УДК 004.738.52

Ю.В.Рогушина, І.Ю.Гришанова

ПОШУК МУЛЬТИМЕДІЙНИХ ОБ'ЄКТІВ ЗА КОНТЕКСТОМ ТА МЕТАОПИСАМИ ГІПЕРМЕДІЙНИХ ІНФОРМАЦІЙНИХ РЕСУРСІВ

Аналізуються типи та формати подання мультимедійної інформації, її розташування в гіпермедійних ресурсах. Розглядаються питання, пов'язані з пошуком мультимедійних об'єктів, поданих в Інтернеті, на основі аналізу контексту і метаописів інформаційних ресурсів, що містять посилання на такі об'єкти. Для підвищення релевантності пошуку пропонується порівнювати контекст і метаописи мультимедійних об'єктів з онтологією предметної області, що цікавить користувача.

Вступ

Значна частина інформаційних ресурсів (ІР), доступ до яких забезпечує глобальна мережа Інтернету, містить елементи мультимедіа. Такі ресурси називаються гіпермедійними. Гіпермедіа – формат даних, подібний до гіпертексту, у якому текст, звук, зображення або інші об'єкти, пов'язані з інформацією, відображуваною на екрані, може бути виведений на дисплей за допомогою гіпертекстового посилання [1]. Пошук мультимедійної інформації являє собою серйозну теоретичну проблему, яка ускладнюється різнорідністю представленої інформації, розходженнями у форматах та стандартах її подання.

Аналіз наукових публікацій показує, що світове наукове співтовариство вважає пошук мультимедіа важливою проблемою сьогодення і приділяє їй значну увагу. ACM Multimedia Special Interest Group [2] займається, зокрема, дослідженням імен мультимедіа і пошуком, збереженням і використанням мультимедійних інформаційних ресурсів (MIP). Moving Picture Experts Group розробляє стандарти подання та опису аудіовізуальної інформації. Synchronized Multimedia Working Group Консорціуму W3C пропонує механізм створення документів, що містять синхронізовану мультимедійну інформацію SMIL [3].

Користувач, що шукає мультимедіа, зазвичай не може навести фрагмент потрібної інформації у формі, придатній для автоматичної обробки. Параметри опису мультимедіа вкрай різноманітні та важко формалізовані й у значній мірі суб'єктивні.

Постановка задачі

Для того щоб забезпечити ефективний пошук мультимедіа, потрібно визначити, що саме є об'єктом пошуку; які параметри може задавати користувач для ідентифікації інформаційних потреб; які відомості про мультимедіа можна отримати внаслідок аналізу ІР та як порівнювати опис запиту з описом ІР.

1. Визначення мультимедіа

Термін "*мультимедіа*" має багато визначень, але в більшості випадків це поєднання текстової, звукової та відео інформації. Під мультимедіа розуміють електронний носій інформації, що включає кілька її видів (текст, зображення, анімація тощо [4]), інтегроване представлення інформації в кількох формах, наприклад, відео, голос, музика або дані [5]. Іноді під мультимедіа розуміють також потоки сигналів від віддаленого обладнання (телескопів, датчиків тощо).

Обробка мультимедіа засобами обчислювальної техніки почалася ще в 60-х роках, коли зображення та текст поєднували в одному документі. Дослідження мультимедіа базується на таких дисциплінах, як обробка сигналів, комп'ютерна графіка, бази даних, когнітивна психологія, інтерфейс користувача. Для подальшої роботи з аналізу мультимедіа введемо більш чітке поняття мультимедійного об'єкта. Це об'єкт, відтворений за допомогою комп'ютера і здатний впливати на одне або кілька людських почуттів – зір, слух тощо. Слід зазначити,

ISSN 1727-4907. Проблеми програмування. 2005. № 3

що основна частина інформації надходить через зір і слух, тому, розглядаючи мультимедіа, враховують вплив тільки на ці почуття (хоча в окремих випадках тактильна інформація також може відтворюватися за допомогою технічних пристроїв, наприклад, тактильного монітора).

Під *мультимедійним об'єктом* (МО) розумітимемо інформацію, подану в електронній формі, яка не є символьно-текстовою та може інтерпретуватися спеціалізованим програмним забезпеченням і периферійними пристроями виведення таким чином, щоб впливати на органи почуттів користувача.

Гіпермедійний інформаційний ресурс (ГІР) – це ІР, який містить МО або посилається на них (приміром, через гіперпосилання).

Контекст МО складається з контенту (вмісту) та метаопису ГІР, який містить посилання на цей МО, та метаопису самого МО. Метаописи як МО, так і ГІР можуть бути відсутні.

Пошук мультимедіа виконується існуючими інформаційно-пошуковими системами (ІПС) з низькою релевантністю. Важливо відмітити, що у випадку, коли пошук здійснюється у такому динамічному середовищі, як Інтернет (на відміну від, приміром, електронних бібліотек), кожен ресурс потрібно індексувати знову при звертанні до нього. Тому семантичний аналіз контента мультимедіа, що потребує великих обчислювальних ресурсів, не ефективен.

Існуючі ІПС індексують МО за контекстом та назвами, вони не враховують семантичну спрямованість ІР, які на них посилаються. Пошук здійснюється за ключовими словами.

Більш корисною для пошуку є обробка вмісту ГІР, у якому зустрічається посилання на МО в текстових документах (сучасні ІПС дозволяють знаходити html-документи, документи у форматі MS Word та PDF, презентації MS PowerPoint тощо), та метаописи як самого цього документа, так і мультимедіа. Щоб аналізувати цю інформацію, потрібно з'ясувати, які типи МО подані в Інтернеті, в яких структурних елементах html-сторінки розміщуються посилання на них, які формати та елементи використовуються для їх метаописів і знання про семантику МО здобуття з цієї інформації.

Мультимедійна інформація може бути подана у різних формах: у вигляді зображень, графіки, 3D-моделей, аудіо, відео тощо (рис.1). Загальноприйнятої класифікації МО на сьогодні немає, хоча цьому питанню приділяється значна увага. Оскільки МО різняться між собою за типом, фізичною сутністю і форматом, на семантичному рівні необхідно, абстрагуючись від фізичної сутності об'єктів, виділити характеристики, які є загальними для всіх ГІР або специфічними для кожного окремого типу.

Джерелами інформації, до яких система переадресує запит користувача, можуть бути локальними і глобальними ІПС. Надалі розглядається пошук МО тільки серед тих ІР Інтернету, доступ до яких здійснюється за протоколом http (а не ftp).

2. Стандарти подання та класифікація ГІР

Зазвичай користувач, який хоче найти певний МО, може визначити його тип (вибрати з набору запропонованих). Проаналізуємо найбільш поширені підходи до класифікації ГІР.

Стандарт подання мультимедійної інформації MPEG. Експертна група Moving Picture Experts Group Об'єднаного комітету зі стандартизації запропонувала сімейство стандартів подання мультимедійної інформації MPEG. MPEG-7 (Multimedia Content Description Interface – Інтерфейс для опису контента мультимедіа ISO/IEC) [6] – це стандарт, орієнтований на семантичне осмислення МО. Він базується на стандартах MPEG-1, MPEG-2 та MPEG-4 і використовує синтаксис XML Schema. Розробники MPEG-7 враховували інші стандарти (TV Anytime, Dublin Core [7], SMPTE Metadata Dictionary тощо), створені для досить вузьких доменів, і запропонували більш загальну модель. Зараз MPEG-7 доробля-

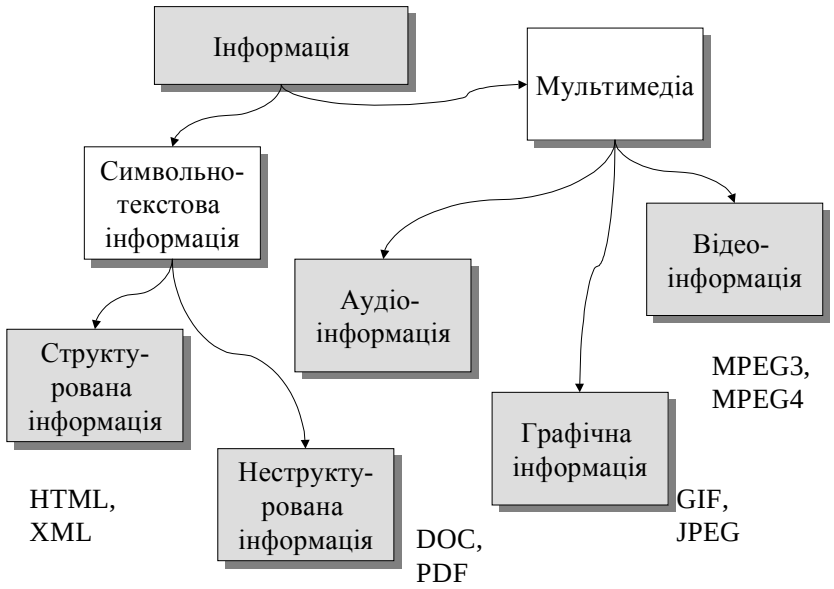

Рис. 1. Види інформаційних ресурсів Інтернету

ють у напрямку інтероперабельності з компонентами проекту Semantic Web.

MPEG-7 надає стандартизований опис різних типів мультимедійного матеріалу, підтримуючи певний рівень інтерпретації змісту інформації, що може бути передана для обробки ЕОМ, для забезпечення ефективного і швидкого пошуку. Він визначає стандартний набір дескрипторів для різних типів МО, стандартизує спосіб визначення своїх дескрипторів і їхнього взаємозв'язку, вводячи для цього *DDL* (Description Definition Language) – мову опису визначень.

Розробники MPEG-7 вважають аудіовізуальними даними статичні зображення, графіку, 3D моделі, аудіо, мову, відео і композиційну інформацію про те, як ці елементи комбінуються в мультимедійному поданні (сценарії). Засіб опису, запропонований MPEG-7, не залежить від того, яка інформація описується (аналоговий відеозапис чи оцифрований контент).

Описові можливості мають однозначно і повністю інтерпретуватися в контексті застосування, тому вони залежать від доменів користувачів і сфери застосування, тобто той самий матеріал може бути описаний через різні типи властивостей, що відповідають специфіці і можливостям застосування.

Приміром, графічне зображення на найнижчому рівні абстракції може бути описане через форму, розмір, текстуру, кольори, палітру, траєкторію руху та положення; а аудіо – через тональність, зміни темпу, положення в звуковому ряд, тоді як на верхньому рівні буде подана семантична інформація типу «*Це сцена з зеленим автомобілем, який їде дорогою, що знаходиться ліворуч, і людиною в білому, яка переходить дорогу праворуч, у супроводі фонового звуку дощу*». Можуть існувати також проміжні рівні абстракції. Рівень абстракції пов'язаний зі способом здобуття інформації: багато низькорівневих властивостей можуть бути витягнуті автоматично, тоді як високорівневі властивості вимагають втручання людини.

Крім опису контента, для пошуку МО можуть використовуватися й інші відомості про нього: *Форма* – формат кодування (JPEG, MPEG-2 і т.п.), розмір даних; *Умови доступу до матеріалу* – умови реєстрації, вартість доступу, угоди про права користування тощо; *Класифікація* – оцінка походження матеріалу і тип його вмісту (обмежена кількість категорій задається заздалегідь); *Посилання на інші релевантні матеріали*; *Контекст* – обставини, за яких створено матеріал.

Стандарт передачі інформації в Інтернеті MIME. Стандарт Multipurpose

Internet Mail Extensions (MIME) [8] призначений для передачі інформації в Інтернеті з урахуванням її типу. Цей стандарт виділяє сім базових типів медіаінформації, для кожного з яких визначені параметри, що характерні саме для цього типу інформації. У стандарті передбачена можливість розширення набору типів та їх параметрів.

Основні дискретні медіатипи MIME:

- *Текст* – текстова інформація. Підтип "plain" – це текст, що не містить команд форматування і не використовує спеціальне програмне забезпечення (ПЗ), за винятком підтримки зазначеної кодової таблиці. Інші підтипи існують для збагаченого тексту (enriched text), що використовують спеціальне ПЗ для показу тексту, однак основний вміст контента можна отримати без цього ПЗ;
- *Зображення* – вимагає для перегляду інформації пристрій візуалізації. Основний підтип – формат зображень JPEG (jpeg), що використовує кодування JFIF;
- *Аудіо* – потребує для інтерпретації аудіопристрій, який відтворює звукову інформацію. Основний підтип – "basic", контент якого – одноканальне аудіо, закодоване з використанням 8bit ISDN mu-law [PCM] з частотою 8000 Hz;
- *Відео* – зображення, що змінюється у часі і супроводжуються узгодженим звуком. Основний підтип – "mpeg", контент якого – відео, закодоване відповідно до стандарту MPEG;
- *Прикладні дані* – бінарні дані, які обробляє спеціалізоване ПЗ. Підтипи: "octet-stream" інтерпретується як бінарні дані, "PostScript" визначає PostScript-дані.

Вводяться також два композиційних базових медіатипа:

- *multipart* – це дані, що складаються з кількох частин, які представляють собою незалежні типи даних. Основним підтипом є "mixed", що визначає загальний комбінований набір частин, "alternative" використовується для даних у різних форматах, "parallel" – для частин, які необхідно переглядати одночасно, "digest" – для багаточасткових фрагментів, у яких кожна частина має тип "message/rfc822";
- *message* – інкапсульоване повідомлення.

Стандарт інтеграції синхронізованої мультимедійної інформації SMIL. *SMIL* (Synchronized Multimedia Integration Language) рекомендований консорціумом W3C (Synchronized Multimedia Working Group) стандарт, який описує механізм створення документів, що містять синхронізовану мультимедійну інформацію. SMIL-презентації являють собою набір інструкцій, що описують текстові, відео- і аудіодані і визначають послідовність їх відтворення.

За допомогою SMIL можна сполучати в html-документі з метою синхронізованого відтворення елементи таких типів:

- текст;
- нерухомі зображення (формати *jpeg, gif, png* тощо);
- відео (формати *mpeg, avi, mov* тощо);
- аудіо (формати *mp3, wav, au* тощо);
- анімацію, що використовує векторну графіку (*svg, swf, vml*, тощо);
- текстовий потік, синхронізований у часі з іншою інформацією (*sub, rt, sami* тощо), приміром, стрічка новин або титри.

Кожний медіаоб’єкт у SMIL має індивідуальне ім'я і використовує атрибут *src* для вказівки імені і місця розташування файлу, що містить медіа-інформацію.

SMIL має атрибути: *alt*, запозичений з html, *longdesc* (зміст, як і в html-елементі *object*) і *readIndex*, які дозволяють описувати медіа наступним чином:

- *alt* – короткий текстовий опис файлу, найчастіше є навіть неповним реченням;
- *longdesc* – посилання (uri) на детальний опис (набір категорій для типів описів ще не встановлено);

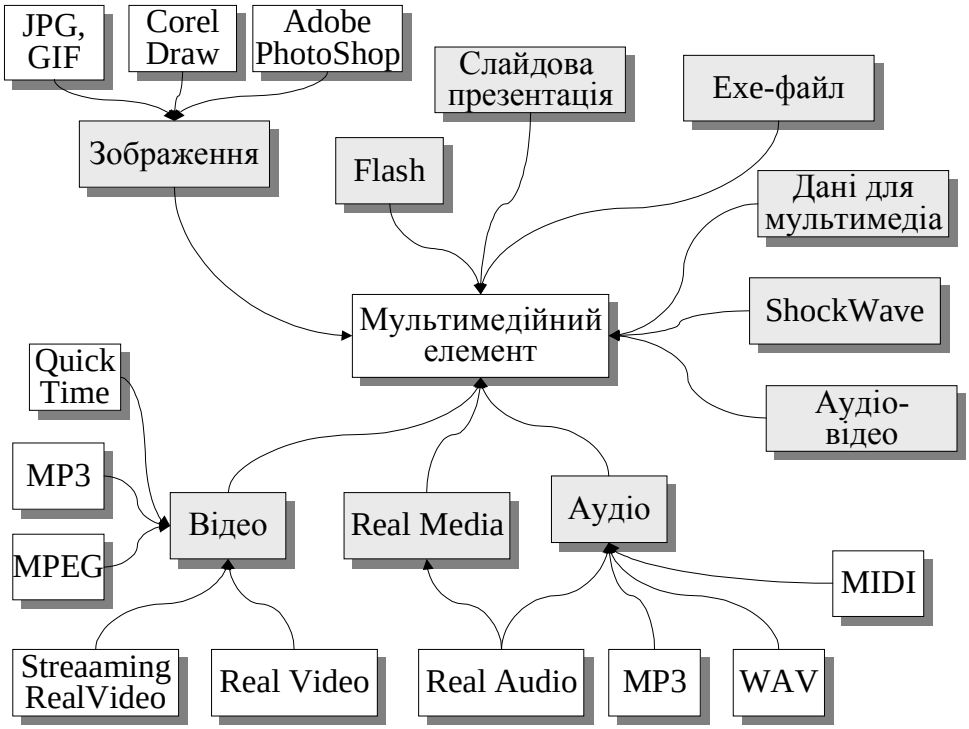

Рис.2. Онтологія МО в SERIF

- *readIndex* – порядок презентації альтернативного тексту, що визначається з використанням елементів *par*, *seq* і *excl*.

Класифікація мультимедійних об'єктів у проекті SERIF. У виділенні характерних ознак мультимедійної інформації цікаві роботи з класифікації МО, виконані в рамках проекту організації доступу до інформації про дослідницькі ініціативи Європи SERIF [9].

Предметна область даної розробки – наукові дослідження, тому базовими поняттями є проекти, персоналії тощо. Важлива роль приділяється метаданим, що зберігаються у базі SERIF. У рамках даного дослідження розроблена онтологія МО (рис.2). Виділяються такі ознаки МО, як розмір (у мегабайтах) та тривалість (у хвилинах), жанр (наприклад, лекція, доповідь), цільова аудиторія та рівень підготовки (студенти, учні, вчені, інженери тощо). Проект орієнтований на наукові дослідження, тому розробники окремо виділяють презентації PowerPoint як об'єднання слайдів, тексту і анімації.

Наведений вище огляд показує, що розповсюджені класифікації МО розроблялися для різних цілей і тому порізному виділяють підкласи МО (табл.1).

3. Таксономія мультимедійних об'єктів Інтернету

Враховуючи специфіку задачі, для якої проводився аналіз засобів подання мультимедіа – пошук інформації в Інтернеті, та інтегруючи розглянуті вище підходи до класифікації мультимедійних об'єктів, пропонуємо наступну таксономію МО (рис.3).

Ця таксономія виділяє окремо потоки даних, які потребують специфічного програмного забезпечення для їх інтерпретації, тому що формат подання таких даних невідомий і користувач не може їх описати (у тому випадку, коли в нього є така інформація, користувач може отримати доступ до цих даних значно простішими шляхами). В окремі групи виділено презентації SMIL та Power Point, тому що наявна в них текстова інформація дозволяє ідентифікувати МО, які входять до їх складу.

Необхідно також відзначити розходження між статичними (зображення, текст) і динамічними (відео, аудіо, анімація, текстовий потік) МО. В безперервних МО слід враховувати ще й такий параметр, як часову послідовність окремих елементів (відеокадрів, символів стрічки

Таблиця 1. **Стандарти подання мультимедіа**

	MPEG	MIME	SMIL	SERIF
Статичні зображення, графіка	*	* (підтипи – JPEG, інші)	*	*, формати jpg, gif, CorelDraw, Adobe PhotoShop
3D моделі	*	-	-	-
Аудіо	*	*	*	*, RealAudio, MP3, Wav, MIDI
Мовлення	*	-	-	-
Відео	*	*	*	*, QuickTime, MP3, MPEG, real Video
Сценарії (композиційні типи)	*	*, 2 підтипи – *multipart* та *message*	-	-
Прикладні дані	-	*	-	-
Анімація	-	-	*	-
Синхронізований текстовий потік	-	-	*	-
Flash	-	-	-	*
ShockWave	-	-	-	*
RealMedia	-	-	-	*
Слайдова презентація	-	-	-	*
Аудіо-відео	-	-	-	*

новин тощо). Однак необхідно пам'ятати, що деякі графічні файли є анімованими і, хоча й представлені з використанням елемента <*img*>, можуть розглядатися як безперервні. Серед графічних об'єктів доцільно виділити векторні та растрові зображення, тому що користувачеві досить легко визначити, які саме зображення він хоче знайти.

4. Пошук мультимедійних об’єктів в Інтернеті

Розглянемо, за якими параметрами здійснюють пошук МО найбільш

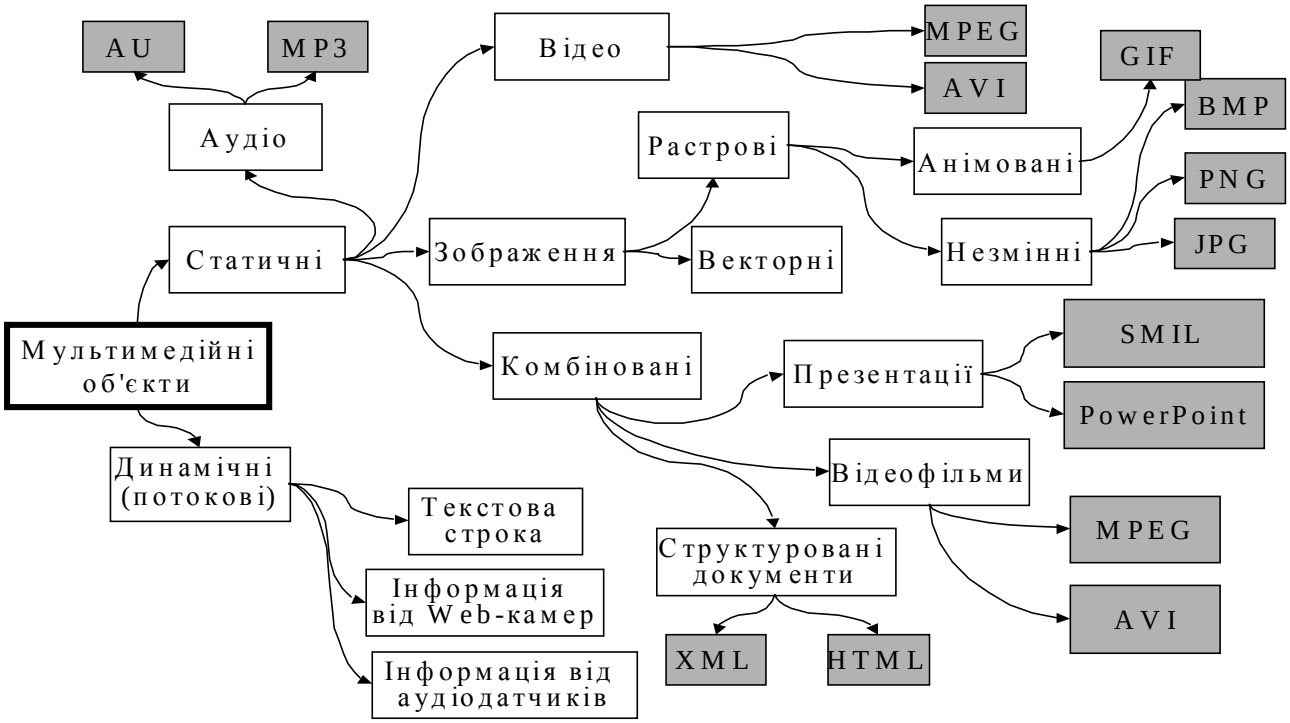

Рис.3. Таксономія мультимедійних об’єктів Інтернету

поширені ІПС Інтернету. Не всі ІПС підтримують пошук мультимедіа, а ті, які містять такі режими роботи, зазвичай дозволяють знаходити тільки деякі підтипи мультимедіа (у більшості випадків – пошук зображень). Крім того, розширений пошук у різних ІПС слабо стандартизований та дозволяє уточнювати різні параметри МО (табл.2).

Пошукова система AltaVista (www.altavista.com) при індексації зображень, крім ключових слів та метатегів, враховує текст, ім'я та шлях до файлу, які містяться в тегу посилання на зображення, а також абзац тексту, в межі якого це посилання попадає. Врахування RDF- та інших не вбудованих в текст сторінки мета-описів не здійснюється. Крім розміру та вирішення зображення, результати пошуку містять інформацію про розташування зображення та абзац тексту сторінки, до якого воно відноситься. Аналогічно здійснюється індексація аудіо та відео-файлів. В результатах пошуку видається тип файлу, його місцезнаходження (тобто адреса сторінки), розмір в байтах та тривалість відтворення.

Найбільш незручний та некоректний пошук здійснює пошукова система українських ресурсів search.com.ua, яка використовує спрощений механізм Google для індексації мультимедіа. Індексації підлягає тільки текст, розміщений в тегах посилання на цей ресурс.

Порівняння можливостей різних

Таблиця 2. **Порівняння можливостей найпоширеніших ІПС у пошуку мультимедіа**

Параметр	Пошукові системи			
	Google	AltaVista	Search. com.ua	Aport.ru
Зображення	+	+	+	+- (тільки листівки та фотографії – в окремих каталогах)
Градація за розміром файлу	невеликий, середній, малий, вирішення в точках	великий перелік вибору розміру зображення	-	-
Вибір кольорової палітри	кольорові та чорно-білі, напівтонові та повнокольорові	повнокольорові, кольорові та чорно-білі	-	-
Тип	-	фото, графіка та кнопки-банери	-	-
Область пошуку	весь Web, окремі ресурси	весь Web, новини, фільми тощо	-	-
Розрізняє формати	gif, jpg, png	-	-	-
Аудіо	-	+	-	+
Уточнення типу файлу	-	MP3, WAV, WindowsMedia, Real тощо	-	-
Формати	-	MP3, WAV, WindowsMedia		MP3 та MID
Тривалість звучання	-	більше або менше хвилини	-	-
Відео	+	+	+	-
Уточнення типу файлу	-	MPEG, AVI, Quicktime, WindowsMedia, Real тощо	-	-
Формати	-	MPEG, AVI, Quicktime, WindowsMedia, Real,	-	-
Тривалість	-	більше або менше хвилини	-	-
Область пошуку	спеціальний оператор *movie*	-	-	-

ІПС у пошуку мультимедіа можна продовжити, аналізуючи такі системи, як Lycos, AllTheWeb, Rambler, Апорт тощо. В результаті проведеного огляду можливостей пошуку мультимедійної інформації існуючими пошуковими системами, можна зробити наступний висновок: пошуку усіх типів мультимедійних ресурсів жодна з пошукових систем не реалізує. Жодна з розглянутих ІПС не виокремлює такі важливі типи, як презентації PowerPoint (презентації індексуються багатьма ІПС як текстові документи, однак вони не розглядаються як джерела мультимедійної інформації) та SMIL (найбільш перспективний напрямок розвитку публікації мультимедійної інформації в мережі), пошук векторних зображень, динамічної інформації.

Слід відмітити, що сьогодні в Інтернеті представлена велика кількість спеціалізованих ІПС, призначених для знаходження окремих видів мультимедіа. Приміром, пошукова система SingingFish (search.singingfish.com) дозволяє шукати за ключовими словами відео- та аудіодані, вказуючи їх категорію (спорт, новини, фінанси тощо) та формат. Але більшості користувачів вони не відомі і тому не використовуються

Недоліком більшості систем є те, що градації оцінок параметрів МО або фіксовані (у зовсім незручному варіанті – наприклад, "менше хвилини", “більше 3 хвилин”), або зовсім нечіткі і тому незрозумілі користувачеві (приміром, "невеликий”, “середній”, “малий" розмір). Крім того, слід зауважити, що значна кількість користувачів цікавляться пошуком мультимедіа у більш прагматичному аспекті та прагнуть застосовувати зовсім інші параметри для опису своїх інформаційних потреб. Приміром, у [10] стверджується, що 71% користувачів WWW цікавляться об’єктами, а не розміткою, кольорами, текстурою та іншими абстрактними характеристиками мультимедійних об’єктів. Для опису об’єктів можна використовувати онтології, які задають, приміром, відношення між базовими класами об’єктів, що не мають конкретних рис, та їх підкласами, які легше співвіднести з МО.

Значно підвищити релевантність пошуку МО дозволяє урахування у процесі пошуку ПрО, до якої він відноситься, та її співвіднесення зі сферою інформаційних інтересів користувача.

5. Використання формалізованого подання знань у пошуку мультимедійних об’єктів

Метаописи інформаційних ресурсів Інтернету. Крім підходів, спрямованих на обробку саме мультимедіа, необхідно враховувати універсальні засоби подання метаінформації, які дозволяють описувати семантику як текстових, так і мультимедійних ІР. Сьогодні найбільше поширення знайшов перспективний підхід до проблеми семантичного розпізнавання інформації – стандарт опису IP RDF (Resource Description Framework) [11] Консорціуму W3C. Мета його створення – стандартизувати визначення і використання метаданих, які описують IP Інтернету. Практично реалізовувати цей підхід почали в 2002 р. на базі Open Directory [12] в рамках проекту автоматичного створення репозиторію RDF-описів ресурсів Інтернету.

RDF використовує базову модель даних «*об'єкт — атрибут — значення*» та має XML-синтаксис. Важливою особливістю стандарту RDF є розширюваність: на RDF можна задати структуру опису джерела, використовуючи і розширюючи вбудовані поняття RDF-схем, такі як класи, властивості, типи, колекції.

Стандарт RDF підтримують багато провідних виробників ПЗ і постачальників контента. Розроблено ряд програмних продуктів, які дозволяють створювати RDF-описи для різного роду джерел (наприклад, RDFPic [13] створений для додавання RDF-опису до зображень). Передбачаються можливості інтеграції існуючих сховищ інформації в загальну базу семантичного опису та інтеграція концепції RDF-бази з форматом MPEG.

Щоб спростити та уніфікувати створення метаописів ресурсів, користувачам потрібно надати певні шаблони та

стандарти опису типових ресурсів. З таких засобів найбільш ґрунтовно розроблено набір елементів для створення метаданих "Dublin Core Metadata Elements", що складається з 15 базових елементів [14], які можна умовно розбити на три групи:

- *Content* (контент) – елементи, які відносяться до контента ресурсу;
- *Intellectual Property* (інтелектуальної власності) – елементи, які відносяться інтелектуальної власності;
- *Instantiation* (реалізація) – елементи, які описують конкретний екземпляр ресурсу.

Деякі елементи основного комплекту опису потребують більш детального розкриття через можливість різних інтерпретацій. Щоб зберегти сумісність з найпростішим описом з 15 елементів, але у той же час збільшити деталізацію і складність описів, різні організації розробляють розширення та додаткові кваліфікатори для базових елементів. Приміром, елемент *Subject* (Тема) визначається за допомогою двох тезаурусів: *предметного* та *функціонального*. Предметний тезаурус містить поняття ПрО і відображає зміст документа, а функціональний – його роль в людській діяльності.

Елемент *Type* (Тип) відображає жанр та категорію ресурсу. Можна обрати один зі стандартних типів: *text, image, sound, dataset, software, interactive, event* або *physical object*. Цей список може бути розширений (кожен елемент поділяється на піделементи), наприклад, *event* (подія) може бути конкретизований як *Конференція, Семінар, Круглий стіл, Виставка* або *Проект*.

Елемент *Формат* (Format*)* відображає середовище, формат даних ресурсу, матеріал, з якого складається ресурс (якщо це фізичний об'єкт), і, можливо, його фізичні розміри. Якщо ресурс подано в електронному вигляді, тоді його формат рекомендується вибирати зі списку вже вищевказаного стандарту MIME. Приклади електронних форматів: *text/xm*l – текст у форматі XML; *text/plain* – текст без форматування; *image/gif* – малюнок у форматі GIF. Для інших ресурсів формат рекомендується вибирати зі списку фізичних об'єктів.

Аналізуючи RDF-опис IP, можна визначити, до якої ПрО він відноситься, хто є його автором, якою мовою та в якому форматі подано інформацію. Фізичні розміри IP визначаються через кількісні показники і можуть порівнюватися з будь-якими вимогами користувача.

Онтологічний опис ПрО, яка цікавить користувача. Традиційні механізми пошуку в Інтернеті, як правило, виконують запити користувача на пошук інформації тільки за переліком ключових слів. Значно підвищити ефективність пошуку дозволяє його *персоніфікація*, тобто використання відомостей про сферу інформаційних інтересів користувача. Враховуючи інформацію про користувача та його інтереси, можна отримувати більш релевантні результати.

Як показує аналіз публікацій, один з перспективних підходів до опису ПрО, що цікавить користувача, ґрунтується на онтологіях, які містять перелік основних термінів, зв'язки між ними і правила виведення (так, у проекті Semantic Web, спрямованому на аналіз семантики IP, саме онтологічний підхід є основою для подання знань про різні ПрО) [15].

Проблема інформаційного пошуку ускладнюється тим, що різні групи людей, які займаються збором і пошуком інформації, застосовують для спілкування з ІПС як власні спеціальні терміни, так і терміни, широко використовувані іншими співтовариствами в іншому розумінні. Поряд із глобальними онтологіями, що описують досить широкі ПрО і для створення яких необхідні значні зусилля експертів ПрО та інженерів зі знань, існують онтології, що дозволяють формально представити знання конкретного користувача щодо ПрО. Такі онтології можуть створюватися і модифікуватися користувачами самостійно. Хоча, можливо, деякі подання користувача ПрО є помилковими, але така онтологія описує ПрО, яка відповідає його інформаційним інтересам (наприклад, якщо користувач помилково вважає дельфіна

рибою і, запросивши зображення якої-небудь риби, отримує зображення дельфіна, тоді його інформаційна потреба буде задоволена).

Створюючи інформаційний запит, користувач визначає через онтологію ПрО сферу своїх інформаційних інтересів.

6. Пошук МО з використанням їх контексту та семантики ПрО

При необхідності вичерпного пошуку МО обов'язковою вимогою є звертання не тільки до спеціалізованих функцій "пошук зображень" у різних системах, але і безпосередній перегляд сторінок, змістовно пов'язаних з ними. Проведені експерименти показують, що пошук МО за допомогою ІПС надає посилання на значно меншу кількість ресурсів порівняно з пошуком серед текстових ресурсів, який дозволяє знайти документи, що містять посилання на МО. Це пов’язано з тим, що в багатьох спеціалізованих виданнях імена файлів ілюстрацій мають числове позначення, а підписи до ілюстрацій взагалі не робляться, тому що електронна версія конвертується з оригінал-макету друкованого видання, у якому ця інформація відсутня. Крім того, імена файлів найчастіше мають скорочену форму, що також не дозволяє зробити їхній пошук з використанням спеціальних функцій.

Інформаційний пошук являє собою процес зіставлення *запиту користувача* з відомостями про ІР, що відомі ІПС, до якої надійшов цей запит. Запит користувача – це опис інформації, доступ до якої він хоче одержати. У загальному випадку такий запит може, наприклад, містити ключові слова, пов'язані логічними операторами; документ-зразок; тип документа і його тему за класифікатором; списки рекомендованих чи заборонених користувачем інформаційних джерел; обмеження на час або обсяг пошуку тощо. Чим складніше форма подання запиту, тим вище релевантність пошуку (релевантність пошуку – це співвідношення між кількістю знайдених документів, що задовольнили користувача, тобто відповідали його запиту, і загальною кількістю знайдених документів). Проте ускладнення форми запиту призводить до ускладнення процедури його обробки, і, отже, до збільшення часу пошуку.

Метод пошуку МО на основі порівняння їх контексту з онтологією ПрО, яка цікавить користувача. Спочатку потрібно за допомогою ІПС знайти множину ГІР, що містять відповідні ключові слова, а після цього – перевірити ці ГІР на наявність посилань на МО тих форматів, що цікавлять користувача (з урахуванням фільтрації банерної реклами та МО менше заданого обсягу, що призначені для службових цілей – кнопки, символи, роздільні елементи тощо). Наступним етапом роботи є порівняння онтології користувача з контекстом МО. Контекст МО – це текстова інформація, яка міститься в ГІР (приміром, в html-або в sgml-документі), в якому зустрічається посилання на цей МО, та в його метаописі.

Пропонуємо метод пошуку МО, який складається з наступних етапів:

- користувач створює запит $Q = \langle\langle z_1,..., z_n\rangle, \langle o_1,..., o_m\rangle, \{t_1,..., t_l\}\rangle$, який складається з таких елементів:
  - ключових слів z_k – довільних слів або словосполучень природної мови;
  - термінів онтології $o_k \in O$, що описує ПрО, до якої відносяться інформаційні інтереси користувача;
  - множини типів МО $t_k \in T$, потрібних користувачеві;
- за типами МО визначається множина форматів подання інформації, що відповідають цьому типу – $\{t_1,..., t_l\} \xrightarrow[F]{} \{f_1,..., f_p\}$;
- до зовнішньої ІПС передається запит з ключових слів та форматів подання інформації $Q` = \langle(z_1 \vee ... \vee z_n) \wedge (f_1 \vee ... \vee f_l)\rangle$;
- від ІПС поступає перелік ГІР L, що містять вказані ключові слова та посилання на МО у відповідних форматах, які супроводжуються додатковими описами, створеними ІПС (звичайно – фрагментами вмісту ГІР, у

яких зустрілися ключові слова, або анотаціями);

- здійснюється фільтрація цього переліку ГІР, для чого описи ГІР порівнюються з переліком термінів онтології ПрО $o_k \in O$, створюється $L_O \subseteq L$;
- ГІР з $L_O \subseteq L$ перевіряється на доступність і повтори, а потім отримується контекст МО – вміст та метаописи відповідних ГІР, створюється $L_{MO} \subseteq L_O$;
- здійснюється фільтрація переліку МО, для чого контекст МО порівнюються з переліком термінів онтології ПрО $o_k \in O$, створюється $L_{MO,O} \subseteq L_{MO}$;
- підраховується коефіцієнти релевантності запиту для кожного МО з $L_{MO,O} \subseteq L_{MO}$;
- МО з $L_{MO,O} \subseteq L_{MO}$ впорядковуються за релевантністю, отриманий список разом з описами передається користувачеві.

Основна відмінність запропонованого методу від аналогічних – здійснювати за допомогою ІПС пошук не МО, а ГІР та потім виконувати фільтрацію його результатів з урахуванням онтології ПрО, що цікавить користувача (рис.4).

Основні переваги запропонованого методу пошуку. Вони полягають у наступному:

- для підвищення релевантності пошуку використовується онтологічний опис ПрО, що цікавить конкретного користувача;
- попередній пошук за допомогою зовнішніх ІПС здійснюється серед не МО, а ГІР, що забезпечує значно більшу повноту пошуку;
- використання метаописів ГІР та МО, на які вони посилаються, дозволяють точніше враховувати їх семантику;
- проведення фільтрації інформації у два етапи дозволяє значно зменшити час пошуку, оскільки ІР, що відфільтровуються на першому етапі, не потрібно копіювати на сервер для подальшої обробки.

Пошук МО можна здійснювати по тексту html-сторінок, які містять мульти-

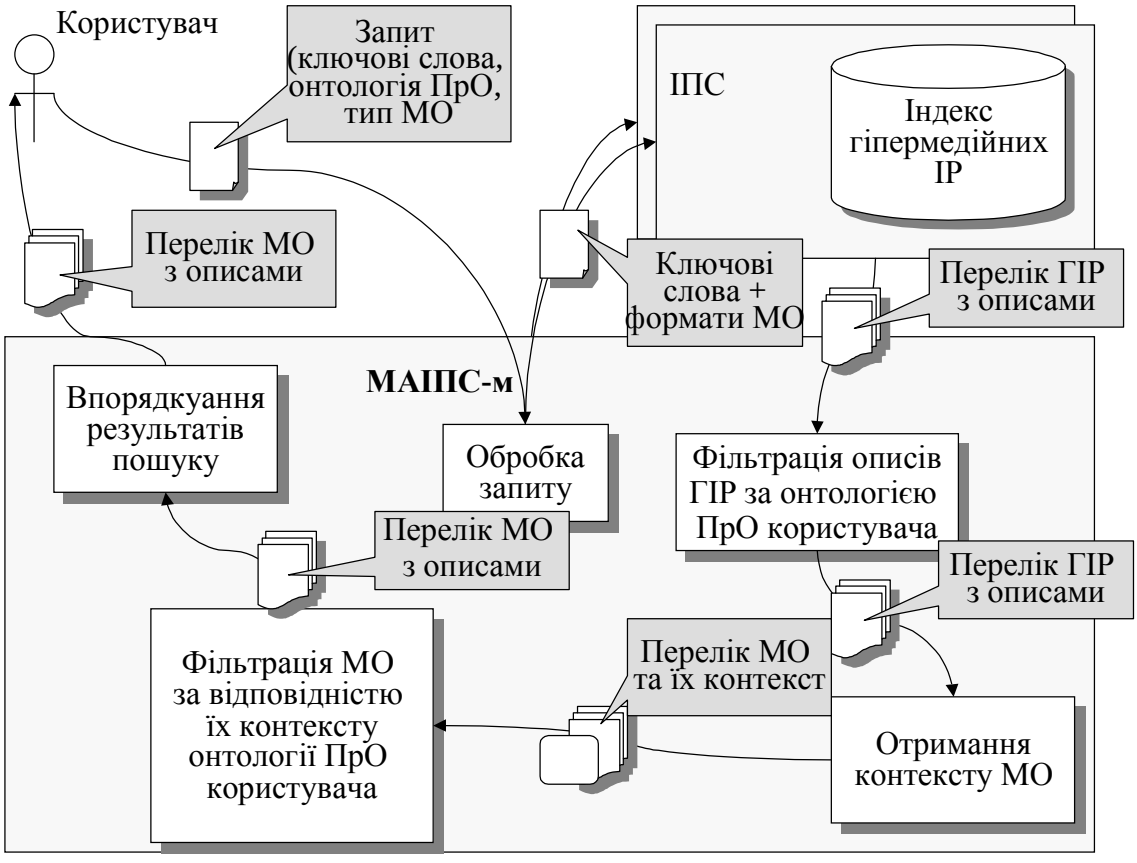

Рис.4. Пошук МО з використанням контексту та онтології ПрО, що цікавить користувача

медіа, по тексту метаописів цих сторінок та текстах метаописів МО. Для пошуку використовують ключові слова, які задає користувач у пошуковому запиті. Однак, з огляду на специфічність подання семантичної інформації про мультимедіа, а також враховуючи сталі інтереси користувача, для підвищення релевантності цей запит необхідно доповнити онтологією користувача – описом ПрО, а також деякими іншими параметрами пошуку.

Місцезнаходження інформації про мультимедійні об'єкти в ГІР. При визначенні коефіцієнту релевантності МО запиту доцільно враховувати вагу різних структурних одиниць ГІР (приміром, заголовки різних рівнів).

В Інтернеті найчастіше використовуються html-сторінки – текстові документи у певному стандартному форматі. Семантична інтерпретація ІР Інтернету (розробки Semantic Web) потребує, щоб документ містив додаткову метаінформацію (у форматах *xml, rdf, owl* тощо). Гіпертекстове подання дозволяє розміщувати на сторінках різні типи мультимедіа – рисунки, фонове звучання музики тощо. Стандарт HTML v.4.0 за допомогою механізму *object* розширює ці можливості, дозволяючи динамічно завантажувати відео- та аудіоролики в окремі місця сторінки. Для того щоб підвищити релевантність пошуку, доцільно враховувати структуру html-сторінки, надаючи словам, що зустрічаються у різних структурних елементах, різну вагу, яка відображає їх відносну важливість для запиту [16].

Для визначення коефіцієнта релевантності ГІР запиту ключові слова запиту та терміни онтології порівнюються з множиною слів, використовуваних у ГІР та його метаописах. При цьому вага слів у ГІР визначається кількістю їх входжень та розташуванням [17], а також загальним обсягом контексту. Так, наприклад, термін, який зустрічається як в RDF-описі мультимедіа, так і на html-сторінці, більш вагомий, ніж якщо він знаходиться тільки в тексті html-сторінки. Можна виділити такі області розташування слів:

- заголовки;
- метатеги;
- метаописи;
- посилання на інші документи або домени;
- параграфи, які містять посилання на МО потрібного типу.

Рішення щодо релевантності ГІР приймається на підставі всієї наявної текстової інформації про них, наприклад підписів і анотацій. Коефіцієнт релевантності K ГІР I запиту Z визначається за формулою:

$$K(I,Z)=\sum_{k=1}^{p}(\sum_{i=1}^{n}n_k(z_i,I)*q_k+\sum_{i=1}^{m}n_k(o_i,I)*q_k),$$ де

$n_k(z_i,I)$ – кількість входжень ключового слова z_i до k-го структурного елементу ГІР I (приміром, до заголовку першого або другого рівня, посилання тощо), $n_k(o_i,I)$ – кількість входжень терміна онтології ПрО o_i до k-го структурного елемента ГІР I, а q_k – вага цього структурного елемента. Значення коефіцієнтів q_k визначаються користувачем залежно від його потреб та специфіки ПрО.

Принциповою відмінністю запропонованого методу пошуку є те, що він дозволяє враховувати метаінформацію навіть у тих випадках, коли вона не підтверджується вмістом сторінки. Приміром, пошукова система Google здатна обробляти описи в форматі RDF, але якщо слова або словосполучення з цього опису не зустрічаються в тексті сторінки, до якої відноситься цей опис, то система ці слова ігнорує. На практиці досить часто така ситуація не є помилкою розробників ресурсу або прикладом некоректної PR-компанії. Наприклад, адреси розробників або власників ресурсу можуть не вказуватися на сторінці, а тема, до якої відноситься ресурс, взагалі майже ніколи ніде явно не вказується, окрім метаопису. Пропонується враховувати такі терміни, але давати їм відносно низьку вагу. В такому разі ці МО будуть запропоновані користувачеві тільки в тому разі, якщо ресурсів, що містять такі ж терміни і в метаописі і в контенті сторінки, не існує або вони не задовольняють його з якихось інших причин (наприклад, входять до списку небажаних джерел).

Проаналізувавши описи МО, що можна отримати шляхом обробки html-сторінки, яка містить посилання на цей МО, зведемо їх у наступні групи (рис.5):

- *RDF-опис ГІР, який містить посилання на МО* – найбільш чіткий і конкретний короткий опис семантики ГІР, представлений в структурованій формі (наприклад, у форматі Dublin Core). Автори або власники IP створюють його вручну або автоматизовано, але в разі випадків використання для реклами (спам) RDF-опис може містити інформацію, яка не має ніякого відношення до контенту ГІР, тому враховувати його необхідно, але з низькою вагою.
- *RDF-опис МО* (приміром, ідентифікатори Dublin Core), створюваний авторами або власниками МО вручну чи автоматизовано.
- *Текстовий опис МО* (*longdescr*, потоковий текст) – зв'язний текст, що докладно описує ГІР та потребує семантичного розбору. Він створюється цілеспрямовано й осмислено автором, тому його істинність висока. Однак і в цьому випадку необхідно враховувати випадки некоректного використання.
- *Короткий опис МО*, який знаходиться в самому html-документі (атрибути *title*, *alt* тощо). Створені автоматично тексти можуть не відповідати семантиці МО.
- *Текстовий вміст ГІР*. Припустимо вважати, що текст ГІР який містить посилання на МО, на семантичному рівні пов'язаний з цим МО (приміром, ГІР та МО відносяться до однієї ПрО, МО є прикладом того, що описується в ГІР і т.д.), тобто між об'єктами, описаними на сторінці та відображеними у МО, існують зв'язки та відповідності. При цьому необхідно розділяти сам текст сторінки та його метаопис, який в свою чергу складається як з метатегів, так і з RDF-опису контента сторінки.

Можна виділити наступні структурні елементи html-сторінки, у яких можна розміщувати посилання на МО:

- гіперпосилання – <A …>;
- зображення <*IMG*…>;
- фонові зображення та звук <*body*…>, <*table*…>;
- елемент об'єкта <*OBJECT*…> з посиланням на тип МО;
- аплет <*applet*…> (деякі аплети можуть бути МО);
- карти < *Map*…>;
- кнопки вводу форм <*input*…>;
- функції javascript <*script* …>;
- елемент *LINK* з *REV=MADE*, який іноді використовується для ідентифікації автора документа, вказує адресу

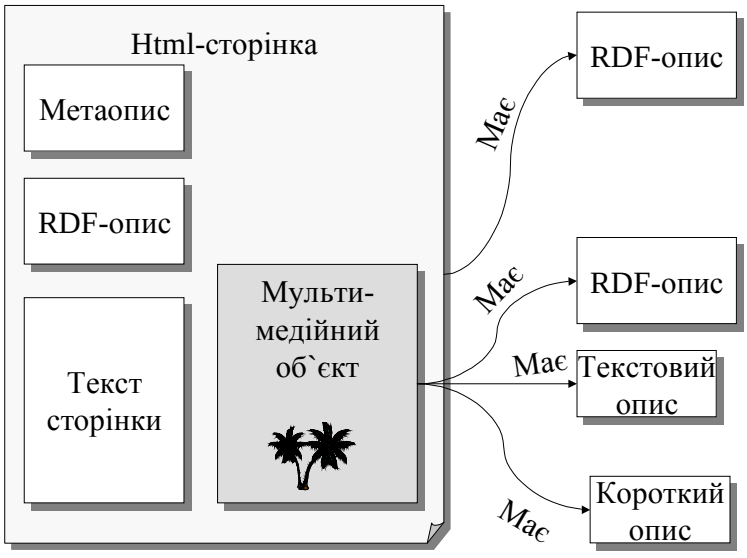

Рис.5. Інформація про МО, яка міститься в html-сторінці

його електронної пошти або посилання на його домашню сторінку.

Висновки

Проаналізувавши поширені визначення мультимедійної інформації, підходи до її класифікації та можливості програмних засобів, що застосовуються для пошуку мультимедіа в Інтернеті, вважаємо доцільним використовувати контекст, у якому зустрічаються мультимедійні об'єкти, їх метаописи та онтологічне подання знань про сферу інтересів користувача для пошуку на семантичному рівні. Крім того, доцільно враховувати структуру ГІР, які містять посилання на МО та їх метаописів.

1. *American* National Standard for Telecommunications. Telecom Glossary 2000. - http://www.its.bldrdoc.gov/projects/telecomglossary2000.
2. *ACM* SIGMM Retreat Report on Future Directions in Multimedia Research., 2004. – http://www.sigmm.org/Events/reports/retreat03/.
3. *Hoschka P.* Synchronized Multimedia Integration Language (SMIL) 1.0 Spec., 1998. – http://www.w3.org/TR/REC-smil/.
4. *Большой* энциклопедический словарь. Современная энциклопедия. – http://dic.academic.ru.
5. *American* National Standard for Telecommunications. Telecom Glossary 2000. - http://www.its.bldrdoc.gov/projects/telecomglossary2000.
6. *MPEG-7* Overview, ISO/IEC, July 2002. – http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm
7. *Dublin* Core – http://dublincore.org/.
8. *Freed N., Borenstein N.* Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types. – http://www.ietf.org/rfc/rfc2046.txt.
9. *SERIF*. – http://derpi.tuwien.ac.at/ ~ andrei/Metadata_Science.htm.
10. *Schomaker L*. Image Search and Annotation: From Lab to Web // Proc. of CIDE, ISBN 2-909285-17-0, 2001. – P.373-375.
11. R*DF/XML* Syntax Specification (Revised), W3C Working Draft, 2002. – http://www.w3.org/TR/rdf-syntax-grammar/,
12. *Open* Directory Project. – http://dmoz.org/.
13. *Describing* and retrieving photos using RDF. – http://www.w3.org/TR/photo-rdf/.
14. *Dublin* Core Metadata Elements. – http://www.faqs.org/rfcs/rfc2413.html.
15. *Овдій О.М., Проскудіна Г.Ю.* Онтології у контексті інтеграції інформації: представлення, методи та інструменти побудови // Проблемы программирования. – №4, 2004. – С.353-366.
16. *Интеллектуальный* семантический поиск с привлечением средств метапоиска / Г.С.Осипов, О.С.Завьялова, И.В.Смирнов, И.А.Тихомиров // 5 Международ. Конф. "Интеллектуальный анализ информации ИАИ-2005". – К.: Просвіта, 2005. – С.214-223.
17. *Боровикова О.И., Загорулько Ю.А., Сидорова Е.А.* Автоматизация сбора онтологической информации в Интернет-портале знаний // Там же. – К.: Просвіта, 2005. – С.82-91.

Отримано 04.04.05

Про авторів:

Рогушина Юлія Віталіївна,
кандидат фіз.-мат. наук,
старший науковий співробітник,
Гришанова Ірина Юріївна,
молодший науковий співробітник.

Місце роботи авторів:
Інститут програмних систем
НАН України,
Київ, пр.Акад.Глушкова, 40,
тел. (044)526 5139,
e-mail: _jjj_@ukr.net,
i26031966@yahoo.com.