ІНФОРМАЦІЙНА СИСТЕМА АВТОМАТИЗОВАНОГО ФОРМУВАННЯ ЛЕКСИКОГРАФІЧНИХ РЕСУРСІВ

УДК 519.7:007.52

БОРИСОВА Наталя Володимирівна
асистент кафедри інтелектуальних комп'ютерних систем
Національного технічного університету «Харківський політехнічний інститут».
Наукові інтереси: автоматизована обробка природної мови, видобування знань, створення електронних словників.

ЯМШАНОВ Ігор Сергійович
к.т.н., старший викладач кафедри автоматизованих систем управління
Національного технічного університету «Харківський політехнічний інститут».
Наукові інтереси: скорингові системи, обробка знань, стеганографія.

ВСТУП

Вирішення багатьох задач автоматизованої обробки природномовних об'єктів, наприклад, машинний переклад, літературно-наукове редагування, повнотекстовий пошук, реферування та анотування, потребує наявності електронних лексикографічних ресурсів, які значно підвищують ефективність та якість такої обробки. Крім того необхідність автоматизованого створення лексикографічних ресурсів виникає тому, що зараз з'являються та розвиваються нові предметні області й виникає проблема формалізації та моделювання цих областей знань. Створення лексикографічних ресурсів, у тому числі і електронних, є складним та трудомістким процесом, тому задача автоматизації цього процесу є актуальною та потребує вирішення.

Метою даної роботи є розробка високорівневої архітектури інформаційної системи автоматизованого формування лексикографічних ресурсів.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Інформаційна система (ІС) автоматизованого формування лексикографічних ресурсів – це інформаційна система, яка оперує як даними (у традиційному розумінні), так і знаннями, представленими у належній формальній формі. При цьому під «знаннями» ми будемо розуміти сукупність фактів, видобутих з даного тексту і представлених у формальному вигляді, яка відображає смисл, закладений у вихідному тексті [8]. Знання мають такі властивості: структурованість, зв'язність, активність, семантична метрика, внутрішня інтерпретованість.

Процеси обробки знань, до яких відносяться вилучення і видобування, представлення і маніпулювання, потребують інтелектуалізації, а автоматизація цих процесів потребує використання інтелектуальних інформаційних технологій та систем. У загальному вигляді процес інтелектуальної обробки природномовних об'єктів (ПМО) та знань можна представити у такому вигляді (рис. 1).

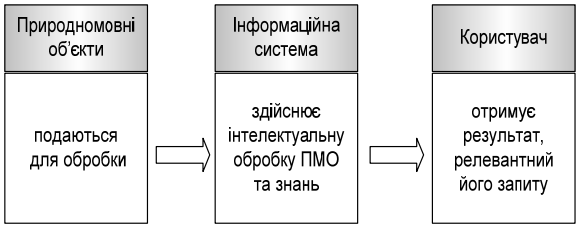

Рисунок 1 – Процес інтелектуальної обробки природномовних об'єктів та знань

Як бачимо, інтелектуальна інформаційна система використовується для задоволення інформаційних потреб користувача, здійснюючи інтелектуальну оброб-

ку природномовних об'єктів з використанням моделей, методів та інформаційних технологій такої обробки.

За допомогою ІС автоматизованого формування лексикографічних ресурсів передбачається вирішення таких завдань:

- автоматизоване створення термінологічних словників для нових предметних областей та предметних областей, які динамічно розвиваються;

- видобування термінів та ключових слів з текстів предметної області [4] з метою подальшого їх використання, наприклад, при індексуванні документів ключовими словами, для покращення повнотекстового пошуку [6] тощо;

- автоматизоване створення словника синонімів [3];

- автоматизоване створення словників сполучуваності слів предметної області [2].

Виходячи із вищезазначеного та загальних вимог до сучасних ІС [5], архітектурна будова зазначеної ІС має відповідати наступним основним принципам:

- відповідність поточним і перспективним цілям, а також функціональним стратегічним завданням створюваної ІС;

- універсальність ІС;

- забезпечення прийнятної для користувачів структурованості даних та достатньої глибини їх опису;

- забезпечення необхідної оперативності пошуку інформації та виконання запитів аналітико-синтетичного характеру;

- гнучкість і можливість розвитку та нарощування функцій і ресурсів ІС відповідно до розширення сфер і завдань її застосування;

- забезпечення віддаленого авторизованого доступу кінцевих користувачів та їх груп до застосування ІС і результатів її функціонування на основі сучасних графічних засобів і наочних інтерфейсів;

- реалізація властивих для таких систем технологічних функцій (забезпечення цілісності, несуперечності, мінімізація надлишковості даних, їх захист від некомпетентних дій та можливість відновлення).

Для створення ІС автоматизованого формування лексикографічних ресурсів було обрано системний підхід, який полягає у комплексному вивченні об'єкта як цілого з представленням його частин як цілеспрямованих систем і вивчення цих систем та відношень між ними. При системному підході об'єкт розглядається як сукупність взаємопов'язаних елементів однієї складної динамічної системи, яка перебуває в стані постійних змін під впливом багатьох внутрішніх і зовнішніх факторів, пов'язаних з процесами перетворення вхідних ресурсів на вихідні. Системний підхід базується на таких принципах:

1) кінцевої мети – абсолютний пріоритет кінцевої мети;

2) єдності – розгляд системи як цілого, і як сукупності елементів;

3) зв'язності – розгляд будь-якого елементу разом з його зв'язками з оточенням;

4) модульної побудови – виділення модулів в системі та розгляд її як сукупності модулів;

5) ієрархії – введення ієрархії елементів та/або їх ранжування;

6) функціональності – спільний розгляд структури і функцій з пріоритетом функцій над структурою;

7) розвитку – врахування змін системи, її здатності до розвитку, розширення, заміни елементів, накопичення інформації;

8) децентралізації – управління централізацією і децентралізацією;

9) невизначеності – врахування невизначеностей та випадковостей у системі.

Характерними ознаками системного підходу є: одночасне охоплення проектуванням великої кількості задач; максимальна типізація та стандартизація рішень; багатоаспектне уявлення про структуру ІС як про систему, що складається з кількох класів елементів, та відносна автономна їх розробка; ключова роль баз даних; локальне впровадження та збільшення функціональних задач.

Задачею системного підходу щодо створення ІС є розробка всієї сукупності методологічних і соціально-наукових засобів обстеження (опис, аналіз, синтез, реалізація) систем різного типу. У методологічному відношенні системний підхід базується на ідеях цілісності, цілеспрямованості, організованості об'єктів, що вивчаються, їх внутрішній активності та динамізмі [1].

Оскільки при системному підході, як вже зазначалося вище, об'єкт розглядається як сукупність взаємопов'язаних елементів однієї складної динамічної систе-

ми, ІС можна вважати сукупністю функціональних підсистем та зв'язків між ними. Функціональна підсистема – це частина ІС, виділена за спільністю функціональних ознак. Функціональна декомпозиція ІС визначає призначення підсистем, тобто для якої сфери діяльності вона призначена і які основні цілі, задачі та функції виконує. В залежності від складності об'єкта кількість функціональних підсистем може бути різною [11].

При виділенні функціональних підсистем ІС необхідно дотримуватися таких вимог:

- межі задач, які утворюють підсистему, не повинні перетинатися між собою;

- задачі, що вирішуються у підсистемах, мають бути тісно пов'язані між собою в інформаційному плані, тобто при їх вирішенні має використовуватися єдина вхідна інформація, а результати вирішення одних задач мають використовуватися для вирішення інших;

- результати вирішення повинні мати єдиного споживача [11].

При виділенні функціональних підсистем мають бути визначені їх параметри: мета функціонування підсистеми, вид ресурсів, особливості показників, що розраховуються у підсистемі.

Для експлуатації функціональних підсистем потрібні відповідні ресурси, які створюють забезпечувальні підсистеми ІС: математичну, алгоритмічну, інформаційну, програмну, організаційну, методичну, технічну, лінгвістичну, правову, ергономічну. Розглянемо більш детально деякі з них для розробленої ІС автоматизованого формування лексикографічних ресурсів, яка розробляється.

Математичне забезпечення ІС – це сукупність математичних методів та моделей, що використовуються в ІС [1]. У якості математичного забезпечення ІС використано моделі відношень між природномовними об'єктами різних рівнів та моделі формалізації знань, розроблені з використанням апарату алгебри скінченних предикатів і предикатних операцій, а також метод компараторної ідентифікації, використаний для опису інтелектуальних функцій людини. На базі цих моделей були побудовані алгоритми процесів, які здійснюються в ІС. Ці алгоритми складають *алгоритмічне забезпечення* ІС. На рис. 2 представлено алгоритм процесу створення термінологічного словника певної предметної області. Спочатку на множині науково-технічних текстів визначаються та зберігаються у вигляді шаблонів дискурсивні маркери дискурсивної операції «визначення», а також формується набір правил розташування дискурсивних маркерів по відношенню до поняття та його визначення. Далі на вхід системи подаються тексти предметної області, з яких за допомогою заданих шаблонів та правил видобуваються поняття та їх визначення. Обрані таким чином поняття та їх визначення піддаються процедурі нормалізації та за допомогою заданих правил розмітки перетворюються на словникові статті, з яких власне і формується термінологічний словник.

Спочатку за допомогою словника предметної області здійснюється аналіз термінів цієї предметної області та створюється набір шаблонів цих термінів. Далі задається набір правил зміни шаблонів за відмінками та числами (якщо це можливо). На наступному кроці за заданими шаблонами та правилами здійснюється пошук термінів-кандидатів у текстах предметної області. Отримана множина термінів-кандидатів піддається статистичній обробці та аналізується експертом. У результаті отримуємо множину термінів предметної області, яка в подальшому може бути використана для покращення повнотекстового пошуку, індексування документів ключовими словами та ін.

Інформаційне забезпечення – це сукупність форм документів, нормативної бази та реалізованих рішень щодо обсягів, розміщення і форм існування інформації, яка використовується в інформаційній системі при її функціонуванні. Інформаційне забезпечення поділяється на позамашинне і внутрішньомашинне. Усю інформацію, що обробляється в ІС, можна поділити на вхідну, проміжну та вихідну [1]. В процесі розробки інформаційного забезпечення визначається [9]:

- склад інформації (перелік інформаційних одиниць, необхідних для вирішення комплексу задач);

- структуру інформації та закономірності її перетворення (правила формування показників і документів);

- характеристики руху інформації (обсяг та інтенсивність потоків, маршрути, часові характеристики);

- характеристики якості інформації (систему кількісних оцінок значущості, повноти, своєчасності, вірогідності інформації);

- способи перетворення інформації.

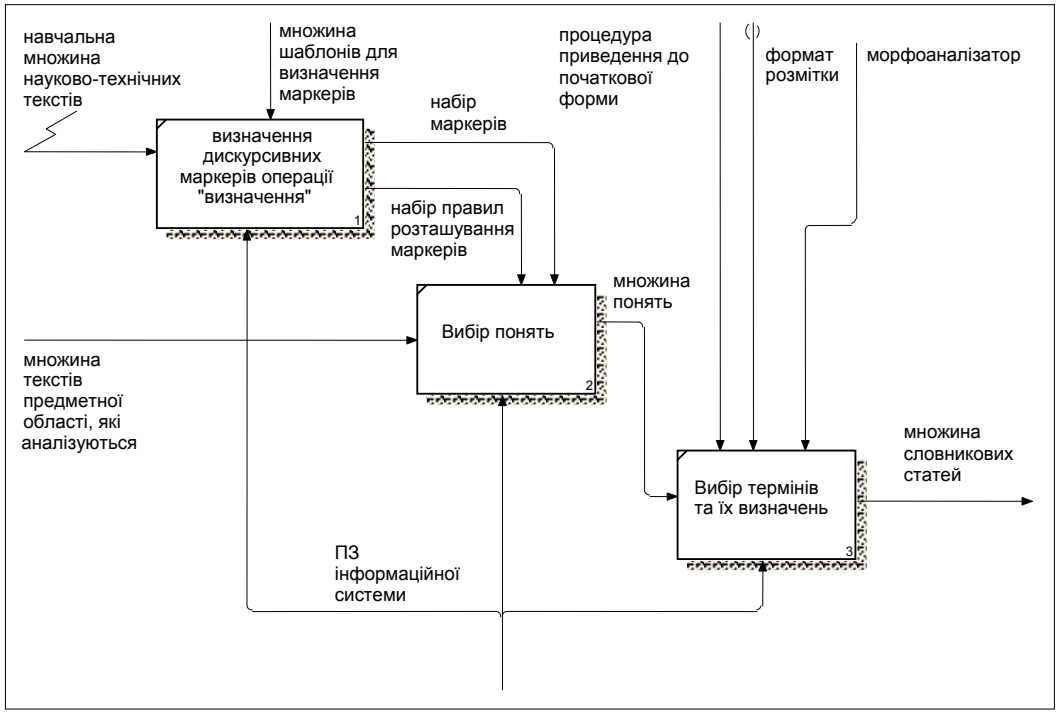

Рисунок 2 – Алгоритм процесу створення словника предметної області

На рис. 3 представлено алгоритм процесу видобування термінів предметної області з текстів.

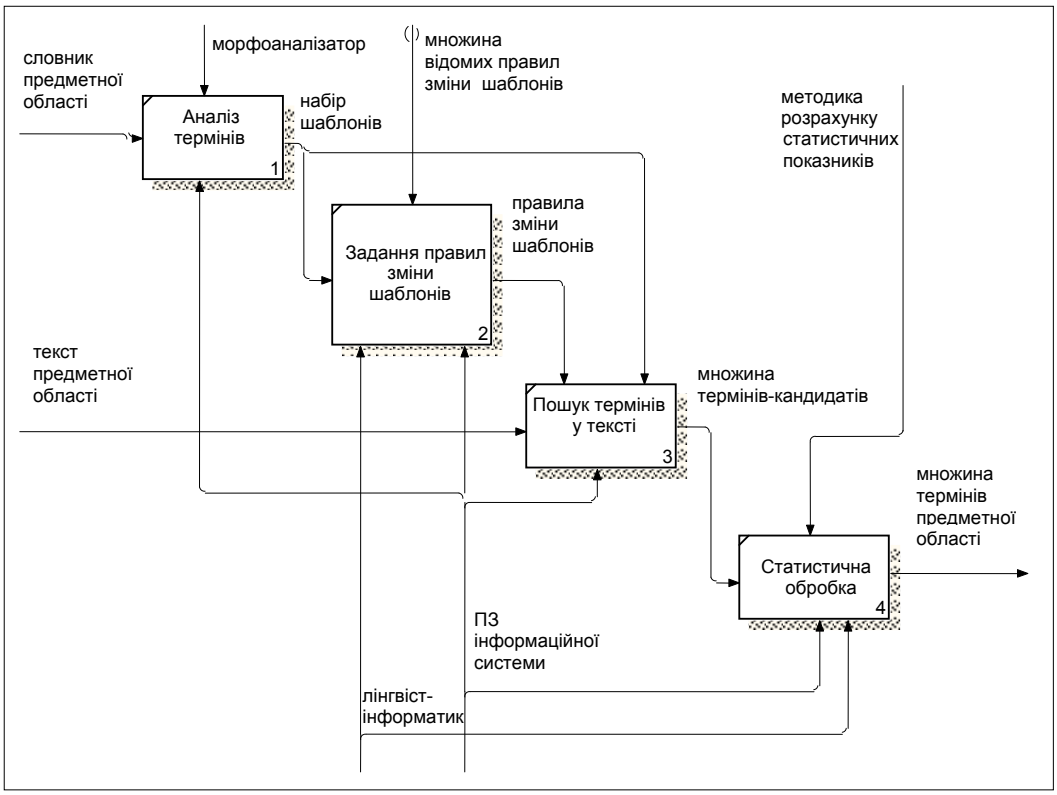

Рисунок 3 – Алгоритм процесу видобування термінів предметної області з текстів

Основу інформаційного забезпечення ІС автоматизованого формування лексикографічних ресурсів складає інтелектуальна інформаційна технологія видобування та обробки предметних знань (рис. 4).

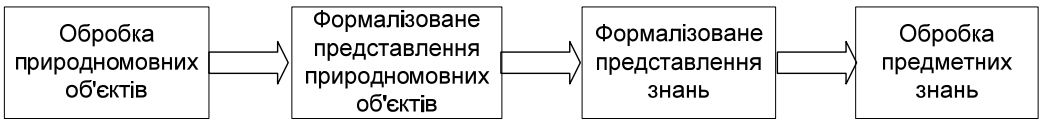

Рисунок 4 – Інтелектуальна інформаційна технологія видобування та обробки предметних знань

У нашому випадку вхідною інформацією в залежності від потреб, запитів користувача є слабкоструктуровані або структуровані природномовні об'єкти. До перших відносяться повнотекстові документи, віднесені до певної предметної області на попередніх етапах їх обробки, до других – тексти електронних словників предметної області. У якості проміжної інформації виступають формалізовані представлення природномовних об'єктів та знань. Вихідною інформацією є оброблені предметні знання у вигляді, релевантному потребам або запитам користувача.

Програмне забезпечення ІС – це сукупність програм на носіях даних і програмних документів, призначених для налагодження, функціонування та перевірки працездатності ІС. Програмне забезпечення складається із системи програм, до яких входять програмні компоненти для організації обробки даних та інструктивно-методичні матеріали щодо застосування засобів програмного забезпечення [1].

Для розробки програмного забезпечення ІС було обрано ітеративний підхід, при якому виконання робіт здійснюється паралельно з постійним аналізом отриманих результатів і корегуванням попередніх етапів роботи. Розробка при цьому підході в кожній фазі проходить такий цикл: Планування – Реалізація – Тестування – Оцінка [11]. Загальна схема ітераційного підходу до розробки програмного забезпечення представлена на рис. 5.

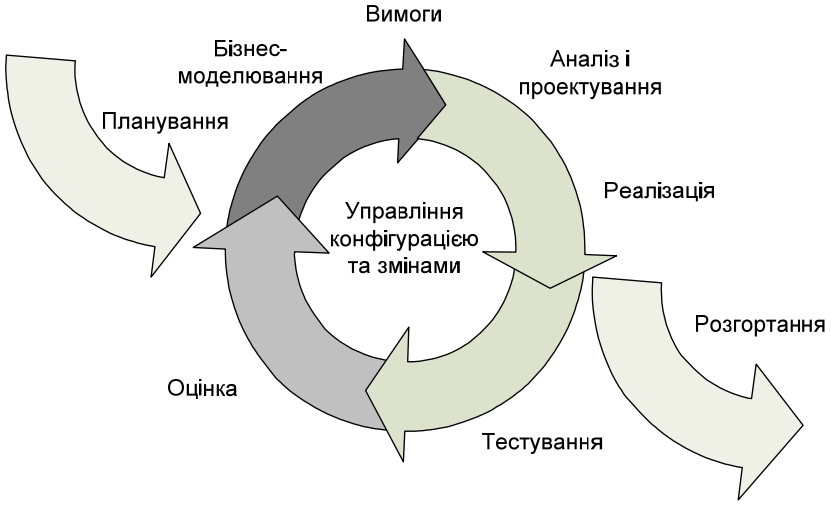

Рисунок 5 – Ітераційна розробка програмного забезпечення

При виборі підходу враховувалася задача, яка вирішується. Вибір підходу для вирішення нашої задачі був зумовлений такими перевагами ітеративного підходу:

- акцентування уваги та зусиль на найважливіших та найскладніших вимогах;
- постійне тестування, що дозволяє оцінити успішність виконання етапів вирішення задачі в цілому;
- реальна оцінка поточного стану вирішення задачі в цілому [7].

На стадії планування визначаються вимоги до ІС, будується її архітектура [11]. Архітектура ІС включає

наступні апаратно-програмні компоненти: бази даних; засоби обробки інформації; засоби доступу до інформаційних ресурсів; засоби організації роботи користувачів; засоби адміністрування; засоби транспортування інформації. Архітектура повинна відповідати характеристикам застосування ІС. Архітектура ІС реалізована на основі Internet/Intranet-технології з елементами безпосередньої взаємодії в локальній мережі Windows. Варіант узагальненої апаратно-програмної структури ІС представлений на рис. 6.

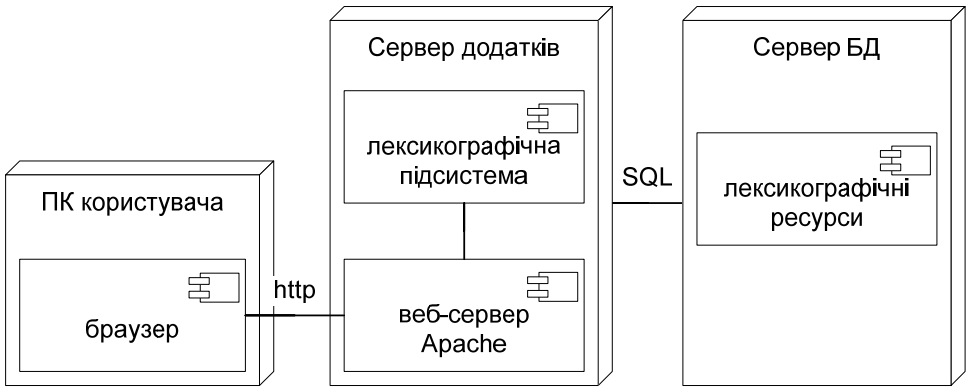

Рисунок 6 – Апаратно-програмна структура ІС

Функціонування ІС забезпечується програмним рішенням, яке реалізує середовище зберігання і обробки даних, інтерфейс доступу до даних і оболонку роботи з ними. Середовищем зберігання даних є мережева система управління базами даних Microsoft SQL Server. Інтерфейс обміну реалізується з використанням web-сервера Apache. Така побудова системи дозволяє:

- забезпечити простий доступ до неї засобами Internet/Intranet;
- істотно скоротити витрати на експлуатацію ІС за рахунок виключення необхідності встановлення і супроводу клієнтських робочих місць (для забезпечення роботи будь-якого робочого місця необхідно встановити тільки браузер);
- застосувати єдину технологію управління серверами і організації процедур резервного збереження/відновлення даних;
- забезпечити масштабованість і переносимість системи.

У межах даної статті не розглядається, але слід зауважити, що також окремо виділяється [10]:

1. *Організаційне забезпечення* ІС як сукупність документів, які встановлюють організаційну структуру, права і обов'язки користувачів в умовах функціонування, перевірки та забезпечення працездатності ІС.

2. *Методичне забезпечення* ІС – сукупність документів, що описують технологію функціонування ІС, методи вибору і використання користувачами технологічних прийомів для отримання конкретних результатів при функціонуванні ІС.

3. *Технічне забезпечення* ІС – сукупність всіх технічних засобів, що використовуються при функціонуванні ІС.

4. *Лінгвістичне забезпечення* ІС – сукупність засобів і правил для формалізації природної мови, що використовуються при спілкуванні користувачів ІС з комплексом засобів автоматизації при функціонуванні ІС. До складу лінгвістичного забезпечення входять:

- інформаційні мови для опису структурних одиниць інформаційної бази ІС;
- мови управління і маніпулювання даними інформаційної бази ІС;
- мовні засоби інформаційно-пошукових систем;
- система термінів і визначень, що використовується у процесі розробки і функціонування ІС тощо.

5. *Правове забезпечення* ІС – сукупність правових норм, які регламентують правові відношення при функціонуванні ІС та юридичний статус результатів її функціонування.

6. *Ергономічне забезпечення* ІС – сукупність реалізованих в ІС рішень щодо погодження психологічних, психофізіологічних, антропометричних, фізіологічних характеристик і можливостей користувачів ІС з техніч-

ними характеристиками комплексу засобів автоматизації ІС і параметрами робочого середовища на робочих місцях користувачів ІС.

ОСНОВНІ РЕЗУЛЬТАТИ ТА ВИСНОВКИ

ІС автоматизованого створення лексикографічних ресурсів є системою, яка забезпечує задоволення інформаційних потреб користувача щодо обробки лексикографічної інформації, а також лексикографічної обробки інформації. Призначення ІС реалізується через її функції: автоматизоване створення лексикографічних ресурсів різного призначення, автоматизований збір, обробку, зберігання лексикографічної інформації, інформаційну підтримку користувачів та ін.

Архітектура ІС включає бази даних, засоби обробки інформації, засоби доступу до інформаційних ресурсів, засоби організації роботи користувачів, засоби адміністрування, засоби транспортування інформації. Програмне забезпечення ІС є набором окремих компонент. Кожна компонента реалізує сукупність тісно взаємопов'язаних задач, що забезпечує реалізацію необхідного набору операцій над даними та знаннями і послідовності їх виконання.

Функціонування ІС забезпечується програмою, яка є сукупністю програмних засобів, що реалізують середовище зберігання і обробки даних, інтерфейс доступу до даних і оболонку роботи з ними. Середовищем зберігання даних є мережева система управління базами даних Microsoft SQL Server. Інтерфейс обміну реалізується з використанням web-сервера Apache. Оболонка роботи з даними представляє собою програму, що реалізує основні функції системи управління даними. Програма працює через web-інтерфейс.

Застосування ІС за переставленою архітектурою забезпечуватиме оптимальну організацію роботи користувачів з інформаційними ресурсами ІС.

ЛІТЕРАТУРА:

1. Bereza A.M. Osnovi stvorennya Informatslynih sistem: Navchalniy poslbnik / A. M. Bereza. – 2-ge vidannya, pereroblene I dopovnene. – K.: KNEU, 2011. – 205 s.
2. Borisova N.V. Formirovanie slovarya sochetaemosti terminov predmetnoy oblasti / N. V. Borisova, O. V. Kanischeva // Shldno-Evropeyskiy zhurnal peredovih tehnologly. – Harklv : PP «Tehnologlchniy Tsentr», 2013. – #5(65). – S. 16-19
3. Borisova N.V. Avtomatizirovannoe formirovanie slovarya sinonimov / N. V. Borisova, O. V. Kanischeva, E. N. Yurchenko // Vestnik HNTU. – Herson : HNTU, 2013. – # 1(44). – C. 91-95
4. Borisova N.V. Avtomatizovane vidobuvannya termlnologlchnih odinits z naukovo-tehnlchnih tekstlv / N. V. Borisova, S. S. Reshetilo // Materlali III VseukraYinskoYi naukovo-praktichnoYi konferentslYi "Intelektualnl sistemi ta prikladna llngvlstika" (m. Harklv, 17 kvltnya 2014 r.). – Harklv : NTU "HPI", 2014. – S. 43
5. Gaydamakin N.A. Avtomatizirovannyie informatsionnyie sistemyi, bazyi i banki dannyih: Uchebnoe posobie / N. A. Gaydamakin. – M.: Gelios ARV, 2012. – 368 s.
6. Kochueva Z. A. Indeksirovanie polnotekstovyih dokumentov dlya zadachi intellektualnogo poiska infor-matsii po klyuchevyim slovam / Z. A. Kochueva, N. V. Borisova // Shldno-Evropeyskiy zhurnal peredovih teh-nologly. – Harklv : PP «Tehnologlchniy Tsentr», 2014. – # 1/2 (67). – S. 4-8.
7. Makkonnell S. Doskonaliy kod / S. Makkonel. – S-Pb: Piter, 2005. – 896 s.
8. Petrenko M.G. Metodi ta zasobi pobudovi znannya-orlEntovanih komp'yuternih sistem z ontologo-kerovanoyu arhltekturoyu : avtoref. dis. na zdobuttya nauk. stup. doktora tehn. nauk: spets 05.13.05 – komp'yuternl sistemi ta komponenti / M. G. Petrenko. – KiYiv : Institut klbernetiki lm. V.M. Glushkova, 2014. – 40 s.
9. Tomashevskiy O.M. Informatslynl tehnologlYi ta modelyuvannya blznes-protseslv / O. M. Tomashevskiy: [Elekt-ronniy resurs]. – [Rezhim dostupu] : http://pidruchniki.ws/13601004/informatika/ informatsiyni_tehnologiyi_ta_modelyuvannya_biznes-protsesiv_-_tomashevskiy_om
10. Yamshanov I.S. Opisanie informatsionnoy tehnologii otrabotki na tehnologichnost strukturyi ob'ektov sborochnogo protsessa / I.S. Yamshanov // Radloelektronnl I komp'yuternl sistemi. Naukovo-tehnlchniy zhurnal. – Harklv : HAI, 2010. – #3 (44) lipen – veresen. – S. 135–140.
11. Yamshanov I.S., Gorchenok O.V. Razrabotka mnogourovnevoy sistemyi otsenki elektronnoy korrespondentsii na prinadlezhnost ee k spamu s primeneniem algoritma skoringa / I. S. Yamshanov, O. V. Gorchenok // Tezi dopovldey UnlversitetskoYi naukovo-praktichnoYi studentskoYi konferentslYi maglstrantlv. – Harklv : NTU «HPI», 2007 : [Elektronniy resurs]. – [Rezhim dostupu] : http://www.kpi.kharkov.ua/archive/Conferences/ /2007/S1.pdf