НАЦІОНАЛЬНА АКАДЕМІЯ ПЕДАГОГІЧНИХ НАУК УКРАЇНИ
Інститут обдарованої дитини

Сухий О. Л., Міленін В. М., Тарадайнік В. М.

АЛГОРИТМИ ПОШУКУ В ІНФОРМАЦІЙНИХ СИСТЕМАХ

Методичні рекомендації

Київ
2015

Рекомендовано до друку Вченою радою Інституту обдарованої дитини НАПН України (протокол № 11 від 25.11.2015 р.)

Рецензенти:
Тименко Володимир Петрович, доктор педагогічних наук, професор, вчений секретар відділення професійної освіти і освіти дорослих НАПН України.
Поліхун Наталія Іванівна, кандидат педагогічних наук, старший науковий співробітник, завідуюча відділом підтримки обдарованості та міжнародної співпраці ІОД НАПН України

Сухий О. Л. Алгоритми пошуку в інформаційних системах : методичні рекомендації / О. Л. Сухий, В. М. Міленін, В. М. Тарадайнік. – К., 2015. – 2,0 д.а.

У методичних рекомендаціях представлено основні концепції інформаційного пошуку, наводиться огляд базових моделей пошуку інформації та загальних методів оцінки якості пошуку інформації, що розглядаються в межах теорії інформаційного пошуку. Окрім того наведені основні етапи та узагальнений алгоритм пошуку інформації, загальні рекомендації щодо організації структури повнотекстових баз даних.

Окремо надаються рекомендації щодо детальної оптимізації веб-сайтів для значного покращення розпізнавання змісту пошуковими системами та взагалі користувачами.

Проводиться детальний екскурс в історію розвитку формату збереження даних XML, визначається теза використання наведеного формату як найкращого формату для збереження неструктурованих документів, оптимального для пошукових систем.

Для спеціалістів у галузі інформаційного пошуку, студентів, аспірантів, також стане корисним при підготовці навчальних курсів з теоретичних і практичних питань інформаційного пошуку.

Зміст

Вступ

Сучасну цивілізацію неможливо уявити без новітніх засобів передачі інформації. З появою персональних комп'ютерів набули широкого розповсюдження комп'ютерні мережі. Сьогодні ми є свідками того, як міцно вони входять в наше повсякденне життя, поступово стаючи невід'ємною приналежністю. Вже зараз є сфери людської діяльності, які не можуть існувати без комп'ютерних інформаційних мереж. Комп'ютерна мережа – ідеальний засіб швидкого обміну інформацією.

Комп'ютерна мережа – це комплекс програмних і апаратних засобів, за допомогою яких ЕОМ, накопичувачі даних та електронні офісні пристрої поєднуються в систему для загального використання інформаційних ресурсів та обміну даними, а також є однією з тих сфер культури майбутнього, в якій активно реалізуються багато напрямків розвитку особистості. Причини цього кореняться не тільки в безсумнівно побутовій та науково допоміжній користі дигітальної (цифрової) техніки.

1 Сучасні мережі та інформаційно-пошукові системи

Величезні можливості у галузі обробки та пошуку інформації відкрилися з появою Інтернет мережі. Інтернет являється глобальною комп'ютерною мережею, що складається з багатьох десятків та сотень тисяч дрібних та більш крупних мереж, таких як корпоративні, наукові, урядові та домашні, частини яких взаємопов'язані.

Початок історії мережі Інтернет має витоки з того часу, коли було створено мережу ARPANET у 1969 році на замовлення Міністерства оборони США. Мережа ARPANET стала енергійно розростатися, вчені різноманітних галузей науки розпочали її використовувати. Вже в 1973 році до мережі підключились одні з перших іноземних організацій Великобританії та Норвегії. У 1984 році Національний фонд науки США започаткував велике між університетське мережеве сполучення NSFNet з набагато більшою пропускною здатністю ніж можливі 56 Кбіт/с у ARPANETа. Проект ARPANET було закрито у 1990 році у зв'язку з великою конкурентною перевагою на користь NSFNet.

Завдяки створеним протоколам TCP/IP (Transmission Control Protocol/Internet Protocol), єдиний адресний простір та принцип маршрутизації пакетних даних відкрилась можливість поєднання мереж з різною архітектурою і топологією. Створення незалежного протоколу IP дало змогу будь-якій Ethernet, що передає цифрові дані в свою чергу передавати дані Інтернет. Виходячи з цього мережі (Ethernet) на основі протоколу IP складають загальний адресний простір у розмірах всього світу, але в кожній окремій Ethernet може існувати свій власний адресний простір, обраний відповідно до класу мережі. Завдяки такій організації маршрутизатори, що автоматично перенаправляють пакети даних, однозначно визначають наступний напрямок для кожного окремого взятого пакету даних.

Стрімкий розвиток і широке розповсюдження веб-технологій («всесвітньої павутини») призвело до появи нової спільноти, оскільки велика кількість людей, що працювала з WWW, не могла себе називати дослідниками і розробниками мережі. Була створена нова організація – Консорціум Всесвітньої павутини (W3C). Консорціум W3C взяв на себе відповідальність за розробку різноманітних протоколів і стандартів, пов'язаних із «всесвітньою павутиною».

Всесвітня павутина – WorldWideWeb (або скорочено, веб) являє собою глобальний інформаційний простір, заснований на фізичній інфраструктурі мережі Інтернет і протоколі передачі даних HTTP.

WorldWideWeb об'єднує мільйони веб-серверів, підключених до Інтернет. На початку розвитку WorldWideWeb існувала невелика кількість веб-сайтів, на сторінках яких окремі автори або веб-майстри публікували інформацію для відносно великої кількості відвідувачів. Сьогодні ситуація різко змінилася. Самі відвідувачі веб-сайтів, споживачі інформації, беруть активну участь і в генерації

контенту, що призвело до різкого зростання обсягів інформації, динаміки веб та виникнення нової методики проектування систем і сучасної форми функціонування інтернет-ресурсів, яка отримала назву Web 2.0.

Для перегляду інформації, отриманої від веб-серверів на комп'ютерах користувачів використовуються спеціальні-програми – веб-браузери, основна функція яких полягає у відображення гіпертексту, що є основним методом представлення інформації в Інтернет.

Для візуалізації гіпертекстової інформації була створена мова HTML (HypertextMarkupLanguage), яка являє собою стандартну мову розмітки текстових документів, за допомогою якої створюються всі веб-сторінки. Останньою актуальною версією є стандарт HTML 4.01, прийнятий у 1999 році. У листопаді 2007 Консорціумом W3C був представлений чорновий варіант специфікації п'ятої версії HTML. Паралельно проводиться робота по подальшому розвитку HTML під назвою XHTML (eXtensibleHTML), який базується на XML і в 2000 році був схвалений в якості Рекомендацій W3C.

Для передачі в мережі Інтернет гіпертекстової інформації використовується протокол HTTP (HyperTextTransferProtocol), який спочатку використовувався виключно для передачі HTML-документів, а вже на сьогоднішній день можна передавати будь-яку інформацію, в тому числі зображення, звук, відео а також будь-які файли.

Оскільки HTML на початку призначався для візуалізації інформації, він лишається незручним для автоматизованої обробки інформації, в тому числі й для організації пошуку і це є основним його недоліком. На різних сайтах подання інформації суттєво відрізняється як за оформленням, так і за розміщенням. Тому, направлений на показ окремих сайтів, WWW є погано пристосованим для автоматизованого збору, класифікації та аналітичної обробки інформації.

Для уніфікації представлення контенту при вирішенні задач обміну інформації між сайтами дані надаються не в HTML, а у вигляді XML, призначеного для обміну даними та їх інтеграцією.

1.1 Основні поняття та терміни

Перед подальшим розглядом ІПС коротко ознайомимося з основними поняттями та термінами.

Інформація – абстрактне поняття, що має різні значення залежно від контексту. Походить від латинського слова *«informatio»*, яке має декілька значень. В даному методичному посібнику інформація розглядається як:

- сукупність даних, зафіксованих на матеріальному носії, збережених і поширених в часі і просторі;

- це усвідомлені відомості про навколишній світ, які є об'єктом зберігання, перетворення, передачі і використання.

Основні властивості інформації

Цінність інформації – визначається корисністю та здатністю її забезпечити суб'єкта необхідними умовами для досягнення ним поставленої мети.

Достовірність – здатність інформації об’єктивно відображати процеси та явища, що відбуваються в навколишньому світі. Як правило достовірною вважається насамперед інформація, яка несе у собі безпомилкові та істинні дані. Під безпомилковістю слід розуміти дані які не мають, прихованих або випадкових помилок. Тоді як під істинними слід розуміти дані зміст яких неможливо оскаржити або заперечити.

Часові властивості – визначають здатність даних передавати динаміку зміни ситуації (динамічність):

- *актуальність* – здатність інформації відповідати вимогам сьогодення (поточного часу або певного часового періоду);
- *оперативність* – властивість даних, яка полягає в тому, що час їхнього збору та переробки відповідає динаміці зміни ситуації;
- *ідентичність* – властивість даних відповідати стану об’єкта.

Повнота інформації – інформацію можна назвати повною, якщо її достатньо для розуміння і прийняття рішень. Неповна інформація може призвести до помилкового висновку або рішення.

Точність інформації – визначається ступенем її близькості до реального стану об’єкта, процесу, явища тощо.

Корисність (цінність) інформації– корисність може бути оцінена стосовно потреб конкретних її споживачів і оцінюється за тими завданнями, які можна вирішити за її допомогою.

Найцінніша інформація – об’єктивна, достовірна, повна, і актуальна. При цьому слід враховувати, що і необ’єктивна, недостовірна інформація (наприклад, художня література), має велику значимість для людини. Соціальна (суспільна) інформація має ще й додаткові властивості:

- має семантичний (смисловий) характер, тобто понятійний, оскільки саме в поняттях узагальнюються найбільш істотні ознаки предметів, процесів і явищ навколишнього світу;
- має мовну природу (крім деяких видів естетичної інформації, наприклад образотворчого мистецтва). Один і той самий зміст може бути виражено на різних природних (розмовних) мовах, записано у вигляді математичних формул тощо.

Властивість недоступності – при розгляді захищеності даних можна виділити технічні аспекти захисту даних від несанкціонованого доступу та соціально-психологічні аспекти класифікації даних за мірою їхньої конфіденційності та секретності (властивість конфіденційності).

Старіння –головною причиною старіння інформації є не сам час, а поява нової інформації, з надходженням якої попередня інформація виявляється невірною, перестає адекватно передавати явища та закономірності матеріального світу, людського спілкування та мислення.

Розсіювання – існування у багатьох джерелах.

Види інформації

Основні види інформації по її формі вистави, способам її кодування і зберігання, що має найбільше значення для інформатики, це:

- *графічна або образотворча* – перший вид, для якого був реалізований спосіб зберігання інформації про навколишній світ у вигляді наскальних малюнків, а пізніше у вигляді картин, фотографій, схем, креслень на папері, полотні, мармурі та ін. матеріалах, що зображують картини реального світу;
- звукова – світ навколо нас сповнений звуків і завдання їх зберігання і тиражування була вирішена за винахід звукозаписних пристроїв в 1877 р.; її різновидом є музична інформація – для цього виду був винайдений спосіб кодування з використанням спеціальних символів, що робить можливим зберігання її аналогічно графічної інформації;
- *текстова* – спосіб кодування мови людини спеціальними символами – буквами, причому різні народи мають різні мови і використовують різні набори букв для відображення мови;
- *числова* – кількісна міра об’єктів і їх властивостей в навколишньому світі; особливо велике значення набула з розвитком торгівлі, економіки і грошового обміну; аналогічно текстовій інформації для її відображення використовується метод кодування спеціальними символами – цифрами, причому системи кодування (числення) можуть бути різними;
- *відеоінформація* – спосіб збереження «живих» картин навколишнього світу, що з’явився з винаходом кіно.

Пошук – в інформаційному сенсі: прагнення знайти що-небудь, дія що направлена на отримання нових даних, знань, закономірностей. Пошуком так само називається один із способів навчання, забезпечення корисних знань про природу, мовою, суспільстві. Мається на увазі отримання знань через самостійні розумові дії для вирішення завдань.

Документ визначається як засіб закріплення будь-яким способом на спеціальному матеріалі будь-якої інформації про факти, події, явища об’єктивної дійсності і розумової діяльності людини. Документи мають різну

форму подання. В автоматизованих документальних ІПС це насамперед текстова інформація на природних мовах у зрозумілій для ЕОМ формі.

Запит являє собою інформаційну потребу, сформульовану на природній мові. Результат «перекладу» інформаційного запиту на інформаційно-пошукову мову називають пошуковим образом запиту (ПОЗ) або пошуковим приписом (ПП). Синтаксис і семантика мов запитів визначається структурою і наповненням документів і загальними завданнями системи.

Інформаційна потреба – потреба у залученні додаткової інформації для досягнення мети, яка стоїть перед користувачем в процесі його професійної діяльності або в соціально-побутовій практиці.

Пертінентность (*лат. Pertineo*– торкаюся, ставлюся) – відповідність знайдених інформаційно-пошуковою системою документів інформаційним потребам користувача, незалежно від того, як повно і як точно ця інформаційна потреба виражена в тексті інформаційного запиту.

Релевантність (*лат. Relevo*– піднімати, полегшувати) в інформаційному пошуку – семантичне відповідність пошукового запиту і пошукового образу документа, тобто не тільки оцінка ступеня відповідності, але і ступеня практичного застосування результату для рішення задачі.

Інформаційна система – організована сукупність програмно-технічних та інших допоміжних засобів і технологічних процесів, функціонально-визначених груп працівників, що забезпечує збір, подання та накопичення інформаційних ресурсів у певній предметній області, пошук і видача відомостей необхідних для задоволення визначених потреб.

Інформаційно-пошукова система – впорядкована сукупність документів та інформаційних технологій призначених для зберігання і пошуку інформації, текстів або даних.

Інформаційний пошук (ІП) – процес пошуку неструктурованої документальної інформації, що задовольняє інформаційні потреби, та наука про цей пошук.

1.2 Склад і структура автоматизованих інформаційно-пошукових систем

Доступ користувачів до сучасних інформаційних мереж та ефективне задоволення їх інформаційних потреб можливо тільки за допомогою розвинених засобів навігації в цих мережах. Основним інструментом при цьому виступають інформаційно-пошукові системи, що забезпечують пошук в гігантських обсягах текстової інформації.

Перші реально функціонуючі повнотекстові інформаційно-пошукові системи (ІПС) з’явилися на початку комп’ютерної ери. Велика кількість університетів і публічних бібліотек почали використовувати ІПС для забезпечення доступу до бібліотечних каталогів, архівів, масивів документів,

таких як статті, нормативні акти, реферати, брошури, дисертації, монографії. Широке розповсюдження отримали ІПС з появою мережі Інтернет і розвитком «всесвітньої павутини».

З розвитком ІПС широкого вжитку набув і термін «інформаційний пошук», який вперше був введений Кельвіном Муерсом у 1948 році в його докторській дисертації і почав використовуватися в літературі з 1950 року.

Основними функціями інформаційно-пошукових систем на початку були:

- зберігання великих обсягів інформації;
- швидкий пошук необхідної інформації;
- додавання, видалення і зміна збереженої інформації;
- видача інформації в зручному для користувача вигляді.

ІПС складається з баз даних, в яких накопичується інформація, джерел інформації, апаратної частини, програмної частини, споживача інформації. Деякі автори включають до складу ІПС також і персонал, що її експлуатує.

Програмна частина ІПС включає в себе *логіко-семантичний апарат* (пошукові мови – одна чи декілька, правила індексування та критерії видачі інформації) , пошуковий масив (визначену множину документів, забезпечених пошуковими образами, в якій відшукуються необхідні).

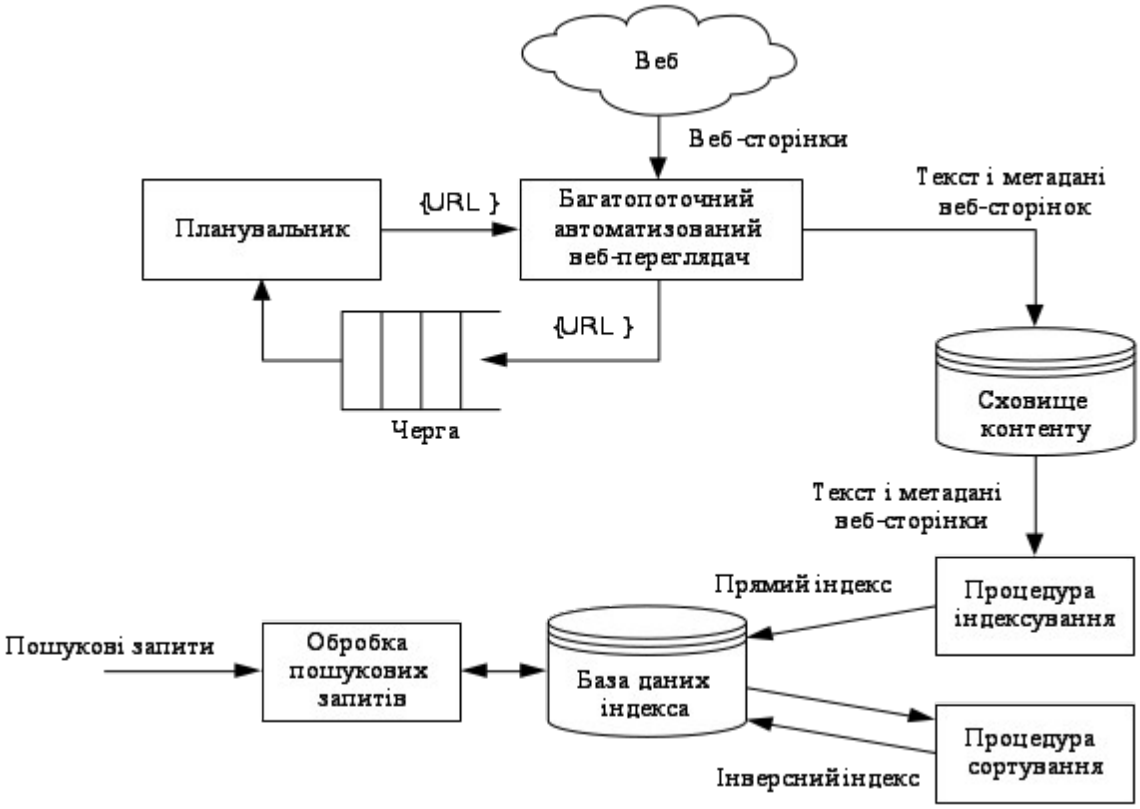

Рис. 1. Архітектура типової пошукової системи

Логіко-семантичний апарат ІПС складається з трьох основних блоків:

- інформаційно-пошукової мови (ІПМ);

- системи індексування (перекладу на інформаційно-пошукову мову);
- логіки, що забезпечують пошук, які в свою чергу можуть бути деталізовані та реалізовані різними способами.

1.3 Характеристика інформаційно-пошукових систем

Опишемо основні характеристики ПС.

Повнота – одна з основних характеристик пошукової системи, що представляє собою відношення кількості знайдених за запитом документів до загального числа документів в базі даних, які відповідають даному запиту.

Точність – ще одна основна характеристика пошукової машини, яка визначається ступенем відповідності знайдених документів запиту користувача. Чим точніше пошук, тим швидше користувач знайде потрібні йому документи, тим менше різного роду «сміття» серед них буде зустрічатися, тим рідше знайдені документи не будуть відповідати запиту.

Актуальність – не менш важлива складова пошуку, яка характеризується часом, що проходить з моменту публікації документів в мережі Інтернет, до занесення їх в індексну базу пошукової системи. Завдяки існуванню у великих пошукових систем так званої «швидкої бази», яка оновлюється кілька разів на день, основні документи індексуються і стають доступні для пошуку.

Швидкість пошуку тісно пов’язана з його стійкістю до навантажень. Значна завантаженість вимагає скорочення часу обробки окремого запиту. Тут інтереси користувача і пошукової системи збігаються: відвідувач бажає отримати результати якомога швидше, а пошукова машина повинна відпрацьовувати запит максимально оперативно, щоб не гальмувати обчислення наступних запитів.

Наочність представлення результатів є важливим компонентом зручного пошуку. По більшості запитів пошукова машина знаходить сотні, а то й тисячі документів. Внаслідок нечіткості складання запитів або неточності пошуку, навіть перші сторінки видачі не завжди містять тільки потрібну інформацію. Це означає, що користувачеві часто доводиться виробляти свій власний пошук всередині знайденого списку. Різні елементи сторінки видачі пошукової системи допомагають орієнтуватися в результатах пошуку.

1.4 Класифікація інформаційно-пошукових систем

Класифікація ІПС за ступенем автоматизації

- *Ручні інформаційні системи* характеризуються відсутністю сучасних технічних засобів переробки інформації та виконанням всіх операцій людиною.
- *Автоматизовані інформаційно-пошукові системи (АІПС)* – найбільш популярний клас ІПС. Припускають участь в процесі накопичення, обробки інформації баз даних, програмного забезпечення, людей і технічних засобів.

- *Автоматичні інформаційні системи* виконують всі операції по переробці інформації без участі людини. Прикладом автоматичних інформаційних систем є деякі пошукові машини Інтернет, наприклад Google, де збір інформації про сайти здійснюється автоматично пошуковим роботом і людський фактор не впливає на ранжирування результатів пошуку.

Класифікація ІПС за архітектурою

- *Локальні ІПС* працюють на одному електронному пристрої, не мають взаємодії з сервером або іншими пристроями.
- *Клієнт-серверні ІПС* працюють в локальній або глобальної мережі з єдиним сервером.
- *Розподілені ІПС* – децентралізовані системи в гетерогенній мережі з великою кількістю серверів.

Класифікація ІПС за пошуковими технологіями

- *Тематичні каталоги* передбачають обробку документів і віднесення їх до однієї з декількох категорій, перелік яких заздалегідь заданий. Фактично це індексування на основі класифікації. Індексування може проводитися автоматично або вручну за допомогою фахівців, які переглядають популярні веб-вузли і складають короткий опис документів (ключові слова, анотація, реферат).
- *Спеціалізовані каталоги або довідники* створюються по окремих галузях і темах, по новинах, по містах, адресам електронної пошти тощо.
- *Пошукові машини* (найрозвиненіший засіб пошуку в Інтернеті) реалізують технологію повнотекстового пошуку. Індексуються тексти, розташовані на опитуваних серверах. Індекс може містити інформацію про кілька мільйонів документів. Наприклад, в індексі популярної ІПС «AltaVista» більше 56 млн. URL-адрес.
- *Ресурси метапошуку*, при використанні якого запит здійснюється одночасно кількома пошуковими системами. Результат пошуку об’єднується в загальний, упорядкований за ступенем релевантності список. Кожна система обробляє тільки частину вузлів мережі, що дозволяє розширити базу пошуку.

Бази інформаційних даних можуть містити практично будь-які види інформації, у тому числі в будь-якій комбінації. Інформаційний пошук здійснюється як за існуючими в повнотекстових електронно-інформаційних ресурсах термінам, так і за спеціальними елементами, що входять до складу інформаційно-пошукових мов, які служать для формування пошукових запитів.

ІПС фактично є системами інформаційного забезпечення і являють собою бази і банки даних.

2 Моделі і структури даних інформаційно-пошукових систем

В попередньому розділі було зазначено, що однією зі складових частин ІПС являється банк даних, який, в свою чергу, має у своєму складі обчислювальну систему, одну або декілька баз даних (БД), систему управління базами даних (СУБД) і сукупність прикладних програм.

База даних забезпечує зберігання інформації і являє собою сукупність даних, організованих за певними правилами, які мають загальні принципи опису, зберігання і маніпулювання даними.

Система управління базами даних являє собою пакет прикладних програм і сукупність мовних засобів, призначених для створення, супроводу і використання баз даних.

Прикладні програми (додатки) у складі банків даних служать для обробки даних, обчислень і формування вихідних документів за заданою формою.

Програмний додаток являє собою програму або комплекс програм, що використовують БД і забезпечують автоматизацію обробки інформації з деякої предметної області. Додатки можуть створюватися як у середовищі СУБД, так і поза СУБД – за допомогою мов та систем програмування, що використовують засоби доступу до БД.

Для більшості ІПС, які функціонують в мережі Інтернет, програмними додатками, які забезпечують зручність роботи з БД, служать Інтернет-браузери, найпопулярнішими з яких є на сьогоднішній день Internet Explorer компанії Microsoft, GoogleChrome, Opera, MozillaFirefox.

2.1 Архітектура інформаційних систем, що основані на концепції баз даних

Ефективність функціонування інформаційної системи багато в чому залежить від її архітектури. В даний час перспективною є архітектура клієнт-сервер. У досить поширеному варіанті вона передбачає наявність комп’ютерної мережі та розподіленої бази даних.

Сервером певного ресурсу в комп’ютерній мережі називається комп’ютер (програма), яка керує ресурсом, клієнтом – комп’ютер (програма), що використовує цей ресурс. В якості ресурсів комп’ютерної мережі можуть виступати, наприклад, бази даних, файлові системи, служби друку, поштові служби. Тип сервера визначається видом ресурсу, яким він керує. Наприклад, якщо ресурсом є база даних, то відповідний сервер називається сервером бази даних.

Перевагою організації інформаційної системи по архітектурі клієнт-сервер є вдале поєднання централізованого зберігання, обслуговування та колективного доступу до загальної інформації з індивідуальною роботою над цією інформацією. Структура розподіленої БД, побудованої за архітектурою клієнт-сервер, показана на рис. 2.

Найважливішим достоїнством застосування БД в інформаційних системах є забезпечення незалежності даних від прикладних програм. Це дозволяє не обтяжувати користувачів проблемами представлення даних на фізичному рівні: розміщення даних у пам’яті, методів доступу до них тощо.

Така незалежність досягається підтримуваним СУБД багаторівневим поданням даних в БД на логічному (користувальницькому) і фізичному рівнях. Іншими словами, завдяки СУБД і наявності логічного рівня представлення даних забезпечується відділення концептуальної (понятійної) моделі БД від її фізичного представлення в пам’яті ЕОМ.

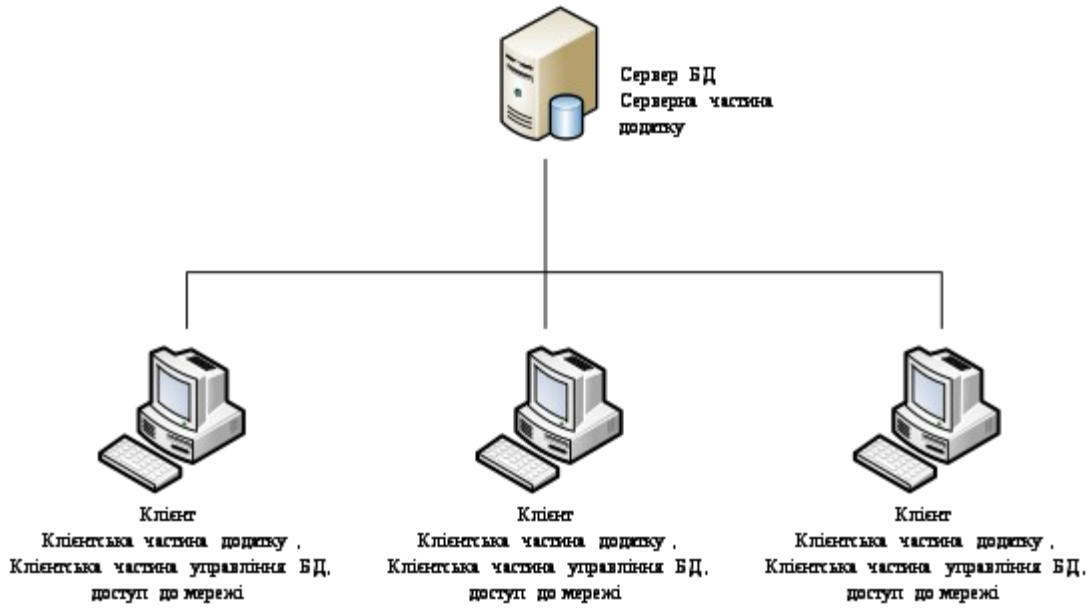

Рис. 2. Клієнт-серверна структура розподіленої БД

Перевагами даної архітектури є:

- можливість, у більшості випадків, розподілити функції обчислювальної системи між декількома незалежними комп’ютерами в мережі;
- всі дані зберігаються на сервері, який, як правило, захищений набагато краще більшості клієнтів, а також на сервері простіше забезпечити контроль повноважень, щоб дозволяти доступ до даних тільки клієнтам з відповідними правами доступу;
- підтримка багатокористувацької роботи;
- гарантія цілісності даних.

Недоліки:

- непрацездатність сервера може зробити непрацездатною всю обчислювальну мережу;
- адміністрування даної системи вимагає кваліфікованого професіонала;

- висока вартість обладнання;
- бізнес логіка додатків залишилася в клієнтському ПЗ.

При проектуванні інформаційної системи, заснованої на архітектурі «клієнт-сервер», більшу увагу слід звертати на грамотність спільних рішень. Технічні засоби пілотної версії можуть бути мінімальними (наприклад, в якості апаратної основи серверу баз даних може використовуватися одна з робочих станцій). Після створення пілотної версії потрібно провести додаткову дослідницьку роботу, щоб з'ясувати вузькі місця системи. Тільки після цього необхідно приймати рішення про вибір апаратури серверу, яка буде використовуватися на практиці.

Збільшення масштабів інформаційної системи не породжує принципових проблем. Звичайним рішенням є заміна апаратури сервера (і, можливо, апаратури робочих станцій). У будь-якому випадку зміни щодо зміни масштабів практично не торкаються прикладної частини інформаційної системи.

Моделі даних

Дані, що зберігаються в базі даних мають певну логічну структуру, тобто представлені деякої моделлю, яка підтримується СУБД. До числа найважливіших відносяться наступні моделі даних:

- ієрархічна;
- мережева;
- реляційна;
- об'єктно-орієнтована.

В ієрархічній моделі дані представляються у вигляді деревовидної (ієрархічної) структури. Вона зручна для роботи з ієрархічно впорядкованою інформацією і громіздка для інформації зі складними логічними зв'язками.

Мережева модель означає представлення даних у вигляді довільного графа. Перевагою мережевий і ієрархічної моделей даних є можливість їх ефективної реалізації за показниками витрат пам'яті та оперативності. Недоліком мережевої моделі даних є висока складність і негнучкість схеми БД, побудованої на її основі.

Реляційна модель даних (РМД) назву отримала від англійського терміна*relation* – відношення. Її запропонував у 70-ті роки співробітник фірми IBM Едгар Кодд. При дотриманні певних умов відношення представляється у вигляді двовимірної таблиці, звичної для людини. Більшість сучасних БД для персональних ЕОМ є реляційними.

Перевагами реляційної моделі даних є її простота, зручність реалізації на ЕОМ, наявність теоретичного обґрунтування і можливість формування гнучкої схеми БД, що допускає налаштування при формуванні запитів.

Реляційна модель даних використовується в основному в БД середнього розміру. При збільшенні числа таблиць в базі даних помітно падає швидкість роботи з нею. Певні проблеми використання РМД виникають при створенні систем зі складними структурами даних, наприклад, систем автоматизації проектування.

Об'єктно-орієнтовані БД об'єднують в собі дві моделі даних, реляційну і мережеву, і використовуються для створення великих БД зі складними структурами даних.

2.2 Представлення, ідентифікація і пошук інформації в реляційній моделі БД

Реляційні бази даних розроблені для швидкого збереження і отримання великих обсягів інформації. Нижче наведені деякі характеристики реляційних баз даних і реляційної моделі даних.

- *Використання ключів:* кожний рядок в таблиці ідентифікується унікальним «ключем», який називається первинним.
- *Відсутність надмірності даних:* за правилами реляційної моделі кожна окрема частина інформації зберігається в одному місці. Це дозволяє уникнути необхідності роботи з даними в кількох місцях.
- *Обмеження вводу:* при використанні реляційної моделі є можливість визначити, який тип даних зберігається у стовпчику.
- *Підтримка цілісності даних:* налаштовуючи властивості полів, встановлюючи зв'язки між таблицями та налаштовуючи обмеження можна значно підвищити надійність збереження даних.
- *Налаштування прав:* більшість СУБД пропонують налаштування прав доступу, які дозволяють задавати певні права окремим користувачам. Таким чином деякі дії можуть бути дозволені або заборонені користувачеві.
- *Структурована мова запитів:* для здійснення операцій над БД використовується мова SQL (StructuredQueryLanguage).
- *Переносимість:* реляційна модель даних стандартна. Дотримуючись правил реляційної моделі даних можна бути впевненим, що дані можуть бути перенесені відносно легко в іншу реляційну СУБД.

Як було зазначено вище, дані в реляційній моделі зберігаються у вигляді таблиць, що містять рядки, або записи. Кожна таблиця відображає набір властивостей певної окремої сутності, а записи в таблиці – набір сутностей в межі однієї таблиці. При проектуванні баз даних важливо на першому кроці визначити кожну окрему сутність (об'єкти, які заслуговують окремих таблиць в БД) та порядок ідентифікації записів.

Для ідентифікації кожного запису таблиці служить первинний ключ, тобто кожний запис завдяки первинному ключу стає унікальним.

Первинний ключ завжди має унікальне значення. Зазвичай первинний ключ – числовий, але може бути й будь-яким іншим типом даних. Часто первинний ключ складається з одного поля, але він може бути й комбінацією декількох стовпців.

На другому етапі необхідно визначити, які зв'язки існують між сутностями.

В реляційній моделі даних існує три типи зв'язків між таблицями.

- *Один-до-багатьох:* виникає коли одному запису з таблиці А може відповідати кілька записів з таблиці В.
- *Багато-до-багатьох:* зв'язок, при якому множині записів з таблиці А можуть відповідати кілька записів з таблиці В. Зв'язок *багато-до-багатьох* створюється за допомогою трьох таблиць, де дві таблиці, А і В – «джерела» і одна «перехресна» таблиця, що їх об'єднує. Вона складається з двох полів, двох зовнішніх ключів, які посилаються на первинні ключі таблиць А і В. Зв'язок *багато-до-багатьох* складається з двох зв'язків *один-до-багатьох*. Таблиці А і В мають зв'язок *один-до-багатьох* з «перехресною» таблицею.
- *Один-до-одного:* зв'язок, при якому один блок сутності А може відповідати лише одному блоку сутності В. Зв'язок *один-до-одного* легко моделюється в одній таблиці, але в рідких випадках зв'язок *один-до-одного* моделюється в двох таблицях. Такий варіант необхідний, коли виникає потреба у подоланні обмежень СУБД або з метою підвищення продуктивності.

Проект реляційної бази даних – це колекція таблиць, що пов'язані між собою первинними та зовнішніми ключами.

Нормалізація даних

Реляційна модель даних включає в себе ряд вимог, яким повинні задовольняти відносини між таблицями. Ці вимоги називаються «нормальними формами».

Нормальна форма – властивість відносин в реляційної моделі даних, що характеризує їх з точки зору надмірності, потенційно призводить до логічно помилкових результатів вибірки або зміни даних.

Процес перетворення відносин бази даних до виду, відповідному нормальним формам, називається *нормалізацією*. Нормалізація призначена для приведення структури БД до виду, що забезпечує мінімальну логічну надмірність, і не має на меті зменшення або збільшення продуктивності роботи або ж зменшення або збільшення фізичного обсягу бази даних. Кінцевою метою нормалізації є зменшення потенційної суперечливості інформації, що

зберігається в базі даних. Загальне призначення процесу нормалізації полягає в наступному:

- виключення деяких типів надмірності;
- усунення деяких аномалій оновлення;
- розробка проекту бази даних, який є досить «якісним» представленням реального світу, інтуїтивно зрозумілий і може служити хорошою основою для подальшого розширення;
- спрощення процедури застосування необхідних обмежень цілісності.

Усунення надмірності виробляється, як правило, за рахунок декомпозиції відносин таким чином, щоб у кожному відношенні зберігалися тільки первинні факти (тобто факти, не виведені з інших збережених фактів).

У створенні та розвитку теорії нормалізації брали участь багато вчених. Однак перші три нормальні форми і концепцію функціональної залежності запропонував Е. Кодд.

Попри тому, що ідеї нормалізації досить корисні для проектування баз даних, вони аж ніяк не є універсальним або вичерпним засобом підвищення якості проекту БД. Це пов’язано з тим, що існує занадто велика різноманітність можливих помилок і недоліків у структурі БД, які нормалізацією не усуваються. Незважаючи на ці міркування, теорія нормалізації є дуже цінним здобутком реляційної теорії і практики, оскільки вона дає науково строгі і обґрунтовані критерії якості проекту БД та формальні методи для удосконалення цієї якості. Цим теорія нормалізації різко виділяється на тлі чисто емпіричних підходів до проектування, що пропонуються в інших моделях даних. Більше того, можна стверджувати, що в усій сфері інформаційних технологій практично відсутні методи оцінки та покращення проектних рішень, що можуть бути порівнянні з теорією нормалізації реляційних баз даних за рівнем формальної строгості.

3 Інформаційний пошук

3.1 Пошукові задачі, засоби та технології інформаційного пошуку

Пошук інформації являє собою процес виявлення в деякій множині документів (текстів) всіх таких, які присвячені зазначеній темі (предмету), задовольняють заздалегідь визначеним умовам пошуку (запиту) або містять необхідні (відповідно до інформаційної потреби) факти, відомості, дані.

Центральна задача ІП – допомогти користувачеві задовольнити його інформаційну потребу. Так як описати інформаційні потреби користувача технічно непросто, вони формулюються як деякий запит, який представляє з себе набір ключових слів, що характеризує те, що шукає користувач.

Класична задача ІП, з якої почався розвиток цієї галузі, – це пошук документів, що задовольняють запиту, в рамках деякої статичної колекції документів. Але список завдань ІП постійно розширюється і тепер включає:

- питання моделювання;
- класифікація документів;
- фільтрація документів;
- кластеризація документів;
- проектування архітектур пошукових систем і користувальницьких інтерфейсів;
- отримання інформації, зокрема анотування і реферування документів;
- мови запитів та ін.

Процес пошуку включає послідовність операцій, спрямованих на збір, обробку та надання необхідної інформації зацікавленим особам.

У загальному випадку процес пошуку інформації складається з наступних етапів:

- формулювання запиту природною мовою, вибір пошукових системи і сервісів, формалізація запиту на відповідній ІПМ;
- проведення пошуку в одній або декількох пошукових системах;
- огляд отриманих результатів (посилань);
- попередня обробка отриманих результатів: перегляд змісту посилань, вилучення та збереження релевантних і пертінентних даних;
- при необхідності, модифікація запиту і проведення повторного (уточнюючого) пошуку з подальшою обробкою отриманих результатів.

Інформаційний пошук має на увазі використання певних стратегій, методів, механізмів і засобів. Поведінка користувача, що здійснює управління процесом пошуку, визначається не тільки інформаційною потребою, а й інструментальним різноманітністю системи – технологіями та засобами, наданими системою.

Стратегія пошуку – загальний план (концепція, перевага, установка) поведінки системи або користувача для вираження і задоволення інформаційної потреби користувача, обумовлений як характером мети і видом пошуку, так і системними «стратегічними» рішеннями – архітектурою БД, методами і засобами пошуку в конкретній ІПС. Вибір стратегії в загальному випадку є оптимізаційним завданням. На практиці в значній мірі він визначається мистецтвом досягнення компромісу між практичними потребами і можливостями наявних коштів.

Метод пошуку – сукупність моделей і алгоритмів реалізації окремих технологічних етапів: побудови пошукового образу запиту (ПОЗ), відбору документів (зіставлення пошукових образів запитів і документів), розширення і реформулювання запиту, локалізації та оцінки видачі.

Пошуковий образ запиту – записаний на ІПМ текст, що виражає смисловий зміст інформаційного запиту і містить вказівки, необхідні для найбільш ефективного здійснення інформаційного пошуку.

Методи пошуку, тобто виділення підмножини документів, які потенційно містять опис рішення задачі відбору документів, є відображенням процесу знаходження рішення і залежать від характеру завдання і предметної області.

Розглядаючи пошук як ітеративний процес, методи скорочення простору перебору (підмножини, що переглядається) утворюють по суті методологічну основу стратегії пошуку і можуть бути розділені на наступні класи – методи пошуку в:

- одному просторі (зазвичай, тематичному);
- ієрархічно упорядкованому просторі;
- альтернативних просторах;
- динамічному (змінюваному в процесі пошуку) просторі.

Метод побудови ПОЗ повинен забезпечувати ефективні способи побудови запиту для досягнення цілей різного типу.

Механізми пошуку – сукупність реалізованих у системі моделей і алгоритмів процесу формування видачі документів у відповідь на пошуковий запит.

Засоби пошуку, з одного боку, – взаємозалежний комплекс інформаційно-пошукових мов (ІПМ) і мов визначення/управління даними, що забезпечує структурні та семантичні перетворення об’єктів обробки (документів, словників, сукупностей результатів пошуку), а з іншого, – об’єкти користувальницького інтерфейсу, забезпечують управління послідовністю вибору операційних об’єктів конкретної ІПС.

Пошукові технології – уніфіковані (оптимізовані в рамках конкретної ІПС) послідовності ефективного використання окремих засобів пошуку в

процесі взаємодії користувача з системою для сталого отримання кінцевого і проміжних результатів.

Навігація як реалізація процесу пошуку за запитом в обраній БД – цілеспрямована, обумовлена стратегією, послідовність використання методів, засобів і технологій конкретної ІПС для отримання та оцінки результату.

Засоби навігації дозволяють користувачеві здійснювати управління процесом пошуку. Вони надаються користувачеві у вигляді інтерфейсу, що дозволяє організувати більш-менш ефективний процес взаємодії з БД. При цьому «дружність» інтерфейсу характеризується не тільки ергономічністю і зрозумілістю, а й варіантністю вибору операційних об'єктів.

Процес пошуку інформації представляє послідовність кроків, що призводять при посередництві системи до деякого результату, і дозволяють оцінити його повноту. Так як користувач зазвичай не має вичерпних знань про інформаційне змісті ресурсу, в якому проводить пошук, то оцінити адекватність виразу запиту, так само як і повноту одержуваного результату, він може, грунтуючись лише на зовнішніх оцінках або на проміжних результатах і узагальненнях, зіставляючи їх, наприклад , з попередніми.

Типологія пошукових задач

За характером і рівня співвідношення у предметі пошуку відомого і невідомого (як ступеня семантичної невизначеності) можна виділити три типи пошукових завдань.

1. *Предметний (або атрибутивний) вид пошуку* - пошук об'єкта, коли відомо, що цей об'єкт існує (наприклад, пошук фактографії або праць конкретного автора). Пошукова модель (логічна ідентифікація об'єкта пошуку) може бути представлена як пошук по атрибутам. Для документального пошуку – це відбір по логічному вираженню над іменами понять, що задаються термінами або їх комбінаціями.

2. *Тематичний вид пошуку* – підбір інформації за деякою темі, наприклад, для пошуку методу вирішення практичного завдання. Тематичний пошук – це знаходження в середовищі ІПС описів актуально існуючих в предметній області основної діяльності об'єктів, властивості яких можуть бути повністю визначені на вже відомій множині атрибутів. Пошукова модель в цьому випадку – це пошук по частині відомого поняття або зв'язкам, частково заданим комбінацією характеристичних ознак. Тематичний пошук реалізується як послідовність атрибутивних пошуків.

3. *Форма проблемного пошуку* – знаходження в інформаційному середовищі описів об'єктів або їх складових, потенційно існуючих в предметній області основної діяльності і в сукупності, можливо, таких, що утворюють ціле, властивості якого будуть більше суми властивостей частин. Тобто цим властивостям в явній формі не відповідають «власні» атрибути, а нова

властивість, наприклад, може бути задана комбінацією вже відомих атрибутів. Логічна пошукова модель для цього випадку – пошук «схожих» документів, зміст яких деяким чином асоціюється із завданням користувача.

Види пошуку

Повнотекстовий пошук – пошук по всьому вмісту документа. Приклад повнотекстового пошуку – будь-яка пошукова система Інтернет, наприклад www.google.com, www.yandex.ru. Як правило, повнотекстовий пошук для прискорення пошуку використовує попередньо побудовані індекси. Найбільш поширеною технологією для індексів повнотекстового пошуку є інвертовані індекси.

Пошук по метаданих – це пошук за деякими атрибутам документа, що підтримується системою – назва документа, дата створення, розмір, автор тощо. Приклад пошуку за реквізитами – діалог пошуку в файлової системі (наприклад, MS Windows).

Пошук по зображенню – пошук за змістом зображення. Пошукова система розпізнає зміст фотографії (завантаженої користувачем або доданий URL зображення). У результатах пошуку користувач отримує схожі зображення. Так працюють пошукові системи Xcavator, Retrievr, PolarRose, PicollatorOnlinebyRecogmission.

На сьогоднішній день існують програмні додатки та Інтернет-сервіси які здійснюють пошук в аудіо файлах по уривку. Таку послугу, зокрема, надає Google.

Компоненти інформаційного пошуку

Для забезпечення процедури інформаційного пошуку в ІПС можна виокремити два рівні розгляду – *абстрактний і конкретний.*

Абстрактною ІПС прийнято вважати сукупність ІПМ, правил індексування та критеріїв видачі, або критеріїв смислової відповідності інформації.

Під конкретною ІПС розуміється практично реалізована, що містить у своєму складі масив документів в якому проводиться пошук, технічні засоби, а також людей, що взаємодіють із системою.

У відповідності до виділення в ІПС абстрактного і конкретного рівнів та з урахуванням особливостей зберігання документальної інформації (бібліотеки, архіви та інші сховища) процедуру інформаційного пошуку можна розділити на дві складові:

- семантичне осмислення запиту і видача адрес відповідних запиту документів;
- знаходження самих документів.

У символічній формі абстрактна ІПС являє собою сукупність ІПМ, правил індексування й логіки, критерії змістовної відповідності.

Рис. 3. Склад логіко-семантичного апарату ІПС

3.2 Організація пошуку, технологія та основні методи

Організація пошуку

Пропонується процедуру пошуку необхідної інформації розділити на дев'ять основних етапів:

- визначення галузі знань;
- вибір типу і джерел даних;
- збір матеріалів необхідних для наповнення інформаційної моделі;
- відбір найбільш корисної інформації;
- вибір методу обробки інформації (класифікація, кластеризація, регресійний аналіз і т.д.);
- вибір алгоритму пошуку закономірностей;
- пошук закономірностей, формальних правил і структурних зв’язків в зібраної інформації;
- творча інтерпретація отриманих результатів;
- інтеграція витягнутих «знань».

Технології пошуку інформації

Пошукові засоби і технології, що використовуються для реалізації інформаційних потреб, визначаються типом і станом задачі основної діяльності, яка стоїть перед користувачем: співвідношенням його знання і незнання про

об’єкт, що потребує дослідження. Крім того, процес взаємодії користувача з системою визначається рівнем знання користувачем змісту ресурсу (повноти уявлення, достовірності джерела і т.д.) і функціональних можливостей системи як інструменту. В цілому ці фактори зазвичай зводяться до поняття «професіоналізму» – інформаційного (підготовлений/непідготовлений користувач) і предметного (професіонал/непрофесіонал) «професіоналізму».

Процес пошуку інформації зазвичай носить емпіричний характер. Він являє послідовність кроків, що призводять при посередництві системи до деякого результату, що дозволяють оцінити його повноту. При цьому поведінка користувача, як організуючий початок управління процесом пошуку, мотивується не тільки інформаційною потребою, а й різноманітністю стратегій, технологій і засобів, що надаються системою.

Зазвичай користувач не має вичерпних знань про інформаційний зміст ресурсу, в якому проводить пошук, тому оцінити адекватність виразу запиту, як і повноту отриманого результату, він може, відшукавши додаткові відомості, або організувавши процес так, щоб частина результатів пошуку могла використовуватися для підтвердження або заперечення адекватності іншої частини.

Операційними об’єктами, що безпосередньо беруть участь у взаємодії користувачів з пошуковою системою є пошуковий образ документа (ПОД) і пошуковий образ запиту (ПОЗ), відповідність яких встановлюється пошуковим механізмом ІПС на формальному рівні. Адекватність образу дійсному змісту документа визначається якістю процесу згортки інформації та рівнем знання суб’єктом засобів відображення – концептуальної схеми предметної області і можливостей ІПМ.

Для проведення пошуку спочатку на комп’ютері користувача завантажується інтерфейс роботи з відповідної БД. Це може бути локальна або віддалена БД. Спочатку слід визначитися з видом пошуку (простий, розширений і т.д.). Потім з набором пропонованих для пошуку полів. ІПС можуть запропонувати для введення одне або кілька полів. При формуванні запиту практично всі системи дозволяють використовувати логічні елементи "AND", "OR", "NOT".

Основні методи пошуку

Адресний пошук – процес пошуку документів за чисто формальними ознаками, зазначеним у запиті.

Для здійснення потрібні наступні умови:

- наявність у документа точної адреси;
- забезпечення строгого порядку розташування документів в пристрої або в сховище системи.

Адресами документів можуть виступати адреси веб-серверів і веб-сторінок, елементи бібліографічного запису, адреси зберігання документів в сховищах.

Семантичний пошук – процес пошуку документів за їх змістом.

Умови для здійснення семантичного пошуку:

- переклад змісту документів і запитів з природної мови на інформаційно-пошукові мови та складання пошукових образів документа і запиту;
- складання пошукового опису, в якому вказується додаткова умова пошуку.

Принципова різниця між адресним і семантичним пошуками полягає в тому, що при адресному пошуку документ розглядається як об’єкт з точки зору форми, а при семантичному пошуку – з точки зору змісту.

При семантичному пошуку знаходиться безліч документів без вказівки адрес. У цьому полягає принципова відмінність каталогів і картотек.

Документальний пошук – процес пошуку в сховищі інформаційно-пошукової системи первинних документів або в базі даних вторинних документів, відповідних запиту користувача.

Існує два види документального пошуку:

- бібліотечний, спрямований на знаходження первинних документів;
- бібліографічний, спрямований на знаходження відомостей про документи, представлених у вигляді бібліографічних записів.

Фактографічний пошук – процес пошуку фактів, що відповідають інформаційним запитам.

До фактографічних даних відносяться відомості, витягнуті з документів, як первинних, так і вторинних і отримані безпосередньо з джерел їх виникнення.

Розрізняють два види фактографічного пошуку:

- документально-фактографічний, полягає в пошуку в документах фрагментів тексту, що містять факти;
- фактологічний (опис фактів), припускає створення нових фактографічних описів в процесі пошуку шляхом логічної переробки знайденої фактографічної інформації.

3.3 Основні моделі пошуку

В даний час інформаційні ресурси тільки веб-простору складають понад двадцяти мільярдів документів, до яких можливий вільний доступ будь-якого користувача. Природно, для того, щоб знайти необхідну інформацію і цієї найбільшої розподіленої повнотекстової бази даних необхідно використовувати

найпотужніші ІПС. Такі системи існують і конкурують один з одним. Сьогодні мільйонам користувачів Інтернет відомі такі інформаційно-пошукові системи, як Google, Yahoo, AltaVista, AllTheWeb, MSN, Яndex, Rambler, які охоплюють мільярди веб-документів. В основу роботи всіх подібних систем покладені спеціальні алгоритми, які є модифікаціями основних підходів – моделей пошуку.

Модель інформаційного пошуку має три ключових аспекти.

1. Формат представлення документа. Під документом розуміється деякий об'єкт, що містить інформацію у зафіксованому вигляді. Документи можуть містити тексти на природній або формалізованій мові, зображення, звукову інформацію тощо.

2. Формат представлення запиту. Під запитом розуміється формалізований спосіб вираження інформаційних потреб користувача системи. Для цього використовується мова пошукових запитів, синтаксис якої змінюється в рамках різних систем.

3. Функція відповідності документа запиту. Ступінь відповідності запита і знайденого документа (релевантність) – суб'єктивне поняття, оскільки результати пошуку, що задовольняють одного користувача, можуть не задовольняти іншого.

В основу традиційних методів покладено три головні підходи, перший з яких базується на теорії множин (булева модель), другий – на векторній алгебрі (векторно-просторова модель), а третій – на теорії ймовірностей (імовірнісна модель). Ці підходи можуть застосовуватися на практиці і в канонічному вигляді, проте у них є спільний недолік, обумовлений припущенням, що зміст документа визначається безліччю слів і стійких словосполучень – термів (*англ.* – *Terms*), які входять в нього без урахування взаємозв'язків, і, більше того, вважаються незалежними. Таке припущення веде до втрати змістовних відтінків, проте воно дозволяє реалізувати пошук і групування документів за формальними ознаками. Відомі такі основні недоліки традиційних моделей:

- *Булева модель* – невисока ефективність пошуку, відсутність контекстних операторів, неможливість ранжирування результатів пошуку.
- *Векторно-просторова* модель пов'язана з розрахунком масивів високої розмірності і в канонічному вигляді малопридатна для обробки великих масивів даних.
- *Імовірнісна модель* характеризується низькою обчислювальною масштабованістю (тобто різким зниженням ефективності при зростанні обсягів даних), необхідністю постійного навчання системи.

Системи, побудована на «рафінованих» пошукових моделях, недостатньо оперативні і володіють слабо розвинутими пошуковими можливостями і засобами узагальнення даних.

Крім представлених нижче, існують і інші моделі пошуку, наприклад, семантичні, в рамках яких робляться спроби організації смислового пошуку за рахунок аналізу граматики тексту, використання баз знань, тезаурусів, онтологій, які реалізують семантичні зв'язки між окремими словами та їх групами. Разом з тим, ефективність систем, що базуються на таких підходах поки, залишається невисокою.

Класична булева модель

Булева модель базується на теорії множин і математичній логіці. Популярність цієї моделі пов'язана насамперед із простотою її реалізації, яка дозволяє індексувати і виконувати пошук у великих документальних масивах.

В рамках булевої моделі документи і запити представляються у вигляді безлічі термів – ключових слів і стійких словосполучень. Тобто розглядається деякий словник $T=\{t_1,\ldots,t_n\}$, де t_i – терми. Термами можуть виступати слова, комбінації цифр, літер тощо. Деякі групи слів також вважаються одним термом, наприклад всі відмінки одного числівника.

Документ, це деяка підмножина словника, набір термів: $D\in\{0,1\}^n$: на *k-й* позиції стоїть одиниця у тому випадку, коли *k-me* слово належить документу і нуль, якщо слово йому не належить. Таким чином вага значення терма в документі приймає лише два значення: $\{0,1\}$.

У булевій моделі запит користувача являє собою логічне вираження, у якому терми зв'язуються логічними операторами кон'юнкції (AND), диз'юнкції (OR) і заперечення (NOT). Відомо, що будь-який логічний вираз можна представити диз'юнкцією деяких висловів, з'єднаних між собою операцією кон'юнкції. Наприклад формула « $t_5 \vee t_7 \wedge NOT\, t_{12}$ » означає, що необхідно знайти документи, які включають п'ятий та сьомий елементи і не включають дванадцятий.

Якщо формула виконана в деякому документі, то вважається, що документ відповідає запиту.

Існує декілька підходів до формування архітектури пошукових систем, відповідних булевій моделі, які знайшли своє втілення в реальних інформаційно-пошукових системах. Однією з реалізацій такої моделі була колись популярна система STAIRS корпорації IBM. База даних цієї системи, що вже стала класичною, складається з наступних основних таблиць:

- текстової, що містить текстову частину всіх документів;
- покажчиків текстів, яка включає покажчики на місцезнаходження документів в текстовій таблиці;

- словникової, що містить всі унікальні слова, що зустрічаються в документах, тобто ті слова, за якими може здійснюватися пошук;
- інверсної, що містить списки номерів документів і координати окремих слів у документах.

Пошук по слову в базі даних системи такої архітектури здійснюється відповідно до алгоритму:

1. Відбувається звернення до словникової таблиці, за якою визначається, чи входить слово до складу словника бази даних, і якщо входить, то визначається посилання в інверсній таблиці на ланцюжок появ цього слова в документах.
2. Відбувається звернення до інверсної таблиці, за якою визначаються номери документів, що містять дане слово, і координати всіх входжень слова в текстах бази даних.
3. За номером документа відбувається звернення до запису таблиці покажчиків текстів. Кожен запис цього файлу відповідає одному документу в базі даних.
4. За номером документа відбувається пряме звернення до фрагмента текстової таблиці – документу, після чого слідує висновок знайденого документа.

Наведений алгоритм охоплює випадок, коли запит складається з одного слова. Якщо ж у запит входить не одне слово, а деяка їх комбінація, то в результаті виконання пошуку по кожному з цих слів запиту формується масив записів, які відповідають входженню цього слова в базу даних. Після закінчення формування масивів результатів пошуку відбувається виявлення релевантних документів шляхом виконання теоретико-множинних операцій над записами цих масивів відповідно до правил булевої логіки.

Така модель іноді використовується у внутрішніх корпоративних системах пошуку, базах даних. Основним недоліком класичної булевої моделі є крайня жорсткість й непридатність до ранжування результатів пошуку за рівнем їх відповідності пошуковим запитам. Якщо слово, що вказане в запиті, присутнє у документі, то він вважається знайденим, в іншому випадку – не знайденим. Не буде знайдений документ, в якому знаходяться лише синоніми слова, у випадку, коли саме слово в документі не зустрічається.

Для усунення цього недоліку Г. Солтон, Э.А. Фоксом и Г. Ву в 1983 році запропонували розширену булеву модель.

Відповідно до цієї моделі, кожному терму приписується вага – значення з інтервалу [0,1]. Замість документа в рамках моделі розглядається вектор з термів (у найпростішому випадку – двох), який також можна розглядати як

точку в квадраті [0,1] х [0,1]. Близькість між документами можна інтерпретувати як деяку нормовану відстань між відповідними точками.

Розширив класичну теорію множин Л.А. Заде і запропонував теорію нечітких множин, за ідеєю якої функція приналежності елемента множини може приймати довільні значення в інтервалі [0,1], а не тільки 0 або 1.

За визначенням, нечіткою множиною на універсальній множині (будь-якої природи) є сукупність пар ступіней приналежності елемента нечіткій множині. Ступінь приналежності – це число з діапазону [0,1]. Чим вище ступінь приналежності, тим більшою мірою елемент відповідає властивостям нечіткої множини. Функцією приналежності називається функція, яка дозволяє обчислити ступінь приналежності довільного елемента універсальної множини до нечіткій множині.

Векторно-просторова модель пошуку

Велика кількість відомих інформаційно-пошукових систем базуються на векторно-просторовій моделі опису даних (*VectorSpaceModel*), яка була запропонована Г. Солтоном в 1975 р і вперше застосованої в системі SMART. Дана модель є класичною алгебраїчною. В рамках цієї моделі документ описується вектором в евклідовому просторі, в якому кожному терму, що використовується в документі, ставиться у відповідність його вагове значення, яке визначається на основі статистичної інформації про його появу як в окремому документі, так і в усьому документальному масиві. Опис запиту, що відповідає необхідній користувачеві тематиці, також являє собою вектор в тому ж евклідовому просторі термів. Для оцінки близькості запиту і документа використовується скалярний добуток відповідних векторів запиту і документа.

В рамках цієї моделі кожному терму в документі відповідає деяка невід'ємна вага.

Кожному запиту, який являє собою також безліч термів, не поєднаних між собою ніякими логічними операторами, також відповідає вектор вагових значень.

Таким чином, кожен документ і запит можуть бути представлені у вигляді *n*-мірного вектора, де *n* загальна кількість термів в словнику моделі. Відповідно до розглянутої моделі, близькість документа до запиту, які як і в попередніх моделях розглядаються як інформаційні вектори, оцінюється як їх скалярний добуток. При цьому вагу окремих термів можна обчислювати різними способами. Один з можливих найпростіших підходів – використовувати нормалізовану частоту входження терма в документ як його вагу в документі.

За формулою визначається матриця M : $M_{ij} = TF_{ij} \cdot IDF_i$, де TF_{ij} (TermFrequency, частота терма) – відносна доля слова *i* в документі *j*, IDF_i (InversedDocumentFrequency) – величина, зворотна кількості документів, що

містять слово *i*. Кожний документ в даній матриці представляється у вигляді стовпчика (*j* фіксоване, *i* міняється).

Для того, щоб підрахувати міру релевантності, представимо запит у вигляді вектора з координатами 0 або 1: $Q=t_3 \wedge t_5=\{0,0,1,0,1,0\ldots,0\}$.

Кожний документ – набір таких координат: багато нульових координат (терми, які не зустрічаються) і декілька ненульових координат.

Мірою релевантності $R(Q,D_j)$ вважаєтьсякосинус кута між вектором запиту Q і документом D_j . Для того, щоб підрахувати це число береться скалярний добуток векторів Q і D_j : $R(Q,D_j)=\cos\alpha=QD_j/|Q||D_j|$.

Нормалізація необхідна для того, щоб врівноважити ваги документів з різною кількістю слів.

Векторно-просторова модель представлення даних забезпечує системам, побудованим на її основі, такі можливості, як:

- обробку запитів без обмежень їх довжини;
- простоту реалізації режиму пошуку подібних документів (кожен документ може розглядатися як запит);
- збереження результатів пошуку з можливістю виконання уточнюючого пошуку.

Разом з тим у векторно-просторового моделі не передбачено використання логічних операцій у запитах, що істотно обмежує її придатність.

Ймовірнісна модель пошуку

В 1977 році Робертсон і Спарк-Джоунз обґрунтували і реалізували ймовірнісну модель. Релевантність цієї моделі розглядається як ймовірність того, що даний документ може виявитися цікавим для користувача.

При цьому мається на увазі наявність існуючого первинного набору релевантних документів, вибраних користувачем або отриманих автоматично при деякому спрощеному припущенні. Ймовірність виявитися релевантним для кожного наступного документа розраховується на підставі співвідношення зустрічаємості термів в релевантному наборі і в решті колекції.

Функціонування моделі базується як на експертних оцінках, що отримуються у результаті навчання моделі, які визнають документи з навчальної колекції релевантними/нерелевантними, так і на подальших оцінках ймовірності того, що документ є релевантним запитом виходячи зі складу його термів.

Документом вважається множина слів без урахування частоти зустрічаємості слова в документі. Можна також представити множину у вигляді звичайного мулевого вектора $D=(d_1,\ldots,d_n)$, де n – кількість всіх термів, а d_i може набувати значень з множини $\{0,1\}$. Запитом вважається множина слів.

Якщо для запиту відомі ці оцінки ймовірностей для всіх документів, то документи можна сортувати за ними і виводити користувачам у спадному порядку. Тобто імовірнісна модель пошуку передбачає визначення ймовірностей відповідності запиту для документів, сортування та надання документів з ненульовою ймовірністю користувачеві.

З самого початку в ймовірнісної моделі використовувалося спрощення, яке допускає незалежність входження в документ будь-якої пари термів (тому такий підхід називається «наївним» Байєсівським).

При цьому в ймовірнісної моделі пошуку передбачається наявність навчальних наборів релевантних і нерелевантних документів, обраних користувачем або отриманих автоматично при якомусь початковому припущенні. Імовірність того, що документ є релевантним, розраховується на підставі співвідношення появи термів в релевантному і нерелевантному масиві документів.

У разі застосування експертних оцінок процес пошуку є ітераційним (у реальних системах, що використовують елементи ймовірності моделі, як експертні оцінки можуть розглядатися, наприклад, переваги користувачів при виборі документів, що їх цікавлять). На кожному кроці ітерації, завдяки режиму зворотного зв'язку, визначається безліч документів, зазначених користувачем як тих, що задовольняють його інформаційним потребам.

3.4 Інформаційно-пошукові мови

Інформаційно-пошукові мови є основними компонентами інформаційно-пошукових систем, за допомогою яких, зокрема, реалізуються інтерфейси між користувачами і системами.

На відміну від реляційних СУБД, у систем повнотекстового пошуку не існує стандартизованої мови запитів. У кожної системи цього типу існує свій спосіб завдання критеріїв пошуку.

Дуже часто мови запитів ІПС наближені до SQL, проте кожній з пошукових систем притаманний ряд індивідуальних особливостей, пов'язаних з такими моментами, як:

- інтерпретація операцій, які задають порядок розташування слів у тексті (операцій контекстної близькості);
- обчислення рівня релевантності знайдених документів запитах для представлення результатів пошуку;
- застосування нестандартних для реляційних СУБД функцій, наприклад, таких як знаходження документів за принципом подібності змісту, побудова дайджестів з фрагментів документів, сніпетів (від англ. Snippet –

фрагмент, уривок), що включаються пошуковими системами в списки знайдених документів і т.п.

У різних повнотекстових інформаційно-пошукових системах застосовуються різні архітектурні рішення, що охоплюють структури даних, алгоритми їх обробки, методи організації пошуку. Разом з тим, у сучасних інформаційно-пошукових систем багато спільних властивостей, наприклад, всі з них забезпечують пошук хоча б по одному слову, більшість подібних систем реалізують граматичний пошук як результат застосування лінгвістичного аналізу (наприклад, в російськомовних Яндекс і Рамблер по терму із запиту «людина» знаходяться не тільки словозміни «людини», «людині», а й множина – «люди»). Більшість із сучасних систем здатні реалізовувати контекстний пошук фрази, укладеної в лапки (Google, Alltheweb, AltaVista, Яндекс тощо), пошук з використанням булевих операторів AND, OR і NOT, а також можливістю вказівки дужок для групування термів і операторів. Функції контекстної близькості свого часу отримали найбільший розвиток в системі Lycos, де були реалізовані за допомогою чотирьох операторів: ADJ, NEAR, FAR і BEFORE.

У найпопулярнішій в світі системі Google використовується досить лаконічний набір операторів (http://www.googleguide.com), основні з яких – це кон'юнкція (використовується за замовчуванням, система видає документи, що містять всі слова запиту), диз'юнкція (OR) і заперечення (–).

Окремо розглядається можливість пошуку за параметрами документів, яка найчастіше дозволяє обмежувати діапазон пошуку значеннями URL, дат, заголовків. У більшій частині систем вийти на можливість пошуку за параметрами можна з режиму розширеного пошуку.

В Google, наприклад, забезпечується пошук по сайту («site:»), визначення посилань на сайт («admissionsite:»), пошук за цінами, наприклад «DVDplayer $ 150..250», пошук по країнам, датам, доменам тощо. У багатьох системах забезпечується пошук не тільки за даними у форматі HTML, а й у форматах PDF, RTF, DOC (MsWord), PS.

Останнім часом набули поширення адаптивні інтерфейси уточнення запитів, найчастіше реалізовані шляхом застосування методів кластерного аналізу до результатів первинного пошуку. З'явилося таке поняття, як метод «папок пошуку» (CustomSearchFolders), що об'єднує безліч підходів, загальна властивість яких– спроба згрупувати результати пошуку і представити групи найбільш пов'язаних документів (кластери) в зручному для користувачів вигляді.

Наприклад, в пошукових серверах Vivisimo (http://www.vivisimo.com), Mooter (http://www.mooter.com) або Nigma (http://www.nigma.ru) застосовується візуальний підхід до подання результатів пошуку шляхом групування релевантних документів за категоріями. В іншому пошуковому сервері iBoogie

(http://www.iboogie.com) результати пошуку відображаються у вигляді, близькому до екрану провідника Windows. Слова і словосполучення в так званих «інформаційних портретах», застосовуваних, наприклад, у корпоративних інформаційно-аналітичних системах Галактика Zoom і InfoStream, також дозволяють адаптивне уточнення первинних запитів.

3.5 Характеристики інформаційного пошуку

Незважаючи на те, що наведені вище моделі є класичними, в чистому вигляді вони застосовуються тільки в моделях систем. Кожна з розглянутих моделей пошуку має недоліки, які перешкоджають їх безпосередньому застосуванні в задачах інтеграції інформаційних потоків.

Основні недоліки наступні:

- *булева модель* – невисока ефективність пошуку, жорсткий набіроператорів, неможливість ранжирування;
- векторно-просторова модель пов'язана з розрахунком масивів високої розмірності, малопридатна для обробки великих масивів даних;
- імовірнісна модель характеризується низькою обчислювальною масштабністю, необхідністю постійного навчання системи.

Наведені класичні моделі спочатку припускали розгляд документів, як безлічі окремих слів, що не залежать один від одного. Така спрощена концепція має назву «BagofWords». У реальних системах це спрощення долається, наприклад, розширеною булевою моделлю, яка враховує контекстну близькість (оператори NEAR, ADJ в багатьох відомих системах).

Системи, що базуються на ймовірнісної моделі, враховують входження словосполучень і зв'язку окремих термінів, хоча більшість з відомих систем боротьби зі спамом, побудовані на ймовірнісної моделі, все-таки базуються на спрощеному підході незалежності окремих слів.

На практиці найчастіше використовуються гібридні підходи, що поєднують можливості булевої і векторно-просторового моделі, часто додають оригінальні методи семантичної обробки інформації. Найчастіше всього в інформаційно-пошукових системах процедура пошуку виконується відповідно до булевої моделі, а результати ранжуються за вагою відповідно до моделі векторного простору.

Яка б не була модель, кожна пошукова система вимагає налаштування, для чого необхідна оцінка параметрів пошуку. Саме завдяки оцінці якості можна говорити про застосовність або не застосовні тієї чи іншої моделі.

Існує багато характеристик пошуку, з яких дві визнані основними – це повнота (*recall*) і точність (*precision*). Багато уваги в даний час відводиться

також такій смисловий характеристиці, як пертинентність. Ця характеристика інформаційно-пошукових систем означає відповідність отриманих в результаті пошуку документів інформаційним потребам користувача, а не формальному відповідності документа запиту. Для обчислення показників якості пошуку прийнято розглядати таблицю, яку заповнюють за результатами пошуку в навчальній колекції документів. Цей підхід був запропонований в рамках створеної Американським Інститутом Стандартів (NIST) конференції з оцінювання систем текстового пошуку TextREtrievalConference (TREC, http://trec.nist.gov). Таблиця результатів пошуку має наступний вигляд:

Документи	Видані	Не видані
Релевантні	***a***	***c***
Нерелеванті	***b***	***d***

Відповідно до таблиці розраховуються основні показники ефективності інформаційно-пошукових систем.

Коефіцієнт повноти (recall) – частина виданих релевантних документів серед всіх релевантних документів:

$$r = a/(a+c) \quad .$$

Коефіцієнт точності (precision) – частина виданих релевантних документів серед всіх виданих:

$$p = a/(a+b) \quad .$$

Коефіцієнт шуму – частина виданих нерелевантних документів серед всіх виданих:

$$l = b/(a+b) \quad .$$

Коефіцієнт осаду – частина виданих не релевантних документів до всіх нерелевантних:

$$q = b/(b+d) \quad .$$

Коефіцієнт специфічності – частина не виданих нерелевантних серед всіх нерелевантних:

$$k = d/(b+d) \quad .$$

Також з наведеної таблиці розраховується коефіцієнт акуратності та помилка:

$$acc=(a+d)/(a+b+c+d) \quad ,$$

$$err=(b+c)/(a+b+c+d) \quad .$$

На практиці значення точності і повноти набагато зручніше розраховувати з використанням матриці неточностей (*confusionmatrix*). У разі якщо кількість класів відносно невелике (не більше 100-150 класів), цей підхід дозволяє досить наочно представити результати роботи класифікатора.

Матриця неточностей – це матриця розміру N на N, де N – це кількість класів. Стовпці цієї матриці резервуються за експертними рішеннями, а рядки за рішеннями класифікатора. Коли ми класифікуємо документ з тестової вибірки ми збільшуємо число, що стоїть на перетині рядка класу який повернув класифікатор і стовпчика класу до якого дійсно відноситься документ.

	0.91	0.96	0.94	0.75	1.00	0.83	0.85	0.97	1.00	0.86	1.00	0.79	1.00	0.75	1.00	1.00	0.96	0.90	0.81	0.89	0.94	0.98	0.86	0.89	0.94	0.92	0.96
0.80		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
0.95	1	94	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
1.00	2	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.29	3	0	0	6	0	0	3	2	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	3	0	2	0
1.00	4	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.50	5	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	1	1
0.92	6	1	0	0	0	0	152	0	0	1	0	0	0	0	0	0	0	1	4	2	3	0	0	0	0	2	0
0.97	7	1	0	1	0	0	0	256	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	2	0
0.33	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
0.97	9	0	0	0	0	0	0	0	0	69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0.82	10	0	0	0	0	0	2	0	0	0	18	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
0.87	11	0	0	0	0	0	0	0	0	0	0	34	0	4	0	0	0	0	0	0	0	0	0	1	0	0	0
1.00	12	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.57	13	0	0	0	0	0	0	0	0	0	0	9	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
0.63	14	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	3	0	0	0	0	0	0	0	0	0
0.50	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1	0	0	0	0	0	0
0.77	16	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	47	0	1	3	4	0	0	2	0	1	0
0.87	17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	69	1	2	5	0	0	0	0	0	0
0.97	18	0	0	0	0	1	4	0	0	1	0	0	0	0	0	0	0	0	197	1	0	0	0	0	0	0	0
0.78	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	35	183	13	0	0	2	0	1	0
0.97	20	0	0	0	0	0	10	3	0	1	0	0	0	0	0	0	0	0	0	4	702	0	0	0	0	6	0
0.93	21	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	0	2	0	0	0
0.29	22	0	0	1	0	0	2	0	0	6	0	0	0	0	0	0	0	0	1	1	1	0	6	2	0	1	0
0.91	23	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	3	6	0	0	115	0	0	0
1.00	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0
0.93	25	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	4	5	0	0	0	1	196	0
0.98	26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	78

Рис. 4. Матриця неточностей (26 класів, результуюча точність – 0.8, результуюча повнота – 0.91)

Як видно з прикладу, більшість документів класифікатор визначає вірно. Діагональні елементи матриці явно виражені. Проте в рамках деяких класів (3, 5, 8, 22) класифікатор показує низьку точність.

Маючи таку матрицю точність та повнота для кожного класу розраховується дуже просто. Точність дорівнює відношенню відповідного діагонального елемента матриці і суми всього рядка класу. Повнота – відношенню діагонального елемента матриці і суми всього стовпця класу.

На практиці виявляється, що неважко побудувати систему, що володіє високою точністю пошуку при низькій повноті або високою повнотою при низькій точності. Тому для адекватної оцінки використовують комбіновану міру, яка враховувала б одночасно і повноту і точність. В якості такого заходу традиційно використовується *F*-міра:

$$F(p,r)=(\beta^2)pr/¿¿2p+r),$$

де β – параметр, що задає пріоритет точності над повнотою та приймає значення $0 < \beta < 1$, якщо пріоритет надається точності, $\beta > 1$, якщо пріоритет надається повноті і $\beta = 1$ для отримання збалансованої *F*-міри.

F-міра являє собою гармонійне середнє між точністю і повнотою. Вона наближається до нуля, якщо точність або повнота наближається до нуля.

F-міра є хорошим кандидатом на формальну метрику оцінки якості класифікатора. Вона зводить до одного числа дві інших основоположних метрики: точність і повноту. Маючи у своєму розпорядженні подібний механізм оцінки набагато простіше прийняти рішення про те чи є зміни в алгоритмі в кращу сторону чи ні.

З наведених параметрів оцінки якості пошуку найчастіше використовують два основних:

- *повнота (recall)*– частина знайдених релевантних документів у загальному числі релевантних документів масиву;
- *точність (precision)*– частина релевантного матеріалу у відповіді пошукової системи.

Саме ці параметри застосовуються на регулярній основі для вибору моделей та їх параметрів в рамках конференції по оцінюванню систем текстового пошуку (TREC). Матеріали цієї конференції вільно доступні за адресою http://trec.nist.gov/pubs.html. Існує також аналогічна російська конференція РОМІП (Російський семінар по Оцінці Методів Інформаційного пошуку – http://romip.ru).

Для оцінки якості пошуку існує ще ряд критеріїв, наприклад, так званий 11-і точковий графік повноти/точності, що показує залежність точності від повноти при кроці повноти в 10%. Чим вище проходить даний графік, тим вища якість інформаційного пошуку, що може надати система.

При оцінці різних інформаційно-пошукових систем за допомогою 11-точкового графіка кращої вважається та система, в якій висока точність досягається при малій повноті, що свідчить про гарне ранжирування результатів пошуку. Крім того, кращою визнається та система, для якої площа під відповідною інтерполяційної кривої є найбільшою.

4 Узагальнений алгоритм пошуку інформації в Інтернет

Інтерес до питання про пошук інформації в мережі Інтернет не слабшає протягом усього часу існування мережі. Пошук може вестися як користувачем-любителем, так і професіоналом. При проведенні пошуку інформації, що задовольняє інформаційним потребам користувача, необхідно знати, від чого залежить успішний пошук, і які проблеми виникають при роботі з інформацією.

Узагальнений алгоритм пошуку інформації в ІПС мережі Інтернет можна представити наступною послідовністю дій:

1. Формування завдання на пошук інформації.
2. Вивчення предметної області завдання на пошук.
3. Підготовка пошукового образу та пошукового припису.
4. Вибір інформаційно-пошукової системи Інтернет.
5. Введення *URL*-адреси.
6. Введення пошукового припису.
7. Аналіз пошукового результату або уточнення пошукового припису, якщо результат відсутній.
8. Завершення пошуку, якщо результат задовольняє інформаційну потребу, або повтор пошуку починаючи з п. 2.

На основі запропонованого найпростішого алгоритму пошуку інформації розроблено модель, що дає можливість перейти до автоматизованого смислового пошуку інформації в розподілених ІПС мережі Інтернет.

Узагальнений алгоритм реалізації моделі пошуку в розподілених ІПС мережі Інтернет виглядає наступним чином:

1. «Зразок документа» (смислова задача), що представляє собою шаблон пошуку, вводиться експертом вручну.
2. З документа виділяється тема запиту, і визначаються пошукові приписи.
3. Розширюється тема запиту за рахунок синонімії та асоціативних запитів.
4. Формується пошуковий образ запиту на основі частотного словника з розбивкою його на окремі пошукові приписи.
5. Проводиться первинний пошук посилань на релевантні документи в існуючих пошукових Інтернет-машинах, загальний результат поміщається в сховищі даних.
6. Здійснюється закачування знайдених документів в сховище даних.
7. Для кожного документа в сховищі даних формується пошуковий образ документа (ПОД).
8. Проводиться ранжування документів відповідно до заданої теми (п.2). Чим точніший зміст в тілі документа, тим він має більш високий рейтинг.

9. Проводиться реферування знайдених документів і здійснюється передача рефератів для ознайомлення і аналізу експерту відповідно до рейтингу.

Запропонований підхід до організації смислового пошуку інформації в розподілених інформаційних системах мережі Інтернет дозволяє якісно поліпшити результати пошукових запитів до пошукових машин мережі Інтернет, дозволяє автоматизувати процес обробки релевантної інформації з ранжируванням смисловий (пертенентної) інформації відповідно до заданої теми, що дає можливість експертам відійти від ручного послідовного перегляду знайдених ресурсів.

5 Рекомендації щодо оптимізації веб-сайту для пошукової системи для веб-майстра.

Оптимізація для пошукової системи найчастіше передбачає внесення незначних змін до змістовних частин веб-сайту. Якщо розглядати такі зміни окремо, вони видаються частковим покращенням, але в поєднанні з іншими заходами оптимізації ці зміни можуть мати помітний вплив на сприйняття користувачами сайту та його відображення в результатах пошуку. Рішення щодо оптимізації необхідно приймати, орієнтуючись насамперед на задоволення потреб відвідувачів сайту. Вони є головними споживачами вмісту та використовують пошукові системи для пошуку інформації. Зосередження на певних прийомах покращення рейтингу в результатах пошукових систем може не призвести до бажаного результату. Оптимізація для пошукової системи – це виокремлення сайту з-поміж інших, коли він з'являється в полі зору пошукових систем.

Оптимізація для пошукової системи – це задача, яка цілком покладається на розробника інформаційного ресурса, веб-майстра. Тому термінологія та деталі цієї роботи потребують технічного розуміння процесів.

Безперечним лідером серед пошукових систем є продукт корпорації Google, тому деякі вимоги, а також приклади отримання результатів посилаються безпосередньо на означену систему.

5.1 Керування тегами веб-сайту

Створення унікальних, точних назв сторінок

Тег <title> (назва) повідомляє користувачам та пошуковій системі про тему кожної окремої сторінки. Тег <title> має міститись всередині тегу <head> документа HTML. В ідеалі, необхідно створити унікальну назву для кожної сторінки вашого сайту. Якщо документ з'являється на сторінці результатів пошуку, вміст тегу <title> зазвичай відображатиметься у першому рядку результатів. Слова з пошукового запиту користувача будуть виділені у назві напівжирним шрифтом. Це допомагає користувачам зрозуміти, чи відповідає сторінка їхньому запиту.

Назва (тег <title>) домашньої сторінки може містити назву веб-сайту та підприємства, а також іншу важливу інформацію, наприклад про місце розташування підприємства чи декілька основних напрямків його діяльності та пропозиції. Теги <title> подальших сторінок сайту мають точно описувати призначення кожної окремої сторінки, але також можуть включати назву сайту чи підприємства.

У випадках із представленням документів, статей тощо, найкращим варіантом буде винесення в зміст тегу повної назви документа та авторів, а також, можливо, знов назви самого інформаційного ресурсу.

Корисні поради з використання тегу <title> сторінки

- **Точний опис вмісту сторінки** –назву, яка дійсно відображає тему вмісту сторінки.

Необхідно уникати:

- використання назв, які не стосуються вмісту на сторінці
- використання стандартних або незрозумілих назв, наприклад "Без назви" чи "Нова сторінка 1"

- **Унікальність тегів <title> для кожної сторінки –** кожна з сторінок, в ідеалі, має мати унікальний тег <title>, який допоможе пошуковим системам розпізнати, чим саме ця сторінка відрізняється від інших на сайті.

Необхідно уникати:

- використання одного й того ж тегу <title> для всіх сторінок вашого сайту чи великої групи сторінок

- **Використання коротких, але описових назв –** короткі назви також можуть бути інформативними. Якщо назва задовга, пошукова система відображатиме в результатах пошуку лише її частину, що може не задовольнити користувачів.

Необхідно уникати:

- використання громіздких назв, які нічим не допомагають користувачам
- нагромадження непотрібних ключових слів у тегах <title>

Використання мета-тегу "description" (опис)

Мета-тег "description" сторінки дає пошуковим системам короткий виклад вмісту сторінки. У той час як назва сторінки може складатися з декількох слів або фрази, її мета-тегом "description" може містити речення чи два, або навіть невеликий абзац. Як і тег <title>, мета-тег "description" знаходиться всередині тегу <head> документа HTML.

Мета-теги "description" важливі, оскільки пошукова система може використовувати їх як фрагменти для представлення сторінок. Додавання мета-тегів "description" до кожної сторінки завжди корисне на випадок, якщо Google не знайде тексту, отриманого як шаблон пошуку, щоб використати його як фрагмент.

Наприклад, в результатах пошуку Google фрагменти відображаються в результатах пошуку під назвою сторінки та над її URL-адресою. Слова, наявні у запиті пошуку користувача, будуть виділені у фрагменті напівжирним шрифтом. Це дозволяє користувачу розпізнати, чи вміст на сторінці відповідає його запиту.

Корисні поради з використання мета-тегів "description"

- **Точний короткий виклад вмісту сторінки** – це опис, який надаватиме інформацію та зацікавить користувачів, коли вони побачать мета-тег "description" як фрагмент сторінки у результаті пошуку.

Необхідно уникати:

- створення мета-тегів "description", які не мають відношення до вмісту на сторінці
- використання базових описів, наприклад "Веб-сторінка" чи "Сторінка документу"
- наповнення тегу "description" лише ключовими словами
- копіювання повного вмісту документа до мета-тегу "description"

- **Використання унікальних мета-тегів "description" для кожної сторінки –** окремий мета-тег "description" для кожної сторінки допомагає як користувачам, так і пошуковим системам, особливо, якщо користувачі відкриють під час пошуку декілька сторінок з одного сайту. Якщо сайт містить тисячі чи навіть мільйони сторінок, навряд чи фізично можливе присвоєння мета-тегів "description" вручну. У такому випадку можна автоматично створювати мета-теги "description" на основі вмісту кожної сторінки.

Необхідно уникати:

- використання одного й того ж мета-тегу "description" для всіх сторінок сайту чи великої групи сторінок

Вдосконалення структури URL-адрес

Створення описових категорій і назв файлів для документів на веб-сайті не лише допоможе краще організувати сайт, але й полегшить сканування документів пошуковими системами. Також це допоможе створити простіші та кращі для сприйняття URL-адреси для тих, хто хоче ознайомитися з вмістом. Задовгі та незрозумілі URL-адреси, які містять лише декілька відомих слів, можуть відлякувати відвідувачів. Такі URL-адреси можуть спантеличувати та відлякувати. У користувачів можуть виникнути труднощі з запам'ятовуванням такої URL-адреси чи створенням посилання на неї. Окрім того, користувачі можуть вирішити, що частина URL-адреси не потрібна, особливо, якщо в ній відображається багато невідомих параметрів. Вони можуть відкинути частину, руйнуючи посилання.

Деякі користувачі можуть переходити на сторінку, використовуючи її URL-адресу, як текст прив'язки. Якщо URL-адреса містить відповідні слова, це надасть користувачам і пошуковим системам більше інформації про сторінку, ніж ідентифікатор чи параметр з дивною назвою. В GoogleURL-адреса документа відображається як частина результату пошуку, під назвою документа та його фрагментом. Як і в назві та фрагменті, в URL- адресі, яка з'являється в

результаті пошуку, слова з пошукового запиту користувача буде виділено напівжирним шрифтом.

Пошукові системи повинні сканувати всі типи структур URL-адрес, навіть дуже складні, але варто витратити трохи часу, щоб зробити URL-адреси якомога простішими як для користувачів, так і для пошукових систем. Деякі веб-майстри намагаються досягнути цього шляхом перезапису своїх динамічних URL-адрес на статичні.

Корисні поради щодо побудови структури URL

- **Використання слів в URL-адресі –** URL-адреси зі словами, що відповідають вмісту та структурі сайту, є зручнішими для відвідувачів, які здійснюють пошук на сайті. Відвідувачі краще запам'ятовують їх і охочіше переходять за ними.

Необхідно уникати:

- довгих URL-адрес із непотрібними параметрами та ідентифікаторами сесій
- використання базових назв сторінок, наприклад "page1.html"
- надмірного використання ключових слів, наприклад "baseball-cards-baseball-cards-baseball-cards.htm"

- **Створення структури простого каталогу,** яка організовує вміст і дає відвідувачам змогу легко орієнтуватись на сайті.

Необхідно уникати:

- складної структури підкаталогів, наприклад ".../dir1/dir2/dir3/dir4/dir5/dir6/page.html"
- використання назв каталогів, які не мають жодного відношення до їх вмісту

- **Надання однієї версії URL-адреси для доступу до документа –** щоб уникнути переходів частини користувачів за однією версію URL-адреси, а решти – за іншою (це може розділити поінформованість про вміст між цими URL-адресами), необхідно сконцентруватись на використанні та посиланні на одну URL-адресу в структурі та внутрішніх посиланнях сторінок. Помітивши, що користувачі отримують доступ до одного й того ж вмісту за допомогою різних URL-адрес, можна позбутись цієї проблеми, налаштувавши переадресацію 301 з небажаних URL-адрес на домінуючі.

Необхідно уникати:

- доступу до одного вмісту зі сторінок підкаталогів і кореневого каталогу (напр. "domain.com/page.htm" і "sub.domain.com/page.htm")
- поєднання версій URL-адрес із www. та без www. у структурі внутрішніх посилань

- використання зайвих великих літер в URL-адресі (багато користувачів звикли до малих літер в URL і краще їх запам'ятовують).

5.2 Реструктуризація сайту

Спрощення пересування по сайту

Питання пересування по веб-сайту важливе, оскільки від нього залежить як швидко відвідувачі знайдуть потрібний вміст. Пересування також допомагає пошуковим системам зрозуміти, який саме вміст веб-майстер вважає важливим.

Усі сайти мають домашню чи "кореневу" сторінку, яка є сторінкою з найвищою відвідуваністю на цьому сайті та початковою точкою пересування для багатьох відвідувачів. Якщо на сайті не лише декілька сторінок, варто подумати про те, як відвідувачі будуть переходити з загальної (кореневої) сторінки на сторінку з більш детальним вмістом. Якщо існує багато сторінок, які висвітлюють певну тему, доцільно створити окрему сторінку з описом таких пов'язаних сторінок (напр. коренева сторінка –> перелік пов'язаних тем –> конкретна тема).

Sitemap – проста сторінка на сайті, яка відображає структуру веб-сайту та зазвичай складається з ієрархічного переліку його сторінок. Користувачі можуть завітати на ці сторінки, якщо у них виникнуть проблеми з пошуком сторінок на сайті. Хоча пошукові системи також відвідуватимуть ці сторінки, щоб краще визначити область для сканування сторінок на сайті, в основному вони зорієнтовані на звичайних відвідувачів. Така сторінка має бути відтворена у форматі XML і її зміст – найкращий спосіб проінформувати пошукову систему про структуру сайту.

Корисні поради для пересування по сайту

- **Створення простої спадної ієрархію –** наскільки це можливо, полегшує користувачам перехід від загального вмісту до більш конкретного, який вони шукають на сайті, за рахунок сторінки для пересування та прив’язки її зі структурою внутрішніх посилань.

Необхідно уникати:

- створення складних вузлів посилань переходів, напр. приєднання кожної окремої сторінки сайту до інших сторінок

- надмірного поділу та нашарування вмісту (щоб дістатись до вмісту в глибині сайту, необхідно здійснити двадцять кліків)

- **Використання переважно текст для переходів –** управління переходами від сторінки до сторінки на сайті за допомогою текстових посилань полегшує сканування та сприйняття вашого сайту системами пошуку. Багато

користувачів надають перевагу такому підходу, особливо на пристроях, які не підтримують Flash чи JavaScript.

Необхідно уникати:

- переходів, які повністю базуються на спадних меню, зображеннях чи анімаціях (більшість, але не всі, пошукових систем можуть віднайти такі посилання на сайті, проте, якщо користувач зможе отримати доступ до всіх сторінок на сайті за допомогою звичних текстових посилань, це покращить доступність сайту)

- **Використання "крокових" переходів** – ряду внутрішніх посилань вгорі чи внизу сторінки, що дозволяють відвідувачам швидко перейти назад до попереднього розділу чи на кореневу сторінку. Багато "крокових" структур мають найзагальнішу сторінку (зазвичай кореневу) як першу, розміщену крайньою зліва в рядку посилань, а вправо від неї перелічено детальніші розділи.

- **Розміщення на сайті сторінку HTMLsitemap, а також використання файл XMLSitemap** – простої сторінки, карти сайту з посиланнями на всі сторінки чи на найважливіші з них. Створення файлу XMLSitemap для сайту допоможе пошуковим системам виявити сторінки на вашому сайті.

Необхідно уникати:

- прострочення терміну дії сторінки HTMLsitemap і руйнування посилань;

- створення HTMLsitemap, де сторінки просто перелічено, а не організовано, наприклад, за темою.

- **Застосування корисної сторінки 404** – інколи користувачі через зламане посилання чи неправильно введену URL-адресу заходитимуть на неіснуючу сторінку на вашому сайті. Значно покращити враження користувачів може стандартна сторінка 404, яка плавно спрямує їх назад на робочу сторінку вашого сайту. На сторінці 404 варто мати посилання назад до кореневої сторінки, а також посилання на популярний або схожий вміст на сайті.

Необхідно уникати:

- індексації пошуковими системами сторінок 404;
- надання лише нечіткого повідомлення "Не знайдено", "404" чи взагалі відсутності сторінки 404;
- використання дизайну для сторінки 404, який відрізняється від решти сайту.

Корисні поради для вмісту

- **Створення легкого для сприйняття тексту** – користувачам подобається вміст, написаний просто та легкий для читання.

Необхідно уникати:

- написання неохайного тексту з багатьма орфографічними та граматичними помилками;
- вставляння тексту в зображення текстуального вмісту (можливо, користувачі захочуть скопіювати текст, а пошукові системи не зможуть зчитати його).

- **Зосередження навколо теми** – завжди варто організовувати вміст таким чином, щоб відвідувачі відчували, де починається одна змістовна тема та закінчується інша. Поділ вмісту на логічні частини чи розділи допомагає користувачам швидше віднайти потрібний вміст.

Необхідно уникати:

- розміщення великої кількості тексту на різну тематику на сторінці без параграфів, підзаголовків чи схематичного поділу

- **Використання відповідної мови -** слів, які може шукати користувач, щоб натрапити на необхідний вміст. Користувачі, обізнані з певною темою, можуть використовувати інші ключові слова у своїх пошукових запитах, ніж ті, хто вперше з нею стикнулись.

- **Створення нового, унікального вмісту** – новий вміст не лише змусить існуючих відвідувачів повертатись, але й залучить багатьох нових.

Необхідно уникати:

- переформулювання (чи навіть копіювання) існуючого вмісту, який не є цінним для користувачів
- дублювання чи використання ідентичних версій вмісту на сайті

- **Ексклюзивний вміст або послуги** – створення нових, корисних послуг, які не пропонує жоден інший сайт. Можна також подати оригінальне дослідження, повідомити захоплюючу новину чи надати певні привілеї для своєї унікальної бази користувачів. Інші сайти можуть мати недостатньо ресурсів або досвіду для цього.

- **Створення вміст у першу чергу для користувачів, а не пошукових систем** – розробка сайту для потреб відвідувачів, не забуваючи про його доступність для пошукових систем – це зазвичай дає позитивний результат.

Необхідно уникати:

- вставляння численних непотрібних ключових слів, які націлені на пошукові системи, але дратують і не мають змісту для користувачів
- текстових блоків, схожих на цей: "часті орфографічні помилки, які призводять до потрапляння на цю сторінку", які зазвичай не мають значення для користувачів
- оманливого приховування тексту від користувачів, але показу його для пошукових систем

Створення кращого текст прив'язки

Текст прив'язки – це текст, що активізується при натисканні та є доступним користувачам у результаті посилання. Він розміщений всередині тегу прив'язки .

Такий текст повідомляє користувачам і пошуковим системам дещо про сторінку, до якої він прив'язаний. Посилання на сторінці можуть бути внутрішніми – які скеровують на інші сторінки на сайті, або ж зовнішніми – які скеровують на вміст інших сайтів. У будь-якому з цих випадків, чим кращим є текст прив'язки, тим легше користувачам пересуватись по сайту, а пошуковій системі – зрозуміти тему сторінки, до якої прив'язаний текст.

- **Потрібно вибирати описовий текст** – текст прив'язки, який використовується для посилання, має давати принаймні основне уявлення про сторінку, до якої він прив'язаний.

Необхідно уникати:

- написання базового тексту прив'язки, як наприклад "сторінка", "стаття" чи "натисніть тут"
- використання тексту, який не відповідає темі чи не пов'язаний із вмістом сторінки, до якої він прив'язаний
- використання в більшості випадків URL-адреси сторінки як тексту прив'язки (хоча існують певні легітимні випадки такого застосування, наприклад реклама чи посилання на нову адресу веб-сайту)

- **Створення стислого тексту** – короткого, але описового тексту, зазвичай, декілька слів або коротка фраза.

Необхідно уникати:

- написання довгого тексту прив'язки, наприклад довгих речень чи короткого параграфу

- **Форматування посилань так, щоб їх було легко помітити** – користувачі мають легко розрізняти звичайний текст і текст прив'язки посилань. Вміст стане менш корисним, якщо користувачі пропустять посилання чи випадково натиснуть на них.

Необхідно уникати:

- використання CSS чи стилізації тексту, які надають посиланням такого ж вигляду, як звичайному тексту

- **Текст прив'язки для внутрішніх посилань** – зазвичай посилання сприймається, як спрямування на зовнішні веб-сайти, але розробка текстів прив'язки, які використовуються для внутрішніх посилань, може допомогти користувачам і пошуковим системам краще здійснювати пошук на вашому сайті.

Необхідно уникати:

- використання задовгого чи переповненого ключовими словами тексту прив'язки лише для пошукових систем
- створення непотрібних посилань, які не допомагають користувачам орієнтуватись на сайті

Правильне використання тегів заголовків

Теги заголовків використовуються на сторінці, щоб представити користувачам її структуру. Існує шість розмірів тегів заголовків, які починаються з <h1>, найважливішого, та закінчуються <h6>, найпростішого.

Оскільки теги заголовків зазвичай роблять текст, який в них міститься, більшим за звичайний текст на сторінці, це візуально підказує користувачам, що цей текст важливий, а також може натякнути їм про вміст, який міститься під текстом заголовку. Декілька розмірів заголовків, що використовуються по порядку, створюють ієрархічну структуру вмісту та полегшують користувачам пошук у документі.

Корисні поради для тегів заголовків

- **Висвітлення як у конспекті** – заголовки відокремлюють нові «теми», основні пункти та підпункти вмісту сторінки.

Необхідно уникати:

- розташування тексту всередині тегу заголовку, який не допомагає визначити структуру сторінки
- використання тегів заголовків, де природніше було б застосувати теги і
- використання тегів заголовків різних розмірів

- **Використання невеликої кількості заголовків на сторінці –** лише там, де необхідно. Забагато тегів заголовків на сторінці може ускладнити користувачам перегляд вмісту та визначення закінчення однієї теми й початку іншої.

Необхідно уникати:

- надмірного використання тегів заголовків на сторінці
- вставляння всього тексту сторінки в тег заголовку
- використання тегів заголовків лише для зміни стилю тексту, а не відображення структури.

Оптимізація використання зображень

Зображення видаються простим компонентом сайту, але можна оптимізувати їхнє використання. Кожне зображення може мати окрему назву файлу й атрибут "alt" – варто скористатись ними обома.

Атрибут "alt" дозволяє визначити альтернативний текст для зображення, якщо його показ неможливий із певних причин. Якщо користувач переглядає

сайт у веб-переглядачі, який не підтримує зображення, або ж використовує альтернативні технології, наприклад програми для читання з екрана, вміст атрибуту "alt" надає інформацію про зображення.

Іншою причиною може бути використання зображення як посилання. Текст атрибуту "alt" такого зображення буде розглядатись як текст прив'язки текстового посилання. Проте, не рекомендується використовувати забагато зображень для посилань у пересуванні по сайту, якщо замість них можна застосовувати текстові посилання. Насамкінець, оптимізація назв файлів зображень й текст атрибуту "alt" полегшують розуміння ваших зображень проектам пошуку зображень.

Корисні поради для зображень

- **Використання коротких, але описових назв файлів і тексту атрибуту "alt"** – як і інші частини сторінки, які підлягають оптимізації, найкраще, щоб назви файлів і текст атрибуту "alt" (для мов ASCII) були короткими, але описовими.

Необхідно уникати:

- використання, якщо можливо, базових назв файлів, як наприклад "зображення1.jpg", "фото.gif" чи "1.jpg" (на деяких сайтах із тисячами зображень варто подумати про автоматичне присвоєння назв)
- написання задовгих назв файлів
- нагромадження в тексті атрибуту "alt" ключових слів чи копіювання та вставляння цілих речень

- **Текст атрибуту "alt", якщо використовується зображення як посилання** – якщо зображення використовується як посилання, заповнення тексту його атрибуту "alt" допомагає пошуковим системам та користувачам краще зрозуміти суть сторінки, до якої воно прив'язане.

Необхідно уникати:

- написання задовгих текстів атрибуту "alt", які вважатимуться спамом
- використання самих лише зображень як посилань для переходів по сайту

- **Збереження зображення в окремому каталозі** – замість того, щоб розкидати файли зображень у різні каталоги та підкаталоги по всьому домену, необхідно проводити збереження в одному каталогу (або у ієрархії, якщо це зумовлено структурою сторінок). Це спрощує доступ до них.
- **Використання звичайних підтримуваних типів файлів** – більшість переглядачів підтримують формати зображень JPEG, GIF, PNG і BMP. Також варто давати назві файлу розширення, яке відповідає типу файлу.

Файл robots.txt

Файл "robots.txt" повідомляє пошуковим системам, чи мають вони доступ до сайту та чи можуть сканувати його частини. Цей файл, який повинен мати назву "robots.txt", розміщений у кореневому каталозі сайту.

Можна відмовитись від сканування певних сторінок сайту, оскільки вони не будуть корисними для користувачів, якщо відображатимуться у результатах пошукових систем. Якщо сайт має піддомени і необхідно заборонити сканування певних сторінок в окремому піддомені, необхідно створити окремий файл robots.txt для цього піддомену.

Існує декілька інших способів перешкодити відображенню вмісту в результатах пошуку – наприклад, додати "NOINDEX" до мета-тегу робота, використати .htaccess для захисту паролем каталогів.

Корисні поради для robots.txt

- **Використання безпечних методів для "делікатної" інформації** – не дуже безпечно використовувати robots.txt для блокування "делікатної" чи конфіденційної інформації. Однією з причин є те, що пошукові системи все одно можуть посилатись на заблоковані URL-адреси (відображаючи лише URL-адреси, без назви чи фрагменту), якщо в Інтернеті зустрічатимуться посилання на них (наприклад журнали джерел переходу). Крім того, несумісні чи зловмисні пошукові системи, які не визнають Стандарт виключення робота, можуть ігнорувати вказівки вашого robots.txt. Насамкінець, зацікавлений користувач може вивчити каталоги та підкаталоги вашого файлу robots.txt і відгадати URL-адресу вмісту, який ви приховуєте. Більш безпечний альтернативний спосіб – шифрування вмісту чи захист паролем .htaccess.

Необхідно уникати:

- сканування сторінок, схожих на результати пошуку (користувачам не подобається переходити з однієї сторінки пошуку та опинятись на іншій, яка не дає їм суттєвої інформації)
- сканування великої кількості автоматично створених сторінок із таким самим чи злегка відмінним вмістом: "Чи насправді необхідно включати до індексу пошукової системи 100000 майже однакових сторінок?"
- сканування URL-адрес, створених у результаті послуг проксі

Тег rel="nofollow" для посилань

Надання значення "nofollow" атрибуту "rel" посилання вказує пошуковим системам, що за деякими посиланнями на сайті не слід переходити чи передавати інформацію сторінкам, на які вона посилається. Команда "не переходу" за посиланням – це додавання rel="nofollow" у тег прив'язки посилання.

Коли корисно використовувати цю команду? Якщо на сайті є блог з увімкненим загальним коментуванням, посилання в коментарях можуть

передати репутацію сторінкам, за які може бути незручно поручитись. Області коментарів блогу на сторінках є доступними для спаму в коментарях. Команда "nofollow" для посилань, доданих користувачами, не дозволяє передавати здобуту тяжко репутацію сторінки небезпечним сайтам.

Більшість пакетів програмного забезпечення блогів автоматично забороняють перехід за коментарями користувача, але ті, які цього не роблять, скоріше за все, можна налаштувати вручну. Ця порада корисна також для інших областей сайту, які можуть включати вміст, створений користувачами, наприклад гостьові книги, форуми, дошки оголошень, переліки джерел переходів тощо. Якщо неохідно поручитись за посилання, додані третьою стороною, тоді немає потреби використовувати команду "nofollow" для посилання. Проте, посилання на сайти, які пошукова система вважає небезпечними, може вплинути на репутацію сайту.

Команда "nofollow" також може бути корисною, якщо створюється вміст і необхідно включити посилання на веб-сайт, але не хочеться передавати йому свою репутацію. Наприклад, створюється публікація блогу на тему спаму в коментарях і необхідно вказати сайт, який нещодавно залишив небезпечний коментар у блозі. Автор прагне попередити інших про цей сайт, тому включаєте посилання на нього у свій вміст, але, звичайно ж, не хоче передавати свою репутацію через посилання. Це саме той випадок, коли можна використати команду "nofollow".

Правильна реклама веб-сайт

Ефективна реклама нового вмісту призведе до швидшого охоплення тих, хто зацікавлений у цій сфері. Надмірне захоплення цими рекомендаціями може навіть погіршити репутацію вашого веб-сайту. Це стосується також більшості інших пунктів, описаних у цьому документі.

Корисні поради для рекламування веб-сайту

- **Створення блогу про новий вміст чи послуги** – публікація у блозі на сайті дозволяє базі відвідувачів довідатись про додані новинки, що є зручним способом повідомити про новий вміст чи послуги. Інші веб-майстри, які посилаються на сайт чи канал RSS можуть також довідатись про новинки.
- **Реклама поза мережею** – корисним може бути також рекламування підприємства чи сайту поза мережею. Наприклад, якщо є сайт підприємства, то необхідно переконатись, що його URL-адресу зазначено на візитних картках, бланках, проспектах тощо. Також можна надсилати по пошті постійні рекламні проспекти для клієнтів, повідомляючи їх про новий вміст на веб-сайті компанії.
- **Соціальні медіа-сайти**– сайти, розроблені для співпраці та обміну між користувачами полегшили донесення певного змісту до зацікавленої групи людей.

Необхідно уникати:

- спроб рекламувати кожну невеличку зміну вмісту, акцент має бути на великих, цікавих позиціях
- включення сайту до схем, де вміст штучно розміщується серед найпопулярніших послуг

- **Звернення до користувачів спільноти, пов'язаної з сайтом** – можливо, є певні сайти, які висвітлюють схожі теми. Зазвичай корисно налагоджувати зв'язки з такими сайтами. Гарячі новини у сфері чи спільноті можуть підказати нові ідеї для вмісту чи закласти основи для гарного соціального ресурсу.

Необхідно уникати:

- запитів небезпечних посилань на всі сайти, пов'язані з тематичною сферою
- купівлі посилань з інших сайтів з метою покращення PageRank замість трафіку

6 Стандарти XML як оптимальний формат структурування даних

Друга половина 90-х років минулого століття стала часом радикальних змін в технологіях WEB. Менш ніж за п'ять перших років свого існування WEB знайшов багато сотень мільйонів користувачів на всіх континентах, в його середовищі сформовані і підтримуються величезні інформаційні ресурси. Ця глобальна інформаційна система інтенсивно вторгається в інші галузі інформаційних технологій, стала одним з важливих ланок інфраструктури інформаційного суспільства.

Разом з тим, ряд обмежень, властивих чинним технологіям WEB першого покоління, став стримуючим фактором його подальшого розвитку. Нові підходи в області технологій WEB, які почали конструктивно втілюватися в життя на порозі нового століття, спрямовані, насамперед, на подолання цих обмежень і створення нової технологічної платформи, здатної забезпечити потенціал для створенняweb другого покоління, званого семантичним WEB, і забезпечення можливостей його розвитку на тривалу перспективу. Основну роль у технологічному переоснащенні WEB став грати розроблений консорціумом W3C нову мову розмітки XML.

6.1 Передумови створення платформи XML

Створення web по праву можна вважати одним з найбільших науково-технічних досягнень останнього десятиліття XX століття. Завдяки реалізації цього проекту народжується цілий ряд нових інформаційних технологій, що мають дуже значимі соціально-економічні наслідки. У короткі терміни WEBстав безпрецедентно інтенсивно розвивається глобальної відкритої нескінченно масштабованої розподіленої гіпермедійного системою. Кількість користувачів і обсяг представлених у ній інформаційних ресурсів продовжують надзвичайно швидко нарощуватися. Така популярність WEB забезпечується багатьма достоїнствами використовуваних технологій: відкритий характер системи (у технологічному сенсі); демократична організація WEB; архітектура WEB, заснована на принципі "клієнт - сервер"; можливість глобального доступу до ресурсів WEB в будь-який час; прозорість для користувача розподілу інформаційних ресурсів WEB в просторі; простота мови розмітки HTML та ін.

Разом з тим, за кілька років інтенсивного розвитку потенціал якісного вдосконалення технологій WEB 1.0 виявився значною мірою вичерпаним.

Стримуючий вплив на подальшу еволюцію технологій WEB та розширення сфер їх застосування стали надавати, насамперед, слабкі сторони мови HTML - основного виразного і структурообразуючого представлення в

WEB-гіпермедійних інформаційних ресурсах, а також обмеження функціональних можливостей середовища підтримки цієї мови в WEB.

Необзідні радикально нові підходи, які могли б забезпечити подальший розвиток WEB. Їх розробка та реалізація стали найважливішим стратегічним завданням подальшого розвитку технологій WEB. Створення розширюваної мови розмітки XML і заснованої на ньому нової технологічної платформи стало важливим результатом цієї діяльності. Роботи з формування цієї модульної платформи з поділом функцій між складовими її стандартами інтенсивно ведуться в консорціумі W3C.

Радикальні зміни в Веб

Базові концепції нової технології WEB були визначені W3C в середині 90-х років, а її практичне втілення почалося з прийняття в 1998 році вже згадуваного нового стандарту - розширюваної мови розмітки ExtensibleMarkupLanguage (XML). Поряд з цим ведуться роботи зі створення його інфраструктури - великого комплексу стандартів, заснованих на XML або сумісних з цією мовою, деякі компоненти якої вже існують, інші знаходяться в процесі розробки.

В даний час робочими групами консорціуму W3C вже підготовлені багато документів, що містять специфікації стандартів платформи XML (далі для стислості іноді званих стандартами XML) і її оточення, які за рішенням консорціуму отримали статус рекомендацій (стандартів W3C). Більше 150 документів знаходиться в даний час у стадії розробки. Серед них не лише документи, що визначають нові стандарти, але й специфікації нових версій вже існуючих стандартів. Робота над ними ведеться досить динамічно. У міру визрівання ідей і попередньої їх опрацювання в дослідницьких центрах та індустріальних компаніях, які співпрацюють з консорціумом, ініціюються нові проекти по стандартизації технологій WEB.

Сьогодні вже цілком правомірно можна говорити про народження нової функціонально розвиненою технологічної платформи WEB другого покоління. Поняття платформи трактується тут як сукупність взаємопов'язаних складових і має єдине функціональне призначення стандартів, розроблених на єдиних концептуальних, архітектурних та інших принципах. Платформа XML володіє потенціалом, достатнім для розвитку середовища WEB на тривалу перспективу.

Процес розробки платформи XML в період, що минув з часу прийняття стандарту її базової мови, був досить динамічним і продуктивним.

Створене функціональне ядро становить її комплексу стандартів і продовжується його розвиток. Кілька десятків стандартів цього ядра визначають мовні специфікації, що забезпечують структурування змісту інформаційних ресурсів (XML-документів) і конструювання з них і їх

фрагментів розподілених в мережевому просторі гіперструктур, розвинення можливостей форматування XML-документів, подання метаданих, що характеризують властивості XML-документів.

Стандарти функціонального ядра платформи визначають також засоби забезпечення інформаційної безпеки, інтерфейси прикладного програмування, мову запитів і інші можливості. Передбачається також використання більш загального порівняно з URL механізму ідентифікації інформаційних ресурсів - URI (UniversalResourceIdentifier) і нового протоколу XMLP (XMLProtocol) обміну XML-ресурсами.

Поряд з розробкою специфікацій для нижніх рівнів інформаційної архітектури WEB 2.0, з осені 2001 року, значно активізувалися роботи по семантичному WEB, ідеї якого народилися ще в середині 90-х років і в більш повному вигляді були сформульовані пізніше. Основу цього напрямку розвитку WEB становить стандарт ResourceDefinitionFramework (RDF), а також кілька створених за останній час дослідними колективами в США і Європі прототипів мови опису онтологій. Основне завдання цього напрямку діяльності W3C в даний час полягає у створенні стандарту мови опису онтологій, заснованого на синтаксисі мови XML.

Слід згадати тут ще один важливий новий напрямок роботи з модернізації WEB на основі технологічної платформи XML, заснований на початку 2002 року. Це - діяльність, мета якої полягає в стандартизації архітектури функціональної надбудови (у тому числі, її програмних інтерфейсів) над інформаційним простором WEB, яка б забезпечувала обмін даними в цьому середовищі. Важливим компонентом зазначеної архітектури є WEB -сервіси.

Одна з принципових установок діяльності консорціуму W3C по стандартизації полягає в неодмінному забезпеченні наступності нової платформи з WEB 1.0, що дозволить зберегти можливість використання і надалі величезних інформаційних ресурсів, представлених засобами мови HTML.

6.2 Організація і функції платформи XML

На відміну від чинної версії WEB, в якій всі основні функції управління інформаційними ресурсами системи базуються на єдиній мові HTML, творці платформи XML обрали інший шлях. Виділено "фундаментальні" стандарти, складові концептуальну і синтаксичну основу платформи. Їх засобами визначається комплекс інших стандартів, кожен з яких виконує власні специфічні функції. І цей комплекс відкритий для його поповнення новими стандартами у разі потреби. Саме така модульність організації платформи забезпечує її відкритий характер, можливості введення нових стандартів, не зачіпаючи вже існуючих. Повна функціональність цієї платформи визначається

цілим комплексом взаємопов'язаних стандартів, частина з яких вже прийнята W3C, інші знаходяться в стадії розробки.

Потрібно відзначити, що спільно зі стандартами платформи XML можуть використовуватися і деякі інші стандарти, які формально до цього комплексу стандартів не відносяться. Деякі з таких стандартів мають статус стандартів консорціуму W3C. Хоча ці стандарти не використовують синтаксис мови XML, вони, однак, функціонально сумісні зі стандартами платформи і використовуються поряд з ними в додатках XML. Інші стандарти оточення засновані на синтаксисі XML, але розроблені не W3C, а різними іншими організаціями (компаніями або консорціумами). Тим не менш, вони отримали достатньо широке визнання і застосовуються на практиці, не маючи офіційного статусу стандартів W3C.

Спільне використання стандартів платформи XML та її оточення має місце не тільки в численних додатках, але і в самих специфікаціях стандартів платформи. Наприклад, стандарт XPath, формально не є стандартом XML, використовується в специфікаціях стандартів XPointer, XSLT, XQuery.

Зазначені обставини служать підставою для обговорення стандартів оточення поряд зі стандартами платформи, і ми не будемо далі виділяти їх в якусь особливу групу.

Для того щоб охарактеризувати функціональні можливості платформи XML та її оточення, введемо функціональну класифікацію складових її стандартів, в якій ми представили не тільки вже прийняті W3C стандарти, але й ряд проектів стандартів, над якими активно ведеться робота.

Основні стандарти платформи XML та її оточення можна розділити на наступні класи:

- фундаментальні стандарти: InfoSet, Namespace, XML;
- структурообразуючі стандарти: XPointer, XLink;
- стандарти форматування і трансформації XML-документів: XSL, XSLT, CSS;
- стандарти представлення метаданих: XML (DTD), XMLSchema, RelaxNG, RDF, RDFS, OWL;
- стандарти мов запитів: XQuery, XPath, XSLT;
- стандарти інтерфейсів прикладного програмування: DOM, SAX;
- стандарти для забезпечення наступності з Веб-1: XHTML, XMLBase;
- стандарти транспорту даних: XML-Protocol, XForms, SOAP;
- стандарти ідентифікації інформаційних ресурсів: URI, URL, URN;
- стандарти інформаційної безпеки: XML-Signature, XMLDecription;
- стандарти архітектури функціональної надбудови Веб: XSDL;
- допоміжні стандарти: XInclude, XFragment, CanonicalXML, XPath;

- стандарти вертикальної сфери: MathML, cXML, CML, WML, GML, UBL, XMI і ряд інших стандартів OMG та ін.

Аналізуючи наведену класифікацію стандартів, неважко бачити, що деякі з них багатофункціональні і, відповідно з цим, віднесені до декількох класифікаційних категорій. Крім того, з назв класів інтуїтивно зрозуміла функціональна структура стандартів платформи XML.

Розширюваність мови і платформи XML

Принципово важливою властивістю мови XML, що забезпечує нові функціональні можливості середовища WEB, є його розширюваність. Це найважливіша його властивість заслуговує більш докладного обговорення.

Досягнення розширюваності XML засноване на двох факторах. Насамперед, він являє собою мову метарівня, підмножину відомої мови SGML, а не конкретну мову, подібну до HTML. Завдяки цьому XML виконує функції мови визначення даних. Використовуючи його синтаксис, можна визначати різні типи елементів, екземпляри яких утворюють зміст конкретних XML-документів, і вводити тим самим адекватний потребам набір тегів розмітки документів. Другий фактор - це використання просторів імен - іменованих множин символів, що використовуються в якості імен типів елементів і атрибутів елементів XML-документів. Простір імен дозволяє також явним чи неявним чином асоціювати потрібну семантику з іменованими елементами документів, їх атрибутами і допустимими для них значеннями.

Важливо підкреслити, що розглянуті принципи забезпечують також розширюваність функціональних можливостей всієї платформи XML. Однак для цієї мети необхідно досягнення консенсусу на рівні консорціуму W3C. Основу кожного доповнюючогоXML стандарту платформи становить деякий набір таких типів елементів XML-документів, синтаксис яких може бути визначений засобами XML і які підтримують необхідні нові функціональні можливості. Крім цього вводиться простір імен із зарезервованим ім'ям, що включає імена нових типів елементів XML-документів та їх атрибутів. Семантика елементів цих типів елементів, їх атрибутів і значень, які вони можуть приймати, визначаються в специфікації нового стандарту.

Наступність з технологіями HTML

За недовгу історію WEB 1.0 в його середовищі були накопичені величезні інформаційні ресурси. Кількість HTML-сторінок видимої частини WEB досягло багатьох мільйонів. Втрата можливості доступу до цих інформаційних скарбам, звичайно ж, неприпустима. Тому необхідною умовою технологічного переоснащення WEB є забезпечення наступності нових технологій з технологіями HTML, збереження доступності ресурсів HTML. Ця вимога була

врахована при розробці нової технологічної платформи WEB, заснованої на мові XML.

По суті, для досягнення зазначеної мети необхідно мати можливість інтерпретувати HTML-сторінки при обробці їх процесором XML в середовищі, що підтримує стандарти заснованої на цій мові платформи, як XML-документи.

Вирішення цієї проблеми досягається завдяки тому, що мова HTML може бути визначена як конкретизація XML. Засобами XML можна побудувати такі визначення типу документів DTD, які будуть специфікувати будь-яку допустиму структуру HTML-сторінок. Якщо асоціювати з HTML-сторінками такі DTD, то вони будуть чітко інтерпретуватися в середовищі, що підтримує XML, як XML-документи. Саме це завдання вирішує згадуваний вище стандарт XHTML. У ньому запропоновані специфікації DTD для трьох рівнів мови HTML, що відрізняються один від одного ступенем повноти використання його функціональних можливостей. Стандарт орієнтований на версію HTML 4.01 - діючу версію стандарту цієї мови.

Таким чином, введення в дію платформи XML в WEB не приведе до втрати накопичених раніше в цьому середовищі інформаційних ресурсів.

Моделювання даних XML

Техніка моделювання даних є таким же необхідним інструментом управління XML-даними, як даними і в традиційних базах даних. Тому поняття моделі даних по необхідності фігурує в документації стандартів XML. Потрібно, однак, зауважити, що автори стандартів XML вживають "старомодне" трактування поняття моделі даних як структури конкретного XML-документа, а не як інструменту моделювання таких інформаційних ресурсів. Один з наслідків такого підходу полягає в тому, що залишаються осторонь питання про стандартизацію операційних можливостей засобів управління XML-даними.

Хоча поняття моделі даних згадувалося в минулі роки в специфікаціях ряду стандартів платформи XML, проблеми моделювання даних не були грунтовно опрацьовані. Єдиної функціонально повної моделі даних, що охоплює як структурні, так і операційні можливості, на якій би базувалися всі стандарти платформи, не існує досі і схоже, що в ближчий час вона навряд чи зможе з'явитися. Ніякої діяльності в цьому напрямку в консорціумі поки не ведеться. Питання моделювання даних обговорюються лише автономно в рамках специфікацій деяких стандартів. При цьому автори мають на увазі лише структурні аспекти моделювання даних. Виняток становить стандарт DOM, що визначає API для репозиторіївXML- і HTML-документів. Зауважимо, що хоча DOM може застосовуватися до XML-даними, він не є стандартом платформи XML (додатком XML), а відноситься до її оточенню. В рамках проекту мови запитів XQuery опубліковано кілька документів. Серед них документи

присвячені специфікації моделі даних. Судячи за найменуваннями цих документів, автори вважають, ймовірно, що до моделі даних має відношення лише перший з цих документів, з чим не можна погодитися.

6.3 Метадані та семантика XML-документів

Однією з найважливіших цілей створення платформи XML є привнесення в середу WEB метаданих, що описують властивості підтримуваних в ній інформаційних ресурсів. Насамперед, метадані дозволяють описувати структури XML-документів та їх смислового змісту (семантики). Завдяки цьому забезпечуються можливості автоматичної перевірки правильності структури XML-документів і зниження рівня інформаційного шуму при пошуку інформаційних ресурсів в WEB за допомогою різних пошукових машин.

Явний опис семантики XML-документів необхідно також для різноманітних просунутих WEB -додатків. Зокрема, стає можливим створення принципово нових додатків високого рівня, заснованих на інтеграції інформаційних технологій і забезпечуючих інтеграцію неоднорідних інформаційних ресурсів. Цей напрямок активно розвивається в багатьох наукових центрах різних країн і пов'язан зі створенням інформаційних систем нового класу, що функціонують в середовищі WEB і які називають електронними бібліотеками.

У стандартах платформи XML передбачено кілька засобів опису та подання метаданих. Як уже вказувалося, для визначення логічної структури XML-документів спеціальні синтаксичні конструкції передбачені в мові XML. Представлені їх засобами метадані називаються визначенням типу документів (DocumentTypeDefinition, DTD).

У DTD XML-документи даного типу описуються як ієрархічні структури, що складаються з елементів документів. Ці елементи можуть бути різних типів, описаних у DTD. Специфікація DTD може бути вбудована в XML-документ або зберігатися де-небудь в WEB. В останньому випадку в документі дається на нього посилання. Для більш витонченого опису структури XML-документів можуть використовуватися властивості стандарту XMLSchema. У порівнянні з DTD, цей стандарт надає для опису XML-документів додаткові можливості, зокрема більш розвинену систему типів значень елементів і атрибутів елементів.

Семантика XML-документа може бути визначена явним чи неявним чином (за замовчуванням). Явна визначення може бути формалізовано в різному ступені. Один з формалізованих способів явного визначення семантики XML-документів забезпечується засобами складається з двох частин стандарту W3C - ResourceDefinitionFramework (RDF). Таке визначення семантики XML-документів називається RDF-специфікацією.

Подальші роботи консорціуму з розвитку засобів представлення метаданих, що визначають семантику XML-документів, проводяться в рамках діяльності по створенню семантичного WEB.

Семантичний Веб

Ще в середині 90-х років задум творців WEB, спрямований на радикальні перетворення цієї вельми значущою для життєдіяльності суспільства системи і перетворення її в систему семантичного рівня, почав активно реалізовуватися в останні роки.

У той час як WEB першого покоління будувався з орієнтацією на обробку інформації, розташованої в ньому людиною, технології WEB нового покоління повинні забезпечувати можливості автоматизованої інтерпретації та обробки інформації, семантичної інтероперабельності інформаційних ресурсів. У цих умовах вже недостатньо розташовувати синтаксичним описом XML-документів за допомогою DTD або XMLSchema. Наприклад, при обміні документами, описаними засобами цих мов, обидві сторони повинні однаковим чином розуміти зміст використовуваних в документах типів елементів і атрибутів елементів, а також розміщених в них гіперпосилань, про що заздалегідь повинні бути прийняті відповідні домовленості, описані вербальним чи іншим чином.

Необхідність вирішення зазначених завдань викликала потреба в таких засобах формального опису семантики XML-даних, які б дозволяли аналізувати і обробляти їх за допомогою програмного забезпечення. При такому підході WEB нового покоління повинен мати багаторівневу інформаційну архітектуру.

Першим кроком консорціуму W3C в розглянутому напрямку було створення стандартів RDF (ResourceDefinitionFramework) і RDFS (RDFSchema).

Опис семантики інформаційних ресурсів XML в термінах виразних засобів стандарту RDF, зване RDF-специфікацією, аналогічно за своїми можливостями концептуальній схемі в системах баз даних і приблизно еквівалентно ER-моделі.

У RDF-специфікації оголошується деякий безліч ресурсів, для кожного з яких визначаються пари "властивість-значення". Інформаційні ресурси в RDF - це ресурси WEB, ідентифіковані унікальним чином за допомогою їх URI. Вони можуть також являти собою колекції інших інформаційних ресурсів або літералів, які називають контейнерами. Допускаються контейнери типу мультимножини, послідовності і альтернативи. Значення властивостей задаються літерально або можуть бути іншими ресурсами, які представляються, у свою чергу, їх властивостями. Таким чином, властивості можуть визначати і зв'язки між ресурсами.

Опис семантики властивостей в RDF називається RDF-схемою. По суті, RDF-схема повинна визначати онтологію предметної області. Онтології

отримали в останні роки широке поширення у вирішенні проблем представлення знань та інженерії знань, семантичної інтеграції інформаційних ресурсів, інформаційного пошуку і т.і.

Під онтологією розуміється "специфікація концептуалізації предметної області". Така специфікація являє собою свого роду словник понять предметної області і сукупність явним чином виражених припущень щодо змісту цих понять. Рівень структурованості опису онтології може змінюватися в широкому діапазоні. У спрощених випадках онтологія представляється як ієрархія понять, пов'язаних відносинами деяких певних видів. Такі визначення онтологій використовуються в різних класифікаціях. Кілька великі можливості забезпечують схеми метаданих, наприклад, широко відоме дублінське ядро. Розвинені визначення онтологій формалізуються засобами мов логіки першого порядку. Вони допускають можливості логічного висновку.

У специфікації стандарту RDF не регламентується спосіб завдання схеми для RDF-специфікації. Достатньо лише представити її як деякий ресурс WEB і використовувати URI цього ресурсу для посилання на неї в RDF-специфікації. У документації стандарту RDF розглядається, наприклад, варіант використання для цих цілей набору елементів метаданих Дублінського ядра.

Один з більш розвинених способів завдання схеми пропонується в проекті другої частини стандарту RDF, званої RDFSchema (RDFS). Цей спосіб заснований на об'єктній моделі, в якій використовуються концепції класів, властивостей і обмежень, асоційованих з класами і властивостями, підтримується ієрархічне відношення "клас-підклас".

Створення розвиненого мови опису онтологійOWL (OntologyWebLanguage) стало останнім часом одним з найбільш важливих ланок робіт по семантичному WEB, що проводяться консорціумом W3C. Мається на увазі мова, що дозволяє описувати структуровані онтології, тобто онтології, які доступні для машинної обробки.

Наприкінці 2001 року для реалізації цього проекту у складі W3C була заснована спеціальна робоча група - WebOntologyWorkingGroup. Розробка OWL починається не з чистого аркуша. Вона ґрунтується на результатах, вже отриманих до цього часу декількома авторитетними дослідницькими колективами, і тому її планується здійснити в досить короткі терміни.

6.4 Сфери застосування стандартів XML

Хоча мова XML і базований на цій мові комплекс стандартів створювалися консорціумом W3C, насамперед, як засоби представлення інформаційних ресурсів WEB 2.0, вони, тим не менш, знайшли широке застосування в різних галузях інформаційних технологій. Багато з них вже набули статусу стандартів де-факто.

Швидке визнання стандартів XML широкими колами фахівців було обумовлено не тільки назрілою необхідністю радикальних змін у WEB, а й розвиненими можливостями, якими володіє платформа XML для представлення інформаційних ресурсів, адаптованістю до умов застосування. Другий фактор полягає в можливості метаописів інформаційних ресурсів з потрібним ступенем структурованості з використанням наданих для цих цілей властивостей, а також у відкритому характері стандартів, що дозволяє інтегрувати засоби користувача в визначене ним середовище. Нарешті, важливу роль відіграють можливості XML як мови, що підтримується в глобальному комунікаційному середовищі WEB. Використання XML як мови-посередника для обміну повідомленнями через WEB дозволяє забезпечити інтероперабельність і взаємодію різного роду систем.

Перерахуємо найважливіші завдання, вирішення яких забезпечує платформа XML:

- створення WEB другого покоління;
- виконання функцій мови-посередника при обміні даними між програмними системами, що реалізовують, можливо, різні технології, і забезпечення тим самим їх інтероперабельності;
- інтеграція неоднорідних інформаційних ресурсів, різних технологій управління даними і додатків;
- створення нової гілки технологій баз даних, званих XML-орієнтованими базами даних;
- поряд з використанням технологій XML-орієнтованих баз даних за їх прямим призначенням - для управління репозиторіями XML-документів, можна очікувати, що вони знайдуть застосування й у вирішенні проблеми "прихованого" WEB; завдяки єдності моделі даних XML-документів, представлених на WEB-сервері, і в доступній через нього XML-орієнтованої базі даних, з'являється можливість "відкрити" "приховані" інформаційні ресурси баз даних для механізмів таких WEB-серверів;
- забезпечення інструментарію для нових сфер застосування WEB, таких як електронний бізнес, електронні бібліотеки, електронні видання тощо

Області застосувань стандартів платформи XML постійно розширюються й охоплюють ряд технологій і стандартів як горизонтальною, так і вертикальної сфери.

У горизонтальній сфері (технології, незалежні від конкретної області додатків) стандарт XML знайшов застосування в ряді стандартів консорціумів ObjectManagementGroup (OMG), MetaDataCoalition (MDC) і WorkflowManagementCoalition (WfMC), в стандартах ISO / IEC та ін.

Планується далі використовувати XML для кодування повідомлень, якими обмінюються клієнт і сервер у відомому стандарті ISO / IEC RDA / SQL (RemoteDatabase Access for SQL) віддаленого доступу до систем SQL баз даних.

У розробленому консорціумом WorkflowManagementCoalition (WfMC) стандарті еталонної моделі потоків робіт визначаються специфікації XML DTD, що дозволяють здійснювати обмін повідомленнями мовою XML між програмними засобами потоків робіт для підтримки їх інтероперабельності.

У зв'язку з успішним просуванням платформи XML в практику, почалися роботи над новим компонентом SQL / XML наступної версії стандарту мови SQL - SQL: 200n. За задумом розробників, він визначатиме можливості спільного використання ресурсів SQL і XML. Зокрема, будуть визначатися уявлення схем і даних SQL у формі XML-документів і навпаки.

Можна згадати також озвучену в колах консорціуму ODMG (ObjectDataManagementGroup) ідею про доцільність використання XML в якості мови обміну об'єктами в рамках розробленого консорціумом стандарту об'єктних даних, натомість визначеного в стандарті ODMG 3.0. мови OIF (ObjectInterchangeFormat).

Важливою сферою застосування стандартів XML стає формування в останні роки нової гілки технологій баз даних - XML-орієнтованих баз даних. У таких системах мова XML використовується в якості мови визначення даних. Мовами запитів служать XPath, XSLT і XQL - одни з ранніх претендентів на роль стандарту мови запитів для платформи XML. Активно ведуться розробки специфікацій стандарту мови запитів XQuery. Є програмні продукти цієї категорії, які забезпечують інтерфейс прикладного програмування, заснований на об'єктній моделі стандарту DOM.

Стандарти XML широко застосовуються також у вертикальній сфері (конкретні області додатків - електронний бізнес, управління виробництвом, транспорт і т.п.). Тут слід, зокрема, згадати технології і стандарти консорціумів OASIS, OMG і OGC (OpenGISConsortia), компаній IBM, Microsoft, Ariba.

6.5 XML і електронні бібліотеки

Однією з важливих галузей застосування стандартів платформи XML стали розробки електронних бібліотек. Розробки і дослідження електронних бібліотек є одним з актуальних напрямків розвитку інформаційних систем в останні роки, що привертає увагу фахівців різного профілю.

Фахівці в галузі бібліотечної справи бачать в електронних бібліотеках нові можливості для вдосконалення автоматизованих бібліотечних систем, перетворення їх у публічні електронні бібліотеки нового покоління з розвиненими можливостями подання різноманітних цифрових інформаційних

ресурсів і доступу до них, платформу для інтеграції видавничих та бібліотечних технологій.

Музейні працівники отримують в нових технологіях можливості збереження національної культурної спадщини та перетворення його в загальнодоступне надбання завдяки забезпеченню глобального доступу в середовищі WEB за допомогою функціонально розвинених сервісів до створюваних ними електронних колекцій цифрових образів музейних експонатів.

Співробітники освітніх установ різних ступенів спільно з бібліотечними фахівцями ведуть велику роботу по створенню принципово нової технології інтерактивного навчання (DigitalLibrariesEducation, DLE), заснованої на новій інформаційній інфраструктурі освітнього процесу, ядром якої мають стати електронні бібліотеки. При цьому передбачається адаптація програм навчання і інформаційної підтримки до потреб і можливостей конкретного учня без будь-яких обмежень на його вік, географічне розташування, розпорядок дня і т.і.

Наукових співробітників, що займаються дослідницькою роботою в різних галузях знань, технології електронних бібліотек привертають можливості забезпечення ефективного поширення результатів досліджень у середовищі наукового співтовариства, підтримки наукового співробітництва колективів дослідників, для якого не є перешкодою адміністративні, географічні та національні кордони. Інформаційні ресурси електронних бібліотек і спеціально розроблені для оперування ними функціональні сервіси стають основою дослідних стендів в різних галузях науки, замінюючи натурні експерименти експериментами з моделями реальних сутностей, процесів або явищ.

Нарешті, фахівці в області інформаційних систем, у свою чергу, розглядають електронні бібліотеки як новий клас інформаційних систем, що базуються на самих передових досягненнях інформаційних технологій і технологій телекомунікацій. Розробки таких систем породжують різноманітні складні теоретичні та технологічні проблеми, які потребують свого дослідження.

Функціональні можливості електронних бібліотек варіюються в досить широкому діапазоні. Передбачається, що вони надають користувачеві глобальний доступ за допомогою різного роду сервісів в середовищі WEB до їх колекцій цифрових інформаційних ресурсів, які можуть бути розподіленими і в різних аспектах неоднорідними. При цьому можуть забезпечуватися різноманітні можливості їх інтеграції на технічному та / або на семантичному рівні, розвинені засоби каталогізації та індексування. У деяких електронних бібліотеках передбачаються інтерфейси з підвищеним рівнем семантики, наприклад, з можливостями семантичного пошуку необхідних інформаційних

ресурсів, багатомовним доступом та візуалізацією даних, засобами персоналізації користувальницьких інтерфейсів і т.і.

Різноманітний характер можуть мати і інформаційні ресурси електронних бібліотек - від традиційних бібліотечних електронних каталогів до складних інтегрованих колекцій інформаційних ресурсів, що включають повнотекстові документи, числові дані, графічні, аудіо та відео ресурси тощо, забезпечених різноманітними метаданими.

Напрями досліджень і розробок в сфері електронних бібліотек охоплюють технологічні, лінгвістичні, економічні, правові, соціальні та інші аспекти систем цього класу, а також методи і інструментарій створення підтримуваних в них колекцій різноманітних цифрових інформаційних ресурсів.

Розроблені нині електронні бібліотеки базуються на передових досягненнях WEB-технологій (платформа XML, роботи зі створення семантичного WEB, зокрема, мови опису онтологій та ін.), Технологій баз даних (об'єктні та об'єктно-реляційні бази даних, XML-орієнтовані бази даних), технологій текстового пошуку (повнотекстовий пошук, моделі семантичного пошуку, нові підходи, орієнтовані на текстовий пошук в WEB), досягнення в галузі методів представлення і виявлення знань, технологій створення та підтримки електронних публікацій, моделювання даних і метаданих.

Найбільш актуальними технологічні проблеми електронних бібліотек є:

- Дослідження архітектурних аспектів таких систем.
- Забезпечення інтероперабельності інформаційного середовища.
- Розвиток методів представлення інформаційних ресурсів електронних бібліотек.
- Визначення складу метаданих, незалежних від застосувань і специфічних для різних сфер програми, а також засобів їх подання.
- Нові походи до каталогізації інформаційних ресурсів.
- Створення функціонально розвинених користувальницьких інтерфейсів (багатомовний доступ, візуалізація даних, персоналізація, підтримка семантичного рівня спілкування користувачів з системою).
- Розробка техніки індексування інформаційних ресурсів різної природи (текст, аудіо, відео тощо), пошуку і виявлення релевантних ресурсів, а також принципів і засобів їхнього аналізу.
- Інтеграція колекцій неоднорідних інформаційних ресурсів, використання для цих цілей адаптерів і семантичних посередників.
- Безпека інформаційних ресурсів електронних бібліотек.
- Розробка методів і засобів автоматичного анотування текстових документів.
- Створення та дослідження прототипів систем електронних бібліотек.

Можна виділити наступні напрями розвитку платформи XML в розробках електронних бібліотек:

- Використання мови XML для представлення колекцій електронних інформаційних ресурсів в електронних бібліотеках. Розмітка опублікованих в WEB наукових праць та експериментальних даних забезпечує нову якість створюваних інформаційних ресурсів за рахунок змісту. Це не тільки покращує їх сприйняття, але й забезпечує автоматичну верифікацію цілісності змістовної структури документів в збережених колекціях.
- Аналогічно технологіям HTML, забезпечення навігаційного доступу до XML-ресурсам електронних бібліотек за допомогою звичних браузерів, що підтримують мову XML.
- Використання представлених за допомогою мови XML інформаційних ресурсів в рамках просунутих WEB-технологій. Так, спеціально розроблені пошукові машини можуть здійснювати пошук документів у поданих колекціях на основі елементів їх змісту. Можлива видача користувачеві не тільки повних документів, що задовольняють критеріям пошуку, але і їх фрагментів, а також похідних документів, що представляють різного роду трансформації документів, знайдених в результаті пошуку.
- Використання XML як мови-посередника для обміну даними між різними додатками або компонентами розподілених додатків, функціонуючими в середовищі WEB або використовують WEB як середовище транспорту даних.
- Використання стандартів платформи XML для представлення метаданих, що описують властивості публікованих в WEB інформаційних ресурсів. Для цих цілей можуть використовуватися не тільки властивості самої мови XML, а й мовні засоби стандартів XMLSchema і RDF. Опис XML-документів засобами XMLSchema дозволяє здійснювати більш тонку верифікацію цілісності представлених XML-документів. Специфікація змісту документів засобами стандарту RDF дає можливість семантичного пошуку інформаційних ресурсів в середовищі, що підтримує такі метадані.
- Зароджується на основі стандартів платформи XML новий клас систем баз даних (XML-орієнтованих баз даних), який надає розробникам електронних бібліотек інструментальні засоби для підтримки колекцій інформаційних ресурсів XML і доступу до них в системах баз даних.
- Використання стандартів платформи XML для забезпечення інтеграції інформаційних ресурсів з різних джерел. Системи інтеграції інформаційних ресурсів вирішують різні завдання. "Технічна" інтеграція передбачає єдине уявлення інтегрувальних інформаційних ресурсів в термінах деякої інтегруючої моделі даних. У багатьох розробках в якості такої моделі використовується мова XML як мова опису даних у поєднанні з яким-небудь з мов платформи XML та її оточення - для маніпулювання даними в XML-поданні. Для цієї мети

найчастіше використовуються мови XPath, XSLT, XQuery. В якості інтегруючої моделі часто використовується також об'єктна модель даних, обумовлена стандартом DOM. Необхідними компонентами архітектури систем інтеграції розглянутого виду є механізми відображення моделі даних джерела ресурсів у інтегруючу модель даних. В інших ситуаціях користувачі систем інтеграції інформаційних ресурсів задовольняються лише інтеграцією метаданих. Прикладом можуть служити корпоративні каталоги розподілених інформаційних ресурсів, представлені засобами мови XML. Дуже складною є задача семантичної інтеграції неоднорідних інформаційних ресурсів. Системи, що забезпечують такі можливості, використовують техніку відображення інформаційних ресурсів в інтегруючу модель за допомогою адаптерів і посередників, онтологічних специфікацій предметної області джерел ресурсів, методів інтеграції онтологій.

- Створення нового інтелектуального середовища подання інформаційних ресурсів електронних бібліотек наступного покоління на основі інструментарію семантичного WEB (стандартів RDF, RDFS, мови опису онтологій та ін.).

Використання технологій XML в електронних бібліотеках є дуже перспективним. На основі XML створені мови розмітки інформаційних ресурсів в математиці, хімії, астрономії, геоінформатиці та в інших областях знань, і вони досить широко застосовуються на практиці. Запропоновано різні підходи до подання бібліографічної інформації за допомогою мови XML і до створення на цій основі електронних бібліотек з пошуковими машинами, що оперують такою інформацією. Є ряд прикладів локальних і розподілених систем такого роду. Зокрема, відомі системи для підтримки електронних ботанічних колекцій, колекцій публікацій в економічній науці і т.і. Методи інтеграції неоднорідних інформаційних ресурсів з використанням технологій XML розробляються в різних дослідницьких центрах.

6.6 Перспективи розвитку платформи XML

Створення платформи XML поклало початок новому більш науковому і технологічно більш досконалому етапу в розвитку WEB. Мова XML і деякі інші стандарти вже стали стандартами де-факто. Всі провідні постачальники програмного забезпечення не тільки WEB, а й систем баз даних, включають у свої програмні продукти підтримку мови XML, створюють засновані на ньому спеціалізовані системи. Просуванням технологій XML в практику поряд з W3C активно займається консорціум OASIS.

Поширенню та активному використанню стандартів платформи XML істотним чином сприяє політика W3C. Консорціум забезпечує вільний доступ до специфікацій розроблених стандартів, підтримує створення ряду вільно

розповсюджуваних синтаксичних аналізаторів для мов платформи та іншого пов'язаного з ними вільно поширюваного програмного забезпечення. Творці стандартів XML приділяють велику увагу забезпеченню наступності для існуючої HTML-платформи і накопичених на її основі інформаційних ресурсів.

Стратегічні перспективи розвитку платформи XML, звичайно ж, пов'язані зі створенням семантичного WEB. Для досягнення цієї мети необхідно вирішити великий комплекс складних науково-технічних завдань. Консорціум цілеспрямовано просувається до цієї мети, охоплюючи у своїй діяльності численні аспекти проблеми.

Ця діяльність має ряд важливих побічних ефектів. Один з них - народження нового напрямку в технологіях баз даних - XML-орієнтованих систем баз даних. Інший побічний ефект, який також не можна не враховувати при оцінці перспектив платформи XML, полягає в тому, що вона починає відігравати суттєву роль в інших широко поширених технологіях - CASE-технологіях, технологіях сховищ даних, потоків робіт, в технологіях баз даних, стає основою інтеграції інформаційних ресурсів WEB і реляційних баз даних. Робляться також кроки, спрямовані на інтеграцію XML-середовища з об'єктними середовищами.

Разом з тим, все ще існують чинники, які стримують енергійне масове практичне використання стандартів платформи XML в середовищі Веб. Насамперед, це - природна інерційність такої масштабної середовища, якою є сьогоднішній WEB. Ця інерція може долатися тільки поступово. Другий фактор – поки ще не завершена робота над засобами визначення глобальної розподіленої гіперструктури, компонентами якої є XML-документи та їх фрагменти. Цю функцію мають виконувати дві важливих стандарту платформи XML - XPointer і XLink. У самій мові XML немає властивостей для визначення гіперпосилань. Специфікації стандарту XLink отримали статус рекомендацій консорціуму наприкінці червня 2001 року. Однак робота над проектом XPointer все ще не завершена, оскільки цей стандарт тісно взаємопов'язаний із технічною характеристикою мови запитів XQuery.

Тим не менш, XML-сайти вже з'являються в WEB. Хоча вони і використовують стандарти платформи зі значними обмеженнями, їх поява свідчить про початок процесу технологічного переоснащення WEB

Список літератури:

1. Введение в системы баз данных [Книга] / Дж. Дейт К. – М : Вильямс, 2005. - 8-е издание. – 1328 с.
2. Интернетика. Навигация в сложных сетях [Книга] / Ландэ Д.В., Снарский А.А., Безсуднов И.В. - Москва : Либроком (Editorial URSS), 2009. – 264 с.
3. Информационный поиск [Електронний ресурс] // Википедия. - 2015 р. – Режим доступу : https://ru.wikipedia.org/wiki/Информационный_поиск.
4. Інформація [Електронний ресурс] // Вікіпедія. - 2015 р. – Режим доступу : https://uk.wikipedia.org/wiki/Інформація.
5. Как работают поисковые системы / Сегалович И.В. – Режим доступу : http://download.yandex.ru/company/iworld-3.pdf.
6. Краткая история сети Интернет // InternetSociety. – Режим доступу : http://www.internetsociety.org/ru/node/10658/что-такое-интернет/история-интернета/краткая-история-сети-интернет.
7. Лекция №3 курса "Алгоритмы для Интернета" / Ю. Лифшиц // Модели информационного поиска. - 2006 р..
8. Математические модели информационного поиска веб-ресурсов [Електронний ресурс] / Кузнецов М.А., Нгуен Т. Т. – Режим доступу : http://www.hi-tech.aspu.ru/files/2(22)/25-30.pdf.
9. Методы классификации и технология Галактика-Зум [Журнал] / А.В. Антонов // Научно-техническая информация. - 2004 р. - 6. - сс. 20-27.
10. Нормальная форма // Википедия. - 2015 р. – Режим доступу : https://ru.wikipedia.org/wiki/Нормальная_форма.
11. Обзор современных методов интеграции данных в информационных системах [Текст] / Шибанов С.В., Яровая М.В., Шашков Б.Д. и др. // НиКа. - 2010 р. – №1. – С. 292-295.
12. Основы интеграции информационных потоков / Д.В. Ландэ. - 04 2015 р. – Режим доступу : http://dwl.kiev.ua/art/monogr-osnov/spusk3.pdf.
13. Основы теории информации и кодирования [Книга] / Кузьмин И.В., Кедрус В.А. - К : Вища школа, 1977.
14. Оценка систем текстового поиска [Журнал] / И. Кураленок, И. Некрестьянов.. - 2002 р. - 4. - С. 226-242.
15. Подходы к формированию смыслового поиска информации в распределенных информационных системах сети интернет / А. Трусов, В. Трусов // Российская ассоциация Электронных библиотек. – Режим доступу :

http://www.aselibrary.ru/digital_resources/journal/irr/irr2725/irr27252785/irr272527852806/irr2725278528062809/.
16. Поиск знаний в Internet [Книга] / Д.В. Ландэ. – М : Диалектика-Вильямс, 2005.
17. Руководство по проектированию реляционных баз данных // Хабрахабр. - 2013 р. – Режим доступу : http://habrahabr.ru/post/193284/.
18. Современные информационные потоки: актуальная проблематика [Стаття] / Брайчевский С.М., Ландэ Д.В. // Научно-техническая информация. - 2005 р. - 11. - С. 21-33.
19. Теорія масової інформації та комунікації [Книга] / В. Партико - Львів : Афіша, 2008. - 292 с.
20. Эффективная методика поиска информации в сети Интернет / А. Попов // CitForum. – Режим доступу : http://citforum.ru/pp/search_03.shtml.
21. Некоторые вопросы применения векторной модели представления документов в информационном поиске// Дубинский А. Г. // Управляющие системы и машины. – 2001. – № 4.
22. Основы поиска информации в Интернете : методическое пособие. // Капустин В. А. – СПб. : Институт "Открытое общество", 1998.
23. Search Engine Optimization Starter Guide // Google, E-Docs, 2010
24. XML / Википедия. - 2015 р. – Режим доступу : https://en.wikipedia.org/wiki/XML
25. Стандарты XML и электронные библиотеки // Когаловский М.Р. – Аналитический обзор 2003.