LaTeX INTERPRETER BASED ON DIFFERENT TYPES OF SYNTAX ANALYSIS

Petr Lebeda
Master Degree Programme (2), FIT BUT
E-mail: xlebed02@stud.fit.vutbr.cz

Supervised by: Alexander Meduna
E-mail: meduna@fit.vutbr.cz

ABSTRACT

The diploma thesis discusses the potential of interpretation of typographical language LaTeX and describes the structure of this language, its functions and their syntax. Also it analyses possibilities of LaTeX interpretation into HTML (HyperText Markup Language), in order to create typographically accurate publications, which could be viewed by common web browser. The solution concept and outlines of possible problems follows.

1 ÚVOD

Jako interpretaci LaTeXu použiji převod zdrojového souboru psaného v LaTeXu do HTML kódu. Hlavní myšlenkou je, aby i na webových stránkách existovaly typograficky korektní, úhledně psané dokumenty, aniž by musely být ve formátu PDF. Tato problematika je sice již poměrně dlouho řešena, bohužel ale většinou na platformě operačních systémů UNIX. Pro operační systém společnosti Microsoft je teoreticky využitelný projekt doc++, tento je ale stále ve fázi vývoje, více viz. [5].

2 LaTeX VS. HTML

Již rozdíly ve využití těchto dvou programovacích jazyků nutí hledat rozumný kompromis, tak aby bylo možné interpretovat LaTeX do HMTL kódu a přitom byl pokud možno identický s PDF souborem, vytvořeným z téhož LaTeXového zdrojového souboru. To s sebou nese nutná omezení velmi robustního nástroje, jakým LaTeX bezesporu je.

Matematické vzorce a obrázky přímo generované LaTeXem nelze snadno převést do HTML. Nezbývá tedy, než z LaTeXu tyto části kódu, kde je daná matematická rovnice, vyčlenit, vytvořit z nich obrázek a tento následně vložit do HTML kódu.

Pro konečné odsazení textu a dodržení definovaného formátu se nabízí použití kaskádových stylů (viz. [3]).

3 SYNTAKTICKÁ ANALÝZA

LaTeX je podobně jako HTML značkovací jazyk, kde ve velkém množství zdrojového textu jsou rozmístěny formátovací značky, podle kterých se dokument vysází. V zadání této práce je uvedeno, že má být použito více typů syntaktických analýz. Kvůli časové náročnosti, obtížnější implementaci a spornému zlepšení syntaktické analýzy bylo od většího počtu typů metod syntaktické analýzy upuštěno a bude použita jediná metoda.

Jako nejvhodnější způsob analýzy se jeví analýza shora-dolů (top-down) a sice analýza pomocí rekurzivního sestupu. Syntaktická struktura jazyka LaTeX umožňuje vícenásobné zanoření různých prostředí a tyto případy jsou z programátorského hlediska nejsnadněji implementovatelné právě pomocí syntaktické analýzy rekurzivním sestupem.

4 INTERPRETACE

Samotná interpretace LaTeXu bude probíhat překladem zdrojového kódu LaTeXu do HTML. Znaky, rovnice, obrázky (vytvářené prostředím picture) a tabulky budou vytvořeny v LaTeXu jako obrázky, které poté budou vloženy do zdrojového kódu HTML. U tabulek samozřejmě závisí na jejich složitosti, v první fázi vývoje ale nejspíše budou takto převáděny všechny. V dalších verzích již budou vytvářeny přímo v HTML.

Bude využit kaskádový styl webových stránek CSS. A to proto, aby bylo dodrženo co nejpřesnější formátování textu. Výsledný HMTL kód by měl být kompatibilní se standardem W3C ([4]), není to ale prioritní podmínka.

Zvláštní důraz bude věnován přesnému fungování křížových odkazů, kdy HTML prohlížeče umožňují pomocí odkazů rychlé přesuny v textu. Určitě tak bude fungovat obsah, seznamy tabulek, obrázků a křížové odkazy do seznamu literatury na jejich čítače.

5 NÁVRH ŘEŠENÍ

Samotný program provede syntaktickou analýzu LaTeXu a interpretuje jej do HTML kódu. Na internetu je dostatek volně šiřitelných textových editorů včetně jejich zdrojových kódů, které umožňují další úpravu dle přání programátora. Interpret by pak mohl být přímo součástí takového textového editoru.

Program bude pracovat následujícím způsobem:

1. Na začátku analýzy program otevře zdrojový soubor, ve formátu `název_souboru.tex`
2. Proběhne syntaktická analýza hlavičky LaTeXového souboru. V něm jsou uvedeny atributy definující vzhled výsledného dokumentu. Tyto atributy se uloží do speciálního objektu. Ten bude určovat základní formát celého výsledného HTML dokumentu, stejně jako u LaTeXu.
3. V samotném těle zdrojového soubotu LaTeXu probíhá syntaktická analýza a již analyzovaný text je zapisován podle definovaného formátování do HMTL souboru.
4. Pokud program narazí na začátek prostředí math, picture, tabular a dalších (více například v [2]), proběhne syntaktická analýza. Kód ale nebude interpretován, nýbrž převeden

do samostatného LaTeXového souboru, který se odešle obslužnému skriptu `textogif`. Ten ke svému fungování ale potřebuje další programy, více viz. [6]. Výsledný obrázek v gif formátu je vložen do HTML kódu. Vzorce, obrázky, tabulky a speciální znaky (například znaky řecké abecedy, matematické znaky) jsou tedy ve výsledku zobrazeny pomocí obrázků.

5. Po návratu z předchozích prostředí opět pokračuje klasická interpretace textu dokud není dosažen konec interpretovaného souboru.

Výsledný soubor bude uložen ve formátu HMTL na pevném disku uživatele. Nabízí se možnost ukládat jednotlivé stránky interpretovaného dokumentu jako samostatné HTML soubory, spíše se ale přikláním k tomu, aby byl celý dokument zobrazen jako jediná webová stránka.

6 ZÁVĚR

Interpretací LaTeXu se již několik projektů zabývalo. Většinou se ale jednalo o nástroje pro platformu UNIX. Pro operační systém Microsoft Windows nástroje tohoto typu chybí. Lze použít i nástroje původně určené pro UNIX, tyto ale vyžadují množství dalších programů, které v operačním systému Windows nejsou.

Mou snahou je vytvořit jednoduchý nástroj pro interpretaci LaTeXu, dostatečně efektivní pro publikování dokumentů ve formátu HTML. Nutnost využití skriptu textogif nehodnotím záporně, v dalších fázích vývoje programu můžu tento skript přepsat do programovacího jazyka C, v kterém chci celý program naprogramovat. Tím odpadne nutnost instalace Perlu.

REFERENCE

[1] Castro, E.: HTML, XHTML a CSS, Brno, ComputerPress 2007, ISBN 978-80-251-1531-2

[2] Rybička, J.: LaTeXpro začátečníky, Brno, Konvoj 2003, ISBN 80-7302-049-1

[3] Cascading Style Sheets [online]: dostupné na: http://www.w3.org/Style/CSS/

[4] World Wide Web Consortium [online]: dostupné na: http://www.w3c.org/

[5] DOC++ project [online]: dostupné na: http://docpp.sourceforge.net/

[6] Walker, J. [online]: skript `textogif.pl`, dostupný na: http://www.fourmilab.ch/webtools/textogif