Міністерство освіти і науки України
Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука

Факультет кібернетики
Кафедра математичного моделювання

Чернявський Олексій Ігорович

Розробка модуля функцій парсингу XML і регулярних виразів як важливої технологічної складової АРМ в автосервісі (на прикладі ТОВ «ДА ТЕХ Рівне»)

8.080201 – „Інформатика"
А В Т О Р Е Ф Е Р А Т
магістерської дисертації на здобуття академічного ступеня магістра з інформатики

Науковий керівник:
Р.М.Літнарович, доцент,
кандидат технічних наук

Рівне – 2012

УДК 004.42 Чернявський О.І. Розробка модуля функцій парсингу XML і регулярних виразів як важливої технологічної складової АРМ в автосервісі (на прикладі ТОВ «Да Тех Рівне»). Автореферат магістерської дисертації на здобуття академічного ступеня магістра з інформатики. Науковий керівник Р.М.Літнарович. МЕГУ, Рівне, 2011.- 28 с.

Робота виконана на кафедрі математичного моделювання Міжнародного економіко-гуманітарного університету імені академіка Степана Дем'янчука

Рецензенти:В.Г.Бурачек, доктор технічних наук, професор
В.О.Боровий, доктор технічних наук, професор
…….Є.С.Парняков, доктор технічних наук, професор
Відповідальний за випуск: Й.В.Джунь, доктор фіз.-мат. наук, професор

Експериментальні алгоритми реалізовані на мові програмування С в середовищі Linux з використанням компілятора GNU GCC, дебагера GDB, текстового редактора VIM. Алгоритми протестовані і показали правильність роботи з вхідними даними різної складності.

Ключові слова: експериментальні алгоритми, модуль функцій, реалізація, тестування.

Экспериментальные алгоритмы реализованы на языке программирования С в среде Linux с использованием компилятора GNU GCC, дебагера GDB, текстового редактора VIM. Алгоритмы протестированы и показали правильность работы со входными данными разной сложности.
Ключевые слова: экспериментальные алгоритмы, модуль функций, реализация, тестирование.

Experimental algorithms programming of C realized in language in the environment of Linux with the use of compiler of GNU GCC, debager of GDB, text editor VIM. Algorithms are tested and showed the rightness of work with the datains of different complication.
Keywords: experimental algorithms, module of functions, realization, testing.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми магістерської дисертації. Мова XML (eXtensible Markup Language) привабила достатньо уваги серед розробників і користувачів середовища Інтернет, щоб задуматись над питанням використання її в якості основного інструменту для створення Web-програм, зберігання даних. Регулярні вирази в свою чергу є мовою пошуку текстової інформації.

XML і регулярні вирази відносяться до області зберігання, передачі, пошуку і обробки текстової інформації. Оскільки XML є найбільш популярним форматом зберігання даних, очевидним є необхідність програмного модуля функцій парсингу XML. Даний модуль може повторно використовуватись в зовнішніх програмних проектах. Регулярні вирази, з іншого боку, це популярна мова опису шаблонів пошуку текстової інформації.

XML і регулярні вирази широко використовуються в області зберігання, передачі і пошуку текстової інформації, що свідчить про необхідність та актуальність даних технологій для індустрії програмного забезпечення.

Мета роботи. Метою даної роботи є розробка алгоритму швидкого парсингу XML і регулярних виразів.

Для досягнення мети в дисертації поставлені завдання:

- Оглянути існуючі програмні бібліотеки парсингу XML і регулярних виразів.
- Висвітлити теоретичні основи технологій XML і регулярних виразів.
- Розробити експериментальні алгоритми парсингу XML і регулярних виразів.

Об'єкт дослідження. Досліджується алгоритм реалізації парсингу XML і регулярних виразів.

Предметом дослідження є стандарт XML, синтаксис мови XML і регулярних виразів.

Методологічною основою написання дисертації є

специфікація стандарту, спеціалізовна література і довідники.

.

Наукова новизна полягає у програмній реалізації розроблених автором алгоритмів парсингу XML і регулярних виразів: високорівнева модель алгоритму, вихідний код на мові С, описується технологічна база розробки і тестування алгоритму.

Практична значимість і реалізація роботи полягає в розробці програмного продукту, який є перевірений, протестований та впроваджений у виробництво. Розроблена високорівнева модель алгоритму відповідає усім вимогам, які були поставлені до даного програмного забезпечення. Перевагою системи є забезпечення можливості її роботи на будь-якому ПК

Апробація роботи. Окремі розділи дисертації були докладені і отримали одобрення на наукових конференціях студентів і аспірантів у 2010 і 2011 роках, а також на науковому семінарі кафедри математичного моделювання.

Публікації. Основні положення дисертації опубліковані в монографії автора : Чернявський О.І. Розробка модуля функцій парсингу XML і регулярних виразів як важливої технологічної складової АРМ в автосервісі (на прикладі ТОВ «Да Тех Рівне»). Монографія. Науковий керівник Р.М.Літнарович. МЕГУ, Рівне, 2012.- 163 с.

Основні положення дисертації, що виносяться на захист:

- повний опис існуючих реалізацій програмних бібліотек парсингу XML і регулярних виразів, аналіз їхніх властивостей, переваг і недоліків;
- опис теоретичних основ технологій XML;
- опис мови регулярних виразів: основні базові принципи, історія виникненя, типи машин регулярних виразів, синтаксис;
- програмна реалізація розроблених автором алгоритмів парсингу XML і регулярних виразів: високорівнева модель алгоритму, вихідний код на мові С;
- опис технологічної бази розробки і тестування алгоритму;

- середовище розробки.

Структура і об'єм роботи:

Магістерська дисертація складається із вступу, п'яти розділів, розбитих на підрозділи, висновків і списку використаних джерел. Обсяг дисертації 163 сторінки. Список використаної літератури із 24 найменувань, в тому числі 10 на іноземній мові.

ВСТУП
ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обгрунтовується актуальність теми, дається короткий огляд результатів, що мають безпосереднє відношення до теми роботи, та загальна характеристика магістерської дисертації.

В першому розділі описується практичне застосування криптографії у сфері інформаційної безпеки, приводяться криптографічні засоби захисту інформації, математичні основи і дається порівняльний аналіз симетричних та асиметричних алгоритмів.

Дисертація складається з п'яти розділів. **Перший розділ** описує основи мови XML та інших технологій тісно пов'язаних з даною мовою. Також представлений огляд існуючих реалізацій парсерів XML. Зокрема в розділі висвітлена слідуюча інформація: що таке XML, для чого потрібна нова мова розмітки, структура документа XML, правила створення XML-документа, конструкції мови, простір імен, розглядаються існуючі бібліотеки парсингу, їхні властивості, переваги і недоліки.

Другий розділ описує основи мови регулярних виразів: історія виникнення, базові принципи, типи машин регулярних виразів, синтакси. Також представлений огляд існуючих реалізацій парсерів регулярних виразів (Regular Expressions).

Третій розділ включає програмну реалізацію розробленого нами модуля парсингу XML: описуються особливості даної реалізації, представлена високорівнева модель алгоритму, вихідний код на мові С, описується технологічна база розробки і тестування модуля.

Четвертий розділ включає програмну реалізацію розробленого нами модуля регулярних виразів: високорівнева модель алгоритму, вихідний код на мові С, технологічна база розробки і тестування модуля.

П'ятий розділ описує реалізацію повнофункціональної прикладної програми в якій в повній мірі застосовані модулі парсингу XML і регулярних виразів. Даний програмний продукт представляє собою реалізацію програми-електронного словника що працює з енциклопедичними і словниковими базами даних в форматі XML. Дана прикладна програма показує в повній мірі ефективність і правильність роботи реалізованих мною модулів парсингу XML і регулярних виразів.

Модуль парсингу XML пройшов тестування з XML-файлами розміром до 300 MB. Тести показали що модуль працює коректно, швидко і не споживає багато пам'яті. Відповідно до результатів тестів можна зробити висновок що розроблений програмний модуль може бути використаний в зовнішніх програмних проектах для парсингу XML і пошуку інформації з використанням регулярних виразів.

Розділ III. Програмна реалізація модуля парсингу XML

3.1 Особливості реалізації

При реалізації модуля парсингу XML були поставлені завдання досягти слідуючі характеристики кінцевої реалізації: швидка робота, толерантне використання пам'яті. Щоб отримати такі характеристики була застосована модель — одне джерело, багато користувачів. Дана модель в контексті парсингу XML представляє слідуючий алгоритм роботи. На початку роботи модуль зчитує в буфер весь файл, що містить XML. Після цього запускається процедура парсингу, якій передається даний буфер.

Парсинг представляє собою процес пошуку і аналізу XML-тегів в буфері. Для кожного знайденого тега створюється структура в якій зберігається вказівник на місце в буфері де даний тег має початок. Дане рішення має ключове значення. Зберігання вказівника на дані, а не копіювання, робить алгоритм парсингу швидким і економічним в використанні пам'яті. Необхідно відзначити, що процедура виділення пам'яті, крім того що з'їдається сама пам'ять, створює додаткові затримки в роботі. Тобто маючи один буфер, що представлає собою одне джерело, ми створюємо багато структур, користувачі, що містять вказівники на місця в буфері де беруть початок відповідні теги.

3.2 Програмна модель

Основою алгоритму парсингу XML є рекурентна функція dexml(). Дана функція працює за слідуючим принципом. На вході вона отримує вказівник на буфер з даними XML. Запускається цикл який послідовно шукає теги на одному рівні ієрархії. На початку циклу перевіряється чи міститься на початку тексту тег XML. Якщо так, то даний текст обробляється як тег, якщо ж ні, то даний текст обробляється як кінцеві дані що містяться в тегу що знаходиться на рівень вище. В першому випадку з тексту виділяється самий зовнішній тег і його вміст. Інформація про виділений тег і його атрибути розміщується в новоствореній структурі DTag:

```
typedef struct _DAttr DAttr;
struct _DAttr {
    wchar_t* name;
       int nlen;
    wchar_t* value;
       int vlen;
       int pos;
    DAttr* next;
};
```

```
typedef struct _DTag DTag;
struct _DTag {
       int type;
       unsigned sid;
       wchar_t *name;
       int nlen;
       wchar_t *value;
       int vlen;
       DAttr* attributes;
       uint32_t pos;
       uint32_t len;

       DTag *parent;
       DTag *body;
       DTag *next;

       wchar_t *source_xml;
};

typedef enum _DTagType {
    DT_Tag,
    DT_Value,
    DT_Attribute,
       DT_Cdata
} DTagType;
```

Тип даних позначається як DT_Tag.

Дана структура відноситься до списочних структур. Це означає що вона вказує на слідуючу структуру яка є такого самого типу [11;266].

Після запису інформації про тег в дану структуру, функція dexml() викликає сама себе (рекурсія), передаючи вміст останнього тегу, а результат її роботи записується в елемент структури body [10;310]. Після того як рекурентний виклик вернув дані, функція dexml() шукає кінець виділеного тега і

зміщує вказівник початку буфера на дані що починаються відразу за кінцевим символом виділеного тега. Контроль вертається на початок цикла.

В кінці функція dexml() вертає вказівник на вже заповнену структуру Dtag.

Якщо на початку чикла функції dexml() виявлені прості дані, а не тег, то ці дані записуються в структуру DTag і тип даних позначається як DT_Value.

Крім самої функції парснгу XML даний модуль предоставляє зручні користувацькі функції перебору і пошуку тегів по структурі DTag, яка містить всі розпарсені дані XML. До таких функцій відносяться: tag_seek() і tag_rnext(). Також в модулі є зручна функція перевірки тега is_tag_name().

3.3 Комп'ютерна реалізація

Модуль парсингу XML був реалізований на мові програмування С в середовищі Linux з використанням компілятора GNU GCC, дебагера GDB, текстового редактора VIM. Модуль був протестований на XML-файлах розміром 300 MB. Тести показали коректність роботи модулі з вхідними даними різної складності. Крім того модуль визначається високою швидкість роботи і малим об'ємом пам'яті що споживається під час парсингу.

Комп'ютерна реалізація складається із слідуючих файлів:

- xml_utf8.c - вихідний файл програми на мові С що реалізує парсер XML. Версія з підтримкою Юнікод (Додаток А).
- xml_utf8.h - вихідний заголовочний файл мови С що містить структури XML і об'явлення функцій модуля XML. Версія з підтримкою Юнікод (Додаток А).
- Makefile - вхідний файл програми make для автоматичного компілювання і збирання модуля програми (Додаток А).

- config.mk - файл що включається в Makefile. Містить загальні параметри для компілятора (Додаток А).
- parse.c - програма для тестування XML парсера

На даний час існують стабільні реалізації парсерів XML і регулярних виразів, сере яких виділяються libxml2 і Expa, PCRE. В Libxml2 реалізований ряд існуючих стандартів, що відносяться для мов розмітки. Дана бібліотека характеризується високою якістю реалізації і підтримкою великої кількості стандартів що є частиною XML. До не доліків даної бібліотеки можна віднести лише її великий розмір що часто ставить під сумнів доцільність її використання для статичного лінкування з невеликими программами. Expat є менш потужною ніж libxml2 що стосується підтримки стандартів XML, але при цьому вона має набагато менший розмір і більшу швидкість роботи.

Найбільш популярною для мови С бібліотекою регулярних виразів є PCRE - Perl Compatible Regular Expressions. Синтаксис регулярних виразів PCRE є значно потужнішим і гнучкішим ніж стандартних регулярних виразів POSIX. Бібліотека сумісна з великою кількістю С компіляторів і операційних систем. Багато людей розробили бібліотеки на основі PCRE щоб зробити її сумісною з іншими мовами програмування. Бібліотека PCRE підтримує лише пошук по регулярному виразу і цю роботу вона виконує досить добре.

Загальними перевагами даних бібліотек є їхня портабельність, тобто можливість скомпілювати для різних операційних систем, і відкритість коду.

XML корисний для автоматизованих програмних засобів, що шукають у Web. Недосконаліть HTML призвела до того, що мережа перетворилася в мішанину тексту, повну різноманітних елементів і тегів, часто використовуваних, що називається Pro Forma і нічого не значущих.

XML має величезний потенціал для удосконалення гіпертекста. Наприклад у HTML для створення зв'язку використовується елемент A, XML же дозволяє створити не

просто посилання, а наприклад, двонаправлений зв'язок.

Перспектива XML полягає в тому, що він буде використовуватися для опису інших мов розмітки, наприклад, JavaScript, що використовується в HTML-документах.

XML розроблений для того, щоб спростити і полегшити використання SGML, при цьому зберігши його великі можливості по створенню, поширенню і публікації Web-документів мережі. XML – підмножина SGML, причому любий дійсний документ XML є дійсним документом SGML. І, як і SGML, XML - це метамова, що визначає інші мови розмітки для специфічних цілей. Наприклад, мова синхронізованої інтеграції мультимедіа (Synchronized Multimedia Integration Language, SMIL) базується на XML.

XML дозволяє визначити формальний синтаксис мови, наприклад правила вкладення елементів. Семантику можна, звичайно, описувати на звичайній англійській мові.

XML використовується для розмітки стандартних документів багато в чому так само, як HTML. Проте XML перевершує його при роботі зі структурованими даними, такими, як результати запиту, метаінформація про вузол Web або елементи і типи схеми.

Регулярні вирази це основна мова пошуку текстової інформації в WEB. Особливо корисні регулярні вирази в програмах, написаних на скриптових (що інтерпретуються) мовах, наприклад, VBScript, JavaScript і Perl.

Найбільший розвиток регулярні вирази отримали в Perl, де їх підтримка вбудована безпосередньо в інтерпретатор. У інших мовах, як правило, використовуються реалізовуючі регулярні вирази доповнення і модулі сторонніх розробників.

Реалізації регулярних виразів розрізняються, проте в цілому вони дуже схожі один на одного, і, як правило, програміст, що одного дня освоїв використання регулярних виразів, надалі практично не зустрічає утруднень.

Синтаксис регулярних виразів до цих пір не повністю стандартизований. Існує POSIX-версия регулярних виразів, що повністю описує стандарт синтаксису для POSIX. Але версія Perl ширша і більш гнучка, чим і пояснюється її широка поширеність.

Бібліотека парсингу XML і регулярних виразів була написана на мові C в середовищі Linux. Дана бібліотека складається з двох модулів. Перший модуль реалізує рекурсивну функцію парсингу XML. Дана функція вертає просту програмну структуру DTag що містить XML дані. В дальшому все що залишається програмісту зробити так це виконати обхід даної структури для розбору XML документа. Операція обходу структури DTag є набагато простішою в порівнянні з парсингом оригінального XML документу.

Модуль регулярного виразу також побудований з використанням рекурсивного алгоритму. Даний модуль реалізує функцію regTest_r, яка отримує два параметри: текст, в якому здійснюється пошук, і регулярний вираз, що також є текстовою стрічкою. Даний алгоритм вирізняється високою гнучкістю в сенсі можливості його розширення для підтримки ширшого набору мета-символів регулярних виразів Perl. Даний модуль був протестований з різними вхідними даними різної складності і довів правильність і ефективність свої роботи.

Розділ IV. Програмна реалізація модуля регулярних виразів

4.1 Програмна модель

Алгоритм парсингу регулярного виразу є рекурсивним. Рекурсивність алгоритму походить від рекурсивної природи самих регулярних виразів, як і більшості інших коп'ютерних мов. Основою реалізації є функція regTest_r. Дана функція отримує 2 параметра: текст, в якому здійснюється пошук, і регулярний вираз, по якому текст сканується. Кожен символ регулярного виразу перевіряється на спеціальний симівол. Якщо зустрічається спеціальний символ що вказує на

множину можливих послідовностей символів, функція рекурсивно викликає саму себе, зменшуючи діапазон тексту пошуку, доти поки якийсь з викликів не верне позитивний результат, що означає що дана частина тексту збігається з даним регулярним виразом, або текст пошуку не співпадатиме з даним піддіапазоном регулярного виразу. В першому випадку функція закінчує роботу і вертає позитивний результат, інакше вертає результат False.

Дана функція є дуже швидкою і її легко розширити для підтримки різних синтаксичних конструкцій регулярних виразів мови Perl.

4.2 Комп'ютерна реалізація

Модуль парсингу регулярних виразів був реалізований на мові програмування C в середовищі Linux з використанням компілятора GNU GCC, дебагера GDB, текстового редактора VIM.

Комп'ютерна реалізація складається із слідуючих файлів:

- regexp_utf8.c - вихідний файл програми на мові C що реалізує парсер регулярних виразів. Версія з підтримкою Юнікод (Додаток Б).
- regexp_utf8.h - вихідний заголовочний файл мови C що містить об'явлення функцій модуля регулярних виразів. Версія з підтримкою Юнікод (Додаток Б).
- Makefile - вхідний файл програми make для автоматичного компілювання і збирання модуля програми (Додаток Б).
- config.mk - файл що включається в Makefile. Містить загальні параметри для компілятора (Додаток Б).
- demo.c - програма для тестування парсера регулярних виразів.

- **Розділ V. Реалізація повнофункціональної прикладної програми з використанням розроблених модулів парсингу XML і регулярних виразів**

5.1 Опис програмного продукту

Для демонстрації практичного застосування модулів парсингу XML і регулярних виразів, а також з метою прикладного застосування, був розроблений програмний продукт — клієнт баз даних XDClient. Дана програма розроблена для реалізації слідуючих задач: пошук інформації в базах даних що зберігаються в форматі XML.

Одним із форматів баз даних є XDXF. Це спеціалізований формат вищого рівня що базується на форматі XML. Формат XDXF розроблений спеціально для зберігання баз даних словників та енциклопедій.

Для XDClient версії 1.0 передбачена повна підтримка баз даних в формат XDXF, тобто програма зможе здійснювати повноцінний пошук і виведення інформації любих баз даних що зберігаються в форматі XDXF. На момент написання даної дисертації XDClient мав версію 0.8. Дана версія повністю вміє читати бази даних XDXF, здійснювати пошук і виведення інформації.

Особливістю даної програми є її здатність виконувати пошук потрібних даних дуже швидко. Для прикладу, при наявності 123 баз даних загальним об'ємом 570 MB, пошук даних триває в середньому 1с. Така безпрецедентна швидкість пошуку була досягнута завдяки двум ключовим технологіям: індексація і бінарний пошук [11;210]. Опис технологій індексації і бінарного пошуку, а також опис реалізації даних технологій в XDClient, виходить за рамки теми даної дисертації. Опис програмної реалізації XDClient також виходить за рамки теми даної дисертації (дивіться Додаток В).

5.2 Комп'ютерна реалізація

Програма XDClient була реалізована на мові програмування С в середовищі Linux з використанням компілятора GNU GCC, дебагера GDB, текстового редактора VIM. В подальшому вона була портована на Windows.

Комп'ютерна реалізація складається із слідуючих файлів:

- xdc.c - вихідний файл програми на мові С що містить точку входу і головні контрольні процедури (Додаток В).
- xdc.h - вихідний заголовочний файл мови С що містить об'явлення внутрішніх функцій і типів XDClient (Додаток В).
- xdxf.c - вихідний файл програми на мові С що містить модуль парсингу баз даних XDXF (Додаток В).
- xdxf.h - вихідний заголовочний файл мови С що містить об'явлення функцій і типів модуля парсингу XDXF (Додаток В).
- index.c - вихідний файл програми на мові С що містить модуль індексації даних (Додаток В).
- index.h - вихідний заголовочний файл мови С що містить об'явлення функцій і типів модуля індексації даних (Додаток В).
- utf8.c - вихідний файл програми на мові С що містить спеціалізовані функції для роботи з даними UNICODE в форматах UTF-8 і Wide Char (Додаток В).
- utf8.h - вихідний заголовочний файл мови С що містить об'явлення спеціалізованих функцій і типів (Додаток В).
- file.c - вихідний файл програми на мові С що містить спеціалізовані функції роботи з файлами і каталогами файлової системи (Додаток В).
- file.h - вихідний заголовочний файл мови С що містить об'явлення спеціалізованих функцій і типів (Додаток В).
- config.h – вихідний заголовочний файл мови С що містить об'явлення глобальних константних даних і конфігураційних

параметрів (Додаток В).

- Makefile - вхідний файл програми make для автоматичного компілювання і збирання модуля програми (Додаток В).

- config.mk - файл що включається в Makefile. Містить загальні параметри для компілятора (Додаток В).

Висновки

На даний час існують стабільні реалізації парсерів XML і регулярних виразів, сере яких виділяються libxml2 і Expa, PCRE. В Libxml2 реалізований ряд існуючих стандартів, що відносяться для мов розмітки. Дана бібліотека характеризується високою якістю реалізації і підтримкою великої кількості стандартів що є частиною XML. До недоліків даної бібліотеки можна віднести лише її великий розмір що часто ставить під сумнів доцільність її використання для статичного лінкування з невеликими программами. Expat є менш потужною ніж libxml2 що стосується підтримки стандартів XML, але при цьому вона має набагато менший розмір і більшу швидкість роботи.

Найбільш популярною для мови С бібліотекою регулярних виразів є PCRE - Perl Compatible Regular Expressions. Синтаксис регулярних виразів PCRE є значно потужнішим і гнучкішим ніж стандартних регулярних виразів POSIX. Бібліотека сумісна з великою кількістю С компіляторів і операційних систем. Багато людей розробили бібліотеки на основі PCRE щоб зробити її сумісною з іншими мовами програмування. Бібліотека PCRE підтримує лише пошук по регулярному виразу і цю роботу вона виконує досить добре.

Загальними перевагами даних бібліотек є їхня портабельність, тобто можливість скомпілювати для різних операційних систем, і відкритість коду.

XML корисний для автоматизованих програмних засобів, що шукають у Web. Недосконаліть HTML призвела до того, що мережа перетворилася в мішанину тексту, повну

різноманітних елементів і тегів, часто використовуваних, що називається Pro Forma і нічого не значущих.

XML має величезний потенціал для удосконалення гіпертекста. Наприклад у HTML для створення зв'язку використовується елемент A, XML же дозволяє створити не просто посилання, а наприклад, двонаправлений зв'язок.

Перспектива XML полягає в тому, що він буде використовуватися для опису інших мов розмітки, наприклад, JavaScript, що використовується в HTML-документах.

XML розроблений для того, щоб спростити і полегшити використання SGML, при цьому зберігши його великі можливості по створенню, поширенню і публікації Web-документів мережі. XML – підмножина SGML, причому любий дійсний документ XML є дійсним документом SGML. І, як і SGML, XML - це метамова, що визначає інші мови розмітки для специфічних цілей. Наприклад, мова синхронізованої інтеграції мультимедіа (Synchronized Multimedia Integration Language, SMIL) базується на XML.

XML дозволяє визначити формальний синтаксис мови, наприклад правила вкладення елементів. Семантику можна, звичайно, описувати на звичайній англійській мові.

XML використовується для розмітки стандартних документів багато в чому так само, як HTML. Проте XML перевершує його при роботі зі структурованими даними, такими, як результати запиту, метаінформація про вузол Web або елементи і типи схеми.

Регулярні вирази це основна мова пошуку текстової інформації в WEB. Особливо корисні регулярні вирази в програмах, написаних на скриптових (що інтерпретуються) мовах, наприклад, VBScript, JavaScript і Perl.

Найбільший розвиток регулярні вирази отримали в Perl, де їх підтримка вбудована безпосередньо в інтерпретатор. У інших мовах, як правило, використовуються реалізовуючі

регулярні вирази доповнення і модулі сторонніх розробників.

Реалізації регулярних виразів розрізняються, проте в цілому вони дуже схожі один на одного, і, як правило, програміст, що одного дня освоїв використання регулярних виразів, надалі практично не зустрічає утруднень.

Синтаксис регулярних виразів до цих пір не повністю стандартизований. Існує POSIX-версия регулярних виразів, що повністю описує стандарт синтаксису для POSIX. Але версія Perl ширша і більш гнучка, чим і пояснюється її широка поширеність.

Бібліотека парсингу XML і регулярних виразів була написана на мові C в середовищі Linux. Дана бібліотека складається з двох модулів. Перший модуль реалізує рекурсивну функцію парсингу XML. Дана функція вертає просту програмну структуру DTag що містить XML дані. В дальшому все що залишається програмісту зробити так це виконати обхід даної структури для розбору XML документа. Операція обходу структури DTag є набагато простішою в порівнянні з парсингом оригінального XML документу.

Модуль регулярного виразу також побудований з використанням рекурсивного алгоритму. Даний модуль реалізує функцію regTest_r, яка отримує два параметри: текст, в якому здійснюється пошук, і регулярний вираз, що також є текстовою стрічкою. Даний алгоритм вирізняється високою гнучкістю в сенсі можливості його розширення для підтримки ширшого набору мета-символів регулярних виразів Perl. Даний модуль був протестований з різними вхідними даними різної складності і довів правильність і ефективність свої роботи.

АВТОРСЬКЕ ПРАВО НА ДИСЕРТАЦІЮ

Назва: Розробка модуля функцій парсингу XML і регулярних виразів як важливої технологічної складової АРМ в автосервісі (на прикладі ТОВ “ДА ТЕХ Рівне”). Дисертація магістра інформатики.

Автори: Чернявський, Олексій Ігорович
Чернявский, Алексей Игоревич
Cherniavskyi, Oleksii Igorovych
Літнарович, Руслан Миколайович
Литнарович, Руслан Николаевич
Litnarovych, Ruslan Mykolaiovych

Ключові слова: модуль функцій
парсинг XML
регулярні вирази
індустрія програмного забезпечення
модуль функций

парсинг XML
регулярные выражения
индустрия программного обеспечения
module of functions
parsing XML
regular expressions
industry of software

Дата публікації: 4-кві-2012

Видавець: Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука

Бібліографічний опис: Чернявський О.І. Розробка модуля функцій парсингу XML і регулярних виразів як важливої технологічної складової АРМ в автосервісі (на прикладі ТОВ “ДА ТЕХ Рівне”). Науковий керівник Р.М.Літнарович. МЕГУ, Рівне 2012.- 163 с.

Короткий огляд (реферат): XML і регулярні вирази відносяться до області зберігання, передачі, пошуку і обробки текстової інформації. Оскільки XML є найбільш популярним форматом зберігання даних, очевидним є необхідність програмного модуля функцій парсингу XML. Даний модуль може повторно використовуватись в зовнішніх програмних проектах. Регулярні вирази, з

іншого боку, це популярна мова опису шаблонів пошуку текстової інформації.
XML и регулярные выражения относятся к области хранения, передачи, поиска и обработки текстовой информации. Поскольку XML является наиболее популярным форматом хранения данных, очевидным есть необходимость программного модуля функций парсингу XML. Данный модуль может повторно использоваться во внешних программных проектах. Регулярные выражения, с другой стороны, это популярный язык описания шаблонов поиска текстовой информации.
XML and regular expressions behave to the area of storage, transmission, search and treatment of text information. As XML is the most popular format of storage of data, obvious there is a necessity of the programmatic module of functions of parsing of XML. This module can be repeatedly used in external programmatic projects. Regular expressions, on the other hand, it is a popular language of description of templates of text information

retrieval.

Опис: XML і регулярні вирази широко використовуються в області зберігання, передачі і пошуку текстової інформації, що свідчить про необхідність та актуальність даних технологій для індустрії програмного забезпечення.

URI (Уніфікований ідентифікатор ресурсу): http://enpuir.npu.edu.ua:8080/123456789/569

Розташовується у зібраннях: Літнарович Руслан Миколайович

Файли цього матеріалу:

Файл	Опис	Розмір	Формат	
Чернявський О.І. Дисертація магістра інформатики.pdf		932,91 kB	Adobe PDF	Переглянути/Відкрити

Показати повний опис матеріалу

Перегляд статистики

<u>ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ «ДА ТЕХ РІВНЕ»</u>

Україна 33013 Рівненська обл. м. Рівне вул. Бахарєва,22 тел./ факс /0362/ 436178, Код ЄДРПОУ 36597842 р/р 26004000051270 у ПуАТ «СЕБ Банк» м. Київ, МФО 300175

м. Рівне
26 березня 2012 року

Довідка про впровадження програмного продукту
Декану факультету кібернетики
Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука
Янчуку Петру Степановичу
ТзОВ "ДА ТЕХ РІВНЕ"
м.Рівне, вул. Бахарєва 22
Патріюк Юрій Ростиславович

Програмний продукт "XDClient" призначений для пошуку інформації в базах даних, автором якого є студент Міжнародного економіко-гуманітарного університету імені академіка Степана Дем'янчука, Чернявський Олексій Ігорович, був представлений на тестування його ефективності і коректності роботи.
Робота даного програмного продукту була проаналізована провідними спеціалістами підприємства та протестована у середовищі Windows XP. Після детальних тестів ПЗ "XDClient" отримав оцінку — відмінно. Від 14.02.2012 року за певною домовленістю із автором, програма була впроваджена на підприємстві і широко використовується працівниками при роботі з технічною документацією на автомобільні вузли, а також

при спілкуванні з іноземними інвесторами та постачальниками. Збереження всіх авторських прав підприємство гарантує.

Директор ***Місце***
печатки_____ Патріюк Ю. Р.

Відгук

наукового керівника, доцента, кандидата технічних наук *Літнаровича Р.М.*
на магістерську дисертацію студента _____*Чернявського Олексія Ігоровича*

(прізвище, ім`я, по батькові)
групи *01КІН-М* спеціальності *прикладна математика*
на тему *Розробка модуля функцій парсингу XML і регулярних виразів як важливої технологічної складової АРМ в автосервісі (на прикладі ТОВ "ДА ТЕХ Рівне").*
Актуальність теми *Мова XML привабила достатньо уваги серед розробників і користувачів середовища Інтернет, щоб дослідити питання використання її в якості основного інструменту для створення Web-програм, зберігання даних, тощо. Регулярні вирази в свою чергу є найбільш популярною мовою пошуку текстової інформації. Це робить актуальним розробку стандарних швидких алгоритмів парсингу XML і регулярних виразів.*
Мета дослідження *Метою даної роботи є розробка алгоритму швидкого парсингу XML і регулярних виразів*._____

Об`єкт дослідження ***Досліджується алгоритм реалізації парсингу XML і регулярних виразів***

Коротка характеристика розділів роботи ***Теоретичні основи першого розділу визначають важливість і актуальність мови XML для сучасних систем зберігання, передачі, пошуку і обробки інформації. Другий розділ лаконічно на прикладах розкриває мову регулярних виразів і показує важливість даної мови для пошуку даних в сучасному світі інформаційних технологій. Алгоритми описані в третьому і четвертону розділах мають чітку і зрозумілу структуру, гарні динамічні характеристики, оптимально використовують пам'ять комп'ютера і їх легко розширяти додатковими функціями. П'ятий розділ демонструє практичне використання розробленого модуля на прикладі повнофункціональної програми.***

Практичне значення роботи ***Розроблений модуль може бути повторно використаний в зовнішніх програмних проектах або взятий за основу для розробки стандартної бібліотеки парсингу текстової інформації. Модуль характеризується невеликим розміром що робить його привабливим для статичного лінкування***

Реалізація результатів дослідження ***Результати досліджень реалізовані у форматі ефективних алгоритмів та відповідних програмних модулів на мові С.***

Зауваження та недоліки ***Спостерігаються деякі диспропорції в змістовому наповненні розділів роботи.***

Висновки та оцінка ***Робота виконана на належному науково-практичному рівні згідно стандартів та вимог до кваліфікаційних робіт і заслуговує на оцінку «Відмінно»***

Науковий керівник Підпис
Р.М.Літнарович (к.т.н., доцент)

"__3"_квітня_20 12 р.

Рецензія

на магістерську дисертацію студента ***Чернявського Олексія Ігоровича***
групи ***01КІН-М*** **факультету** ***кібернетики***
ПВНЗ "Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука"
Тема роботи ***Розробка модуля функцій парсингу XML і регулярних виразів як важливої технологічної складової АРМ в автосервісі (на прикладі ТОВ "ДА ТЕХ Рівне").***

Стисла характеристика розділів роботи ***Розділ І. Огляд технології XML***
Розділ ІІ. Огляд технології регулярних виразів
Розділ ІІІ. Програмна реалізація модуля парсингу XML
Розділ IV. Програмна реалізація модуля регулярних виразів
Розділ V. Реалізація повнофункціональної прикладної програми з використанням розроблених модулів парсингу XML і регулярних виразів

Пропозиції, внесені студентом, рівень їх наукового обгрунтування та ефективність
В реалізації алгоритмів парсингу було застосовано лаконічний рекурсивний алгоритм і застосовані ефективні структури даних. Особливо виділяється малий розмір розробленого модуля що дає можливість його статичного компонування в зовнішніх програмних проектах.

Практичне значення роботи ***Програмний модуль може бути використаний в зовнішніх програмних проектах для зчитування і розбору даних в форматі XML і пошуку інформації з використанням мови регулярних виразів. Алгоритм парсингу регулярних виразів легко розширити для підтримки всіх синтаксичних і лексичних конструкцій регулярних виразів мови програмування Perl.***

Якість оформлення роботи ***Дисертація виконана на професійному рівні.***

Недоліки в роботі ***Незначні орфографічні помилки тексту.***

Загальний висновок ***Студент Чернявський О.І. професійно підготовлений і заслуговує на присвоєння кваліфікації "магістр інформатики".***

Оцінка дипломної роботи ***Відмінно (93 бали –А за шкалою ECST).***

Рецензент Панченко Ігор Михайлович,
кандидат фізико-математичних наук, доцент,
завідувач кафедрою математичних дисциплін
та інформаційних систем Рівненської філії
Європейського університету
"_3_" _квітня__ 2012р.

(Підпис рецензента)

Місце печатки

Підпис засвідчую
Начальник відділу кадрів

(Підпис

Чернявський Олексій Ігорович
спеціаліст системотехнік, магістрант інформаційних технологій

Розробка модуля функцій парсингу XML і регулярних виразів як важливої технологічної складової АРМ в автосервісі (на прикладі ТОВ «ДА ТЕХ Рівне»)

8.080201 – „Інформатика”
А В Т О Р Е Ф Е Р А Т
магістерської дисертації на здобуття академічного ступеня магістра з інформатики

Комп’ютерний набір в редакторі Microsoft®Office® Word 2007 О.І.Чернявський
Редагування, верстка, макетування та дизайн Р.М.Літнарович.
Науковий керівник Р. М. Літнарович, доцент, кандидат технічних наук
Міжнародний Економіко-Гуманітарний Університет ім. акад. Степана Дем’янчука
Кафедра математичного моделювання
33027,м.Рівне,Україна
Вул.акад. С.Дем’янчука,4, корпус 1
Телефон:(+00380) 362 23-73-09
Факс:(+00380) 362 23-01-86
E-mail:mail@regi.rovno.ua
ocherngavskyy@gmail.com