APPLICATION FOR RETRIEVING AND PROCESSING TABLE DATA ON WEB SITES

Jiří Kalus
Bachelor Degree Programme (3), FIT BUT
E-mail: xkalus00@stud.fit.vutbr.cz

Supervised by: Jiří Koutný
E-mail: ikoutny@fit.vutbr.cz

ABSTRACT

The tables are very well-arranged and they are very frequently medium of information on the Internet. This paper describes a Web application which is focusing on the acquisition of these tables. This application provides the possibility to obtain the tables from some web address in periodical intervals and shows these tables clearly. The paper is mainly focused on the detection method of the tables in the document and on the single treatment of these documents. These documents can be written in different markup languages.

1. ÚVOD

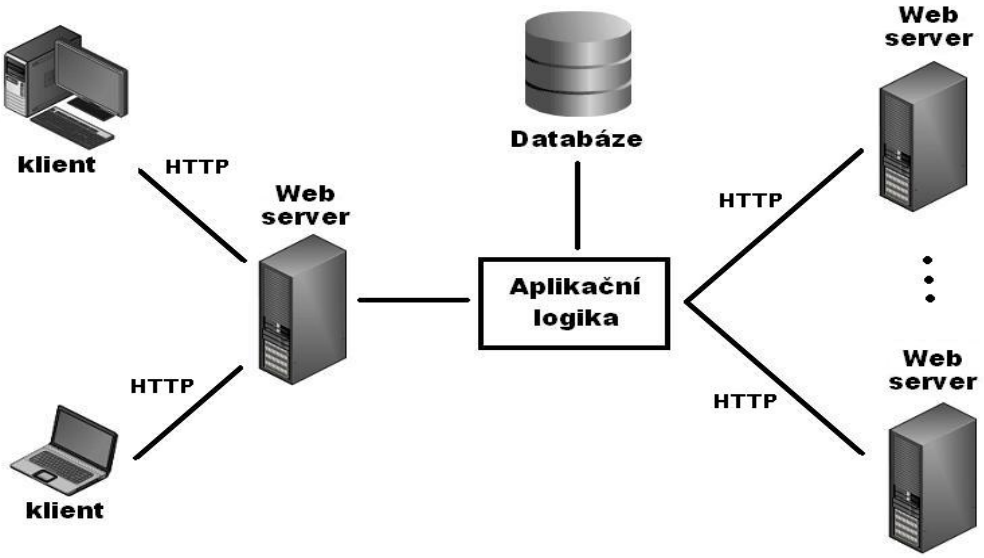
Informace jsou nedílnou součástí dnešního života. Jsou všudepřítomny. Setkat se s nimi můžeme v různé podobě, z různých zdrojů a na různých místech. Nás v tomto článku budou hlavně zajímat informace textové, konkrétně ve formě tabulek umístěných na internetu. Informace ve formě tabulek jsou často používané. Tabulky dokážou přehledně zobrazit údaje, je tedy snadné v nich číst. Mají však také velmi zásadní problém. Údaje se mohou rychle měnit. Je tedy pro uživatele dost obtížné neustále je sledovat a zaznamenávat její změny. V tom by měla pomoci aplikace, jejíž návrh je popsán v tomto dokumentu. Ta by měla udržovat tabulku neustále aktuální a zároveň zaznamenávat změny. Další funkce této aplikace jsou podrobněji popsány v rozboru níže.

2. ROZBOR

Aplikace pro získávání a zpracování dat získaných z webových stránek by měla pracovat jako klasický tabulkový procesor s menšími rozdíly. Tabulky nebudou zadávány ručně, ani nebudou načteny ze souboru. Budou získány z webové stránky zvolené uživatelem. Takto získané tabulky budou uchovávány a pravidelně, v zadaných intervalech, aktualizovány. Nad těmito tabulkami bude uživateli umožněno provádět změny. Například bude možno tabulky vizuálně přizpůsobit a pomocí statistik také vyhodnotit. Uživatel bude moci data seřadit nebo vyhledat. Takto upravenou tabulku si pak může vytisknout nebo nechat vyexportovat do souboru ve formátu CSV nebo do HTML kódu.

Jelikož tato aplikace bude data získávat z internetu, očekává se, že uživatel nebude mít problém s konektivitou. Proto je možné aplikaci navrhnout čistě jako webovou. Tento ná-

vrh přináší řadu výhod i nevýhod, nicméně výhody tohoto návrhu převládají. Aplikace se tak stává zcela multiplatformní a je uživateli přístupná odkudkoli, z libovolného zařízení s přístupem k internetu. Konkrétní návrh aplikace je znázorněn v modelu na obrázku 1.



Obrázek 1: Model aplikace

3. ROZPOZNÁNÍ TABULEK V (X)HTML DOKUMENTU

Veškeré tabulky v aplikaci budou získávány z dokumentů na internetu, zapsaných v jazyce HTML nebo XHTML. Tyto dva jazyky se téměř od sebe neliší, nicméně jazyk XHTML je, na rozdíl od HTML, jazyk striktně řídící se pravidly. Proto je pro strojové zpracování vhodnější. Právě z tohoto důvodu bude tento jazyk při zpracování preferován. Dokumenty zapsány v jazyce HTML budou nejprve převedeny a poté až zpracovány stejně jako dokumenty v XHTML.

Pro zpracování a nalezení všech tabulek v dokumentu byly vypracovány dva způsoby. V prvním případě byl vytvořen algoritmus využívající regulárních výrazů pro nalezení všech počátečních a koncových značek tabulky. Z těchto nalezených značek tento algoritmus vytváří korespondující páry a na základě jejich ofsetu určí, zda tabulka je nebo není vnořená. V druhém případě se detekce provádí zcela odlišně. Je využito parseru na bázi XML a zásobníkového automatu. Pomocí parseru jsou v dokumentu nalezeny veškeré značky, které jsou předány zásobníkovému automatu. Ten vymezí jednotlivé tabulky. Jelikož si zásobníkový automat udržuje kontext, není třeba řešit problém s vnořenými tabulkami. Pomocí něj totiž dokážeme určit, zda je tabulka vnořená a ke které tabulce patří. Nemusíme si tedy uchovávat žádné dodatečné informace, ze kterých bychom poté usoudili, zda je daná tabulka vnořená, či nikoli. Odpadá tak například složitá práce s ofsety ve zdrojovém kódu dokumentu, jak tomu bylo při prvním způsobu. Proto pro detekci tabulek je využito způsobu druhého.

4. IMPLEMENTACE

Při implementaci jednotlivých částí bylo využito prvků OOP a také návrhových vzorů.

4.1. KLIENT

Pro interakci na straně klienta bylo využito jazyka Javascript a knihovny jQuery, která umožňuje jednoduše vytvářet vysoce interaktivní webové stránky. Aby webová aplikace co nejvíce vypadala a chovala se jako desktopová, bylo využito komunikace prostřednictvím technologie AJAX. AJAX umožňuje klientovi komunikovat asynchronně, tzv. na pozadí. Nedochází tedy zbytečně k znovunačtení celé stránky ani k jejímu překreslení. Uživatel tak může s aplikací po provedení operace dále pracovat, aniž by čekal na odpověď. Celá aplikace se tak stává mnohem přívětivější a práce s ní je efektivnější.

4.2. SERVER

Implementace na straně serveru byla provedena prostřednictvím jazyka PHP v. 5.2.0. K tomuto jazyku bylo nutné připojit některá rozšíření, která nejsou jeho základní součástí. Mezi nejdůležitější jmenujme rozšíření TIDY. Toto rozšíření umožňuje převést dokument HTML do takzvaného formátu „well formed“, který se řídí stejnými pravidly jako jazyk XML. Díky tomuto rozšíření je tedy možné využít XML parseru na HTML dokument.

4.3. DATABÁZE

K uchování tabulek a konfigurace uživatele byla zvolena databáze MySQL. Ta poskytuje dostačující operace pro naši aplikaci a také umožňuje bezplatné nasazení do reálného provozu. Aplikaci je tedy možno provozovat s téměř nulovými náklady.

5. ZÁVĚR

V současné době jsem doposud nenarazil na žádný podobný systém, který by byl zcela nezávislý na konkrétním dokumentu a který by poskytoval navrženou funkčnost. Myslím si, že právě nezávislost systému na dokumentu je hlavní výhodou této aplikace. Z tohoto důvodu by aplikace mohla být v praxi velmi užitečná a často užívaná.

LITERATURA

[1] Gutmans A., Bakken S.S, Rethans D.: Mistrovství v PHP 5, Brno: Computer Press, 2007. Vyd. 1. 656 s. ISBN: 80-251-0799-X

[2] Holzner S: Mistrovství v Ajaxu, Brno: Computer Press, 2007. Vyd. 1. 592 s. ISBN: 978-80-251-1850-4

[3] Hauser, M.: HTML a CSS: velká kniha řešení. Brno: Computer Press, 2006. Vyd. 1. 912 s. ISBN 80-251-1117-2