УДК 004.9
О.К.Жигаревич
Луцький інститут розвитку людини ВМУРоЛ «Україна»

ПРОБЛЕМИ ІНФОРМАЦІЙНИХ КОМУНІКАЦІЙ В ІНТЕРНЕТІ

У статті досліджується проблема інформаційних комунікацій в інтернеті. Програми-агрегатори, які дозволяють групувати публікації з різних джерел. Розглядається формат RSS, який забезпечує резюмування вмісту веб-сайту. Переваги роботи RSS для пошуку інформації.

Ключові слова: *програмний продукт, програмне забезпечення, обмін інформацією, веб-сайт, пошукові системи, інтернет - новини.*

Мова HTML описує зовнішній вигляд web-сайтів, їх окремих сторінок, забезпечуючи перш за все візуалізацію. Формат був розроблений, в першу чергу, для вирішення задач відображення змісту на кожному конкретному ресурсі, тому не завжди зручний для автоматичної обробки інформації, у тому числі і організації пошуку. В результаті вся Мережа Інтернет орієнтована, перш за все, на окремі сайти і не дуже пристосована для автоматизованого узагальнення інформації, її класифікації і аналітичної обробки.

Дуже часто виникає необхідність обміну інформацією, наприклад, між декількома сайтами, при цьому завжди постає питання про технологію однотипного представлення їх змісту. Якщо така технологія не використовується, то зміна HTML-оформлення одного сайту приведе до необхідності одночасної модифікації програмного забезпечення на всіх сайтах, які приймають від нього інформацію. Приблизно така ж ситуація виникає при необхідності імпортувати інформацію на один ресурс з декількох інших, припустимо, тематично близьких. Зміни оформлення на кожному з сайтів-експортерів інформації кожного разу вимагатиме модифікації відповідного програмного коду на сайті-імпортері.

Все це зумовило необхідність використання уніфікованого представлення даних. Був потрібен деякий стандарт представлення інформації на сайтах, забезпечуючи однотипний обмін даними в такій складній системі, як Інтернет. Сьогодні як такий уніфікований формат все частіше використовується формат RSS.

Одним з перших проектів, покликаних вирішити задачі уніфікації обміну даними між серверами Великої Мережі, став Semantic Web. В його основу була покладена наступна ідея організації даних в Інтернеті. Сервери повинні були вміти не тільки візуалізувати інформацію, але і використовувати її. Таким чином різні програми різних виробників могли ефективно працювати з даними з Мережі. Справа залишалася за малим — створити правила формування блоків інформації, які змогла б зрозуміти не тільки людина, але і комп'ютер. Саме для проекту Semantic Web були розроблені специфікації XML.[1]

XML є метамовою, тобто мова, на базі якої можна визначати нові мови. Він призначений не тільки для створення програмного забезпечення, також для організації обміну даними в Web, але і для розпізнавання семантики цих даних. На відміну від HTML, XML призначений для представлення інформації в «чистому вигляді», припускаючи структурну, а не розмітку даних.

Разом з тим, XML, будучи необхідну частину рішення задачі обміну інформаційним наповненням сайтів, сам по собі не може дати нічого того, що необхідне для інфраструктури обробки даних. Річ у тому, що формально теги XML відірвані від визначення їх змістовного наповнення. Паралельно з XML була почата розробка стандарту схеми опису джерел (Resource Description Framework, або RDF). Специфікації RDF підтримують теги, дозволяючи визначати будь-які поняття (наприклад, теги PRICE і INVOICE можна використовувати для позначення типів даних, відповідно, («ціна» і «рахунок»). В цьому випадку непотрібно аналізувати всю решту змісту web-сторінки у пошуках потрібної інформації. Даним у форматі RDF присвоюються дескриптори, які можуть визначатися в окремих файлах визначення типів документів Document Type Definitions (DTD). В кожному розділі є свій список DTD, що постійно розширяється. Вузли що знаходяться в мережі метаданих RDF повинні забезпечити значно більш високу якість і швидкість обміну інформацією і пошуку даних в мережі.

На основі XML і RDF був розроблений формат RSS, спеціально призначений для легкого і швидкого обміну контентом між сайтами — організації інформаційної комунікації між серверами. Компанія Netscape створила RSS для свого порталу Netcenter, як один з перших XML-додатків.

Абревіатура RSS припускає неоднозначні, але близькі трактування — Really Simple Syndication, Rich Site Summary, RDF Site Summary. Мається на увазі, що йдеться про простий спосіб узагальнення і розподілу інформаційного наповнення (синдикації) сайтів.

Формат RSS що завоював сьогодні популярність, забезпечує злагоджений спосіб резюмувати вміст сайтів. Завдяки ньому адміністратори сайтів новин, щоденників (weblog) онлайна, форумів і інших web-ресурсів, що часто обновляються, отримали простий і уніфікований метод подачі інформації про події, що відбуваються. Сьогодні RSS розглядається, в першу чергу, як формат, призначений для публікації і забезпечення експорту новин на сайтах новин. Після того, як інформація перетворена у формат RSS, будь-яка програма, орієнтована на даний формат, може завантажувати відомості про оновлення web-сайтів. Залежно від результату, виконувати певні дії, наприклад, автоматично оновляти список актуальних інформаційних повідомлень. [2]

RSS (Really Simple Syndication - дуже проста синдикація) - це формат на базі XML для розподілу контенту. Код RSS створює файл XML з описом каналу RSS (вміст сайту, наприклад, новини або інший вид інформації).

Структура каналу RSS така, що кожний її елемент має заголовок, короткий зміст статті і посилання на саму статтю, що дозволяє швидко проглядати весь потік інформації і максимально швидко знайти відомості або дані що нас цікавлять.

Всі інтернет-сайти можна умовно поділити на дві частини. Перші містять статичний, постійний, довідковий контент, який практично ніколи не змінюється. До них можна віднести енциклопедії, словники, довідники. Інші ж містять динамічну інформацію, яка оновляється через деякий час. Це нові, аналітичні сайти, різного роду авторські проекти і блоги. Звичайно зрозуміло, що безліч сайтів поєднують в собі статичний і динамічний контент. [3]

Як тільки з'явилися перші контент-проекти, відразу виникла проблема обміну даними між ними. Цілком нормальна ситуація, коли сайт створений для продажу компакт-дисків, буде відображати новини кіно і музики. Сайти не конкурують між собою, а гармонійно доповнюють один одного. Припустимо, що власники ресурсів домовилися між собою, і залишилося тільки вирішити технічні питання. Перший варіант, який відразу спадає на думку - це вручну переписувати новинки. Втім, будь-яка розумна людина від цієї думки швидко відмовиться, оскільки недоліки видні неозброєним поглядом. Другий варіант – реалізувати трансляцію інформації з одного сайту на інший автоматизованими способами. Для того, щоб зрозуміти, наскільки це складно, достатньо відвідати десяток різних сайтів. Навіть у людини виникають проблеми з прочитанням інформації на них. Що вже говорити про комп'ютерну програму, яку доведеться настроювати на кожний сайт окремо. І після кожної зміни дизайну сайту джерела доведеться міняти свою програму. Це, зрозуміло, не дуже зручно. Але всі сайти світу не можуть мати однаковий дизайн.

Довгий час цю проблему намагалися вирішити різними способами. Багато сайтів пропонували інформацію для експорту у вигляді текстового файлу з роздільниками, java-скрипта і так далі. На щастя, компанія Netscape одного разу розробила для використання на своєму порталі Netcenter формат RSS. З його допомогою здійснювався імпорт новин на портал з інших сайтів.

На жаль, компанія Netscape досить скоро припинила роботу над порталом, і формат RSS виявився непотрібним. Роботу над технологією підхопила компанія Userland, яка спростила формат і випустила специфікацію на RSS 0.91. Після цього було ще декілька інкарнацій цього формату, але на даний момент найпопулярнішими є версії 0.91 і 2.00. Переважна більшість розробників сайтів використовують RSS 0.91 для трансляції простої інформації, що дозволяє передавати заголовок, дані про мову повідомлення, посилання на нього і короткий опис. Друга версія дає дещо більше свободу дій для розробника сайту - в неї можна включати декілька додаткових полів - таких, як джерело кожної новини або дату її написання. [3]

Всі ці версії відрізняються одина від одної, але об'єднує їх те, що вони орієнтовані на один тип інформації і містять однакові базові поля. Основний блок інформації (channel), що складається з назви (title), посилання (link), даних про мову новин (language) і логотипу (image). Потім йде список самих новин, де в кожному пункті (item) указується заголовок (title), короткий опис (description) і посилання на новину (link). [4]

Отже, RSS — це формат даних і технічний стандарт, який забезпечує інтегрований доступ до нової інформації, представленої на сайтах, спеціально створений для обміну їх контентом RSS.

Ще минулого року переважала думка, що RSS — це формат, що використовується в основному на іноземних сайтах, проте сьогодні ситуація різко змінилася. Наприклад, переважна більшість RSS-фідів російськомовного сегменту Інтернет знаходиться за адресою http://my.yandex.ru/rss.opml.

Якщо підвести підсумок, то потрібно відзначити, що формат RSS в даний момент розвивається і використовується вже на багатьох інтернет - сайтах. Вміти працювати з ним потрібно не тільки веб-майстрам, але і простим користувачам Інтернету. Він досить простий в користуванні, і значно скоротить час перебування в Мережі. Група програмістів вирішила трохи посунути на ринку формат RSS і взялись до розробки нової технології для обміну, архівації і редагування інформації, яка вже отримала назву Atom.[4]

Більшість SEO-фахівців вважає, що сайти, де всі сторінки відповідають певній темі, області або набору ключових слів, рангуються пошуковими системами вище. Якщо сайт продає медичні ліки, то весь контент повинен бути сфокусований на медицині і медичних препаратах. Завдяки таргетованим RSS-фідам на сайті можна регулярно представляти релевантну інформацію. Оскільки пошукові системи, наприклад, Google, віддають переваги сторінкам тісно зв'язаним однією тематикою.

Існує три основні переваги RSS для пошукових систем:

1. RSS - фіди забезпечують поточний релевантний контент. Існують видавці RSS-фідів, які спеціалізуються на контенті певної тематики. Оскільки фіди сильно таргетовані, вони можуть містити ключові словами. Додавання даних ключових слів на сторінки сайту сприятиме високій оцінці при ранжируванні пошуковими системами.

2. RSS-фіди забезпечують свіжий контент, що регулярно обновляється. RSS-фіди від великих видавців обновляються в певний час. Дані зміни також відображаються і на ваших сторінках з RSS-фідами. Тобто у вас буде свіжий релевантний контент кожну годину або щодня.

3. RSS-фіди сприяють для візитів пошукових роботів. Доведено на практиці, сайти з RSS-фідами Googlebot обходить майже щодня. Тобто сайт частіше індексується, значить, всі нові сторінки вашого сайту Googlebot знайде швидше, ніж на сайті без RSS-фідів.

Інтернет є гігантським сховищем інформації, обсяг якої подвоюється щороку. За експертними оцінками, кількість новин тільки в українському сегменті Інтернету перевищує 50 тисяч повідомлень на добу.

Очевидно, що така різноманітність інформації може бути корисною лише при ефективному доступі до неї, що виявляється не просто здійснити на практиці. Так, за оцінками експертів, близько 79 % журналістів звертаються до Інтернет у пошуках новин, і лише 20 % знаходять ту інформацію, яка їм необхідна.

Користувачі, звичайно ж можуть читати RSS-файли за допомогою стандартних Web-браузерів, проте це зв'язано з переглядом XML-розмітки і повною відсутністю всякого оформлення. Для інтерпретації цього формату існує безліч програм, створених в основному в останні два-три роки. Тобто користувачі можуть отримати доступ до даних у форматі RSS за допомогою спеціальних програм.

Ці програми називаються RSS - агрегаторами і в наочному вигляді відображають зміст RSS-фідів.

За допомогою сучасної RSS-технології користувачі Інтернет отримали надійний і простий доступ до ресурсів оперативної інформації з Web-сайтів Мережі. Перспективність і популярність RSS як стандарту обумовлена перш за все його доступністю і простотою. Сьогодні практично всі провідні інформаційні сайти в світі, "живі журнали", що працюють в Інтернет, використовують RSS як інструмент оперативного представлення оновлень своїх ресурсів. [5]

Електронна пошта приваблива і для спамерів. Нерідко списки електронних адрес користувачів новин на сайтах і порталах стають здобиччю для хакерів. Цей чинник робить підписку через e-mail достатньо ризикованим заняттям. Тому можна припустити, що на зміну розсилкам прийде використовування RSS-фідів.

Програма-агрегатор дозволяє збирати RSS-файли з Web-сайтів, одночасно стежити за появою на них новин і читати їх зміст, цих новин. Програми-агрегатори (їх ще називають RSS-парсерами) виконують синтаксичний розбір даних, представлених у форматі RSS, після чого можуть реалізовувати будь-які дії по відношенню до цих даних, наприклад, посилати їх по

електронній пошті або відображати на певному Web-сайті. Сьогодні найбільш популярні агрегатори, що дозволяють збирати RSS-дані з різних Web-сайтів разом.

RSS-агрегатор, повинен мати інтерфейс, який нагадує інтерфейс поштових програм. У користувача, знайомого з поштовими клієнтами, робота з програмою не повинна викликати труднощів. На відміну від розсилок по електронній пошті, де доставка ініціюється адміністраторами сайтів, після того, як користувач залишив свою адресу, у випадку з RSS користувач сам вводить адресу необхідного йому RSS-фіду в програму-агрегатор. Ця програма періодично перевіряє, чи не змінився зміст RSS-фіду, і при наявності змін автоматично закачує його вміст. Головною перевагою RSS-технології являється однобічний зв'язок - користувач сам приймає рішення про отримання кожного конкретного повідомлення. Популярність RSS-технології у власників Web-ресурсів набирає все більшу популярність ще і завдяки своїй економічності - не вимагається ніяких засобів боротьби із спамом, фільтрації листів, управління розсилкою. При цьому все, кому це необхідне одержують необхідну інформацію про важливі події, корпоративні анонси, оновленнях Web-сайтів. Системи синдикації Інтернет-новин вирішують проблему знаходження необхідної інформації, але залишають без уваги такі задачі, як узагальнення даних - їх обробку і аналіз. Одним з найперспективніших напрямів узагальнення інформаційних потоків в даний час є метод "глибинного аналізу текстів" (Text Mining). Стосовно потоків новин, його ідеологію можна сформулювати як постійне відтворення в часі виконання їх змістовного аналізу. Саме безперервна аналітична обробка повідомлень є найхарактернішою межею цього методу, який дозволяє формувати автоматичні дайджести, виявляти нові поняття і їх взаємозв'язки, розраховувати різноманітні рейтинги.

Переваги роботи RSS для пошуку. Власники веб-сайтів вже давно усвідомили, що нова інформація допомагає привертати і утримувати відвідувачів, а тому кількість джерел нової інформації в Мережі постійно зростає, ускладнюючи тим самим пошук конкретних даних. Можна сказати, що проблема пошуку інформації сьогодні отримала нове значення: пошук інформації в необмеженому неоднорідному динамічному інформаційному середовищі.

Традиційні пошукові системи пропонують лише часткове рішення цієї проблеми. Періоди індексації у них складають від тижнів до декількох місяців. І не дивлячись на те, що практично всі відомі пошукові портали (Yahoo!, AltaVista, Lycos і ін.) мають розділи новин, вони, самі по собі, вже багато кого не влаштовують. Традиційним підходам до організації пошуку мережної інформації властиві такі недоліки, як низька оперативність, залежність від набору джерел і обмеженість спектру цих джерел, середні пошукові можливості, відсутність засобів повідомлення про появу нових даних.

Одна з проблем знаходження інформації в Мережі зумовлена основним форматом, в якому представлена ця інформація – HTML. Цей формат був розроблений, в першу чергу, для вирішення задач відображення змісту на кожному конкретному веб-ресурсі, тому не завжди зручний для автоматичної обробки інформації, у тому числі і організації пошуку. В результаті інформація в Інтернет виявилася орієнтована, перш за все, на окремі сайти і дуже слабо пристосована для автоматизованого узагальнення, класифікації і аналітичної обробки. [6]

При імпортуванні у веб-ресурс інформації з іншого сайту (включення повідомлень новин і т. п.) виникає питання однотипного представлення їх змісту (контента). Якщо це питання не розв'язується, то зміна HTML-оформлення сайту-джерела приводить до необхідності одночасної модифікації програмного забезпечення на всіх сайтах, які приймають від нього інформацію. Оптимальне рішення, здатне допомогти орієнтуватися в інформації Мережі, в даний час надають інформаційні служби нового типу - системи синдикації новин. Під синдикацією в даному випадку розуміються технології збору інформації в Інтернеті і подальше розповсюдження її фрагментів відповідно до потреб користувачів.Служби синдикації забезпечують одночасну публікацію одних і тих же даних на різних сторінках, сайтах і мобільних пристроях (у тому числі, в кишенькових комп'ютерах і мобільних телефонах), а також доставку інформації користувачам.

Технологія синдикації Інтернет-новин включає (навчання) програм збору інформації структурі вибраних джерел, сканування інформації, її нормування, приведення до загального формату (RSS), класифікацію, кластеризацію і доставку користувачам різними каналами (e - mail, WWW, Wap, SMS і ін.). [6]

Формат RSS забезпечує злагоджений спосіб резюмувати вміст веб-сайтів. Якщо є необхідність оперативно відстежувати зміни на сайті (містить фід), то пропонується робити це за допомогою програми-агрегатора не відвідуючи самого сайту. На рис. 1. зображено основні риси RSS технології.

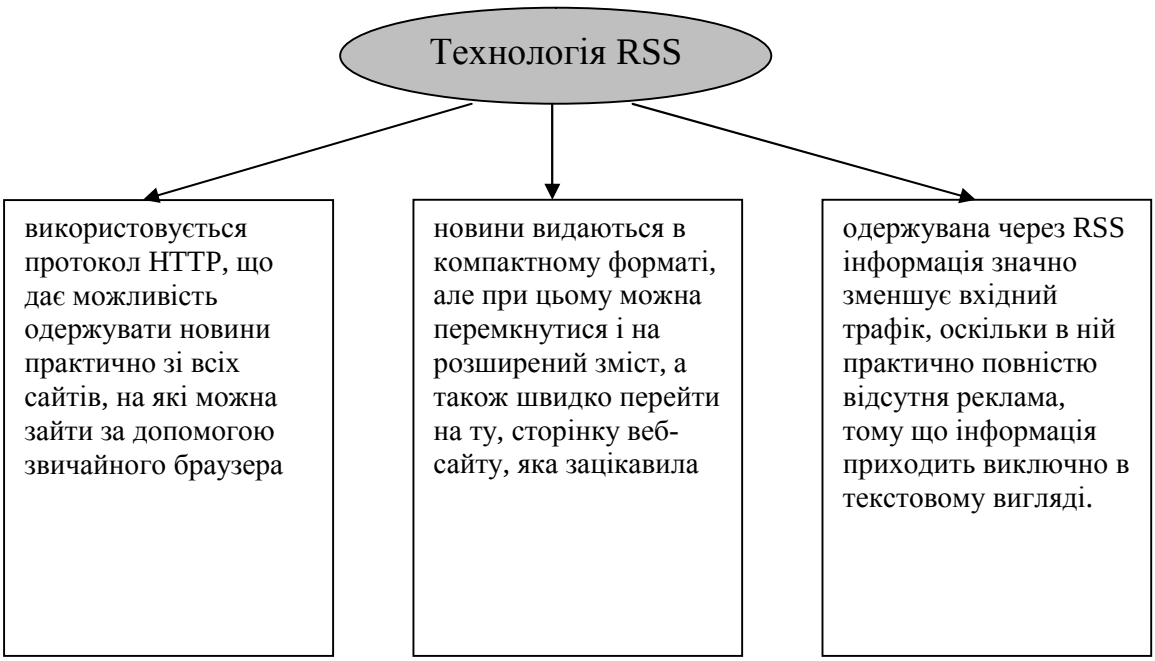

Рис. 1. Основні риси технології RSS

Програма - агрегатор дозволяє збирати всі публікації що цікавлять користувача сайтів разом, одночасно стежити за появою новин на всіх сайтах відразу і читати їх короткий зміст, не відкриваючи кожний сайт окремо.

Загальні функції, що забезпечують користувача головним - можливістю читання RSS-каналів, у кожної програми або сервісу доповнені своїми. Одні програми краще структурують прочитані новини, інші дозволяють зручно шукати і зберігати найцікавіші записи. Сервіси онлайна для читання RSS-стрічок забезпечують рішення іншої важливої задачі - читати новини можна з будь-якого комп'ютера, що є актуально для тих, хто хоче бути в курсі новин і на роботі, і вдома. Можливості RSS – агрегатору можна представити графічно на рис.2.

Рис. 2. Можливості RSS – агрегатору

Висновок. На даний час проблема інформаційних комунікацій в інтернеті є актуальною. Користувачі хочуть, як найшвидше отримати, або відправити інформацію. Програми-агрегатори дозволяють групувати публікації з різних джерел. Таким чином з'являється можливість одночасно відстежувати появу новин на всіх сайтах, без відкриття кожного ресурсу окремо. Формат RSS забезпечує злагоджений спосіб резюмувати вмісту веб-сайтів, що є дуже зручно і практично.

[1] С.Браун "Мозаика" и "Всемирная паутина" для доступа к Internet: Пер. з англ. - М.: Мир: Малип: СК Пресс, (1996), с.167.
[2] Б.Ігер Работа в Internet / Під ред. А. Тихонова; Пер. з англ. - М.: БІНОМ, (1996), с. 313.
[3] Эд.Крол Все об Internet: Руководство и каталог / Пер. з англ. С.М. Тимачева. - Киев: BNV, (1995), с. 591.
[4] А.В.Фролов , Г.В.Фролов Глобальные сети компьютеров. Практическое введение в Internet, E-mail, FTP, WWW, и HTML, программирование для Windiws Sockets. - Діалог - МІФІ, (1996), с. 283.
[5] R.Brachman, J.Schmolze An Overview of the KL-ONE Knowledge Representation System // Cognitive Science, ─ Vol. 9, ─ No.2, ─ (1985), P.171-216.
[6] T.Bray, J.Paoli, C.Sperberg Extensible Markup Language (XML) 1.0. W3C Recommendation. ─ Feb (1998).