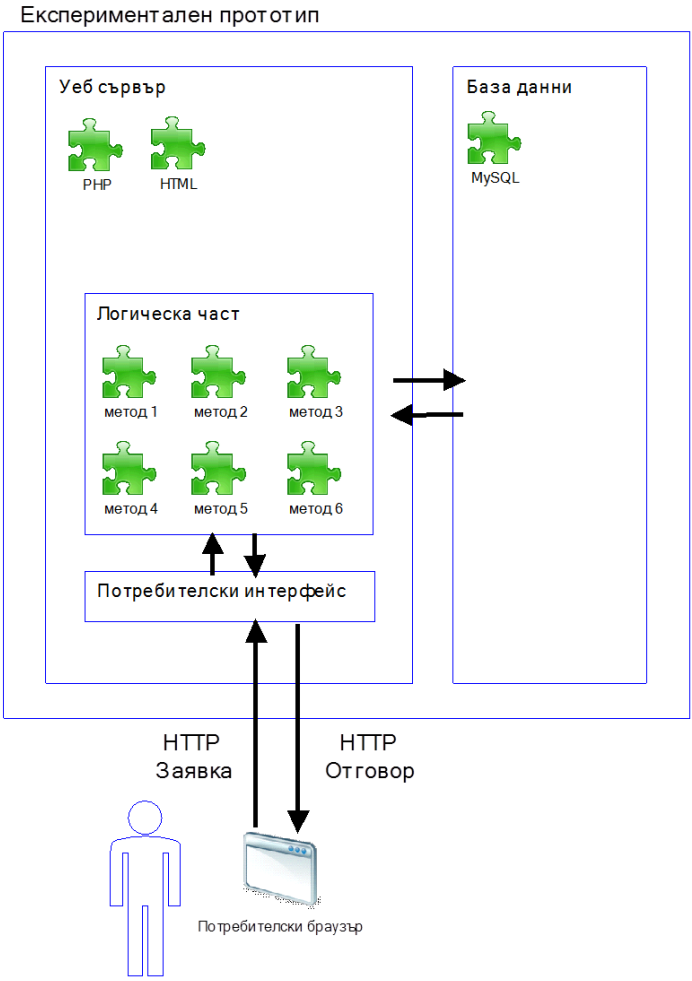
По отношение на начина си на работа, експерименталният прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, е от тип клиент – сървър. Включва в себе си уеб сървър, който поддържа сървърния език за програмиране PHP[75] и различни файлови формати, като HTML, както и система за управление на бази от данни. В областта на уеб сървъра е разположена програмната реализация на експерименталния прототип, която е разделена на две отделни части, а именно:

- Логическа част;
- Потребителски интерфейс.

Към логическата част се включват програмните средства, с които са реализирани софтуерно шестте метода от системата от методи (компоненти, класове, програмни библиотеки и др.) и всички изчислителни процедури и алгоритми, които изграждат цялостната функционалност на експерименталния

[75] https://www.php.net/docs.php

прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.



Фиг. 11 Експериментален прототип – компоненти и взаимодействия

Потребителският интерфейс играе ролята на мост за управление между потребителя и логическата част, и представлява организация на визуални програмни контроли (бутони, етикети, таблици), чрез които се управлява и наблюдава работата

на експерименталния прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

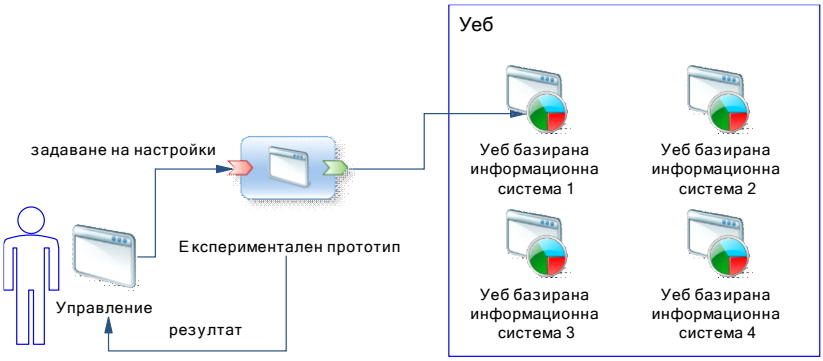
Системата за управление на бази от данни представлява софтуерно решение за управление и съхранение на колекции от логически свързани данни, които в контекста на експерименталния прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма могат да представляват настройки и практически резултати. В текущата реализация системата за управление на бази от данни е MySQL[76].

Чрез уеб браузър потребителите на експерименталния прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма достъпват потребителския интерфейс и изпълняват определени операции (задават се команди във вид на HTTP заявки). Това води до взаимодействие на потребителския интерфейс с логическата част и по този начин се установява управление на експерименталния прототип – задават се определени входни параметри при прилагането на системата от методи на подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма. Работата на логическата част е свързана и с взаимодействие с базата от данни, което включва операции на зареждане и записване на съответните данни, които се създават при работа с експерименталния прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

На фиг. 12 е представена схема на приложение на експерименталния прототип върху уеб базирани информационни системи.

При провеждане на експеримент потребител следва да зададе съответните настройки. Експерименталният прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма получава текстови данни (на база установени настройки) от дадена уеб базирана информационна система. Обработката на тези данни обуславя жизнения цикъл на експеримента, като започва с метода за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма и приключва с метода за визуализиране на представените в структурирана форма текстови данни. Изпълнението на всеки метод води до получаването на съответен резултат към потребителя.

[76] https://www.mysql.com/



Фиг. 12 Приложение на експерименталния прототип върху уеб базирани информационни системи

Преминаването през различните етапи от жизнения цикъл на експеримента на практика представлява приложение на системата от методи от разработения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма. Съгласно описанията във втора глава, всеки един от тези методи използва шаблони и фрагменти, чрез които се осъществява представянето на текстовите данни в структурирана форма. Тези шаблони и фрагменти са представени абстрактно при описанието на съответните методи, но за нуждите на експерименталния прототип е необходимо да се изготви начин за тяхното описание. Това описание се осъществява по определени правила, които са дефинирани чрез разширяемия маркиращ език XML[77]. Шаблоните и фрагментите се състоят от едно или повече описания на групирани елементи. За провеждането на експеримента дефинираме два типа групирани елементи – тип препратка и тип детайлна страница. В контекста на уеб базираните информационни системи, за целите на настоящото изследване, с понятието препратки се визират области от дадена уеб страница , които обикновено са повече от една, имат подобни или еднакви характеристики и се намират на едно и също ниво в дървовидната HTML структура. Тези области могат да представляват хиперлинкове или други HTML елементи, които съхраняват в себе си хиперлинкове. С тях се представят множества от уеб адреси, препращащи към категории със списъци или към детайлни страници с описателно съдържание. Такъв тип области се описват в шаблони и фрагменти чрез дефинирания групиран елемент от тип препратка. В една уеб страница може да има и области, които съхраняват множество текстови данни, свързани помежду си. Типичен пример за такава област представлява уеб публикация, която се състои от отделни текстови фрагменти, като

[77] https://www.w3.org/TR/xml11/

заглавие, дата, автор, описание и др. Този тип област в шаблоните и фрагментите се описва чрез дефинирания групиран елемент от тип детайлна страница. Двата групирани елемента от тип препратка и от тип детайлна страница представляват съответно елемент от тип препратка и елемент от тип детайлна страница. Тези елементи в XML формат се дефинират чрез тагове и представляват описания на дървовидни структури, които се състоят от множества йерархично организирани поделементи. Таговете се характеризират с имена и атрибути.

Провеждането на експеримент с реализирания експериментален прототип има за цел получаването на резултати от приложението на подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма. За провеждането на експеримента следва да бъдат изпълнени следните задачи:

- Практическо описание на приложението на подхода за представяне на текстови данни от уеб базирани информационна система в структурирана форма;
- Определяне на множество от уеб базирани информационни системи, върху които ще се приложи експерименталният прототип;
- Индивидуално приложение на експерименталния прототип върху всяка от избраните уеб базирани информационни системи;
- Обобщаване на резултатите от индивидуалните приложения на експерименталния прототип.

След създаването на експерименталния прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, се преминава към практическото приложение на прототипа върху петдесет избрани за целите на настоящото изследване уеб базирани информационни системи, достъпни в Интернет. Те са подбрани на база на следните критерии:

- Класификация – избраните уеб базирани информационни системи представляват медии, които са определени като регионални за Р България, т.е. по тематичност се отнасят към конкретна област от страната, напр. област Бургас, област Варна и др. Характерното за този тип уеб базирани информационни системи е, че най-голям дял от уеб публикациите им засягат регионални теми, като на следващо място се нареждат теми с национален или международен характер;
- Технология – избраните уеб базирани информационни системи съдържат публична информация, основно представена в текстов вид. Всяка платформа се състои от взаимно свързани уеб страници, всяка от които съответства на уникален уеб адрес;

- Отличителни характеристики на данните – за всеки от избраните уеб базирани информационни системи е характерно, че представят данни във вид на уеб публикации, които имат поне две обособени текстови части – заглавие и описание, като често уеб публикациите могат да включват и допълнителни данни – дата на публикуване, автор, снимки и др.;
- Организация на данните – за всяка от избраните уеб базирани информационни системи е характерно, че данните, представени във вид на уеб публикации, са организирани в категории, които в уеб страниците представляват списъци с уеб адреси и наименования. Тези списъци от своя страна могат да препращат към уеб страници, съдържащи описания на списъци с уеб публикации или директно към уеб страници, съхраняващи самите уеб публикации.

В таблица 2 са представени избраните за провеждането на експеримента уеб базирани информационни системи.

Таблица 2 Уеб адреси на уеб базирани информационни системи за провеждане на експеримента

Уеб базирана информационна система	**Начален уеб адрес**
100 Вести	https://www.stovesti.info
Arena media	http://www.arenamedia.net
Bourgas.org	https://www.burgaslargo.com
Burgas news	http://www.burgasnews.com
Dgpazar eu	https://dgpazar.eu
Dupnicanews.eu	http://dupnicanews.eu
Maritime.bg	https://www.maritime.bg
PA media	https://pa-media.net
Pa1.bg	https://pa1.bg
Sandanski1	https://sandanski1.com
Благоевград 24	https://www.blagoevgrad24.bg
БургасИнфо	https://www.burgasinfo.com
Варна утре	https://varnautre.bg
Велико Търново нет	https://veliko-tarnovo.net
Вестник Десант	http://www.desant.net
Вестник Посредник	http://www.posredniknews.com

Виделина	http://www.videlinabg.com
Градвелин.ком	http://gradvelin.com
Грамофона	https://www.gramofona.com
Добруджа	http://dobrudjabg.com
Златоград	https://www.zlatograd.com
Икономически портал на регион Стара Загора	https://www.chambersz.com
Казанлък	https://www.kazanlak.com
Козлодуй онлайн	https://kozloduy-bg.info
Компас	https://www.kompasbg.com
Конкурент	https://www.konkurent.bg
Кърджали Хабер	https://kircaalihaber.com/bg
Кюстендил инфо	http://www.kustendil.info
Марица	https://www.marica.bg
Модерна Враца	http://modernavratza.com
Монт Прес	https://www.mont-press.com
Народно дело	http://narodnodelo.bg
Несебър news	http://www.nessebar-news.com/news
Паралел 44	http://www.paralel44.com
Пловдив online	https://plovdiv-online.com
Пловдив медия	https://plovdivmedia.bg
Посоки Плевен	https://www.posoki.com
Родопи 24	https://rodopi24.blogspot.com
Сакар нюз	https://sakarnews.info
Север.бг	https://sever.bg
Седмица	https://sedmica.sliven.net
Сливен нюз	https://sliven-news.com
Сливенски новини	https://sl-news.sliven.net
Смолян днес	https://www.smolyandnes.com
Стара загора днес	https://stzagora.net
Старият мост	https://www.stmost.info
Струма	https://www.struma.com
Телевизия Стара Загора	https://tvstz.com

Хасково нет	https://www.haskovo.net
Черно море	https://www.chernomore.bg

За извършването на оценка на получените резултати от проведения експеримент следва да бъдат изпълнени следните задачи:

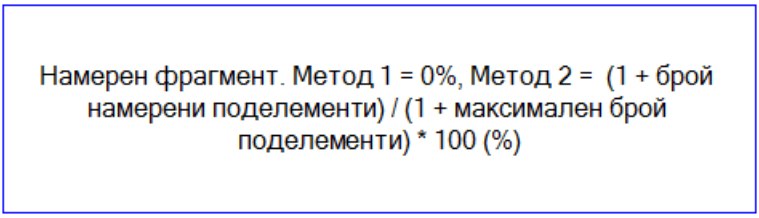
- Определяне на методика за оценка на резултатите от индивидуалното приложение на експерименталния прототип;
- Изготвяне на индивидуални оценки на приложението на експерименталния прототип на база на резултатите за всяка уеб базирана информационна система в рамките на проведения експеримент;
- Изготвяне на комплексна оценка на приложението на експерименталния прототип на база на всички получени индивидуални оценки;
- Изводи от получената комплексна оценка.

За оценяване на получените резултати от практическото приложения на експерименталния прототип, е необходимо да бъде създадена и приложена методика за оценка на резултатите от изпълнението за всеки един от представените уеб адреси. Оценката от индивидуалното изпълнение на всеки адрес следва да се формира на база на успешното или неуспешното изпълнение на метода за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма и метода за разпознаване на фрагменти в структури от текстови данни, резултатите от които се получават автоматизирано, на база на представените описания на шаблони. След това се прави обобщена оценка на база на оценките от изпълнението на всеки един уеб адрес. Следва описание на методиката, по която следва да се оцени крайният резултат от практическото приложение на експерименталния прототип по предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма:

1. Крайните индивидуални резултати от изпълнението на метода за съпоставяне с множество шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, метода за разпознаване на фрагменти в структури от текстови данни и обобщеният резултат на база на двата метода представляват коефициент в интервала от 0 до 100 в проценти, , който определя нивото на успешно изпълнение на съответния метод. Стойност 0 означава абсолютен неуспех, стойност 100 означава абсолютен успех, а

всички междинни стойности между 0 и 100 се определят като частичен успех;

2. Резултатът от изпълнението на метода за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма винаги е или 100 (разпознат е шаблон) или 0 (не е разпознат нито един шаблон). Когато резултатът е 100, съгласно подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, методът за разпознаване на фрагменти в структури от текстови данни се пропуска, т.е. неговият резултат е празен и се приема за стойност 0;
3. В случай че резултатът от метода за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма е 0, се преминава към изпълнение на метода за разпознаване на фрагменти в структури от текстови данни, за който резултатът се определя от това дали е разпознат динамично генериран шаблон и каква част от неговите фрагменти са идентифицирани. Формулата за определяне на процента за успешно изпълнение е представена на фиг. 13;
4. Ако резултатът от изпълнението на метода за разпознаване на фрагменти в структури от текстови данни е 0, имаме наличие на празен резултат от изпълнението на двата метода;
5. Крайната оценка от приложението на двата метода се определя от по–голямата стойност от резултатите им. Например, ако резултатът от изпълнението на методите е съответно 0% и 33%, крайната оценка ще бъде 33%;
6. Обобщените оценки от резултатите за всички уеб адреси на двата метода и крайната оценка представляват техните средни аритметични стойности.



Фиг. 13 Резултат при успешно изпълнение на метода за разпознаване на фрагменти в структури от текстови данни

Съгласно формулата, изобразена на фигурата, за успешно изпълнение се приема идентифицирането на поне един динамично генериран елемент от тип препратка. Генерирането на динамични елементи е базирано на всички отделни фрагменти, изграждащи наличните в експерименталния прототип шаблони. Всеки един елемент за групиране се характеризира с максимален брой поделементи. Така например, за елемент от тип препратка, максималният брой поделементи представлява броят на дъщерните елементи (тагове link и title) на тага target. Това означава, че максималният брой поделементи за този тип елемент е 2. Формулата, представена на фигурата, се прилага при успешно идентифициране на главния елемент, т.е. има наличие на поне един идентифициран фрагмент. Определена е еднаква степен на тежест за всеки един от трите фрагмента (главния и двата дъщерни). По тази причина процентът за определяне на нивото на идентифициране на главния елемент представлява процент на броя на всички намерени поделементи и главния елемент (1 брой) спрямо максималния брой поделементи (2 броя) и главния елемент (1 брой). Ако разгледаме случай, при който е намерен елемент от тип препратка, но не е идентифициран нито един от неговите поделементи (link, title), резултатът ще бъде ((1 + 0) / (1 + 2)) * 100 = 33%. Получава се така, че положителните потенциални резултати от изпълнението на метода за разпознаване на фрагменти в структури от текстови данни за елемент от тип препратка са 33%, 66% и 100%. Важно е да се отбележи, че е възможно да бъде намерен повече от един главен елемент. По тази причина за краен резултат от изпълнението на този метод се приема най-високият резултат от резултатите на всички намерени главни елементи.

На фиг. 14 е представен графично крайният резултат от проведения експеримент за всички петдесет уеб базирани информационни системи, където се вижда, че в проведения експеримент 18% от крайния резултат се основава на успешно изпълнение на метода за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, а 0,67% от крайния резултат е на база успешното изпълнение на метода за разпознаване на фрагменти в структури от текстови данни. Останалите 81,33% представляват нулев резултат следствие на неуспешно изпълнение на методите. Крайният успех от изпълнението на двата метода е 18,67%. Този резултат представлява сумата от резултатите от двата метода, тъй като съгласно подхода за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, всеки индивидуален резултат се базира или на метода за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, или на метода за разпознаване на фрагменти в структури от текстови данни.

Фиг. 14 Графично представяне на резултатите от проведения експеримент

Тъй като за провеждане на експеримента е използвана ограничената база от осем шаблона, крайният резултат от 18,67% може да бъде основание да се счита, че предложеният в дисертационния труд подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма има добавена стойност и принос в изследваната предметна област.

IV. НАУЧНИ И НАУЧНО-ПРИЛОЖНИ ПРИНОСИ

1.Извършен е задълбочен литературен преглед и анализ на предметната област, включващ сравнителен анализ на софтуерни решения за извличане на данни от уеб базирани информационни системи и представяне на извлечените текстови данни в структурирана форма.

2.Предложен е цялостен концептуален подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, състоящ се от система от шест метода: метод за съпоставяне с множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, метод за разпознаване на фрагменти в структури от текстови данни, метод за създаване на множество от шаблони за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, метод за изчистване на излишни фрагменти в структури от текстови данни, метод за съхранение на конфигурации за представяне на текстови данни от уеб базирани информационни системи в структурирана форма, метод за визуализиране на представените в структурирана форма текстови данни.

3.Дефинирани са пълни проектни спецификации на методите в състава на предложения подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

4.Разработен е експериментален прототип, който представлява технологична реализация на предложения концептуален подход за представяне на текстови данни от уеб базирани информационни системи в структурирана форма.

5.Експерименталният прототип е практически приложен върху множество от уеб базирани информационни системи и на база на постигнатите резултати е анализирана добавената стойност в предметната област.

V. СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ТЕМАТА НА ДИСЕРТАЦИОННИЯ ТРУД

1.Табов, Я., Милев, П., Възможности за извличане на информация от неструктурирани данни, Сборник доклади от Национална научна конференция "Насоки и проблеми на приложението на статистиката, информационните технологии и математиката в социално-икономическата област", УНСС - София, 2017, стр. 275 - 282, ISBN 978-619-232-023-2.

2.Tabov, Y., Milev, P., Methodological features of web data structuring, International Journal INFORMATION CONTENT & PROCESSING, Sofia, 2019, Vol. 6, Number 2, pp. 42 - 59, ISSN 2367-5128.

3.Tabov, Y., Technological Issues of Retrieving Information from Unstructured Data, Proceedings of the 7th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE – 2017), UNWE - Sofia, 2020, pp. 562 - 566, ISSN 2367-7635.

4.Tabov, Y., Technological Issues of Analyzing Unstructured Data, Proceedings of the 8th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE – 2018), UNWE - Sofia, 2020, pp. 305 - 308, ISSN 2367-7635.

5.Tabov, Y., Modern Trends of Retrieving Information from Internet, Proceedings of the 8th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE – 2018), UNWE - Sofia, 2020, pp. 309 - 313, ISSN 2367-7635.

6.Tabov, Y., Approach for Automatic Identification of Unstructured Data, Proceedings of the 9th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE – 2019), UNWE - Sofia, 2020, pp. 261 - 265, ISSN 2367-7635.

7.Tabov, Y., Application of Concept for Structuring of Web Data, Proceedings of the 9th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE – 2019), UNWE - Sofia, 2020, pp. 266 - 270, ISSN 2367-7635.