УДК 004.415

В.Ф. Запольський, магістр
А.М. Ковальчук, к.т.н., доц.
Житомирський державний технологічний університет

РОЗРОБКА ІНФОРМАЦІЙНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ СИСТЕМИ АВТОМАТИЗОВАНОГО ПОШУКУ ТА ОБРОБКИ ІНФОРМАЦІЇ У МЕРЕЖІ ІНТЕРНЕТ

Проведено аналіз основних існуючих систем пошуку інформації та розглянуто класичну архітектуру пошукової системи. Сформульовано основні вимоги щодо пошуку понятійно-чисельної інформації, сформовано концепцію пошукової системи понятійно-чисельної інформації, розроблено архітектуру системи та основні її компоненти. Визначено взаємозв'язки між підсистемами та побудовано повну функціональну схему системи. Проаналізовано математичне забезпечення, придатне для розв'язку задач категорізації та фільтрації інформації, а також розроблено алгоритм фільтрації для понятійно-пошукової машини.

Постановка проблеми. Інформаційні ресурси мережі Інтернет на даний час складають більше десяти мільярдів документів (Web-сторінок), які є загальнодоступними звичайному користувачу мережі. Зростання кількості послуг мережі Інтернет в усьому світі швидко набирає темпи, подекуди витісняючи традиційні інформативні послуги з різних галузей. Станом на 2004 рік кількість повідомлень у розділі новин, наприклад, складає понад мільйон на добу. Для того, щоб знайти потрібну інформацію у цьому, напевно, найбільшому розподіленому масиві даних доводиться використовувати потужні інформаційно-пошукові системи (ІПС). Однак існуючі найбільші мережеві інтегратори новин встигають обробити всього декілька десятків тисяч повідомлень за добу, а ІПС – близько 35000. Дана ситуація призводить до цікавого феномену мережі Інтернет: інформації в Інтернет стає все більше, проте віднайти потрібну інформацію з часом все важче і важче. Явище різкого росту об'ємів інформаційних мас породило нові та підвищило актуальність існуючих проблем, а саме: непропорційний ріст інформаційного шуму (звісно, в бік збільшення), багатократне дублювання інформації, виникнення паразитуючої інформації, внаслідок чого, при використанні традиційних методів, втрачається швидкість та зменшується якість пошуку. Як результат, виникає проблема розробки та впровадження нових технологій та методів пошуку інформації. Розробка нових прийомів та методів обробки даних з глобальної мережі Інтернет, здатних ефективно працювати в нових реаліях часу, спроможних давати раду існуючим проблемам, зокрема проблемі виділення потрібної інформації, ігнорування дублювань та паразитної інформації є актуальною задачею.

Мета і задачі роботи. Метою роботи є розробка інформаційного забезпечення для автоматизації пошуку, обробки, виділення потрібної інформації з результатів пошуку традиційних ІПС. Основне спрямування досліджень на алгоритмізацію та реалізацію процедур пошуку інформації стосовно глобальної світової мережі Інтернет, використання апарату математичної статистики для аналізу результатів пошуку. Результат пошуку подається у стиснутому фіксованому форматі, що відповідає вимогам адміністратора запиту, тобто результат, що не потребує додаткових перетворень.

Серед цілей роботи потрібно відзначити такі:

– автоматизація пошуку інформації з використанням існуючих ІПС;

– підвищення швидкості та якості пошуку потрібної інформації серед результатів пошуку, що виконаний ІПС;

– зменшення присутності людини-оператора в процесі пошуку та відбору потрібної інформації.

Для досягнення поставленої мети було виділено та сформульовано такі задачі:

– розробка теоретичних засад системи: межі функціональності, математичний апарат;

– реалізація математичного апарату системи відносно понятійно-числової категорії інформації;

– розробка архітектури програмної системи;

– розробка середовища зберігання отриманих звітів з метою повторного використання;

– реалізація підсистеми статистичної обробки інформації та підсистеми формуванням звітів.

Наукова новизна роботи. Розробка засобів автоматизації пошуку інформації та методів обробки результатів пошуку інформації виконується з урахуванням необхідності зменшення ступеню присутності людини в процесі пошуку та обробки даних. Наукова новизна роботи сформульована у таких пунктах:

– модифіковано метод Бейєсівської фільтрації як засобу доведення змістовної єдності та аналізу елементів мовних конструкцій;

– розроблено структуру даних для зберігання понятійно-числової категорії інформації із можливістю здійснення моніторингових поповнень.

Практична цінність роботи:

– Можливість використання розробленої системи для пошуку понятійно-числової категорії інформації.

– Пошук та збереження поодиноких понятійних величин та формування статистичних оцінок.

– Високий рівень автоматизації системи, що призводить до необхідності втручання користувача в роботу системи тільки на стадії формування параметрів пошуку та налагоджування системи фільтрації.

Аналіз досліджень і публікацій. Серед ІПС зарубіжжя [1] потрібно, в першу чергу, відзначити загальновідомі системи SemioMap, Northern Light Technology, Oracle Text, SAS Text Miner, MARRI. Далі детальніше розглянемо архітектурні особливості побудови систем та технологій, які покладено в основу механізмів обробки інформації.

Робота системи SemioMap відбувається в декілька етапів: індексування неструктурованого тексту; кластеризація понять та побудова лексичної мережі; візуалізація карт зв'язків.

Одним з перших серйозних інтеграторів новин у мережі Інтернет стала служба Northern Light Technology. Нею створена і постійно поповнюється "спеціальна колекція", що включає статті з більш ніж 7000 джерел. Northern Light Technology вважається одним з найбільших в Інтернеті пошуковим механізмом.

Oracle Text – це програмний комплекс, інтегрований у систему керування базами даних, що дозволяє ефективно працювати як з запитами, що відносяться до неструктурованих текстів, так і з запитами до реляційних баз даних на мові SQL. Основним завданням Oracle Text є задача пошуку документів за змістом, словами чи фразами.

Програма Text Miner дозволяє визначати, наскільки правдивий той чи інший текстовий документ. Для пошуку таких змін використовується принцип, що полягає в пошуку аномалій та трендів серед записів баз даних, текстів без з'ясування їхнього змісту.

Система MARRI розроблена для пошуку Web-сторінок, релевантних запитам у визначеній предметній області. Для розв'язання поставлених задач система використовує знання, представлені як безліч концептів та зв'язків між ними. Базисне припущення розроблювачів полягає в тому, що релевантні тексти складаються зі значимих для предметної області пропозицій, що містять фрагменти, "порівняні" з онтологією предметної області.

Серед вітчизняних систем [2–4] потрібно відзначити систему, що базується на потоковій технології інтеграції новин InfoStream, розроблену в інформаційному центрі "ЭЛВИСТИ". Система, побудована на базі цієї технології, включає три типові складові: stream-центри збору й обробки інформації, stream-центр інтерактивного доступу до інформаційних баз даних і stream-центр моніторингу змісту інформації.

Розробка архітектури системи автоматизованого пошуку та обробки інформації у мережі Інтернет. При проектуванні архітектури системи враховано вимогу того, що система автоматизованого пошуку та обробки інформації у мережі Інтернет повинна містити підсистеми пошуку, обробки, зберігання та інтерфейсу з такими властивостями:

– підсистема пошуку інформації повинна мати засоби використання існуючих ІПС як систем швидкого та якісного завантаження інформації з мережі Інтернет;

– підсистема обробки інформації повинна мати засоби статистичної фільтрації потрібної користувачеві понятійно-числової інформації;

– підсистема зберігання повинна компактно зберігати інформацію; забезпечувати високу швидкість доступу до інформаційних масивів.

Як методи статистичної фільтрації підсистеми обробки даних запропоновано використати методику бейєсівської фільтрації, яка полягає у використанні умовних статистичних даних як критерію визначення оцінки умовності щодо довільних даних. Використовуючи властивості "природних" текстів, перш за все, їх умовнофіксовану форму, можна припустити одноманітність їх змісту за умови подібності форм, тобто те, що подібні тексти мають подібні статистичні дані, в першу чергу, умовні статистичні дані.

Структурна схема системи. На структурній схемі (рис. 1) зображено загальний вигляд структурних елементів системи пошуку, що розробляється. Дана схема відображає ключові елементи системи та акцентує увагу на направленості зв'язків між елементами.

Фактично, вся система складається з чотирьох підсистем досить значних за об'ємами виконуваної роботи. Підсистема управління виконує пошук інформації в Інтернеті за допомогою зовнішніх ІПС, використовуючи керуючу інформацію з бази даних настройок. Пошук документів в Інтернеті відбувається з використанням існуючих технологічних засобів (огляд характеристик та способів пошуку існуючих ІПС в даній роботі не проводиться). Результати пошуку перенаправляються до підсистеми аналізу даних. Вона виконує глибинний аналіз текстів згідно з додатковими умовами, які зберігаються в базі параметрів пошуку. Віднайдена понятійно-числова категорія інформації зберігається за допомогою підсистеми зберігання результату.

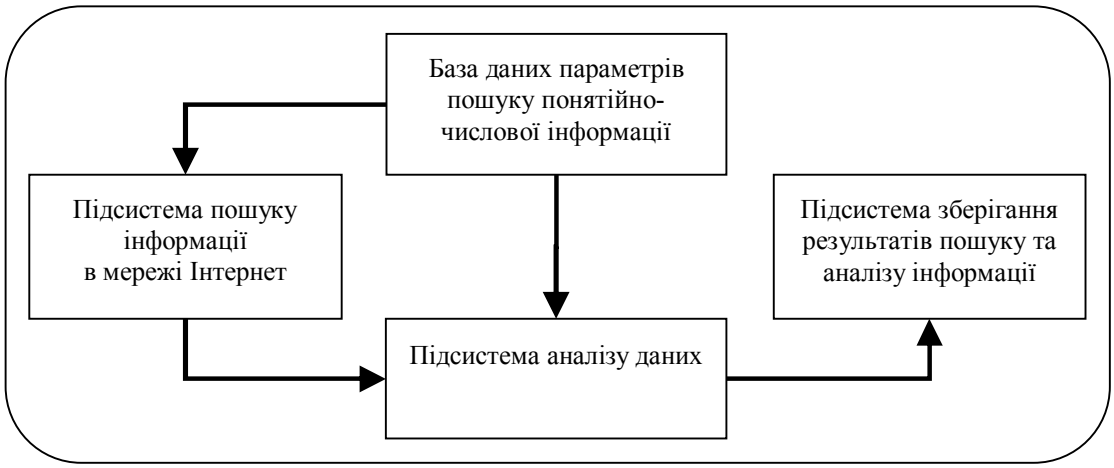

Рис. 1. Структурна схема ядра системи пошуку та обробки даних з мережі Інтернет

Функціональна схема системи. Головне завдання функціональної схеми (рис. 2) полягає не тільки у більш детальному зображенні ключових елементів розробки, але й у розкритті сутності зв'язків між ними.

За логікою функціонування системи, основним елементом, який дає можливість користувачеві спілкуватися з програмним забезпеченням, є підсистема інтерфейсу з користувачем. З рис. 2 можна зробити висновок про те, що функціональна здатність підсистеми інтерфейсу полягає у такому:

– отримання параметрів пошукового запиту від користувача;

– попередня обробка, верифікація та трансляція параметрів пошукового запиту до системи управління;

– формування та відображення звітів про результати роботи.

Підсистема управління, отримавши дані від користувача, остаточно формує та зберігає параметри пошукового запиту. З певною регулярністю ця підсистема проводить моніторинг бази параметрів пошукових запитів, обробляє їх та, власне, ініціює подальший пошук і фільтрацію інформації. Вище згадані дії підсистема управління виконує шляхом впливу на інші підсистеми. В основному, цей вплив має бінарний характер – ввімкнути/вимкнути, – але які саме підсистеми ввімкнути, а які вимкнути підсистема управління вирішує на основі параметрів запиту користувача. Так, наприклад, активна дія з боку користувача щодо створення запиту, тобто оформлення предмету пошуку, повинна призвести до активації підсистем інтерфейсу, зберігання параметрів; запит на пошук (мається на увазі дія) повинен призвести до активації підсистем пошуку, аналізу, зберігання результатів пошуку; запит щодо формування звіту, ініційований підсистемою управління або користувачем, повинен призвести до активації підсистем зберігання та інтерфейсу.

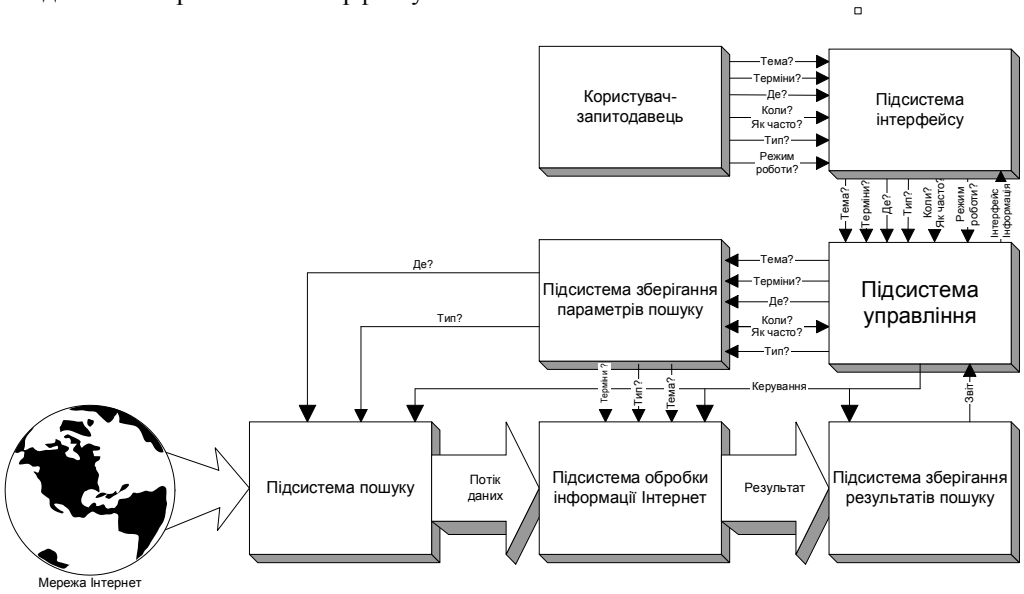

Рис. 2. Функціональна схема системи пошуку та обробки інформації в Інтернет

З вище сказаного зрозуміло, що підсистема зберігання параметрів пошуку необхідна для зберігання форм запитів, які надалі будуть основою для виконання пошуку. Параметри пошуку, які заносяться до бази даних системи, повністю формуються користувачем на етапі створення запиту. Зокрема в цій базі містяться відомості про час виконання активації засобів пошуку, оброки та зберігання інформації, які необхідні для підсистеми управління.

Підсистема пошуку інформації в мережі Інтернет адрес Інтернет-ресурсів, типи документів отримує від підсистеми зберігання параметрів та відповідно до них виконує пошук та завантаження даних, які прямують до підсистеми обробки інформації. Як підсистема пошуку може бути використана будь-яка з існуючих ІПС, що надає власні ресурси для виконання автоматизованого пошуку.

Підсистеми обробки інформації виконують паралельно-послідовну обробку інформації з застосуванням декількох етапів фільтрації. Сумарний результат фільтрації аналізується та на його основі робиться висновок про подальшу долю завантаженого документа. У випадку, коли документ дійсно містить корисну інформацію, вона зберігається у базі даних підсистеми зберігання результатів пошуку. В фіналі циклу пошуку та обробки підсистема управління отримує відгук про поновлення бази результатів пошуку.

Формати обміну даних між складовими системи. На вхід підсистеми пошуку потрапляє перелік адрес, за яким буде виконуватися пошук. Адреса має бути представлена у вигляді доменного імені. Далі підсистема формує запит до пошукової системи на основі доменного імені та параметрів пошуку. Результатами пошуку є завантажені з мережі Інтернет документи у текстовому форматі, які зберігаються у вигляді текстових файлів. Текстові файли мають ANSI кодування, аналогічне кодовим сторінкам Windows 1250 та 1252. Тому, фактично, підсистеми системи пошуку та обробки інформації мережі Інтернет ведуть обмін даними у текстовому форматі.

Підсистема зберігання параметрів пошуку здатна сприймати SQL-запити на пошук інформації, а також SQL-команди поповнення бази даних. У відповідь підсистема зберігання параметрів пошуку, також на мові SQL, повертає запитувані структури даних.

Підсистема аналізу обробляє текстові файли, які створені підсистемою пошуку інформації, результати їх аналізу представляються у вигляді масиву об’єктів спеціального вигляду, які, в свою чергу, складаються з полів, що зазначені в базі даних зберігання результату. Основні поля такі: ім’я терміну, його числова величина, назва одиниць виміру та час знаходження числової величини.

Підсистема зберігання результату, окрім сприймання попередньо визначеного формату даних (застосовується для поновлення та зберігання), розуміє також мову запитів SQL для виконання пошукових робіт. Результати запиту повертаються у вигляді розширеного запису, який включає усі поля, що означені в базі даних зберігання результату.

Процес обміну з підсистемою інтерфейсу протікає за стандартами html-конструкцій як в одному, так і в іншому напрямі.

Підсистеми зберігання параметрів та результатів пошуку. Під час проектування бази настройок виконується розробка концепції бази зберігання параметрів пошуку. Зроблено вибір в бік реляційних баз даних, враховуючи достатньо високу швидкодію обробки запитів, що є дуже суттєвим, оскільки передбачається часте звертання до цієї бази. База настройок складається з 15 таблиць, 6 з яких зберігають буквенно-числову інформацію, тобто конкретні дані. Інші створені для того, щоб підтримувати концепцію зв’язків між таблицями, тобто забезпечують єдність і несперечливість бази даних. Більша частина таблиць зберігання даних стосується параметрів запиту.

Специфіка бази зберігання результатів пошуку полягає в незвичайному форматі даних та їх загальної структури, які необхідно зберігати. Тут розглянуті властивості понятійно-чисельних структур та знайдено компроміс щодо зберігання таких структур у реляційних базах даних. База зберігання результатів складається з 6 таблиць, 4 з яких повністю відповідають таблицям із бази настройок. Для відсутності дублювань створено зв’язок бази зберігання результатів пошуку із базою настройок.

Розробка підсистеми статистичної оцінки інформації. В роботі детально проаналізовано теореми Бейєса та принцип критерію максимальної правдоподібності. Можливість використання формул Бейєса для виділення понятійно-числової категорії інформації полягає в тому, що існує можливість розрахунку ступеня достовірності існуючих даних до понятійно-чисельної категорії, а критерію максимальної правдоподібності – в точковій остаточній достовірності такого припущення відносно змісту ланцюжка ключових слів. Теоретичні основи для обґрунтування можливості використання статистичних методів представлено в [5].

Далі наведено запропоновану реалізацію підсистеми статистичної оцінки інформації у вигляді мовного опису алгоритму обробки вхідної інформації.

Користувач формує зв’язані ланцюжки ключових слів (понять, величин, одиниць виміру, дати, часу, місця народження інформації), які характеризуються відношенням до пріоритетності до інформації щодо якої виконується пошук. Ланцюжки ключових слів згуртовуються у відповідні словники.

При автоматичній обробці формування словника взаємопов'язаних ключових слів виконується в кілька етапів, під час яких формується умовна статистична картина відносно ключових слів, яка потім виконує керуючу функцію при опрацюванні невідомих текстів. Виконується пошук слів у тексті, які відносяться до словника пріоритетних слів D_1, з яких будуються ланцюжки ключових слів. Недобудовані ланцюжки або ланцюжки, елементи яких не пройшли фільтрацію, руйнуються, а їх елементи втрачають статус ключового слова. Фільтрація відбувається на основі порівняння частот зустрічності ключового слова, яке тестується, в межах абзацу з його аналогом зі словників, а також частот навколишніх слів і загальна картина частот усіх слів абзацу.

Формування словника виконується в кілька етапів, під час яких формується умовна статистична картина відносно ключових слів, яка потім виконує керуючу функцію при опрацюванні невідомих текстів. Завершальний етап доказу виконується за методом максимальної правдоподібності. Маємо закон розподілу слів абзацу, в якому знайдено ланцюжок ключових слів, для якого відомий статус "відповідає запиту" (належить до словника D_1) та ланцюжок ключових слів в "невідомому" абзаці (належить до словника D_3). Використовуючи останню вибірку, виконуємо оцінку "невідомого" параметра закону розподілу вибірки з відомим статусом. Тут (тимчасово) робиться припущення, що параметр закону розподілу невідомий. Отриману оцінку порівнюємо зі справжнім значенням параметра та виконуємо висновок щодо подібності "невідомої" вибірки з відомою. В ролі "невідомого" параметра було вибрано математичне очікування $M[P(\omega)]$. Запишемо функцію правдоподібності:

$$l\{D_3 \mid M[P(D_3)]\} = \prod_{i=1}^{n} P(S_i)\{S_i \mid M[P(D_3)]\} = (2\pi)^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(S_i - M[P(D_3)])^2},$$

де S – це поточне, щойно знайдене в тексті, слово; D – відповідно до словника згуртовані вибірки слів.

Часто рекомендують виконувати розрахунок в логарифмічній системі; використавши операцію логарифмування, отримуємо:

$$l\{D_3 \mid M[P(D_3)]\} = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(S_i - M[P(D_3)])^2$$

звідки, після диференціювання, матимемо:

$$\frac{d}{dM[P(D_3)]}\{\ln(l\{D_3 \mid M[P(D_3)]\})\} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(S_i - M[P(D_3)]).$$

Згідно з методикою пошуку екстремуму, яка входить до методу максимальної правдоподібності, припускаємо, що обидві частини рівняння дорівнюють нулю. В результаті оцінка максимальної правдоподібності буде мати вигляд:

$$\hat{M}[P(D_3)] = \frac{\sum_{i=1}^{n} S_i}{n} = \overline{S}.$$

Jul

OPEC oil prices reunified at $12.70 per barrel as Saudi Arabia and UAE fall into line, then official oil price rises to $13.66 per barrel.

Oil prices began dropping this week after U.S. government data showed that supplies of crude oil and gasoline are growing. Oil prices fell by nearly $2 a barrel after the International Energy Agency forecast slower demand growth this year.

The U.S. industry's price has been heavily regulated through production or price controls throughout much of the twentieth century. In the post World War II era oil prices have averaged $19.61 per barrel adjusted for inflation in 2000 dollars. Through the same period the median price for domestic crude oil was $15.25 in 2000 prices. That means that only fifty percent of the time from 1947 to 2003 have oil prices exceeded $15.25 per barrel. Until the March 28, 2000 adoption of the $22-$28 price band for the OPEC basket of crude, oil prices only exceeded $22.00 per barrel in response to war or conflict in the Middle East.

Over the same period world oil prices averaged $1.51 higher at $21.12 per barrel. The median world oil price of $15.89 was only slightly higher than the U.S. median of $15.25. (See note in box on right.)Crude Oil Prices 1947-2003

The very long term view is much the same. Since 1869 US crude oil prices adjusted for inflation have averaged $18.43 per barrel compared to $19.20 for world oil prices. Fifty percent of the time prices were U.S. and world prices were below the median oil price of $15.28 per barrel.

From 1974 to 1978 world crude oil prices were relatively flat ranging from $12.21 per barrel to $13.55 per barrel. When adjusted for inflation the prices were constant over this period of time.

Events in Iran and Iraq led to another round of crude oil price increases in 1979 and 1980. The Iranian revolution resulted in the loss of 2 to 2.5 million barrels of oil per day between November of 1978 and June of 1979. In 1980 as a result of the Iran/Iraq War, Iraq's crude oil production fell 2.7 MMBPD and Iran's production fell 600,000 barrels per day. The combination of these two events resulted in crude oil prices more than doubling from $14 in 1978 to $35 per barrel in 1981. U.S. and World Events and Oil Prices 1973-1981

...

Рис. 3. Приклад фрагменту обробленого тексту

Переглянувши початкові тексти та результати їх аналізу, можна зробити висновок про те, що понятійно-пошукова машина знаходить далеко не всі потрібні значення, але якість її роботи, як для системи-прототипу цілком задовольняє розробників.

Результати роботи. Для перевірки можливостей запропонованих технологій обробки текстової інформації та виділення постійної числової категорії інформації було проведено низку тестів. Один з результатів, як типовий та характерний, наведемо далі. Для перевірки було проведено тестування розробленої програмної системи на предмет визначення вартості нафти з інформаційних повідомлень мережі Інтернет. Тексти були попередньо переглянуті адміністратором запиту та поділено на такі, що містять потрібну інформацію, та такі, що її не містять. Далі на вхід понятійно-пошукової машини подавалися тести для аналізу та виділення з них потрібних даних. На рис. 3 показано приклад фрагменту вхідного тексту, а в таблиці 1 – приклад фрагменту результату аналізу.

Таблиця 1

Приклад фрагменту результату аналізу

№	Назва	Поняття	Значення	Од. виміру	Дата	Джерело інформації
1	oil	prices	12.70	$ per barrel	Jul	Saudi Arabia
2	oil	prices	13.66	$ per barrel		
3	oil	prices	2	$ a barrel		The International Energy Agency
4	oil	prices	19.61	$ per barrel		
5	oil	prices	15.25		March 28	
6	oil	prices	22.00			
7	oil	prices	21.12			
8	oil	prices	18.43			

Висновки. В ході даної роботи проаналізовано основні сучасні системи пошуку текстової інформації, розглянуто класичну архітектуру пошукової системи, основні способи та формати зберігання різноманітних даних. Сформульовано основні вимоги щодо пошуку понятійно-чисельної інформації, окреслено середовище та концепцію виконання системи пошуку, розроблено структурну схему системи та основні її компоненти, визначено спектр задач, що стоїть перед системою пошуку, та виконано розподіл його по підсистемах. Визначено взаємозв’язки між підсистемами та побудовано повну

функціональну схему системи. Розроблено математичні засоби щодо розв'язку задач категорізації та фільтрації інформації, а також виконано пристосування алгоритмів фільтрації до проблем реалізації понятійно-пошукової машини.

ЛІТЕРАТУРА:

1. *Гаврилов П.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. – СПб: Питер, 2001. – 384 с., ил.
2. *Ландэ Д.В.* Добыча знаний // СНІР. – № 10. – 2003. – С. 76–82.
3. *Ландэ Д.В.* Интернет для людей // Бизнес регистр. – Київ, 2002. – № 5 (11). – С. 3–5.
4. http://www.visti_InfoStream.net/~dwl/art/iua/index.html
5. *Джонсон Н., Лион Ф.* Статистика и планирование эксперимента в технике и науке: Методы обработки данных: Пер. с англ. / Под ред. Э.К. Лецкий – М.: Мир, 1980. – 610 с.

ЗАПОЛЬСЬКИЙ Владислав Францович – магістр Житомирського державного технологічного університету.

Наукові інтереси:
– комп'ютерні інформаційні технології;
– архітектура програмних систем;
– системи автоматизації інтелектуальної обробки інформації.

КОВАЛЬЧУК Андрій Михайлович – кандидат технічних наук, доцент, доцент кафедри АіКТ Житомирського державного технологічного університету.

Наукові інтереси:
– комп'ютерні інформаційні технології;
– архітектура програмних систем;
– графічні системи та візуалізація даних;
– програмне забезпечення математичного моделювання технічних та природничих систем;
– використання обчислювальної техніки в навчальному процесі.

Подано15.09.2005

Запольський В.Ф., Ковальчук А.М. Розробка інформаційного забезпечення для системи автоматизованого пошуку та обробки інформації у мережі Інтернет.

Запольський В.Ф., Ковальчук А.М. Разработка информационного обеспечения для системы автоматизированного поиска и обработки информации в сети Интернет.

Zapolskyy V.F., Kovalchuk A.M. Software design for the automated information search and analysis in the Internet.

УДК 004.415

Разработка информационного обеспечения для системы автоматизированного поиска и обработки информации в сети Интернет / Запольський В.Ф., Ковальчук А.М.

В работе проведен анализ основных существующих систем поиска информации, рассмотрена классическая архитектура поисковой системы. Сформулированы основные требования относительно поиска понятийно-численной информации, сформирована концепция поисковой системы понятийно-численной информации, разработана архитектура системы и основные ее компоненты. Определенно взаимосвязи между подсистемами и построена полная функциональная схема системы. Проанализировано математическое обеспечение пригодное для решения задач категоризации и фильтрации информации, а также разработан алгоритм фильтрации для понятийно-поисковой машины.

УДК 004.415

Software design for the automated information search and analysis in the Internet / Zapolskyy V.F., Kovalchuk A.M.

The analysis of the basic existent systems of information retrieval and architecture of the searching system is considered. The basic requirements are formulated in relation to the concept-numeral information retrieval. The conception of the searching system of concept-numeral information and architecture of the system is developed. The complete functional diagram of the system and the procedure of intercommunications between subsystems is created. The algorithm of categorization and filtration of information is developed for a concept-searching machine.