Návrh softwarového projektu

Základní údaje

Zkratka: XBW
Název: Burrows-Wheelerova transformace pro XML nad abecedou slabik.
Vedoucí: Mgr. Jan Lánský (zizelevak@gmail.com)
Řešitelé: Stanislav Kovalčin (sorbac@gmail.com)
Radovan Šesták (radovan.sestak@gmail.com)
Petr Uzel (petr.uzel@centrum.cz)
Tomáš Urban (tomas.urban@gmail.com)
Mária Szabó (majka.sz@email.cz)
Pavol Kumičák (pavol.kumicak@gmail.com)

Popis

Cílem projektu je kompresní program XBW [3], který bude reálně použitelný pro kompresi metadat fulltextového systému EGOTHOR [2]. Tato metadata se uchovávají v souborech o velikosti cca 15 – 20 MB, každý soubor obsahuje 1000 webových stránek Program bude napsán v jazyce C a bude muset být přeložitelný jak pod Windows tak i pod Linuxem. Program bude primárně optimalizován na účinnost komprese, sekundárně na rychlost komprese a dekomprese při zachování rozumné paměťové náročnosti.

Program XBW bude založen na Burrows – Wheelerově transformaci (BWT) pracující nad abecedou slabik [1] a bude využívat toho, že vstupem jsou špatně formované XML [5] dokumenty.

Program XBW bude obsahovat parser, který převede dokument na posloupnost elementů (slabiky, XML značky a atributy). Zkomprimovaný slovník použitých elementů [4] musí být součástí výstupního souboru. Na posloupnost elementů je aplikována BWT a její výstup bude zpracován nejen klasicky pomocí MTF + RLE + kanonické Huffmanovo kódování, ale budou vyzkoušeny i jiné kompresní metody (např. LZW, LZ77, LZSS, PPM), které se v současné době pro kompresi výstupu BWT nepoužívají. Bude otestován i vliv nahrazení Huffmanova kódování za aritmetické kódování.

Protože program bude prakticky nasazen, musí být pro libovolný vstupní soubor stabilní a zaručit správnost komprese i dekomprese.

Literatura

1) Lánský, J., Žemlička, M.: Text Compression: Syllables. V: Richta, K., Snášel, V., Pokorný, J. (Eds.): DATESO 2005. ČVUT, Praha, 2005. ss. 32-45.

2) Galamboš, L.: EGOTHOR, http://www.egothor.org/

3) Leo Galamboš, Jan Lánský, Katsiaryna Chernik: Compression of Semistructured Documents. In: International Enformatika Conference IEC 2006, Enformatika, Transactions on Engineering, Computing and Technology, Volume 14, August 2006, pg. 222-227, ISBN 975-00803-3-5, ISSN 1305-5313

4) Jan Lánský, Michal Žemlička: Compression of a Dictionary. In: Snášel, V., Richta, K., and Pokorný, J.: Proceedings of the Dateso 2006 Annual International Workshop on DAtabases, TExts, Specifications and Objects. CEUR-WS, Vol. 176, pg. 11-20, ISBN 80-248-1025-5.

5) World Wide Web Consorcium: Extensive Markup Language XML, http://www.w3.org/XML/

======= doplněno projektovou komisí: ======================

DOBA: 9 měsíců