РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2014 102 136** (13) **A**

(51) МПК
G06F 17/00 (2006.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ЗАЯВКА НА ИЗОБРЕТЕНИЕ

(21)(22) Заявка: **2014102136/08, 24.01.2014**

Приоритет(ы):
(22) Дата подачи заявки: **24.01.2014**

(43) Дата публикации заявки: **27.07.2015** Бюл. № **21**

Адрес для переписки:
119270, Москва, Фрунзенская наб., 38/1, кв. 136, Коваленко В.В.

(71) Заявитель(и):
Закрытое акционерное общество "РИВВ" (RU)

(72) Автор(ы):
Нагорный Алексей Сергеевич (RU)

(54) СПОСОБ ИЗВЛЕЧЕНИЯ ПОЛЕЗНОГО КОНТЕНТА ИЗ УСТАНОВОЧНЫХ ФАЙЛОВ МОБИЛЬНЫХ ПРИЛОЖЕНИЙ ДЛЯ ДАЛЬНЕЙШЕЙ МАШИННОЙ ОБРАБОТКИ ДАННЫХ, В ЧАСТНОСТИ ПОИСКА

(57) Формула изобретения

1. Способ извлечения полезного контента из установочных файлов мобильных приложений для дальнейшей машинной обработки данных, в частности поиска, содержащий этапы на которых:

- загружают из Интернета на сервер установочный файл приложения неизвестного формата;
- подбирают к нему разархиватор;
- разархивируют загруженный установочный файл в каталог с файлами;
- анализируют полученный каталог, составляют список файлов, содержащихся в нем;
- выбирают из списка файл для дальнейшего анализа;
- подбирают программное обеспечение для чтения файла;
- анализируют выбранный файл на предмет поиска первичного контента;
- формируют список адресов внутреннего размещения первичного контента в виде набора строк;
- переходят к анализу следующего файла, до тех пор, пока в каталоге есть файлы;
- проводят анализ текстового содержимого списка адресов внутреннего размещения первичного контента, и разделяют текст каждой строки на набор символов, идентифицирующих способ хранения соответствующей единицы контента, набор символов, идентифицирующий документ, к которому относится данная единица контента, и набор символов, идентифицирующий тип этой единицы контента;
- разделяют строки адресов внутреннего размещения единицы контента по способу хранения на служебный контент и полезный контент;
- служебный контент удаляют;
- выделяют в оставшемся списке группы строк с адресами внутреннего размещения единиц контента, имеющие полностью совпадающие по месторасположению и тексту

RU 2014102136 A

RU 2014102136 A

группы символов, отражающие способ хранения контента;

- проводят статистическую фильтрацию выделенных групп;

- проводят анализ текстового содержимого строк списка адресов по набору символов, идентифицирующих документ, и формируют документы путем реферирования;

- выкачивают из приложения полезный контент;

- формируют описания приложений;

- сохраняют в базе данных название приложения, ссылку на приложение и описание приложения;

- загружают установочный файл нового приложения, и повторяют все описанные последовательности;

- производят машинную обработку полученной базы данных;

- хранят созданный индексируемый массив базы данных на сервере;

- используют для поисковых запросов пользователей, поступающих через Интернет

2. Способ по п.1, в котором разархиватор подбирают из заранее созданного расширяемого и модифицируемого набора всех известных архиваторов.

3. Способ по п.1, в котором список файлов составляется путем формирования строк содержащих полный путь к файлу в каталоге и название файла.

4. Способ по п.1, в котором программное обеспечение для чтения файла подбирают из заранее созданного расширяемого и модифицируемого набора программного обеспечения.

5. Способ по п.1, в котором анализ выбранного файла на предмет поиска первичного контента производят путем следующих действий: проводят поиск внутрифайловых адресов всех единиц контента самого нижнего уровня, затем проверяют эти единицы контента на соответствие первичному контенту, при этом, если контент является не первичным, повторно подбирают программное обеспечение, раскрывают вложенную структуру данных, проверяют все внутриструктурные адреса контента нижнего уровня и так до тех пор, пока во внутрифайловых адресах самого нижнего уровня не будет найден первичный контент.

6. Способ по п.1, в котором каждая строка из набора строк списка адресов содержит информацию о местонахождении файла в каталоге, полный внутрифайловый адрес каждой единицы первичного контента с указанием всех этапов по извлечению этой единицы первичного контента и полного перечня программного обеспечения, с помощью которого открывают эту единицу на каждом этапе.

7. Способ по п.1, в котором анализ текстового содержимого списка адресов внутреннего размещения первичного контента производят путем выделения перебором комбинаций или на основе эмпирических правил наборов символов, а затем присвоения этому набору символов смыслового значения на основе данных о его месторасположении и повторяемости в списке.

8. Способ по п.1, в котором строки адресов внутреннего размещения единицы контента разделяют по способу хранения, характерным для хранения служебного контента и адреса со способом хранения, характерным для хранения полезного контента с помощью заранее созданного расширяемого и модифицируемого набора правил.

9. Способ по п.1, в котором статистическая фильтрация выделенных групп производится исходя из условия, что если полезный контент хранится одинаковым способом, то он должен быть в большинстве своем однотипным, т.е. если количество однотипного контента превышает заранее назначенный порог, то этот тип контента назначается всей группе, если количество однотипного контента ниже заранее назначенного порога, то из дальнейшего рассмотрения отбрасывается вся группа, как не удовлетворяющая условию хранения полезного контента.

10. Способ по п.1, в котором проводят анализ текстового содержимого строк списка

RU 2014102136 A

RU 2014102136 A

адресов по набору символов, идентифицирующих документ, и формируют документы путем следующих действий: производят анализ текстового содержимого оставшихся строк списка на предмет поиска строк с различными наборами символов, идентифицирующих способ хранения и совпадающими наборами символов, идентифицирующих документ, к которому относится контент, адресом которого является эта строка, если в списке адресов внутреннего размещения таких строк нет, то каждая единица контента определяется как отдельный документ, если есть, то выделяют группы строк, в которых совпадают идентификаторы документа и различаются идентификаторы способа хранения, затем контент, хранящийся по адресам, выделенным в эти группы, объединяют в документы путем реферирования.

11. Способ по п.1, в котором полезный контент из приложения выкачивают в виде набора документов, пригодных для дальнейшей машинной обработки.

12. Способ по п.1, в котором описания приложений формируют путем объединения текстового содержимого документов в единый текст, отражающий набор документов, содержащихся в приложении.

13. Способ по п.1, в котором установочный файл нового приложения загружают и повторяют все описанные последовательности до тех пор, пока в глобальных компьютерных сетях и маркетах имеются новые необработанные приложения.

RU 2014102136 A

RU 2014102136 A