FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2017

# Automatická extrakce příspěvků z diskusních fór

Jakub Sido[1]

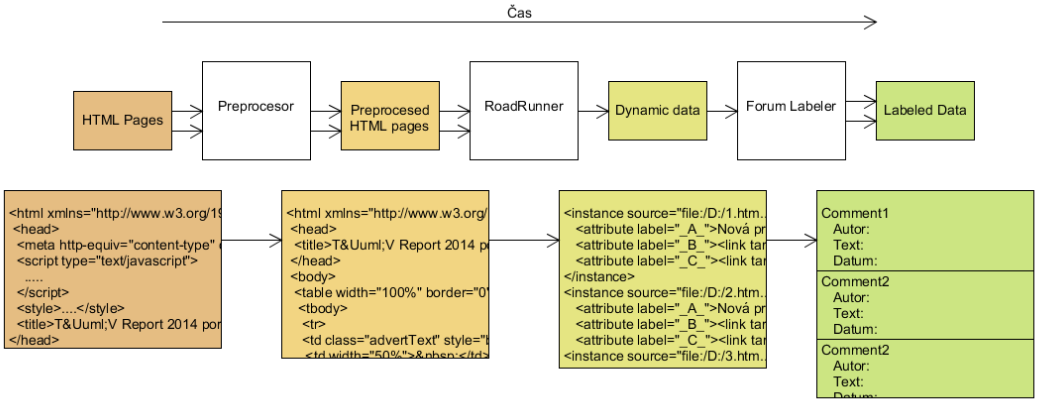
## 1 Úvod

Internet je velice rychle rostoucí médium. Stává se více žádané data na něm obsažená zpracovávat automaticky. Tato práce se zabývá extrakcí informací z webových zdrojů, především z webových diskuzních fór. Pojednává o tomto oboru a zkoumá existující systémy. Následně jsou tyto poznatky aplikovány a je navrhnut systém, který tento úkol plní bez zásahu člověka. Dále jsou použity metody strojového učení a analýzy přirozeného jazyka k označení významu získaných dat.

## 2 Hlavní aspekty realizace

Existují způsoby, jak vytvořit nástroj, který bude extrahovat žádaná data z konkrétních webových stránek. Je však vždy potřeba optimalizovat systém pro určitý zdroj. Cílem této práce je však vytvořit prostředek, kterým bude možné automaticky získávat data z velkého množství malých webových diskuzí.

Byl použit existující systém na extrakci dynamických dat z webových stránek a následně byla provedena analýza možností hledání významu těchto dat, které budou označeny pro pozdější použití.




**Obrázek 1:** Data Flow

[1] student navazujícího studijního programu Aplikované vědy a informatika, obor Inženýrská informatika – Softwarové inženýrství e-mail: sidoj@students.zcu.cz

## 3 Závěr

Bylo prozkoumáno několik systémů, které se věnují extrakci dat z webových stránek obecně, i těch, které se zabývají konkrétně webovými diskuzemi RR (2005) EE (2012). Tato práce kombinuje několik ověřených přístupů, avšak navrhuje a aplikuje v této oblasti nové postupy.

Byly použity statistické metody, strojové učení a analýza přirozeného jazyka na webové stránky obsahující zmíněná data. Také se objevilo několik nedostatků, které se týkají jednotlivých částí procesu. Především to byl problém s extrakcí dynamických dat pomocí šablony.

Tato data mohou být použita různými způsoby, od cílených reklam přes analýzu názorů po vyhledávání nevhodných činností ve virtuálním světě, jako je například šikana, zneužívání dětí nebo extrémistické chování.

## Literatura

Crescenzi, V. (2005) *Roadrunner: Towards automatic data extraction from large web sites.*. In VLDB, 1, s. 109 118, 2001.

Machová, K. Penzés, T. (2012) *Extraction of web discussion texts for opinion analysis.*.International Symposium on, s. 31 35. IEEE, 2012.