O. Pyechkurova

LOCALITY-SENSITIVE HASHING AND ITS SCOPE

To date, there are many technologies for working with data of small sizes; they perform their tasks quickly and qualitatively. However, they are ineffective for large amounts of data, especially when it comes to finding similarities. To compare billions or even trillions of sets requires a method or technology that would focus on those pairs of sets that can be very similar to each other, while ignoring the vast majority of other pairs. To do this, a locality-sensitive hashing was invented, which is capable of pointing to such pairs without getting through the swamp of all other pairs.

The method does not examine the whole set of elements; rather, it considers the elements that are likely to be similar. That is, locality-sensitive hashing focuses its attention on similar pairs of elements (candidate pairs) without exploring each pair. However, if the goal is to study the similarity of each pair, then locality-sensitive hashing does not work for this.

Keywords: locality-sensitive hashing (LSH), hash function, shingle.

Матеріал надійшов 12.09.2017

УДК 004:002.1-028.27

Борозенний С. О.

ОСОБЛИВОСТІ ВИКОРИСТАННЯ ФОРМАТУ EPUB 3 ДЛЯ СТВОРЕННЯ ЕЛЕКТРОННИХ ПУБЛІКАЦІЙ

У статті викладено підходи до створення електронних публікацій технічних текстів у форматі epub3, наведено приклади створення елементів публікацій з використанням HTML5, MathML, крім того, розглянуто проблеми, що виникають при їх створенні, та шляхи вирішення.

Ключові слова: EPUB 2, EPUB 3, MathML, електронна публікація.

Вступ

EPUB 2 (англ. electronic publication) – відкритий стандарт формату електронних книг, що відповідає стандарту International Digital Publishing Forum (IDPF) [1].

EPUB 2 став офіційним стандартом IDPF у вересні 2007 року, замінивши старий стандарт Open eBook. Може скластися враження, що EPUB 2 використовується лише для публікації художньої літератури, але насправді EPUB є загальноприйнятим форматом електронного документа і може бути використаний для представлення багатьох видів публікацій: журналів, газет і т. п., тобто будь-який тип документа, який ви хочете розповсюдити в електронному вигляді, може бути представлений як EPUB 2.

EPUB 2 забезпечує всі можливості форматування та макетування HTML 4 та CSS 2, що цілком достатньо для текстових публікацій. Проте відомо, що EPUB 2 не найкраще рішення для створення мультимедійних книг, книг зі складним макетом, математичних публікацій та інтерактивних документів.

Інструментом, що дає змогу ефективно розв'язувати такі проблеми, є новий стандарт електронних публікацій EPUB 3.

EPUB 3

Згідно з визначенням [3], EPUB 3 – формат розповсюдження та обміну цифрових публікацій та документів. Формат EPUB 3 забезпечує можливість представлення, упакування та кодування структурованого та семантично вдосконаленого веб-вмісту (зокрема HTML, CSS, SVG та інших ресурсів) для розповсюдження в контейнері одним файлом.

Ключові технології, які було додано в EPUB 3:

- XHTML 5
  (для представлення текстового
  та мультимедійного вмісту, що включає
  підтримку MathML);
- SVG 1.1
  (для представлення графіки);
- CSS 2.1 та 3
  (для візуального відображення
  та відтворення вмісту документа);
- JavaScript
  (для інтерактивності та автоматизації);
- SSML / PLS / CSS 3 Speech
  (для впровадження озвучування тексту);
- SMIL 3
  (для синхронізації відтворення тексту
  та звуку);
- RDF словники
  (для вбудовування семантичної інформації
  про публікацію та зміст).

Важливою ознакою справжньої електронної книги, що відрізняє її від файлу, підготовленого для друку, є адаптивний дизайн. Електронна книга розбивається на сторінки прямо на пристрої для читання, залежно від розміру екрана і того, якого розміру шрифт вибрав читач. Для звичайної художньої книги проблема форматування тексту є не надто гострою. Є правила хорошого тону, яких хотілося б дотримуватися (відсутність висячих рядків, нерозривні пробіли і т. п.). Їх реалізація порівняно проста, а для багатьох читачів не надто й важлива. Інша річ – «складні тексти», наприклад, ті, що містять математичні формули.

У «класичному» ePub2 єдине вирішення «проблеми формул» – вставляти зображення. Це незручно для авторів, часто негарно виглядає, до того ж такі формули ніяк не реагують на зміну розміру шрифту. Класичним інструментом для роботи з математичними формулами є TeX/LaTeX. Але проблема в тому, що TeX потрібно компілювати, щоб отримати готовий текст. І цей текст експортується у формат PDF (або не більше зручних DVI і PostScript), що добре для друку, але не зовсім зручно для адаптивних електронних публікацій.

Одним зі способів вирішення проблеми є використання MathML. MathML – це низькорівнева специфікація для математичного та наукового контенту в електронному форматі на базі XML. Для візуалізації використовується таблиця стилів css.

Наприклад, ось таку математичну формулу

$$\boldsymbol{atan}(\boldsymbol{x}) = \int_0^{\boldsymbol{x}} \frac{\boldsymbol{dt}}{\boldsymbol{t}^2 + 1}$$

можна записати таким чином:

```
<math mode="display">
 <mi>a</mi>
 <mspace width="0"/>
 <mi>t</mi>
 <mspace width="0"/>
 <mi>a</mi>
 <mspace width="0"/>
 <mi>n</mi>
 <mo lspace="0" rspace="0" stretchy="false">(</mo>
 <mi>x</mi>
 <mo lspace="0" rspace="0.278em" stretchy="false">)</mo>
 <mo lspace="0" rspace="0.278em">=</mo>
 <msubsup>
  <mo lspace="0" rspace="0.167em" stretchy="false">∫</mo>
  <mn>0</mn>
  <mi>x</mi>
 </msubsup>
 <mfrac>
  <mrow>
   <mi>d</mi>
   <mspace width="0"/>
   <mi>t</mi>
  </mrow>
  <mrow>
   <msup>
    <mi>t</mi>
    <mn>2</mn>
   </msup>
   <mo lspace="0" rspace="0">+</mo>
   <mn>1</mn>
  </mrow>
 </mfrac>
</math>
```

Проблема полягає в тому, що на сьогодні візуальні редактори, що дають змогу створювати та зберігати математичні формули у форматі MathML, перебувають на стадії розробки, крім того, формат підтримує обмежена кількість програм для читання EPUB 3.

Одним зі способів розв'язання цієї проблеми може стати використання **MathJax. MathJax** – це JavaScript-рушій з відкритим кодом, що дає змогу візуалізувати тексти, написані мовами LaTeX, MathML та AsciiMath, та який підтримується всіма сучасними браузерами.

Висновки

У роботі розглянуто найпопулярніший на сьогодні формат для створення електронних публікацій EPUB 2 та його розширення EPUB 3. Розглянуто можливість використання формату EPUB 3 для створення повноцінних публікацій технічних текстів.

Список літератури

1. International Digital Publishing Forum [Електронний ресурс]. – Режим доступу: http://www.idpf.org.
2. The MathJax Consortium [Електронний ресурс]. – Режим доступу: https://www.mathjax.org.
3. EPUB 3 Community Group [Електронний ресурс]. – Режим доступу: https://www.w3.org/community/epub3/
4. Garrish M. What is EPUB 3? [Electronic resource] / Matt Garrish. – O'Reilly Media, Inc., 2011. – Mode of access: https://www.safaribooksonline.com/library/view/what-is-epub/9781449318710/. – Title from the screen.

S. Borozennyi

FEATURES OF USING EPUB 3 FORMAT FOR THE ESTABLISHMENT OF ELECTRONIC PUBLICATIONS

Two formats of electronic publications are considered in this paper: EPUB 2 (electronic publication) – an open standard for the format of electronic books, which corresponds to the International Digital Publishing Forum standard, and the format to be replaced by EPUB 2 – EPUB 3. EPUB 3 is a format for the distribution and sharing of digital publications and documents that enables presenting, packaging, and encoding structured and semantically enhanced web content, including HTML, CSS, SVG and other resources for distribution in a single file in a container. The EPUB 3 format has been substantially expanded compared to EPUB 2 and consists of a number of new technologies, including: SSML / PLS / CSS 3 Speech (for voice overwriting), SVG 1.1 (for graphic representation), XHTML5 (for presenting text and multimedia content that includes MathML support). Consequently, technical texts can be published as an adaptive design e-book that distinguishes it from a file that is prepared for printing.

Also, the paper describes approaches to the creation of electronic publications of technical texts in the format of EPUB 3 and examples of creation of elements of publications using HTML5, and presents MathML. Besides, it considers the problems that arise in their creation and the ways of solving the problems. The article describes MathML – a low-level specification for mathematical and scientific content in an XML-based electronic format. According to the author, this direction of creating and distributing scientific and mathematical texts has prospects for development in the direction of the distribution of these texts in the global network.

Also, the article outlines the issues that need to be solved; namely, at present, visual editors that allow you to create and save mathematical formulas in the MathML format are under development, besides, the format supports a limited number of EPUB 3 readers, and the ways to solve them. These problems, in particular, one of the ways to solve this problem may be to use MathJax. MathJax is an open source JavaScript engine that allows to visualize texts written in languages LaTeX, MathML and AsciiMath, which are supported by all modern browsers.

Keywords: EPUB 2, EPUB 3, MathML, electronic publication.

Матеріал надійшов 01.11.2017