VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INTELLIGENT SYSTEMS

METODY DOLOVÁNÍ RELEVANTNÍCH DAT Z PROSTŘEDÍ WEBU S VYUŽITÍM SOCIÁLNÍCH SÍTÍ

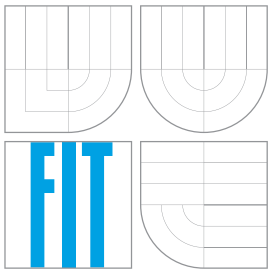
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. JAKUB SMOLÍK

BRNO 2013

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INTELLIGENT SYSTEMS

METODY DOLOVÁNÍ RELEVANTNÍCH DAT Z PROSTŘEDÍ WEBU S VYUŽITÍM SOCIÁLNÍCH SÍTÍ

DATAMINING OF RELEVENAT INFORMATION FROM WWW WITH USING SOCIAL NETWORKS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. JAKUB SMOLÍK

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. JAN SAMEK, Ph.D.

BRNO 2013

Abstrakt

Tato práce se zabývá řešením problémů spojených s hledáním relevantních informací v prostředí internetu. V textu představujeme možné východisko ve formě aplikace, která bude umožňovat s využitím automatizovaného zpracování a agregace dat z webového prostředí přehledně zobrazovat relevantní informace vzhledem ke hledaným klíčovým slovům. Za tímto účelem byly prostudovány a popsány možnosti automatické extrakce dat z tří vybraných datových formátů používaných pro přenos dat na internetu. Zároveň jsme se zaměřili na možnosti dolování dat ze sociálních sítí. Výsledkem je popis návrhu, implementace, realizace a testování vytvořené aplikace umožnující snadné hledání, zobrazování a přistupování ke hledaným informacím.

Abstract

This thesis focuses on solving problems related to searching of relevant data on the internet. In text is presented possible solution in form of application capable of automated extraction and aggregation of data from web and their presentation, based on input key words. For this purpouse there were studied and discribed possibilities of automated extraction from three chosen data types, mainly used as data storages on the internet. Furthermore it focuses on ways of data mining from social networks. As a result it pressents planning, implementation, realization and testing of created appliation which can easily find, display and let user easy access searched informations.

Klíčová slova

Web mining, data mining, sociální sítě, automatická extrakce dat, získávání dat, datové zdroje, analýza dat, html, xml, rss, nette.

Keywords

Web mining, data mining, social networks, automated data extraction, data acquisition, data sources, data analysis, html, xml, rss, nette.

Citace

Jakub Smolík: Metody dolování relevantních dat z prostředí webu s využitím sociálních sítí, diplomová práce, Brno, FIT VUT v Brně, 2013

Metody dolování relevantních dat z prostředí webu s využitím sociálních sítí

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením Ing. Jana Samka, Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Jakub Smolík
21. května 2013

Poděkování

Především bych chtěl poděkovat vedoucímu mojí práce, panu doktoru Samkovi, za jeho aktivní pomoc a čas strávený při konzultacích této práce. Velké díky také patří mé rodině a přátelům za vytrvalou podporu během celého mého studia.

Obsah

Kapitola 1

Úvod

Hlad po informacích lidstvo provázel celým jeho vývojem. Podle vědců je právě tato naše vlastnost klíčová pro přizpůsobení a přežití, ať už z hlediska krátkodobého či evolučního. U žádného jiného organismu nedosahuje potřeba hledání, shromažďování, sdílení a konzumace informací takového stupně, jako právě u lidí. Ale v dnešní době, kdy už nebojujeme o přežití druhu, se snažíme maximalizovat zisk ze získaných informací. Informace totiž také mají svoji cenu, čas a pozornost, jež jim musíme věnovat. A právě tyto dvě komodity se stávají čím dál cennějšími, částečně i díky tomu, že dostupných informací kolem nás neustále přibývá. Často se tak setkáváme s trendem dnešní doby, jímž je konzumace informací bez jasně stanoveného cíle. Příklady lze vidět dnes a denně, když se ztrácí hodiny a hodiny drahocenného času při bezcílném brouzdání po internetu. Mezi nejčastější zloděje času patří zábavné obrázky, videa a články. Přitom častokrát zapínáme internetový prohlížeč s konkrétním cílem, ale vše možné reklamy a doporučená videa nás od původního cíle odvádí hlouběji a hlouběji do temných zákoutí internetu.

V počátečních kapitolách představujeme pojmy spojené s dolováním dat na internetu a zkoumáme možnosti získávání datových zdrojů z internetu. Kapitola 3 je zaměřena na možnosti získávání dat ze sociálních sítí. V kapitole 4 rozebíráme vybrané datové formáty zdrojů na internetu a možnosti jejich automatického zpracování. Všechny výše uvedené kapitoly tvoří teoretický základ k vytvoření aplikace, jež agreguje zdroje diskutované v kapitole 5. K aplikaci se rovněž váží všechny následující kapitoly, přičemž kapitola 6 obsahuje návrh aplikace, 7. použité implementační postupy, 8. testování, zhodnocení a možná rozšíření a poslední 9. kapitola obsahuje závěr práce.

Kapitola 2

Web mining

Web mining znamená v češtině dolování informací z webů. V praxi se však spíše používá anglická varianta tohoto termínu, proto se jí v následujícím textu přidržíme. Web mining posouvá prostředí world wide web (dále jen www) směrem k použitelnější a dostupnější variantě, kdy mohou uživatelé rychle a snadno nacházet informace, které potřebují. Zvyšuje tak aktivně efektivnost při jejich získávání, protože zvyšuje relevantnost dat, stejně jako snižuje dobu potřebnou pro jejich nalezení. Tyto postupy v sobě zahrnují jak sběr, tak jejich následnou analýzu, stejně jako analýzu různých typů dokumentů a multimédií, dostupných z www. Hlavním cílem je tedy získat data, která lze posléze analyzovat, například pomocí technik data miningu. Data mining je obecné označení analytické metodologie získávání netriviálních, skrytých nebo potenciálně užitečných informací z dat. Hledané informace se mohou týkat jak obsahu dokumentu, tak jeho struktury, nebo přidružených statistik [13].

Web mining lze rozčlenit do tří fází, které na sebe postupně navazují. První fázi web miningu tvoří sběr dat v podobě procházení webových stránek s využitím klíčových slov. Druhá fáze je hledání odkazů na podobné odkazy nebo zdroje, z kterých bylo čerpáno. Tím se vytváří propojená struktura, která přidává vyhledávání hloubku. Třetí fází představuje vyhledání a analýza logů předchozích hledání nebo přístupů. Další analyticky využitelné hledisko, na které se můžeme zaměřit, je porozumění uživateli. Sem patří obecně jeho návyky a chování na internetu, které lze určit podle různých faktorů, jako jsou historie hledání, preference, reakce na vrácené výsledky, a další [13].

Web mining je proto z podstaty více úrovňová činnost, která zasahuje mezi několik vědních oborů. Využívají se techniky zpracování přirozeného jazyka, extrakce informací, strojového učení, databází, získávání a skladování dat, nebo vizualizace. V praxi jsou pak tyto postupy aplikovatelné v e-commerce, m-commerce, e-goverment, e-learning, managementu znalostí a digitálních knihovnách. V následném textu si právě tyto termíny blíže objasníme.

2.1 Pojmy z oblasti využití web miningu

- **M-commerce (Mobile commerce).**
  Tento pojem vznikl současně se začátkem konsorcia Global Mobile Commerce Forum (Londýn, 1997), jež se v současnosti sestává z více než 100 firem. Význam pak souvisí s nakupováním nebo placením přes mobilní telefon, ať už prostřednictvím placených SMS zpráv, stahováním placených tónu vyzvánění, kupování lístků na vlak nebo letenek přes mobil. Obecně se jedná o služby pro mobilní zařízení, které produkují

zisk za poskytnuté služby. S příchodem chytrých telefonů se z velké části upustilo od systému SMS, kvůli jeho nízké zabezpečenosti a přestoupilo se na využívání internetových aplikací pro mobilní zařízení, která tyto funkce podporují. M-commerce dnes tvoří most mezi nakupováním přes internet (e-commerce), mobilním zařízením a nákupem v kamenném obchodě. Oproti fyzickému nákupu sféra m-commerce přináší různé rozšiřující funkce, jako například přímý přístup k recenzím produktů, srovnání cen, atd. [14]

- **E-commerce (Electronic commerce).**
  Je celosvětové používaná zkratka pro nákup zboží s využitím elektronických zařízení a počítačové sítě. Tento postup v sobě zahrnuje spoustu pokročilých technologií, jako jsou elektronický přesun financí, řetězec dodavatelů zboží, internetový marketing, online zpracování transakcí, elektronická výměna dat nebo systémy pro automatický sběr dat. Moderní e-commerce využívá ve většině případů internet, i když je schopen využívat i technologie jako email, mobilní zařízení a telefony. Jeho vývoj započal už v roce 1979 [14].

- **E-government (Electronic government).**
  Značí pojem pro modernizaci vládních institucí, zavedením informačních technologií. Ty jsou následně využity pro doručení informací od vlády občanům příslušné země. Informační technologie sebou přinášejí zvýšení efektivity komunikace a šíření informací od oficiálních představitelů do běžných domácností. Kromě této činnosti slouží i k reprezentaci obyvatel, volební kampaně, veřejné diskuze s obyvateli nebo jejich začlenění do správního procesu.

- **E-learning (Electronic learning).**
  Jedná se o současný pojem, představující využití informačních technologií pro podporu učení a vzdělávání. Tento trend se stále více prosazuje ve školách jako jeden z nástrojů výuky.

Oblasti využití web miningu lze vskutku nalézt v mnoha společenských i vládních strukturách. Vzhledem k uvedení čtenáře do kontextu předložíme definice základních pojmů z oblasti web miningu, se kterými pracujeme, jako jsou data, informace, apod.

2.2 Obecné pojmy a metody využívané při

- **Data.**
  Diskrétní fakta reprezentována v nějaké symbolické podobě, která jsou zaznamenána v rozličných formátech. Jsou vyjádřena fyzickým nosičem. Jako taková mají data pouze malý význam nebo využití, i když mají vypovídající schopnost [9].

- **Informace.**
  Data použitá a interpretovaná osobou. Takováto data jsou relevantní a mají význam, který ovlivňuje nebo mění rozhodovací proces u osoby, jež je využila [9].

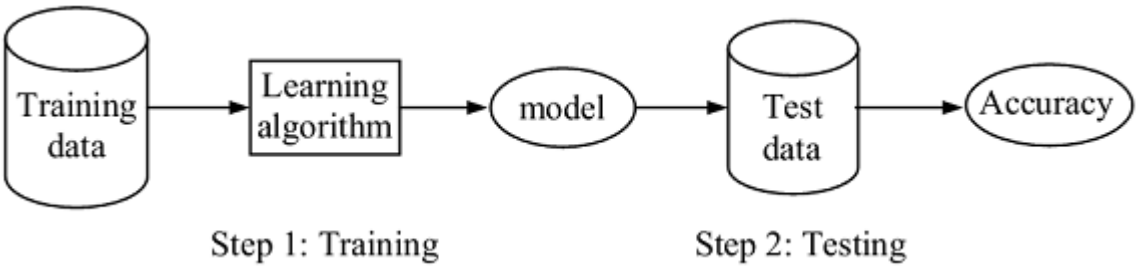
- **Znalosti.**
  Vlastní nebo skupinová zkušenost, hodnota poznání získaná aplikací informace. Použitelná informace je taková, která je relevantní a dostupná ve správný čas, na správné místě a v kontextu osoby, jež ji použije v jejím rozhodovacím procesu. Přispívá k procesu učení, kdy se systém dokáže rozhodovat autonomně bez potřeby dalších informací [9].

- **Proces rozhodování (Decision making).**
  Na základě různých kritérií se využívají různá data s cílem vybrat ideální strategii pro dosažení požadovaných cílů. Toto se jeví jako kritický požadavek pro společnost, jejíž prvořadým cílem je provádět včasná, přesná a zásadní rozhodnutí, na základě ověřených faktů. Jedná se komplikovanou oblast vědeckého zkoumání lidského chování, kde figuruje velké množství činitelů ovlivňujících procesu rozhodování. Pro účely práce není potřeba zabíhat do větších detailů.

- **Metadata.**
  Metadata, čili data o datech, představují doprovodné informace, jejichž potřeba neustále vzrůstá. Jejich hlavní funkcí je definice obsahu nezbytná pro odhalování relevantních dat. Jako taková jsou metadata naše znalosti o datech, která jsme interpretovali jako informaci v konkrétní rozhodovací situaci.

- **Associacion rule mining (Dolování na principu asociačních pravidel).**
  Technika využívaná při hledání zajímavých vzorů, vztahů a příbuzností mezi různými množinami prvků. V prostředí webových stránek to mohou být například časté návštěvy mezi dvěma servery, jejichž obsah spolu přímo nesouvisí. Takto lze odhalit dříve skryté vztahy a skupiny lidí, které využívají podobný vzorec, což je přímo použitelná informace pro oblasti e-commerce [2].

- **Klasifikace.**
  Každý záznam obsahuje několik odlišných atributů, kde jedním z jich je právě jeho třída (anglicky class). Klasifikace tedy představuje proces řazení záznamů do jedné z několika předdefinovaných tříd. Klasifikace pomáhá například vyhledávačům při hledání relevantních výsledků na základě zadaných dotazů. Problém klasifikace se dá vztáhnout do oblasti umělé inteligence, konkrétně učení s učitelem, kde se snažíme inteligentní prvek (klasifikátor) naučit pomocí testovací množiny správně řadit vstupní záznamy do příslušných tříd. Následně pak necháme tento prvek automaticky provádět naučenou činnost na dalších vstupních záznamech. Pro řazení se využívají různé metody, ať už na základě rozhodovacích stromů, předdefinovaných pravidel, paměti, neuronových sítí, nebo jiných. Parametry, na základě kterých třídění funguje, se mohou vztahovat také ke konkrétním informaci o stránce. Příkladem může být téma stránek, role stránek, názor autora na konkrétní věc, klasifikace spamu, atd. [2].



Obrázek 2.1: Příklad procesu klasifikace dat [2]

.

- **Shlukování (Clustering).**

Metoda používaná pro setřiďování prvků s podobnou charakteristikou. V našem případě to mohou být například uživatelé nebo webové stránky.

2.3 Problémy spojené s web mining

I přes veškeré snažení se i v prostředí www setkáváme s překážkami na naší cestě za získáváním relevantních informací. Ty tvoří například:

- Problémy šíření redundatnich dat s různými metadaty.
- Malá orientace na konzistenci metadat mezi různými zdroji.
- Chybí softwarová podpora pro výše zmíněné.

Z těchto důvodů vznikaly a stále vznikají různé frameworky a postupy s cílem zvýšit konzistenci a určit hodnotu metadat v závislosti na konkrétních datech (pro správu dat, informací, organizaci znalostí, hodnotící standardy pro redundanci a kvalitu webové obsahu, nástroje a standardy pro hodnocení sémantické integrity webu, vývoj chytrého softwaru schopného nový obsah začlenit do znalostní báze pod správná metadata) [13].

Problémy centralizovaných vyhledávacích systémů jako jsou Google nebo Yahoo! vidíme v tom, že internet je příliš velký (navíc se rozrůstá) a příliš rychle se mění, než aby bylo možné ho zmapovat a udržovat o něm aktuální databázi. Intenet jako takový je navíc systém decentralizovaný, což z něj činí entitu špatně postihnutelnou jakoukoliv centralizovanou formou. Často tak dochází k situaci, kdy i přes snahu poskytnout uživateli co nejpřesnější informace na základě jejich vyhledávacího dotazu, nemají příliš velkou úspěšnost. Výsledků se mohou vyskytovat desítky i stovky tisíc, což činí z hledaní konkrétních informací složitou operaci. Odpovědí na tento problém mohou být konkrétní, tématicky zaměřené, vyhledávače (pro vědu, literaturu, cestování, linux, ...), ale na ty se lze spolehnout pouze tehdy, pokud uživatel ví přesně, do které kategorie jeho dotaz spadá. To sebou nese další úskalí, jako je obtížná automatizace postupu pro správné řazení indexovaných stránek do kategorií. To vše dohromady přesto nezaručuje požadovanou kvalitu informací, protože i stránky z jiné kategorie mohou obsahovat odkazy na velmi relevantní a specializované zdroje, jež mohou být hledaným cílem [13].

Kapitola 3

Data mining v sociálních sítích

Sociální sítě jsou již nějaký ten rok fenoménem, který nemá v prostředí internetu obdoby. Milióny lidí každý den vyplňují osobní informace do prostoru veřejného internetu, kde si je může prohlížet kterýkoliv zaregistrovaný uživatel. Oblíbenému zaškrtávacímu políčku „Četl jsem licenční podmínky a souhlasím s nimi“ přitom málokdo věnuje pozornost. Kdyby tomu tak nebylo, pak by si možná spousta uživatelů dobře rozmyslela, zda chce předat svoje osobní informace do rukou *third party* společností. Za poslední roky se možnosti získávání informací v mnohém vyvinuly do vysoké úrovně jak v pokrytí, tak v sofistikovanosti [10].

Tento fenomén vychází z lidské povahy, konkrétně z lidské pýchy, kdy se snažíme pochlubit se před světem vlastními úspěchy. Doveden do extrémů, může tento jev vést k případům, kdy se lidé neváhají svěřovat v sociálních sítích s celým svým životem, čehož se dá ze strany obchodních subjektů snadno využít. Informace jsou v současnosti jednou z nejžádanějších komodit na trhu a prostředí sociálních sítí je jejich stále se zvětšující studnicí.

Mohlo by se zdát, že cena informací klesá, protože se dají poměrně snadno najít po několika kliknutích myší. Dokonce by se mohlo jevit, že do budoucna budou informace kompletně zdarma dostupné pro všechny. Co se však učí každý se základními marketingovými vzděláním je, že šíření zdrojů zdarma, obzvláště pokud je po nich poptávka, musí vyústit do nekontrolovatelné situace. Proto i zde je potřeba regulace a kontrola, která spočívá v přístupu ke kanálům nezbytným pro šíření a extrakci informací. A právě cena těchto kanálů zvyšuje i cenu informací [10].

Když nepočítáme cenu jako konkrétní peněžní obnos, pak lze najít i jiné ceny, které si nalezení požadované informace účtuje. Jedním z nich je „payment of kind“ [8], kdy jsou vyhledávače a hostující weby posety reklamami na nejrůznější zboží, slevy a akce. Jedná se metodu jak podstrčit druhé osobě prostřednictvím reklamy informace, které nevyžaduje, ale které jsou její platbou za poskytnutí služby. Marketingová oddělení navíc mají snahu cílit reklamu na konkrétní subjekty, které z analýzy sociálních dat vykazují velkou pravděpodobnost na přechod mezi různými značkami (anglicky brands). Spousta průmyslových odvětví využívá takzvané věrnostní bonusy, které zákazník dostává za delší setrvání u výrobků nebo služeb daného výrobce. Cílem pak je zmapovat a zahrnout takovéto subjekty do svých věrnostních programů, aby měli ještě větší šanci na dlouhodobé setrvání daného zákazníka u své značky. Druhou přímou aplikací představuje možnost vytvoření virální reklamy. Pokud se nám podaří identifikovat v dané sociální síti nejvlivnější subjekty, můžeme na ně s úspěchem aplikovat marketingovou strategii, která se poté jako virus může rozšířit po velké oblasti sítě, aniž by si vyžádala vysoké náklady na její vytvoření nebo udržování. Funguje zde jev takzvaného „word of mouth“ [6], kdy se šíří povědomí o konkrétním produktu mezi uživateli na základě jejich ohlasů. Tato strategie se s úspěchem využívá v řadě

marketingových oddělení po celém světě [3].

Že se jedná o důležitou oblast pro dolování informací, reflektuje i zájem ze strany vládních složek velmocí, jako je USA. Vláda prezidenta Obamy využívá průzkum dat z Facebookových stránek své strany, aby mohla lépe posoudit preference voličů a jejich ohlasy na současné dění [10].

3.1 Monetizace dat

Monetizací rozumíme proces převedení položky negenerující profit na položku, která profit generuje. V podstatě se jedná o operaci, která přetvoří jinak ekonomicky nezajímavou položku na produkt, jež bude mít svoje místo na trhu, stejně jako vlastní cenu, za kterou se najdou kupci ochotní danou věc koupit. V našem případě jde o formu monetizace, která zahrnuje maximalizaci potencionálního zisku z konkrétních dat tím, že zařídí jejich záznam, uložení, analýzu a případnou aplikaci ve zvoleném kontextu. Kromě toho se může tento proces zaměřit i na vylepšení uživatelského zážitku (anglicky user experience), nebo získávání stálých uživatelů. Souhrnně se tedy snaží vytvořit alternativní kanály přijmu z jednoho zdroje. Data, jež vlastníme a třeba nám někdy pomohla k obchodu a zisku, každá jednotlivá informace, mohou být využita ke snížení našich nákladů a zvýšení přijmu. Lze ji konkretizovat, rozebrat, zařadit a prodat dál. Existují totiž lidé, kteří naše data potřebují a my jim je můžeme se ziskem prodat.

Výhody plynoucí z takto upravených dat mohou představovat:

- Lepší rozhodování, které vede k vyšším ziskům, sníženým nákladům a menšímu riziku.
- Častější tvorbě rozhodnutí, například místo měsíčních se zvýší na týdenní atd.
- Včasnějším rozhodnutím, kdy se zvýší šance na provedení nejvýhodnějších obchodů a rozhodnutí.

Jak je vidět, tak sběr a analýza dat může mít další využití, jež nepřímo plynou z jejich vlastnictví. Tím dostáváme další ukázku důležitosti a výhodnosti data miningu v praxi.

3.2 Získávání zdrojů ze sociálních sítí

Abychom mohli analyzovat data ze sociálních sítí, je nejdříve potřeba je nějak získat. Z tohoto hlediska lze použít několik různých postupů, z nichž si ty nejpoužívanější stručně představíme. Příkladem internetové aplikace, jež uchovává z našeho pohledu zajímavá data, může být sociální síť Facebook. Uživatelé zde na svém profilu vytvářejí a sdílejí obsah, který mimo jiné poskytuje i spoustu dalších zajímavých sémantických informací, jako jsou jejich aktivita, členství ve skupinách nebo oblíbené produkty. My bychom tyto informace rádi extrahovali ve velkém měřítku, na což se zaměřuje i početné množství vědeckých článků. Facebook stejně jako podobné sociální sítě neposkytuje prostředky pro automatický přístup k uživatelským informacím z veřejně dostupných účtů, alespoň ne ve zdarma dostupné variantě. Děje se tak díky skrývání soukromí svých uživatelů pro komerční zájmy, což nakonec tvoří hlavní příjmy dané služby – poskytování osobních informací obchodním subjektům (samozřejmě s dodržením zákonných stanov o osobních informacích). Z těchto důvodů je tedy extrakce dat obtížná, ale přesto proveditelná [13].
V praxi se využívají 3 základní přístupy pro získávání dat ze sociálních sítí:

- Analýza provozu na síti.
- Ad-hoc aplikace.
- Procházení (anglicky crawling) uživatelských grafů.

3.2.1 Analýza provozu na síti

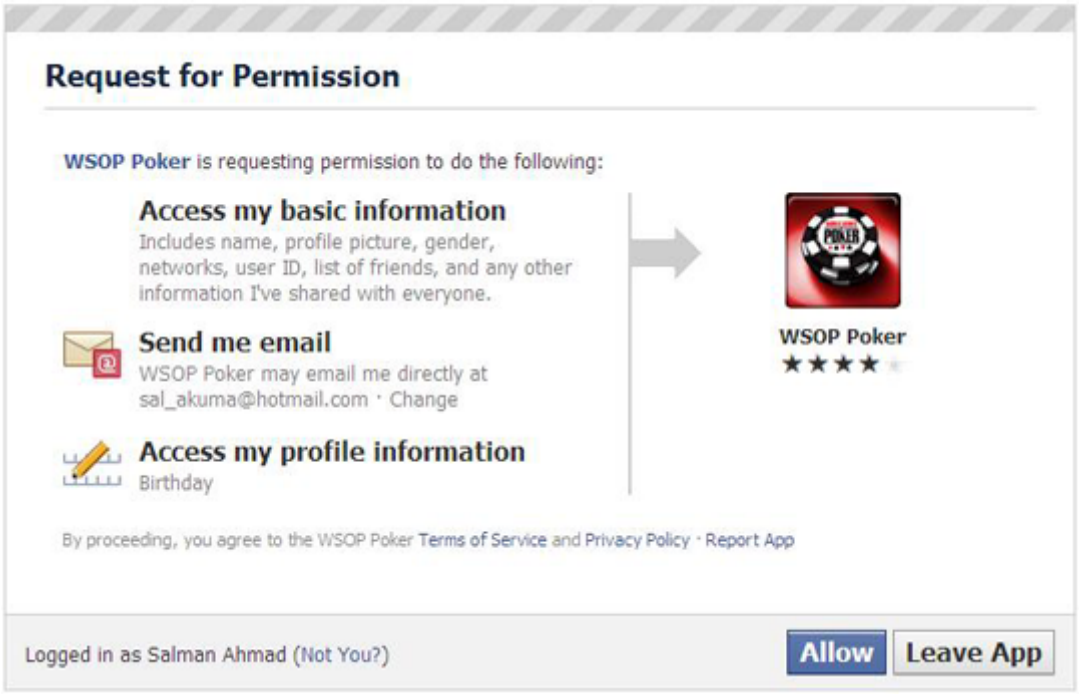
Jedná se o klasickou metodu sledování příchozích a odchozích packetů ze síťového toku (network stream). Konkrétně se zajímáme o TCP žádosti a příslušné odpovědi, které nám poskytují představu o komunikaci na síti. Odtud lze odvodit informace o procházení webového obsahu, akcích na sociálních sítích a obsahů stránek ze sekce payload příslušných paketů. Z teoretického hlediska tak tato metoda poskytuje použitelný a přesný prostředek pro získávání dat vhodných k analýze, avšak sledování provozu na síti sebou přináší technická i legální omezení [13]. Ty tvoří následující trojice:

- Všechny země aplikují omezení na sledování a analýzu síťového provozu, především z důvodů ochrany osobních údajů svých obyvatel. I když lze v omezeném měřítku analýzu přesto aplikovat, většinou je pak zapotřebí uzavřené prostředí, kde budou zároveň seznámeni s účastí na experimentu, což nemusí být pro obecnou analýzu žádoucí.
- Technická omezení, týkající se především velkého množství paketů, které mohou na vysokorychlostních linkách přetížit mechanismus na sledování paketů. Po aplikaci vzorkování nám naopak data způsobují nekonzistentnost díky absenci příslušných párů žádost/odpověď.
- Data obsažená v sekci payload jsou obvykle ve formátu HTML, který je sám o sobě nekonzistentní. Je proto problém použít unifikovaný prostředek pro syntaktickou analýzu (dále parsování) obsahu, jenž by nám umožnil jeho automatické zpracování. I když většina sociálních sítí interně využívá jednotnou formu, tak mezi různými sociálními sítěmi se bude struktura lišit.

3.2.2 Ad-hoc aplikace

Jedná se o aplikace třetí strany (third-party applications), které využívají sadu API (např. Facebook Developer Platform, OpenSocial, ...), aby dovolily vytvářet nové služby a hry uživatelům sociální sítě. V této architektuře tak uživatel nepřistupuje přímo k aplikačnímu serveru, ale k vrstvě propojující uživatele a aplikaci. Vývoj ad-hoc aplikací za účelem získání dat o uživatelích sociální sítě dovoluje sbírat informace dvěma kanály. Prvním je přístup k profilovým informacím o uživatelích, kteří se do aplikace zaregistrovali. Druhým je pak analýza logů, která poskytuje informace o jejich chování v rámci aplikace. Z důvodů nutné registrace uživatelů, rozšířená o potvrzení možného využití jejich osobních údajů pro statistické účely, pak řeší legální problém se sledováním a analýzou osobních informací. [13]

Omezení tohoto přístupu pak tvoří nutnost existence sociální sítě, která podporuje aplikační rozhraní (API) pro aplikace třetí strany. Navíc musí být sledovaná aplikace dostatečně populární, aby nashromáždila co největší množství registrovaných uživatelů, jejichž data budou pro analýzu využita. To celé pak dává dohromady otázku na téma rozsáhlosti investice, kterou je nutno vynaložit, a zda se takový podnik vyplatí, protože výsledky založené na těžko předvídatelných preferencích uživatelů bývají při nejmenším nejisté.



Obrázek 3.1: Příklad výzvy s potvrzením o sdílení osobních údajů aplikací třetí strany.

3.2.3 Crawling

Jedná se nejpopulárnější a nejpoužívanější metodu pro získávání informací ze sociálních sítí. Využívají se přitom veřejně dostupné informace, které nejsou omezovány legálními záležitostmi. Navíc se dá tento postup využít na většinu existujících sociálních sítí (Twitter, Flickr, YouTube, ...). V principu se pak vytváří graf, popisující strukturu sociálně sítě na základě uživatelských vztahů. Celý systém funguje iterativně, kdy jsou nový uživatelé přidáváni v každém dalším kroku. Existuje celá řada vědeckých článků s popisem algoritmů, jak daný graf vytvořit. Často se využívá algoritmů pro průzkum stavového prostoru, proto jsou nejoblíbenější přístupy například BFS (breadth-first search), DFS (depth-first search), atd. [13]

Nevýhodami tohoto přístupu jsou pak nutnost velkého výpočetního výkonu, který bude provádět průzkum stavového prostoru, stejně jako paměťová náročnost, kde obsáhlé sociální sítě s miliony uživatelů budou narážet na technická omezení. Průzkum celé struktury sociální sítě může trvat i měsíce, což nedovoluje udržovat průběžně aktualizovanou verzi celého prostoru. Navíc sociální sítě používají prostředky pro boj s hloubkovými crawlery (roboty vykonávající indexační činnost), jako je banování IP adres, omezení na počet dat ve výsledcích apod. Tyto prostředky jsou využívány proti útokům typu odmítnutí služby (denial of service), jež může činnost crawleru v mnohém připomínat.

Kapitola 4

Formáty dat na webu

V rámci hledání nejvhodnějšího formátu, který by bude sloužit jako datový zdroj pro následnou implementaci naší aplikace, jsme ze zaměřili na tři nejpoužívanější datové formáty, jež se v prostředí internetu vyskytují. Těmi jsou XML, HTML a RSS. Každý z formátů byl prostudován, přičemž jsme se zaměřili na rozšířenost a dostupnost datových zdrojů, stejně jako náročnost automatického zpracování souborů z odlišných zdrojů.

Pojďme si tedy každý z uvedených formátů blíže rozebrat.

4.1 XML

4.1.1 Proč právě XML?

Jak se množství informací na internetu zvyšuje, tak se rozrůstá i potřeba po unifikovaném formátu, který budeme moci použít pro jejich uchovávání, analýzu a dolování dat. Historicky se nejvíce využívaly relační databáze, avšak internet jako médium pro přenos analytických dat upřednostňuje právě formát XML. XML formát může vystupovat jako data centrický, kdy čímž zdůrazňujeme fakt, že informace v dokumentu přenášené budou využity právě pro automatické zpracování strojem. XML se využívá proto, že dokáže snadno uchovávat strukturovaná data, což však neznamená že by byl jediným používaným formátem pro tyto účely [11].

XML je otevřený formát, což sebou nese značné výhody, ale i četné nevýhody. XML využívá vlastní definované značky (tagy) pro popis dat a jejich vztahů v rámci dokumentu. XML značky tak popisují sémantický a strukturní význam informací v textu dokumentu, což tvoří XML dokumenty polo-uspořádané a sebe-popisující. XML se již dříve stal standardem pro prezentování a výměnu dat [11].

4.1.2 Pojmy spojené s XML

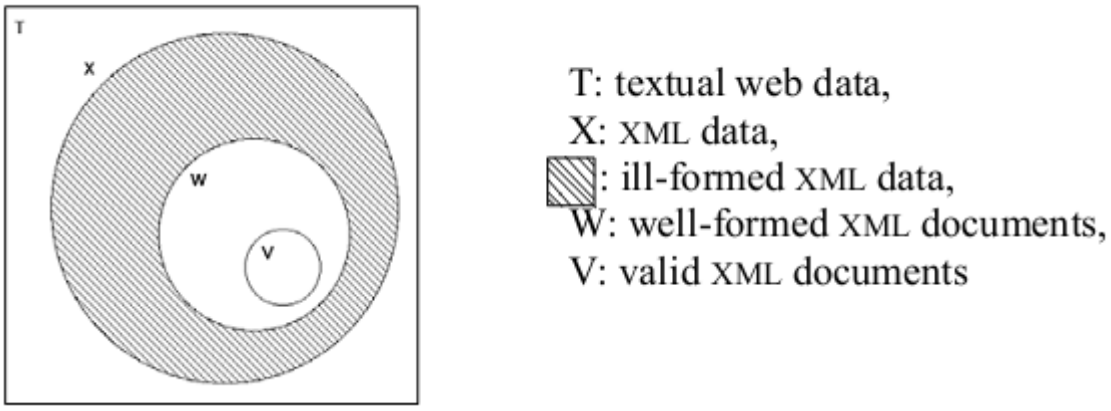
- **Shlukování XML dokumentů (Clustering).**
  Seskupení dokumentů, u kterých se vyskytují podobnosti. Ty mohou mít jak obsahový tak strukturní charakter. Pro analýzu a členění existuje velké množství algoritmů a dělí se podle přístupu na strukturní nebo datové.

- **XML dokument.**
  Jedná se o instance příslušného XML schématu. V praxi tak zobrazují jednu verzi možného obsahu pro dané XML schéma. Data v souboru, která jsou považována za XML dokument, musí být tzv. well-formed **??**.

- **XML schéma.**
  Popisuje elementy a strukturu dokumentu. Zahrnuje v sobě povolené elementy, atributy, počet výskytů elementů a mnohá další omezení. Schéma pro dokument může být jak externí tak interní, podle toho zda je obsaženo v samotném dokumentu nebo jako externě přiložený soubor. Existuje několik jazyků pro popis XML schématu. Nejvyužívanějšími z nich jsou DTD (Documenty Type Definition) a XSD (XML Schema Definition). Jazyk DTD se vyznačuje značnými omezeními, mezi které patří omezený počet datových typů nebo omezení obsahu pouze na text. Těmto neduhům se jazyk XSD vyhýbá zavedením jednoduchých a komplexních datových typů, dědičností a rozšířením omezujících pravidel.

- **Well-formed XML.**
  XML data ve formě XML dokumentu, která dodržují správnou XML syntaxi. Mezi takové vlastnosti patří například správné zanořovaní a ukončování tagů, využívání povolených atributů, pouze jeden kořenový element, apod. Podle standardu W3C nelze považovat XML dokument za XML, pokud nesplňuje vlastnost well-formed. Pokud obsahuje XML dokument odkaz na schéma, pak se doporučuje ho validovat.

- **Validní XML.**
  Jedná se o podmnožinu well-formed XML dokumentů takovou, kde odpovídá struktura dokumentu přiloženému schématu nebo DTD.



Obrázek 4.1: Vztahy mezi jednotlivými typy XML souborů [11].

4.1.3 Mining dat z XML

K získávání dat z formátu XML lze použít několik různých technik. Jejich hlavní odlišnost spočívá v přístupu k odlišným částem XML – struktuře dokumentu nebo jeho obsahu. Podle toho dělíme celý proces buď na mining struktury nebo obsahu [11]. Mezi zmíněné metody patří:

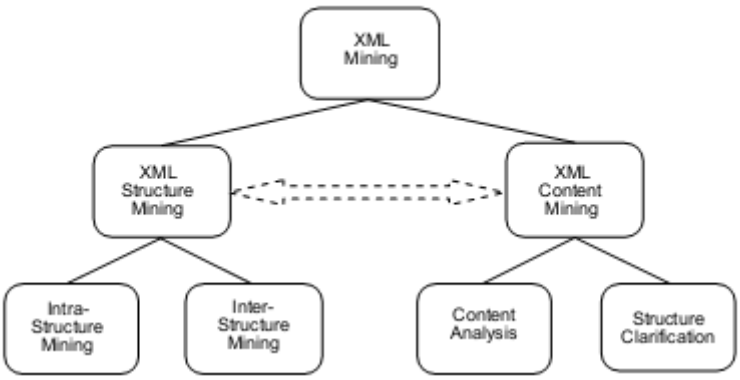
1. **Intra-structure mining.**
   Využívá především schémata, jež se zároveň ukládají pro další využití a lze je použít v případech dokumentů se shodným schématem. Schéma představuje popis třídy XML

dokumentu, kde dokumenty s daty jsou jeho instancemi. Dokument může obsahovat schéma v různých formách (DTD, XSD), ale pokud ho neobsahuje, nezbývá nám než se pokusit dokument parsovat s cílem extrahovat z něj strukturu. Kvalita takto získaného schématu se však může lišit, stejně jako jeho validita. Obecně při tomto typu data miningu využíváme následující metody:

- Klasifikace.
  Srovnávání nového schématu s množinou původních (testovací množina). Klasifikaci lze ideálně provádět nad validními XML dokumenty. S horšími vlastnostmi se šance na úspěšnou klasifikaci snižuje (well-formed, ill-formed, bez schématu).
- Shlukování.
  Identifikace podobností mezi XML dokumenty.
- Asociační pravidla.
  Využívají se pro správnou sémantickou interpretaci zkoumaného dokumentu, například při textově shodných, ale významově odlišných výrazech.

2. **Inter-structure mining.**
   Zabývá se hierarchií schémat dokumentů. Hledá mezi nimi vztahy a snaží se nalézt autora původního schématu. Tvůrci instancí daných schémat se nazývají „huby“. Tato znalost nám může pomoci při hledání dokumentů se stejným schématem (URI) a tím značně zjednodušit proces analýzy struktury zkoumaných XML dokumentů.

3. **Analýza obsahu.**
   Obsah, čili slova mezi počátečním a koncovým tagem. V tomto případě řešíme problémy stejných výrazů s odlišným významem (heterogenita, synonyma), což lze řešit analýzou kontextu. To je výhodné obzvláště tam, kde se vyskytují různá schémata, ale stejný obsah. Analýza je obtížnější, protože záleží na stupni granularity, ve kterém vyhledáváme (celé XML dokumenty, jednotlivé atributy, atd.).



Obrázek 4.2: Typy analýzy XML dokumentů [11].

4.1.4 Shrnutí

Z předchozího textu vyplývá, že dolování dat z formátu XML může být díky otevřenosti schémat a jejich kvality velice obtížnou disciplínou. Je potřeba brát v úvahu velké množství

parametrů a zároveň dodržovat fáze, pro určení správného postupu při parsování souboru. Přesto se jeví XML jako skvělý zdroj pro shromažďování informací z různých zdrojů na internetu, což lze spatřit na jeho stále se udržující oblibě.

4.2 RSS

4.2.1 Představení formátu

Jedná se o představitele z rodiny formátu XML, který se převážně používá pro přenos informací jako například nejnovější články, aktualizace, výsledky hledání, apod. Navíc dovoluje sdílet tyto novinky mezi jednotlivými weby a uživateli, kteří mají možnost tyto informace odebírat [5]. Na rozdíl od obecného standardu XML se RSS vyznačují v rámci jednotlivých verzí jednotnou strukturou (schématem). Tento fakt z nich činí velmi výhodný zdroj z hlediska následné automatické extrakce dat.

4.2.2 Parsování RSS s použitím PHP

Programovací jazyk PHP nám nabízí řadu sofistikovaných řešení, jak snadno dolovat data z unifikovaných zdrojů XML, mezi které patří i RSS. Jednou z metod je použití XSLT, což je jazyk sloužící pro konverzi XML do čitelnějších formátů, jako je například HTML. S využitím funkcí jazyka XSLT lze vytvořit šablonu, pomocí které již lze extrahovat z XML požadovaná data.

S využitím novější verze, konkrétně PHP 5, se nám však nabízí ještě jednodušší metoda. Jedná se o interní rozšíření PHP pod názvem SimpleXML [7] (dostupná jako součást standardního balíku PHP), s jehož pomocí je parsování XML otázkou pár řádek kódu. Jediným omezením těchto funkcí SimpleXML bývá omezení na kódování dat v utf-8. Tento problém však lze v PHP obejít pomocí funkcí na změnu kódování textu s příslušnými parametry.

4.2.3 Další řešení pro parsování

Ačkoliv se většina řešení parserů RSS kanálů vztahuje k oblasti jazyka PHP, lze najít i alternativy pro ostatní programovacích jazyky. Zkusme vypsat několik dalších nástrojů, které by nám s daným úkolem (tj. data mining z RSS) mohly pomoci:

- C++ – knihovna feed-reader-lib, určená pro získání a parsování RSS/Atom z webu.
- Java – projekt ROME dostupný v podobě volně stažitelné knihovny, určený mimo jiné pro parsování novinkových příspěvků různých formátů (RSS, ATOM), využívaný v řadě volně i komerčně dostupných aplikací.

4.2.4 ATOM

Téměř identickým formátem k výše rozebíranému RSS je novější formát ATOM. Hlavní rozdíly lze najít v pojmenování jednotlivých atributů (např. channel – feed, copyright – rights, ...). Pro účely dolování dat je proto téměř identický a dokonce většina nástrojů pro automatické zpracování RSS podporuje zároveň i ATOM [4]. V dalším textu často spojujeme formáty RSS a Atom, protože se jedná v podstatě o totéž.

4.2.5 Shrnutí

RSS představuje pro účely naší aplikace ideální zdroj pro dolování relevantních dat, především díky unifikované struktuře, což vede ke snadné automatické extrakci dat, pro kterou existuje velké množství nástrojů. Jediným úskalím by se mohlo stát hledání správných zdrojů RSS kanálů, tzv. „feedů“, a jejich rozumný počet. Některé weby navíc nemusí poskytovat své články v podobě RSS a u sociálních sítí je tato situace ještě složitější. Problematika zdrojů je diskutována dále v sekci 5.2.

4.3 HTML

4.3.1 Úvod

Většina informací na internetu se nachází v podobě HTML, což z něj činí pravděpodobně největší zdroj informací na světě. Získat z něj užitečné a relevantní data zůstává přesto problémem. Web mining, čili získávání informací z internetu se v případě HTML může dělit do tří různých kategorií [2]. Jsou jimi:

- Mining přístupů – vzory chování uživatelů na webu.
- Mining struktury webu – struktura odkazů (hyperlinks).
- Mining obsahu webu – užitečné informace nebo znalosti.

Toto rozdělení odpovídá i postupům použitým při získávání konkrétních informací z webu. V našem případě se zaměříme na hledání užitečného obsahu, tedy hledaných dat, obsažených v kódu HTML souborů. Postup je proto obdobný jako u dolování textů, kde zdrojový kód představuje prohledávaný text.

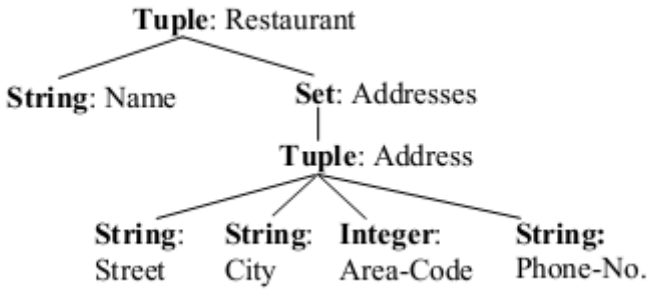
V posledních letech se oblast web miningu značně rozšířila, zejména díky stálému rozrůstání fenoménu, jímž internet rozhodně je. Zároveň se také zdůraznil ekonomický potenciál využití internetu, jako zdroje různých informací užitečných k analýze a zpracování. Problémem však zůstává heterogenita a různá úroveň strukturovanosti webů, což jejich automatické zpracovávání značně znesnadňuje, což dosvědčuje velký vědecký zájem v dané oblasti [2]. Velká část dat se na webu nachází ve standardní strukturované formě, což jsou data pro nás nejzajímavější. Zmiňujeme zde 3 přístupy, jak se k datům lze propracovat:

1. Využitím wrapperů.
2. Automatickou extrakcí.
3. Mining data records (dolování datových záznamů).

4.3.2 Wrappery

Pro vygenerování pravidel pro extrakci dat využijeme strojové učení. Uživatel označí cílové prvky v sadě trénovacích stránek, podle kterých si systém vygeneruje pravidla. Ty jsou následně aplikována na ostatních stránkách pro získání cílového obsahu. Na tomto principu vznikla řada velkých systémů (například komerční Fetch, nebo jeho nekomerční varianta Stalker, oba popisované ve zdroji [2]). Problémem wrapperů však stále zůstává nastavení, verifikace a údržba. Dohližitel (supervisor) má na starosti pracné a časově náročné ruční označování hledaných dat, což je vážný nedostatek obzvláště v datově rozsáhlých dokumentech. Dále se musí kontrolovat, zda pravidla vrací relevantní výsledky, což bývá klíčovým

faktorem u dynamicky se měnících webů. Verifikace bývá taktéž často pracná a časově náročná.



Obrázek 4.3: Příklad wrapperu zobrazeného pomocí stromové struktury [2].

4.3.3 Automatická extrakce

Tento přístup obvykle začíná tím, že systému předáme správné příklady, z nichž se vygenerují regulární výrazy. Vytváří se tak gramatika na základě dané stránky, přičemž se dále rozšiřuje řešením konfliktů s jinými stránkami. Z toho vyplývá, že je nutné dodat na začátku sadu příkladů, které sdílejí stejné schéma, ale přitom tvoří jeho různé instance. Odpadá tím však nutnost ručně označovat cílová data, protože se systém sám dokáže přizpůsobit heterogenitě mezi jednotlivými dokumenty. Nevýhodou může být, že srovnáváme celé stránky, nikoliv jenom data, což může do výsledků extrakce zanést i nechtěné části.

4.3.4 Mining data records (MDR)

Je metoda, která využívá pro extrakci dat i strukturu HTML, takzvaný *document object model* (zkráceně DOM). Pomocí něj dokáže určit regiony, kde by měla být obsažena data, a snižuje tak počet nerelevantních informací extrahovaných ze zdrojového webu. Problémem tohoto přístupu je samotné vytvoření správného stromu, jež má vystihovat strukturu zdroje. Pokud se nejedná o validované dokumenty, může docházet k existenci chybných tagů a dalších nekonzistencí, jež jakoukoliv analýzu znesnadňují nebo přímo znemožňují.

4.3.5 Parsování

K parsování HTML se dají využít opět knihovny nejrůznějších programovacích jazyků. Pro nás nejdůležitější je podpora v programovacích jazycích PHP a Java.

4.3.6 Shrnutí

HTML obsahuje největší množství informací obsažených na internetu, což z něj činí žádaný zdroj pro automatické zpracování a dolování dat. Velkou překážku však představuje nekonzistence jak datová tak strukturní. Automatické zpracování s cílem extrakce obsažených dat se tak jeví jako značně náročná operace, u které zároveň není zaručeno, že přinese očekávané výsledky.

Kapitola 5

Datové zdroje pro výslednou aplikaci

5.1 Volba zdrojového formátu

Z výzkumu provedeného v předchozí kapitole vyplývá, že většina zdrojů dat na internetu se vyznačuje problémy s nekonzistencí nebo obtížným zpracováním. Konečná volba formátu, jež bude použit pro získávání dat v naší aplikaci, padla na RSS. K tomuto rozhodnutí přispěla řada faktorů, které se nyní pokusíme shrnout:

Výhody

- Dobrá strukturovanost vedoucí ke snadné automatizaci získávání hledaných dat.
- Moderní a rozšířený formát, který bývá standardem pro kvalitní webové portály.
- Podpora vytváření RSS kanálů ze sociálních sítí, což dále rozšiřuje informační záběr datových zdrojů.
- Podpora hledání multimediálního obsahu a vytváření datových kanálů z něj.
- Snadná dohledatelnost uživatelem, který poté může odebírat nebo vytvářet vlastní datové kanály.

Nevýhody

- Neúplnost informací, kdy se v sekci *description* uvádí pouze výňatek celého článku, na který se odkazuje.
- Datové toky ze sociálních sítí, především Twitteru, mohou ze své podstaty zanášet velké množství nerelevantních informací.
- Není možné vytvářet RSS kanály uživatelských účtů na Facebooku, pokud se jedná o soukromé subjekty.
- Značky (Tagy) u multimediálního obsahu jsou plně v režii jejich autorů, proto nemusí odpovídat jejich obsahu.

5.2 RSS zdroje

Naše aplikace dovoluje uživateli rozšiřovat zdrojovou základnu odebíraných RSS kanálů, a tak napomáhá všestrannosti při jejím používání. Očekává se aktivita ze strany uživatelů, kteří si budou moci přizpůsobovat aplikaci k obrazu svému. Očekávané zdroje budou patřit do jedné z následujících kategorií:

- Webové portály a weby s podporou vlastních RSS kanálů.
- Kanály uživatelů vybraných sociálních sítí, jejichž aktivita se dá sledovat pomocí dále uvedených postupů.
- Pokročilé RSS kanály, jež jdou vytvořit na základě vyhledávacích dotazů (query) nebo značek (tagů).

V sociálních sítích je situace výrazně složitější, protože většina z nich nadále oficiálně nepodporuje jednoduchý přístup k RSS obsahu jejich uživatelů. Přesto se však dají nalézt ve „skrytých“ adresářích. Při získávání jejich obsahu nám pomohli návody dostupné na internetu [1]. Takto lze vytvořit RSS kanál téměř ke kterémukoliv účtu v nejznámějších sítích jako jsou:

- Twitter,
- YouTube,
- Facebook,
- Instagram,
- Flickr,
- Picasa,
- Tumbler,
- Blogspot,
- Wordpress.

Některé sociální sítě nabízí i pokročilé možnosti, co se tvorby RSS kanálů týče. Některé mají naopak možnosti získávání datových kanálů značně omezené a požadují netriviální znalosti od případného uživatele (například skryté identifikátory, přístupové tokeny, apod.). Rozebereme si proto dále ty největší a pro nás nejzajímavější.

5.2.1 Běžné RSS kanály

Do této kategorie spadají především veřejně přístupné kanály z oblasti internetových webů a magazínů. V dnešní době je standardem, že téměř každý moderní portál poskytuje možnost odebírat novinky o dění v jeho obsahu právě pomocí RSS. Právě proto budou tyto datové kanály tvořit páteř při sestavování uživatelských požadavků na hledání klíčových slov.

Tyto zdroje tak lze logicky rozčlenit do dvou hlavních kategorií:

- Přednastavené datové zdroje

- Uživatelem definované datové zdroje

Rozdíl bude patrný pouze v počáteční fázi fungování aplikace, protože brzy se s přibývajícími uživateli rozroste i základna s uživatelsky definovanými zdroji. V aplikaci přednastavené zdroje tak mají za cíl pouze usnadnit používání aplikace, kdy dáváme uživatelům výběr z předdefinovaných vzorů a není tak nutné pro první hledání definovat i vlastní zdroje hledání.

Při výběru vhodných kandidátů pro zdroje informací se budeme řídit podle oblíbenosti a návštěvnosti webů, jež jsou monitorovány různými internetovými aplikacemi. Příkladem budiž doména „navrcholu.cz“, kde lze nalézt informace z monitoringu českých internetových webů s možností vyhledávání v předem určených kategorií.

5.2.2 Twitter

Twitter je označení pro sociální síť a mikroblog, kde uživatel vystavuje na svou profilovou stránku krátké zprávy (tweety), které připomínají SMS zprávy známé z mobilních telefonů. Ostatní uživatelé pak mají možnost přihlásit si odběr těchto zpráv od konkrétních uživatelů. Právě na tyto vlastnosti se zaměřují i příslušné RSS kanály.

1. Získat 20 nejčerstvějších updatů konkrétního uživatele.
   `https://api.twitter.com/1/statuses/user_timeline.rss?screen_name=<user>`
2. Získat nejoblíbenější tweety určitého uživatele.
   `https://api.twitter.com/1/favorites.rss?screen_name=<username>.rss`
3. Získat zmínky (mentions) konkrétního uživatele Twitteru.
   `http://search.twitter.com/search.rss?q=to:@<username>`
4. Získat RSS kanál pro jakékoliv hledání nebo hashtag na Twitteru.
   `http://search.twitter.com/search.rss?q=<query>`

5.2.3 YouTube

YouTube v současnosti přestavuje největší internetový server a sociální síť postavenou na nahrávání video souborů. Dovoluje uživatelům vytvářet svoje kanály, kde mají možnost sdílet videa s ostatními uživateli. Samozřejmostí je i možnost odebírat novinky z kanálů oblíbených uživatelů, spravovat komentáře videí, členit nahrané soubory do složek nebo vytvářet videoblogy. Následují příklady možných RSS kanálů.

1. Získat nejnovější uploadovaná videa daného uživatele.
   `https://gdata.youtube.com/feeds/api/users/<user>/uploads`
2. Získat RSS kanál videí, která obsahují konkrétní tag.
   `https://gdata.youtube.com/feeds/api/videos/-/<tag>`
3. Získat RSS kanál pro jakýkoliv vyhledávací dotaz na YouTube.
   `https://gdata.youtube.com/feeds/api/videos?q=<query>&orderby=relevance`

5.2.4 Facebook

Facebook nejspíše netřeba příliš detailně představovat. Jedná se o rozsáhlou webovou službu na principu sociální sítě, která registrovaným uživatelům umožňuje vytvořit si osobní, veřejný nebo částečně veřejný profil a komunikovat spolu. Nabízí pro tyto účely nejrůznější nástroje a funkce, jako například sdílení multimediálních dat, vytváření vztahů mezi uživateli, nebo využívání aplikací třetí strany.

Na rozdíl od předchozích sociálních aplikací Facebook nepodporuje vytváření RSS kanálů pomocí identifikátorů jako je *username* nebo *query*, ale využívá unikátní *ID* uživatele. Navíc není tento identifikátor dostupný přímo na veřejné stránce uživatele, ale musí se získat přes službu externí službu tak, jak si nyní ukážeme na příkladu.

Pro získání *ID* musíme zadat jméno za lomítkem z URL adresy uživatele jako parametr v URL zde:
`http://graph.facebook.com/<facebook_name>`


```
{
   "about": "Slune\u010dnice.cz - programy rychle a zadarmo",
   "company_overview": "Des\u00edtky gigabyt\u016f softwaru naj
   "founded": "1997",
   "is_published": true,
   "talking_about_count": 109,
   "username": "Slunecnice.cz",
   "website": "http://www.slunecnice.cz/",
   "were_here_count": 0,
   "category": "Computers/internet website",
   "id": "437146520244",
   "name": "Slune\u010dnice.cz",
   "link": "http://www.facebook.com/Slunecnice.cz",
   "likes": 7858,
   "cover": {
      "cover_id": "10151499653955245",
      "source": "http://sphotos-e.ak.fbcdn.net/hphotos-ak-ash3/
      "offset_y": 0
   }
}
```


Obrázek 5.1: Příklad nalezení příslušného ID subjektu na Facebooku.

Všimněte si políčka ID. Toto číslo slouží jako obecný identifikátor na Facebooku. Stačí ho tedy zadat podobně jako u ostatních sítí do příslušného políčka v následujících URL. Problém spočívá ve faktu, že takto nelze odebírat informace z dění na účtech konkrétních uživatelů, ale pouze od veřejných institucí, jako jsou firmy, organizace, místa, atd.

1. Odebírat novinky z Facebookového účtu pomocí RSS.
   `https://www.facebook.com/feeds/page.php?format=atom10&id=<ID>`

5.2.5 Pinterest

Tato relativně mladá sociální síť umožňuje svým uživatelům vytvářet digitální nástěnky, kde mohou spravovat tematicky laděné kolekce obrázků. V poslední době se Pinterest velice rozšířil, a to především v Americe, kde podle oficiálních průzkumů tvoří třetí nejrozšířenější sociální síť na internetu. Zajímavostí může být, že většinu uživatelů tvoří ženy (podle

průzkumů až 79%). Následuje výčet možných RSS kanálů. Bohužel chybí možnost vytvoření kanálu z vyhledávacího dotazu nebo tagu, což by se u této služby velmi hodilo.

1. Vytvoření kanálu konkrétního uživatele.
   `http://pinterest.com/<user>/feed.rss`
2. Vytvoření kanálu z konkrétního zdi (board) daného uživatele.
   `http://pinterest.com/<user>/<board>/rss`

5.2.6 Obrázkové RSS kanály – Instagram, Picasa, Flickr

Všechny tyto uvedené komunitní weby slouží pro sdílení fotografií mezi uživateli na internetu. Instagram je zaměřena především na mobilní operační systémy, jako jsou iOS a Android a dovoluje uživatelům snadno upravovat jejich fotky, které poté lze nasdílet na jiné sociální aplikace, jako Facebook nebo Twitter. Picasa představuje obdobnou aplikaci, již spravuje společnost Google. Vyznačuje se proto především rychlým a přesným vyhledáním ve velkých databází sdílených obrázků podle tagů a názvu souboru. Posledním uváděným webem je Flickr. Ten podobně jako předchozí aplikace umožňuje kromě sdílení obrázků i jejich následné umisťování do mapy. Jedná se o obdobnou aplikaci jako předchozí jmenované.
Podívejme se na výčet RSS kanálů pro tyto aplikace.

1. Získání RSS kanálu z obrázků nahraných uživatelem Flickru.
   `http://api.flickr.com/services/feeds/photos_public.gne?id=<ID>`
2. RSS kanál obsahující obrázky z Flickru označené konkrétními tagy.
   `http://api.flickr.com/services/feeds/photos_public.gne?tags=<t1>,<t2>`
3. RSS kanál s obrázky z Instagramu označené konkrétními tagy.
   `http://instagr.am/tags/<tag>/feed/recent.rss`
4. RSS kanál obsahující obrázky odpovídající vyhledávacímu dotazu na Picasa.
   `http://photos.googleapis.com/data/feed/base/all?alt=rss&kind=photo&q=<search>`

5.3 Shrnutí

Formát RSS nabízí uspokojivou datovou základnu díky širokého množství dostupných datových zdrojů, včetně částečné podpory sociálních sítí. Proto byl zvolen jako hlavní datový zdroj při vytváření cílové aplikace. Co se strukturovanosti týče, jediné zřejmé problémy souvisí s existencí dvou rovnocenných verzí, RSS 2.0 a Atom. Jejich rozdíly však lze při implementaci snadno překlenout pomocí odpovídajícího programového vybavení a nic proto nebrání unifikaci principů jejich automatického zpracování.

Kapitola 6

Návrh aplikace

Nejviditelnějším výstupem práce se stane výsledná aplikace, jež bude postavena na analýze informačních zdrojů a formátů uvedených v předchozích kapitolách. V této kapitole si shrneme cíle aplikace, funkční i programové zaměření, očekávané vlastnosti a konečně i návrh funkčnosti a strukturu programového členění.

6.1 Funkce aplikace

Abychom mohli začít správně formovat návrh aplikace, je nutné stanovit si cíle, které by měla aplikace ve finálním stavu plnit.

6.1.1 Role aplikace z hlediska zacílení

Z funkčního hlediska chceme uživateli poskytnout alternativní možnosti hledání klíčových slov v prostředí internetu. Není naším cílem konkurovat ústředním představitelům v oblasti internetových prohlížečů, jako jsou Google nebo Yahoo!, ale spíše poskytnout doplněk, či alternativu, která pomůže při hledání specifického druhu dat z vybraných zdrojů. Nejde nám tedy v první řadě o fulltextové hledání a indexování ve většině obsahu vystaveném na internetu, ale o sledování určité, obvykle tematicky nebo obsahově specificky zaměřené části, kde budou hledány reference na klíčová slova nebo hledaný výraz. Takto ohraničené oblasti budou tvořené z informačních zdrojů, jejichž výběr bude záviset výhradně na uživateli aplikace.

Kromě toho se od klasického vyhledávání bude lišit časové hledisko vyhledávání. Běžné vyhledávače se snaží nalézt shodu v dříve uloženém obsahu, což může vést k neaktuálnosti nalezených výsledků. Naše aplikace však počítá s průběžným indexováním aktuálních dat v delším časovém úseku, kde počátek indexace tvoří vložením požadavku na hledání. V praxi tak vlastně od vytvoření úkolu sledujeme vybrané zdroje a hledáme v jejich RSS záznamech zadaná klíčová slova. Tento fakt sebou nese své výhody i nevýhody. Výhody spočívají v aktuálnosti obsažených informací, takže uživatel může z průběžných výsledků hledání získávat nejnovější informace z aktuálního dění. Zároveň z hlediska implementačního není potřeba tolik řešit potřebu na výkon a dostupný datový prostor serverové stanice, jež by indexace veškerého dění v zadaných datových zdrojích vyžadovala, a to především ve fázi s rychle se rozšiřujícím počtem aktivních uživatelů. Zároveň se také vylučuje možnost sledování historických trendů v datových zdrojích, díky absenci staršího obsahu. Na druhou stranu ideální možnost představuje sledování aktuálních krátkodobých trendů, díky

časovému řazení a menšímu množství výsledků, navíc s využitím sociálních sítí, kde se nové zprávy často vyskytují velice rychle po zveřejnění.

Pro využití dlouhodobějších statistik už nyní uvažujeme s rozšířením, které by propojilo tuto aplikaci se systémem pro hodnocení důvěryhodnosti datových zdrojů na internetu a tvorbu sítí mezi jejich subjekty, kde by se využívaly naší aplikací sesbírané informace z delších časových průběhů hledání.

6.1.2 Typ aplikace

Ve fázi návrhu se nám otevírají různé možnosti, jak finální aplikaci pojmout z hlediska typu. Mohli bychom se do jisté míry inspirovat již existujícími variantami, které jsou určeny pro zpracování RSS zdrojů z internetu. Těch se vyskytuje poměrně velké množství, v různých formách, ať jsou integrovány do běžného internetového prohlížeče (např. pluginy pro Mozillu Firefox - RSS nebo SAGE, ...) nebo existují jako samostatné desktopové aplikace (NewzCrawler, FeedDemon, Google Reader, ...). Opomenout bychom neměli i možnost vytvoření aplikace pro mobilní zařízení, které jsou v poslední době čím dál žádanější. Pojďme si tedy shrnout poznatky pro každou z těchto kategorií:

1. **Forma pluginu do existující aplikace**
   Výhody tohoto přístupu jsou zřejmé, není potřeba vytvářet celou aplikaci od začátku, ale lze využít již existující rozhraní a metody pro integraci námi požadované funkcionality. Podobný postup jsme použili v předchozí práci, zabývající se vylepšením použitelnosti a uživatelského rozhraní se zachováním účelnosti webové aplikace. Přesvědčili jsme se tak, že tento přístup dovoluje autorům oprostit se od většiny implementačních detailů a tím lze věnovat více prostoru vlastním inovacím aplikace.

   Značnou nevýhodu však představují omezení, která jsou na množinu pluginů kladena. Ať se již jedná o limitaci z hlediska implementačního, tak z hlediska aplikačního, vždy je nutné se předem přesvědčit, jaké jsou k dispozici pro domovskou aplikaci prostředky a pravidla, jimiž se musí vývoj jejich rozšíření řídit. Právě tento fakt může vážně omezit kreativní složku výsledného produktu, stejně jako jeho potenciál pro případné komerční rozšíření.

2. **Forma vlastní desktopové aplikace (aplikačního softwaru)**
   Jedná se o velmi oblíbenou variantu mezi většinou existujících aplikací, které sdružují hledání informací z internetu do jedné, pro uživatele přehledné, podoby. Instalace do systému na lokálním počítači s sebou obvykle nese možnosti různých nastavení a přizpůsobení se uživateli, jelikož se jedná o standard v oblasti desktopových aplikací. Díky vysokému výkonu cílové stanice stejně jako dostupným paměťovým prostorám se lze uchýlit i k náročným operacím a pokročilým vizualizačním efektům.

   Nevýhodu pak představuje separace každé instalace aplikace, u kterých může být problém s komunikací a sdílení zajímavého obsahu mezi nimi. Vše je závislé na stálosti internetového připojení a nastavení komunikace z uživatelské stanice, na což se může z uživatelské perspektivy nahlížet s jistým podezřením. Uživatel nerad nechává odcházet komunikaci z jeho domovského počítače, aniž by přesně věděl, co obsahuje. Proto se často chrání firewally a aktivními síťovými prvky, které mohou většinu komunikace blokovat. Stejně tak je problém s průběžnými aktualizacemi takového typu aplikací, kdy je nutné upozornit uživatele na změny a vyžádat si stažení aktualizované verze.

3. **Forma webové aplikace** Webové aplikace představují oblíbené řešení díky multiplatformnosti, kterou tato varianta přináší. Většina obsahu pro jednotlivé internetové prohlížeče je vzájemně kompatibilní, stejně jako pro výchozí zařízení, odkud se spouští. Proto není problém používat stejnou webovou aplikaci jak na výkonném stolním počítači, notebooku, či chytrém telefonu. Pokud je navíc cílový obsah příslušně optimalizován, nabízí na všech těchto variantách zařízení jednotný vzhled a funkčnost všech svých částí. Dalším výhodu skýtá jednotná verze aplikace pro všechny její uživatele, s nejnovějšími aktualizacemi a čerstvým obsahem, který lze mezi uživateli snadno sdílet.

   I zde se najdou faktory, ke kterým je nutno přihlédnout. Například výkonnost a spolehlivost zdrojového serveru je naprosto klíčová, protože bez něj cílová aplikace neexistuje. Stejně tak musí být aplikace zajištěna proti napadení z vnějšku, protože v prostředí internetu se nachází nespočet útočníků čekajících na příležitost vyzkoušet si své schopnosti na nové spuštěné aplikaci nebo službě. I jeden úspěšný útočník může zničit veškerou funkcionalitu i pověst programu během krátké chvilky.

4. **Forma mobilní aplikace**
   Vytvářet aplikace pro mobilní zařízení je v dnešní době velice lukrativní záležitostí. Důvodem jsou neustále se rozvíjející technologie, postupující velice rychle kupředu v oblasti výkonnosti zařízení. A právě tyto nová zařízení hledají adekvátní aplikace, které by co nejlépe využili jejich nově nabytého potenciálu.

   Z hlediska ekonomického představují mobilní zařízení dosud nenasycený trh, kde mohou autoři stále nacházet oblasti, jež mohou vyplnit právě svou aplikací a tím dosáhnout značného úspěchu i zisku. Na druhou stranu, o totéž se v současnosti snaží velké množství nezávislých autorů, stejně jako velkých a zaběhnutých značek z oblasti osobních počítačů, což zvyšuje konkurenci v daném sektoru a není snadné se zde prosadit.

   Nevýhodu však představuje především značná heterogenita, jak z hlediska výkonu zařízení, tak z hlediska jednotlivých verzí operačních systémů, jež nemusí poskytovat plnou interoperabilitu se staršími modely. Navíc díky současnému rychlému vývoji v této oblasti může být obtížné udržet krok jak s technologiemi, tak s konkurencí na trhu.

6.1.3 Hodnocení a zpětná vazba k uživateli

Protože bude jedním z poskytovaných nástrojů i hodnocení výsledků a s nimi spojených datových zdrojů, chceme získávat i další data, jež nám pomohou s přizpůsobením aplikace jejímu uživateli. V ideálním případě poskytneme všem uživatelům sdílená data, co se datových zdrojů a jejich hodnocení týče. Uživatelské vstupy by se v takovém případě ukládaly do jednotné databáze a po jejich zpracování by se distribuovaly změny podle konkrétního typu aplikace zpět do uživatelské aplikační instance. Navíc bychom rádi zaznamenávali i další data z jejich aktivity uvnitř našeho systému, takže by se dala následovně rozdělit:

1. Data zaznamenávána z uživatelských akcí na serveru a při používání naší aplikace. Zde se zaměříme na vstupy, výstupy, celkovou orientaci uživatele v rámci ovládacího rozhraní. Rovněž budeme mapovat nejčastěji vykonávané aktivity v systému. Tyto data pak budou mít přímý vliv na hodnocení použitelnosti spuštěné aplikace a dovolí nám ji vylepšovat směrem k lepšímu uživatelskému zážitku. Pro zaznamenávaní

dat nám pomůžou aplikace typu Google Analytics, Piwik, nebo různé softwary pro zaznamenávaní *heatmap* na serveru. Výstup takového analytického softwaru lze vidět na obrázku 6.1.

Obrázek 6.1: Příklad výstupu z analýzy www pomocí heatmap.

2. Hodnocení výsledků hledání, frekvence využití dostupných zdrojů, uživatelské akce na základě vyhledávání. Tyto data budou následně použita pro hodnocení zdrojů z hlediska spokojenosti uživatelů s jejich výsledky, podle nichž lze například řadit datové zdroje podle jejich hodnocení. Dále můžeme tyto informace využít v rámci dalších rozšíření, jako například komplexního hodnocení zdrojů na internetu nebo tvorby struktury vztahů mezi uživateli.

6.1.4 Agregace datových zdrojů

Asi hlavním tématem aplikace je schopnost hledání informací z různých datových zdrojů současně. Naším cílem tedy bude agregovat různá data z různých zdrojů do jednotného toku, který poté poskytne možnost unifikovaného hledání v něm. V tomto bodě nám velice pomůže volba RSS jakožto zdrojového formátu právě díky velkému množství dostupných datových zdrojů, stejně jako jeho unifikované struktuře. Procházením seznamu validních URL adres s RSS kanály můžeme snadno agregovat datové zdroje z různých portálů a webů včetně využití zvláštních zdrojů ze sociálních sítí. Jednou z vlastností aplikace bude rovněž sdílení zdrojů mezi jejími uživateli, čímž poskytneme širokou základnu zdrojů, z nichž může uživatel při sestavování hledání vybírat, aby mu poté mohly být předloženy výsledky hledání v nich.

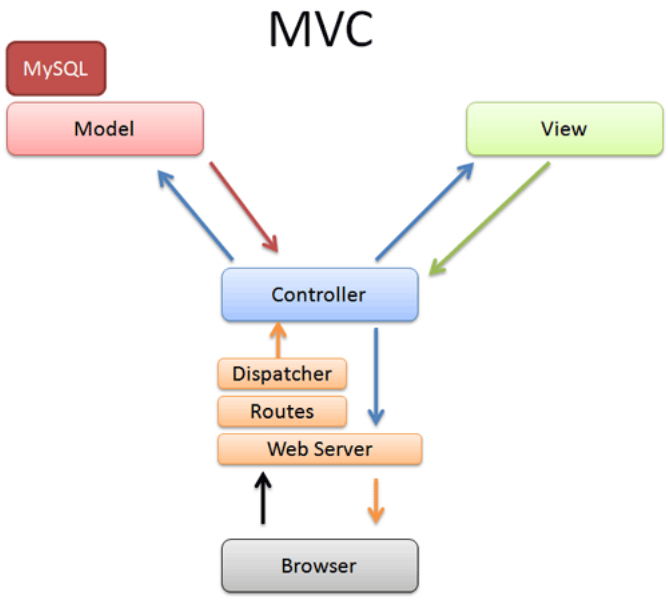
6.2 Implementační hledisko aplikace

Aby bylo možné začít realizovat navržené cíle, jež bude aplikace plnit, je nutné si určit programovací jazyk, strukturu a typ aplikace. Tím silně ovlivníme zacílení aplikace na určitý segment potencionálních uživatelů, což je faktor jenž by neměl být ve fázi návrhu opomenut.

6.2.1 Volba typu aplikace

Po zvážení všech faktorů a vzhledem k charakteru naší aplikace jsme zvolili jako nejvhodnější řešení podobu webové aplikace. Klíčovými faktory se pro nás stala především interoperabilita napříč většinou dostupných zařízení, které dokáží v základní podobě používat moderní internetové prohlížeče. Tím byl určen i programovací jazyk, jímž se stalo HTML s využitím PHP a v kombinaci s databázovým systémem MySQL. Co se PHP týče, jevilo se nám vhodné zvolit některý z dostupných frameworků, především díky úspoře času při implementaci a realizaci aplikace podle návrhových vzorů. Většina frameworků pracuje na modelu objektové orientace, která se k převedení návrhu do programové podoby využívá nejvíce. Stejně tak poskytuje i nejvyšší míru z hlediska čitelnosti a srozumitelnosti kódu, což vede k snadné údržbě a případně rozšiřitelnosti. Na základě praktických zkušeností se nám osvědčil typ implementace MVC (Model View Controler) nebo od něj odvozené varianty (konkrétně MVP - Model View Presenter). Tento způsob vyžaduje striktní oddělení části pro funkce a práci s databází od zpracování a předáváním nezbytných dat šablonám a šablon samotných, které se starají o vykreslení vloženého obsahu do finální podoby. Schéma takového rozložení je pro ilustraci uvedeno na obrázku 6.2. Díky dědičnosti tak lze ušetřit velké části kódu mezi jednotlivými třídami. Dále v sobě frameworky často skrývají i různá vylepšení, například co se ladění při vývoji týče, nebo podporu celé řady pluginů, jež můžou mít zásadní vliv na podobu kódu nebo funkční stránku výsledné aplikace. V zásadě představují frameworky využívající tento návrhový vzor velice efektivní nástroj pro vytváření webových aplikací.

Jediná negativa z hlediska jejich použití lze spatřit jednak v občasných omezeních, na která lze při implementaci narazit, tak v nutnosti naučit se s konkrétním frameworkem efektivně pracovat. Frameworky MVC často nepodporují psaní v „čistém“ PHP kódu, ale používají vlastní metody a notace, jejichž funkcionalita může být na první pohled skryta. Aby byl vývojář schopen porozumět vnitřní stavbě, je nezbytné strávit nějaký čas s dokumentací a ukázkami kódu, jež poté poskytnou klíč k efektivnímu používání nového systému.



Obrázek 6.2: Strukturní rozdělení aplikace v návrhovém vzoru MVC.

6.2.2 Nástroje pro vývoj

V návaznosti na předchozí množinu výhod vyplývajících z použití objektových návrhových vzorů a frameworků na nich postavených, jsme zvolili v České republice velice populární framework **Nette**.

Využití Nette nám poskytne následující výhody:

- Kvalitní a pro vývojáře přizpůsobené prostředí, využívající model objektově orientovaného programování.
- Pokročilé zobrazování chybových hlášení pomocí nástroje „Laděnka“ , která efektivně pomáhá vývoji a testování aplikace.
- Rozsáhlou základnu českých vývojářů, kteří rádi poradí s řešením problémů, jež se mohou při programování v Nette i v PHP jako takovém objevit.
- Různé doplňky a pluginy určené ke snadné implementaci do vyvíjené aplikace.
- Bezpečnou manipulaci se vstupy, především co se přístupů do databáze týče, což značně omezuje napadnutelnost celého systému.
- Vysoký výkon, podporu skriptování, AJAX, znovupoužitelnost kódu.
- Vše s licencí Open-source (BSD licence), jež patří k nejvolnějším a dovoluje framework zdarma používat i v komerčních projektech.

Toto vše ve spojení s předchozími zkušenostmi se zmiňovaným frameworkem z něj učinilo jasnou volbu při rozhodování o zdrojovém frameworku. Pro webový server byla zvolena aplikace **Apache**, konkrétně jedna z jeho nejaktuálnějších verzí 2.2.22. Jak naznačují uvedené průzkumy (obrázek 6.3), jedná se v současnosti o nejvyužívanější produkt v oblasti webových serverů, poskytující kvalitní podporu pro moduly **PHP** a **MySQL**, rozšiřující dále svoje možnosti nastavení pomocí řady doplňkových funkcí. Příkladem mohou být *rewrite_mode*, nezbytný pro takzvané „hezké url“, jež naše aplikace bude využívat.

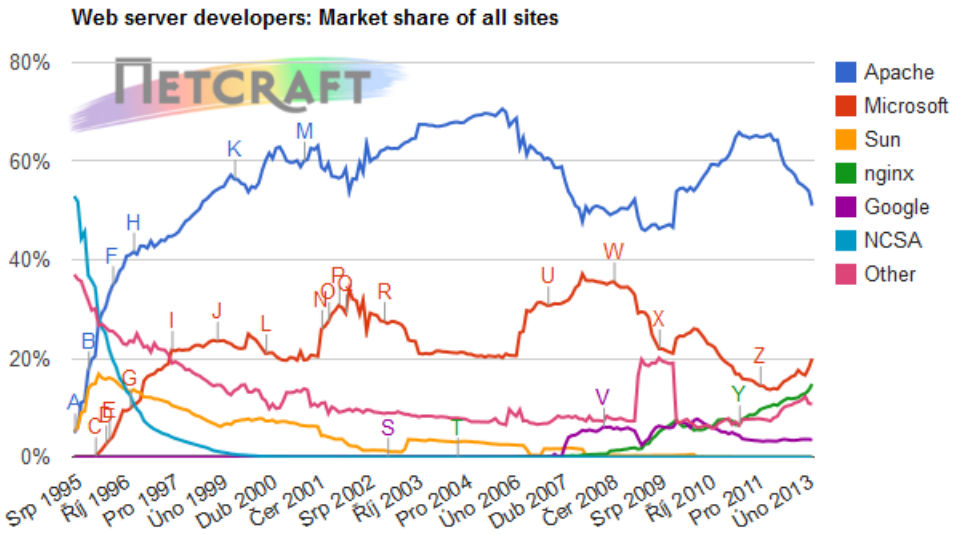
6.2.3 Struktura aplikace

Námi vytvořená webová aplikace jakožto celek se bude skládat z několika logických částí. Těmi jsou:

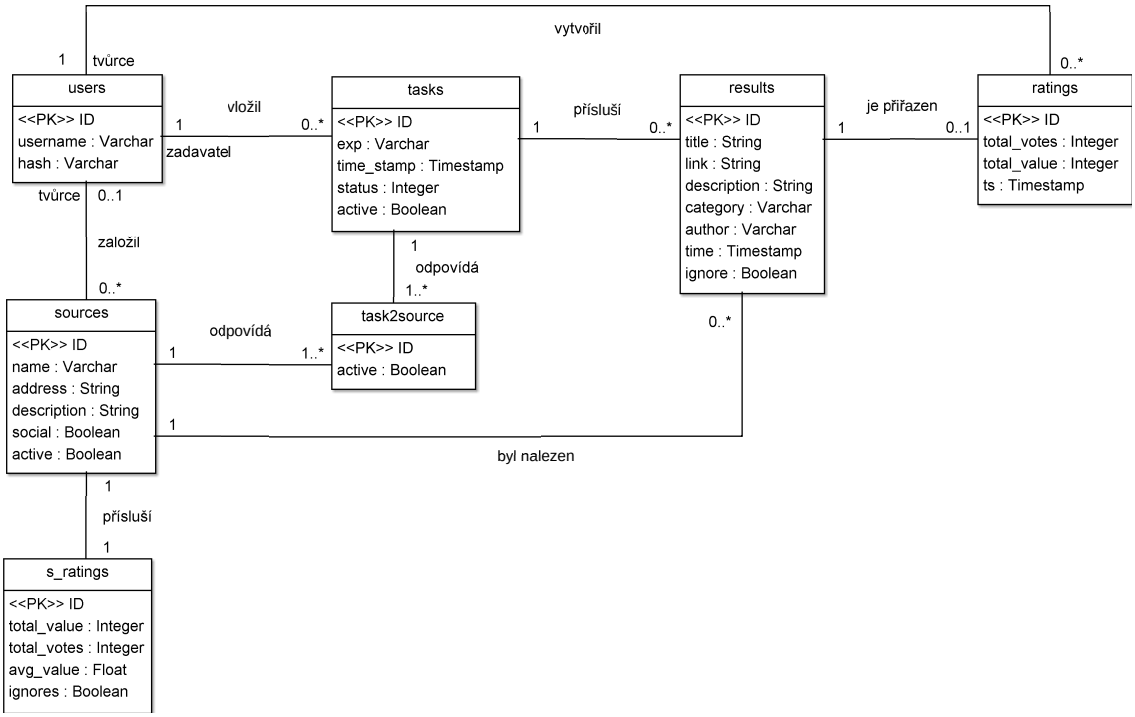
1. Databáze,
2. indexer,
3. uživatelské rozhraní.

Databáze bude ze své podstaty využita pro uchovávání veškerých persistentních dat, která bude aplikace potřebovat. Konkrétně se jedná o uložení uživatelů, zdrojů, úkolů hledání, výsledků, hodnocení a všech potřebných asociací, jež bude nutné při formování databáze vytvořit. Při návrhu databázové struktury jsme postupovali systematicky a vytvořili odpovídající UML model. Konrétně jsme využili ER diagram 6.4.

Při návrhu databáze počítáme i s dalším případným rozšiřováním směrem ke komplexnějšímu hodnocení zdrojů a výsledků, se kterými uživatel pracuje. Prozatím jsou v návrhu jen dvě základní hodnotící tabulky. První se týká samotných výsledků hledání a uchovává



Obrázek 6.3: Průzkum podílu na trhu v oblasti internetových domén [12].



Obrázek 6.4: ER diagram využitý pro návrh aplikace.

uživatelská hodnocení. Odtud se pak hodnocení propaguje do druhé tabulky, jež uchovává statistické informace pro hodnocení zdrojů a tato informace se přenáší zpět k uživateli. Podle hodnocení se pak mohou řadit zdroje podle relevance a úspěšnosti předchozích hle-

dání, stejně jako mohou pomoci uživateli při rozhodování s výběrem zdrojů pro následné aktivity.

Indexer bude tvořit jediný PHP skript, jehož úkolem bude prohledat všechny zdroje vybrané uživateli aplikace k hledání shody na základě klíčových slov. Jeho spouštění pak bude řízeno pomocí plánovače úloh v konkrétním operačním systému, na kterém bude webový server spuštěn. Vycházejí nám tedy dvě hlavní varianty: Cron v případě systému Linux, nebo Plánovač úloh pro variantu Windows. Jejich úkolem bude spouštět skript po uplynutí nastaveného intervalu a tím pádem provedení aktualizací výsledků ukládaných do databáze.

Co se samotné struktury indexeru týká, bude se jednat o kombinaci 3 základních funkcí:

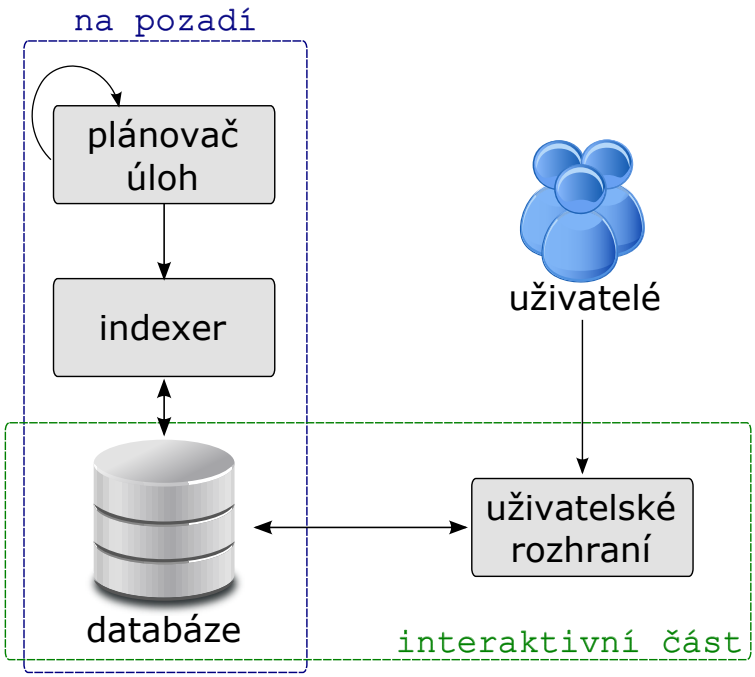
- Stažení zdrojových dokumentů a jejich parsování.
- Prohledání zdrojových dat na výskyt klíčových slov.
- Připojení do databáze a uložení nalezených výsledků.

Zdrojové dokumenty budeme číst jako data, která se poté rozparsují a uloží do příslušné struktury, jež bude představovat jeden RSS záznam. K parsování využijeme interní rozšíření PHP s názvem SimpleXML. Jeho podstatou je umožnit snadné čtení dokumentů s XML syntaxí, umožňující přístup k jednotlivým elementům XML jako částem asociativního pole.

Obsah uložený v datových částech RSS záznamu poté projde cyklem na hledání klíčových slov. Pro tyto účely využijeme PHP funkce na porovnání řetězců pomocí regulárních výrazů. Hlavní výhodou plynoucí z jejich využití jsou rozsáhlé možnosti parametrizace, která může ovlivnit vlastní hledání a tím pádem můžeme uživateli umožnit nastavení hledání podle vlastních preferencí.

Při shodě u hledání uložíme výsledný RSS záznam do databáze, konkrétně do tabulky *results*, která kopíruje jednotlivé atributy RSS záznamů. Kvůli zpětné kompatibilitě převedeme kódování vstupního textu do *utf8_unicode_ci*, tedy výchozího formátování naší databáze. Pro časová razítka využijeme standardní linuxový čas v sekundách, tedy databázového typu *timestamp*.

Uživatelské rozhraní bude vytvořeno pomocí frameworku Nette a s využitím návrhového vzoru MVP popsaného v kapitole 7.3.1. Zaměříme se především na jednoduchost a přehlednost ovládání s cílem udržet co největší míru použitelnosti aplikace. Nejdříve přistoupíme k vytváření funkční části s jednoduchým uživatelským rozhraním, jež bude využito především za účelem ladění a testování funkcí. Po dokončení této stránky aplikace se zaměříme i na sofistikovanější uživatelské rozhraní, kde zapojíme další doplňkové technologie jako jsou *jQeury*, *AJAX* a *CSS* pro lepší interaktivitu a dynamičnost ovládání.



Obrázek 6.5: Schéma rozdělení aplikace.

Kapitola 7

Realizace a implementace

7.1 Server pro aplikaci

Vývoj aplikace probíhal zpočátku na lokálním počítači v prostředí operačního systému Windows, kde byl spuštěn místní webový server. K tomu byl využit software **WampServer**, který v sobě nese všechny nezbytné komponenty a funkce spojené se serverem v jednotném balíčku. WampServer se skládá z následujících částí:

- Webový server Apache2.
- PHP ve volitelných verzích.
- Multiplatformní databázový systém MySQL.
- PhpMyAdmin pro jednoduchou správu databází.
- Jednoduchý instalátor pro operační systém Windows.
- Funkční nastavení pro všechny uvedené služby, takže lze server ihned spustit a začít používat.

Řízení je pak usnadněno pomocí jednoduchého ovládacího rozhraní na přes *tray* ikonu v dolní části obrazovky, kde lze servery vypínat, restartovat nebo spouštět, přistupovat snadno k datovému úložišti webového serveru, nebo ke konfiguračním souborům jednotlivých služeb. Zároveň je WampServer dostupný zdarma pod licencí GPL, proto poskytuje efektivní prostředek při úspoře času spojené s instalací a konfigurací softwaru spojeného s provozováním webového serveru.

Po vytvoření prvních funkčních prototypů aplikace jsme přistoupili k přesunu stabilní verze z lokálního na veřejný webový server. Zvažována byla možnost využití serverů veřejných poskytovatelů, ale vyskytl se problém týkající se možností nastavení systémového plánovače, který aplikace nezbytně potřebuje pro pravidelné spuštění skriptu indexeru. Naprostá většina poskytovatelů neposkytuje přístup ke spouštění vnitřních procesů hostitelského operačního systému, kromě vyhrazených služeb pro webový server. Proto jsme přistoupili k variantě vlastního fyzického serveru, čili počítače sestaveného čistě pro účely vývoje aplikace, financovaného z vlastních zdrojů. Ten byl poté umístěn do demilitarizované zóny, odkud k němu lze přistupovat jak z vnitřní tak vnější sítě. Napájení serveru a souvisejících síťových komponent bylo zajištěno pomocí dodatečných zdrojů energie UPS, aby byla zaručena co největší spolehlivost a dostupnost serveru a na něm běžící aplikace. Zároveň

byl nastaven plánovač operačního systému, konkrétně Windows Scheduler, na spouštění krátkého *batch file* skriptu, který má za úkol aktivovat PHP skript indexeru. Perioda pro opakované spouštění indexeru je aktuálně nastavena na 1 hodinu, počínaje celou hodinou. Tento parametr nelze nastavit prostřednictvím aplikace a ani taková funkce není z hlediska vývoje a bezpečnosti žádoucí.

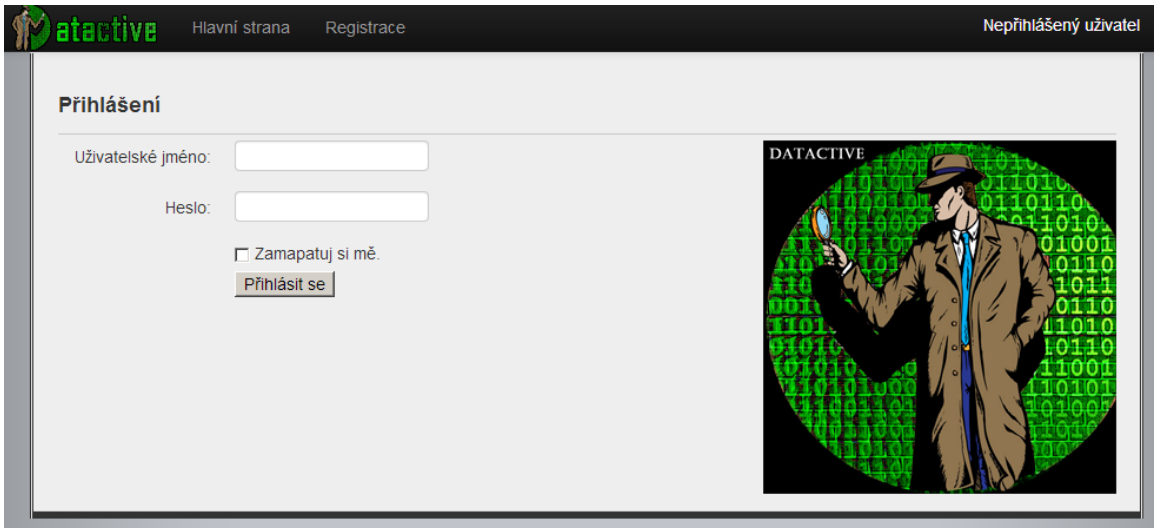
Protože fyzický přístup k serverovému stroji by byl nepraktický, byla mu do softwarové výbavy doinstalována aplikace **TeamViewer**, umožňující snadné vzdálené ovládání počítače a přenos souborů. Funguje na podobném principu jako vestavěná funkce operačního systému Windows s názvem „sdílená plocha“.

7.2 Funkční členění aplikace

Ovládání aplikace bude rozděleno mezi 6 základní části:

- Přihlašovací obrazovka,
- hlavní strana,
- zdroje,
- úkoly,
- výsledky,
- detaily výsledků.

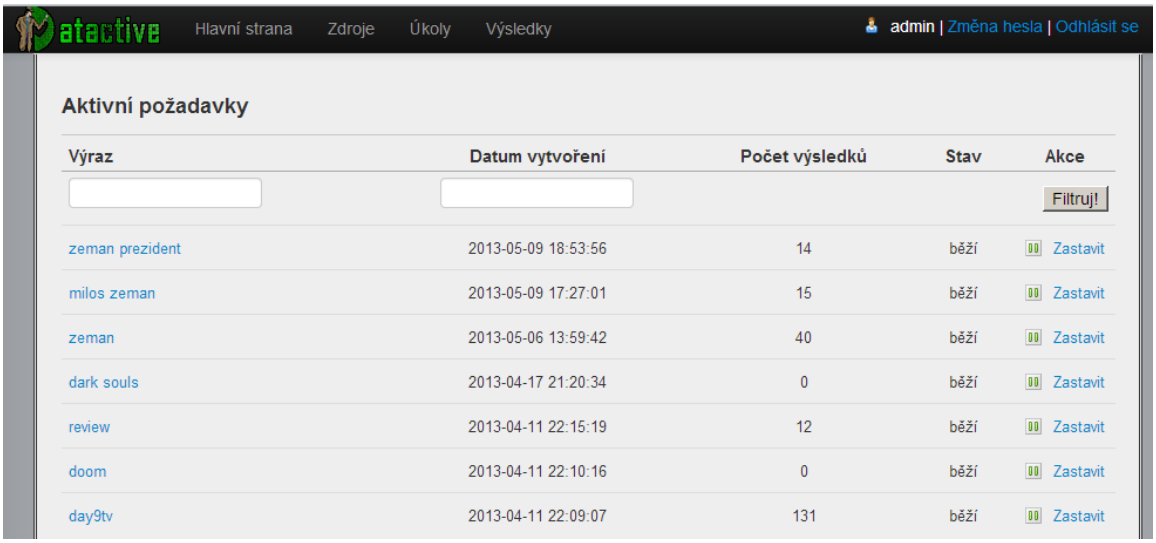
Po přistoupení na zdrojovou IP adresu serveru, na které je aplikace v prostředí internetu veřejně dostupná, se dostane uživatel do přihlašovací obrazovky sloužící ke vstupu do systému. Řešení je veskrze standardní, tedy přihlašovací jméno a heslo, jež musí odpovídat údajům zadaným při registraci. Pokud uživatel nemá dosud založený účet, musí nejdříve navštívit registrační sekci, dostupnou v horním ovládacím menu.



Obrázek 7.1: Přihlašovací obrazovka.

Registrace je řešena tradičně, tedy zadání jména a dvojité zadání hesla, přičemž obě hodnoty hesla musí odpovídat. Po odeslání formuláře je uživatel přesměrován zpět na přihlašovací obrazovku se vstupním formulářem. Zde se rovněž nachází zaškrtávací pole s možností zapamatování uživatele, což aplikace na pozadí řeší založením uživatelské *session* s delší trvanlivostí. Uživatel se pak nemusí ani po delší době nečinnosti opakovaně přihlašovat do systému. Zadávání uživatelského jména a hesla, stejně jako ostatních vstupů v aplikaci, prochází přes ochranu vstupu zprostředkovanou frameworkem Nette, aby se maximálně omezila možnost útoků typu *SQL injection*, *cross-site scripting* a podobně. Heslo se dále předává do hashovací funkce, kde se pomocí algoritmu *MD5* a takzvané *salt* vytváří jeho zakódovaná podoba, jež se nakonec společně s uživatelským jménem ukládá do databáze.

Po přihlášení se dostává uživatel na hlavní stranu aplikace, kde jsou umístěny pro něj nejdůležitější informace. Jedná se přehled aktivních hledání, jež na pozadí aplikace probíhají, a také aktuální počet nalezených výsledků. Jednotlivá hledání mají jako identifikační položku hledaný výraz, jenž zároveň slouží jako přímý odkaz na stránku s výsledky odpovídající danému hledání. Zbytek hlavní strany necháváme s výjimkou horního ovládacího menu a informace o aktivním přihlášení prázdné, abychom zajistili maximální přehlednost a uniformitu uživatelského rozhraní. Není však vyloučeno, že s přibývající funkcionalitou umístíme právě na tuto obrazovku nějaké další ovládací prvky, za účelem zvýšení efektivity a komfortu při používaní nejčastěji využívaných funkcí aplikace.



Obrázek 7.2: Hlavní obrazovka po přihlášení uživatele.

Druhou záložkou v horním ovládacím menu jsou „zdroje“. Jak již název naznačuje, zde probíhá vytváření, procházení a mazání zdrojů, z kterých aplikace při hledání čerpá. Vytváření nového zdroje je podle typu zdroje rozděleno na dvě části. Vytváření klasického RSS zdroje obsahuje tři základní položky, jež musí uživatel před odesláním formuláře vyplnit:

1. Název zdroje - klíčový údaj pro uživatele, protože podle něj lze zdroj hledat nebo filtrovat.

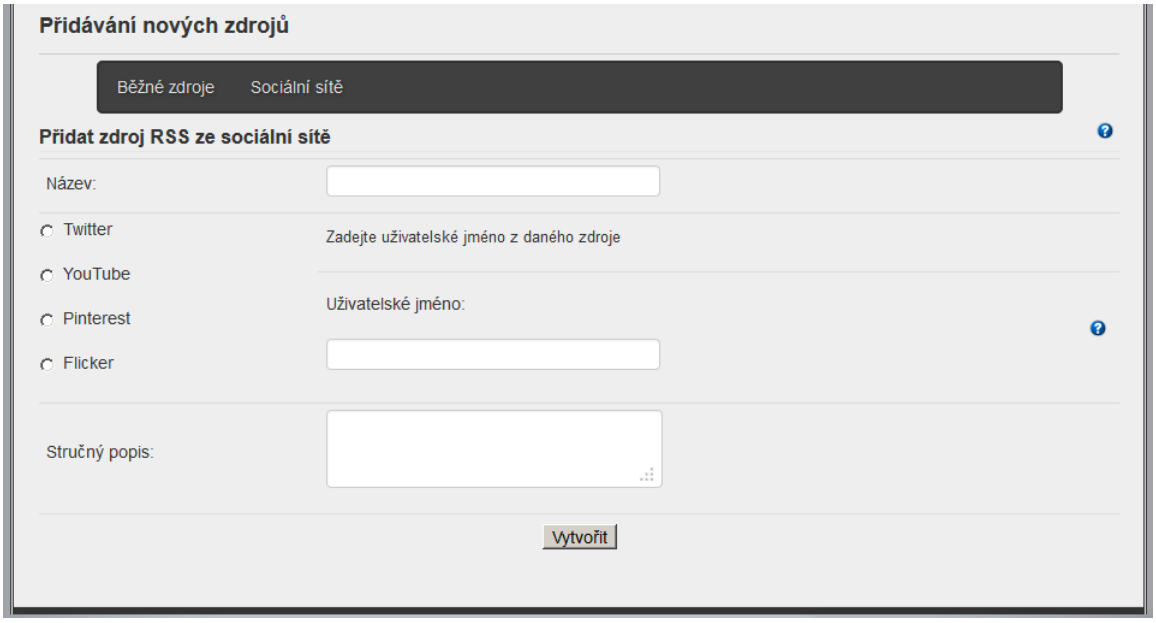
2. Adresa zdroje - nejdůležitější údaj z hlediska aplikačního. Bude využit pro případné hledání v daném zdroji. Je nutné aby byl tento údaj vyplněn správně, pokud má být hledání podle něj funkční. Validitu této hodnoty lze z programového hlediska jen

velmi špatně kontrolovat nebo ověřovat její platnost. Některé webové servery mohou mít dočasné výpadky, proto nemusí být v konkrétně danou chvíli jejich *URL* s RSS kanálem dostupná, ale časem se mohou projevit jako aktivní. Proto uživatel při špatně zadané adrese zdroje nezjistí problém okamžitě, ale upozorní ho nulová hodnota výsledků i po delší době hledání.

Aby nebylo nutné vyplňovat kompletní *URL* adresy, tedy hlavně část obsahující *http://*, tak nabízíme uživateli oba způsoby, tedy s i bez uvedeného protokolu, kdy si ho indexační skript v případě potřeby za běhu doplní. Počítáme samozřejmě i s možností protokolu *HTTPS*, pro které využíváme v nastavení serveru přídavný modul *OpenSSL*.

3. Stručný popis zdroje - tvořený textovým polem. Může obsahovat pro uživatele důležité informace v podobě konkretizace informací, jež se v daném zdroji vyskytují. Zároveň tvoří položku, podle které lze zdroje filtrovat, není proto vyloučena varianta, kdy si uživatel do popisu zdroje vyplní například svoje uživatelské jméno, aby mohl snadno vyfiltrovat velké množství zdrojů pouze na jím přidané.

Formulář pro sociální sítě skrývá odlišnou metodu, co se zadávání zdrojové adresy týče. Uživatel nemusí vědět, pod jakou adresou lze získat konkrétní kanál z vybrané sociální sítě. My mu poskytujeme výběr ze sociálních sítí a jedinou požadovanou informaci pak představuje jméno uživatele, podle kterého se vybere konkrétní RSS datový zdroj. Ve většině sociálních sítí tedy vytváříme RSS kanál z aktivit zadaného uživatele, kde potom hledáme klíčové slovo. Mezi aktivity patří například články vydané uživatelem v sociální síti. Zbývající položky, tedy název zdroje i popis, jsou řešeny stejně jako u varianty klasických zdrojů popsané v předešlém odstavci.



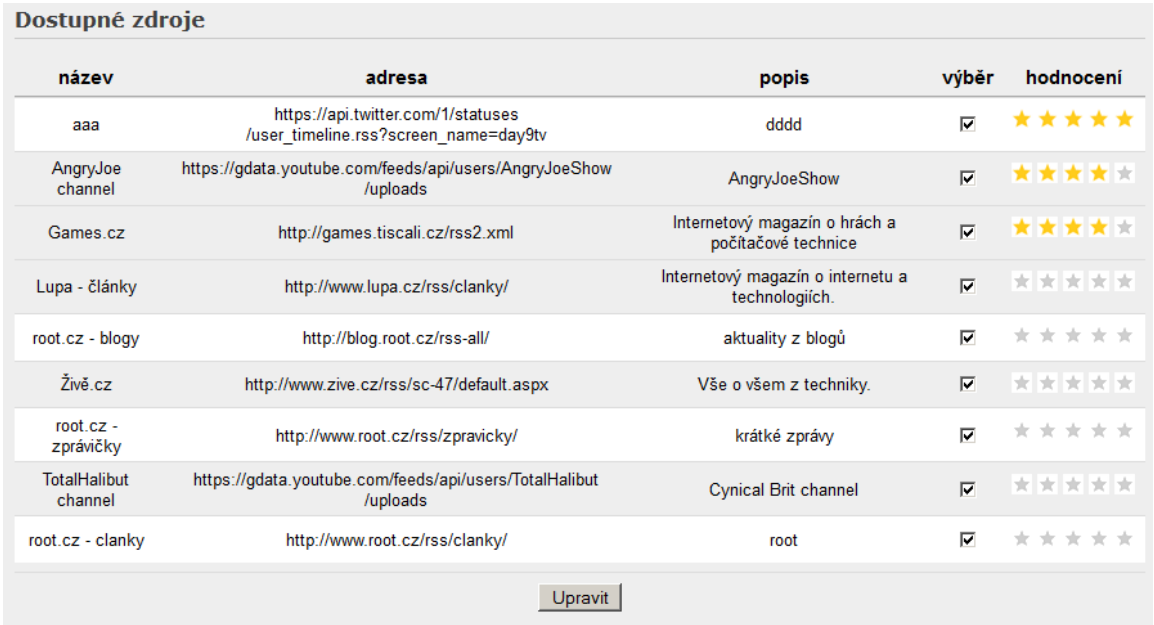
Obrázek 7.3: Formulář pro přidávání zdrojů ze sociálních sítí.

Dále jsou na stránce uvedeny dostupné zdroje s políčky pro příslušné filtrování. V tabulce jsou zobrazeny textové položky, tedy jméno zdroje, jeho adresa a popis. Popis je zkrácen, aby v případě rozsáhlejšího textu nezabíral příliš velký prostor ve vertikální linii

tabulky. U každého zdroje je navíc přidán aktivní prvek, který po stisknutí vyvolá dotaz pro smazání daného zdroje. Poslední položkou viditelnou u každého zobrazeného zdroje je jeho hodnocení ve formě hvězdiček. Jedná se o zaokrouhlenou hodnotu založenou na aritmetickém průměru uživatelského hodnocení výsledků, jež byly v daném zdroji nalezeny. Pod zdroji se nachází stránkování, aby bylo možné listovat ve velkém množství zdrojů. Popis implementace stránkování se nachází ve zvláštní sekci 7.3.4.

Třetí záložku s názvem úkoly bude uživatel využívat zřejmě nejčastěji, protože obsahuje pro něj nejdůležitější funkci celé aplikace. Tu tvoří možnost vytvářet úkoly na hledaní klíčových slov v uživatelem vybraných zdrojích. Základní dělení se opět týká volby, zda se bude vytvářet hledání v již existujících datových zdrojích, nebo zda se má vytvořit hledání jako dotaz (anglicky query) hledání v sociální síti. Taková možnost poté vrací výsledky ve formě RSS kanálu tak, jako by bylo zadáno do vyhledávacího pole přímo v prostředí sociální sítě. Ne všechny sítě tuto možnost poskytují, navíc bývá někdy skryta i pro běžného uživatele. U hledání v sociálních sítích tedy dáváme uživateli na výběr zdrojovou sociální síť a textové pole pro výraz, jež se bude vyhledávat. Hledání v běžných zdrojích poskytuje v prvním kroku pouze textové pole pro zadání klíčových slov. Po jeho vyplnění se uživatel přenese na další obrazovku, kde se mu nabízí výběr zdrojů. Zaškrtnutím příslušných zaškrtávacích polí se určí oblast hledání zadaného výrazu.

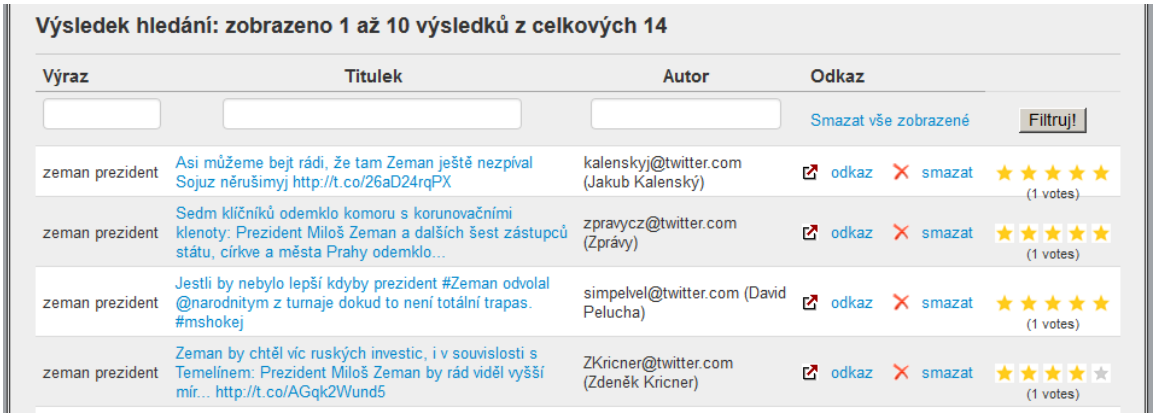
Spodní část obrazovky pak skýtá náhled na aktivní hledání, podobně jak je tomu na hlavní straně. Avšak na rozdíl od ní se zde nachází přídavné tlačítko s volbou editace. Vytvořená hledání tak lze upravovat ve formě hledaného klíčového slova nebo zdrojů, v nichž aplikace hledá. Toto nastavení je tvořeno stejným formulářem jako při vytváření klasického hledání, konkrétně v části s výběrem zdrojů. Jediný rozdíl spočívá ve faktu, že již použité zdroje jsou v základu zaškrtnuty a lze je tedy odškrtnutím z hledání i odebrat, jak znázorňuje obrázek 7.4.



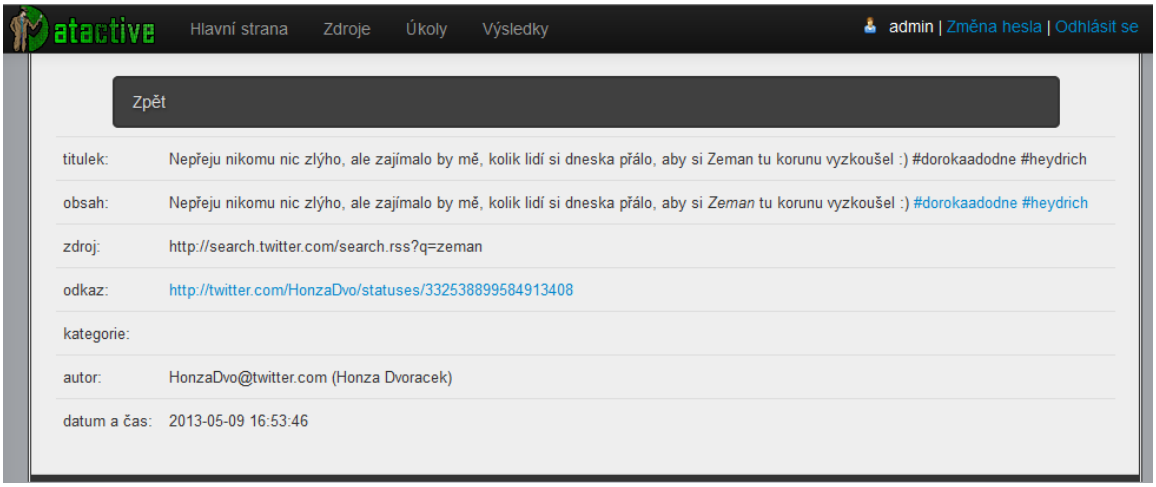
Obrázek 7.4: Ukázka výběru zdrojů pro cílové hledání.

Poslední záložka menu je vytvořena čistě pro účely zobrazování uživatelských výsledků hledání. Po kliknutí se uživateli zobrazí zpráva s informací o celkovém počtu výsledků asoci-

ovaných s jeho hledáními. Zároveň se mu předává informace o rozsahu aktuálně zobrazených výsledků z důvodu jednodušší orientace při velkém počtu výsledků. V záznamech lze také snadno filtrovat pomocí textových polí umístěných nad příslušnými sekcemi tabulky, v níž jsou výsledky zobrazeny. Jelikož se jedná o přehled výsledků, tvoří každý záznam pouze výřez nejdůležitějších položek, přičemž hlavičky záznamů slouží jako odkazy na podrobnější zobrazení. Ty jsou označeny jako detaily, kde se zobrazují výsledky kompletní. Nesmíme opomenout také možnost smazání výsledků buď po jednom, nebo po větším množství záznamů. Takto odstraněné výsledky se poté v databázi označí jako ignorované, tím pádem se znovu nepřidávají při opakovaných hledáních. Každý záznam má také přiřazenu hodnotící stupnici od jedné do pěti hvězd, jež jednak pomáhají při řazení výsledků a zároveň propagují uživatelské hodnocení k příslušným zdrojů, z kterých vzešly.



Obrázek 7.5: Obrazovka s nalezenými výsledky.



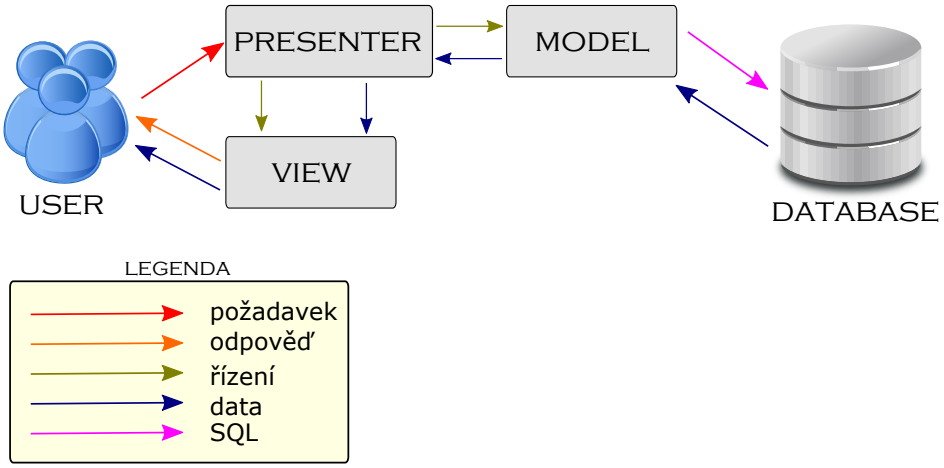
Obrázek 7.6: Detailní zobrazení příspěvku s odkazem na něj a příslušný zdroj.

7.3 Implementace konkrétních částí

V této části se zaměříme na konkretní řešení klíčových částí aplikace, včetně detailního objasnění využitého návrhového vzoru.

7.3.1 MVP

Návrhový vzor *Model View Presenter* (zkráceně MVP) je odvozený od dříve uvedeného *Model View Controler* (MVC). Hlavní oblast jeho využití se nachází ve vytváření uživatelských rozhraní. Na rozdíl od MVC se striktně odděluje komunikace mezi modelem a pohledem (šablonami). Pro tento účel se využívá rozšířená komponenta označovaná jako presenter. Presenter obsahuje veškerou aplikační a prezentační logiku vytvářeného systému. Využívá data z úložišť (v našem případě databáze), jejichž získávání obstarává část model. Následně je zpracovává do požadované podoby a nakonec předává do šablon (view), které se starají o vykreslení a zobrazení výstupu. Navíc oproti MVC se presenter stará i o obsluhu interaktivních prvků, obsahuje tedy funkce reagující na uživatelské vstupy. Schéma modelu včetně komunikačních kanálů jsme znázornili na obrázku 7.7.



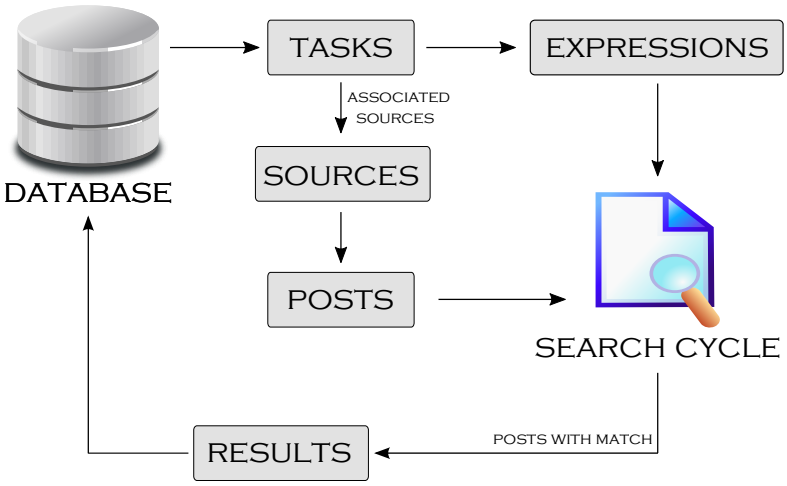
Obrázek 7.7: Schéma návrhového vzoru MVP.

Využití popisovaného návrhového vzoru v naší aplikaci je spojeno se zvoleným PHP frameworkem Nette, jež dané schéma využívá. Výhody plynoucí z uplatnění frameworku a návrhového vzoru při vývoji aplikace jsou popsány v kapitole 6.2.2.

7.3.2 Indexer

U implementace indexeru jsme se drželi původního návrhu. Jedná se tedy o jeden samostatný skript v jazyce PHP, vytvořený bez použití frameworku Nette. K tomuto kroku jsme přistoupili především díky jednoduchosti skriptu, kvůli kterému by se výkonově nevyplatilo zavádět celý framework. Protože skript běží na pozadí bez možnosti zásahu uživatele a využívá již ošetřená uživatelská data uložená v databázi, nehrozí vážnější bezpečnostní rizika, jež by mohla být s vývojem mimo framework spojená. Pro ilustraci funkčnosti indexeru uvádíme schéma 7.8.

Jak je znázorněno na schématu, indexer pracuje pouze s uživatelskými vstupy obsaženými v databázi. Na základě zadaných úkolů hledání se vytváří množina klíčových slov a množina příspěvků vyparsovaných ze zdrojů, ve kterých je zadáno hledání. Jako datové zdroje slouží webové RSS kanály na zadaných URL adresách. Záznamy z obou skupin dat následně vstupují do cyklu, kde jsou vzájemně porovnány pomocí příslušné funkce na základě regulárních výrazů. Pokud dojde ke shodě, je příspěvek uložen do databáze k ostatním výsledkům.



Obrázek 7.8: Schéma programové struktury indexeru.

Celá indexovací operace probíhá periodicky díky spouštění přes systémový nástroj *Plánovač úloh* (*Windows Scheduler*). Časová perioda je v současnosti nastavena na 1 hodinu, ale není vyloučeno, že se bude hodnota na základě naměřených výsledků měnit, aby byl poskytnut co nejlepší poměr pokrytí všech příspěvků novinek a systémového zatížení způsobeného indexací. Rovněž se uvažuje nad rozšířením ve formě většího počtu PHP skriptů, které by byly spouštěny v různých časových periodách a zaměřovaly se na různé datové zdroje.

7.3.3 Filtrování

K filtrování výsledků byla použita PHP funkce zprostředkovávající databázový dotaz. Jako parametry jí jsou předávány uživatelem vyplněné filtrovací fráze a na jejich základě se provádí hledání v databázi. Ukázku kódu lze vidět v příloze A. Nejsou tedy využity žádné zvláštní funkce nad standardní MySQL dotazy do databáze.

Po zavolání filtrování se předávají filtrovací parametry zadané uživatelem do modelu pracujícího s konkrétní tabulkou databáze, odkud se převádějí do odpovídající SQL syntaxe. Společně s nimi sr předávají informace pro stránkování, konkrétně hodnoty *limit* a *offset*. Bližší detaily ohledně stránkování popisujeme v následující sekci. Překreslování stránky podle zadaných filtrovacích parametrů pak využívá technologii AJAX. Aby nemusela být pokaždé překreslována celá, překreslují se pouze prvky spojené s hledáním a stránkováním. Filtrování je persistentní, přetrvává i při listování mezi stránkami filtrovaných výsledků, dokud uživatel nezadá jiné, nebo nesmaže stávající.

7.3.4 Stránkování

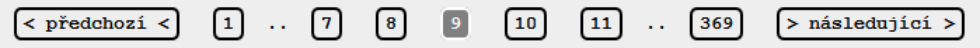
Aby bylo možné ve velkém množství výsledků nebo zdrojů snadno listovat, vytvořili jsme sofistikovaný systém stránkování. Z počátku se jevilo jako dostačující řešení, kdy bylo na výběr uživateli poskytnuto pořadí čísel představující jednotlivé stránky v řadě za sebou. Tento způsob však znesnadňoval orientaci pořadí, tedy která je aktuální, následující a předešlá strana. Proto byla přidána dvě k tomu uzpůsobená tlačítka pro pohyb zpět a vpřed.

Následně bylo přepracováno množství zobrazených stránek pro případ opravdu velkého množství výsledků. Výsledné řešení je vidět obrázcích 7.9, 7.10, 7.11 a 7.12.

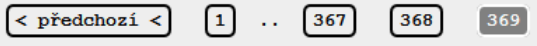
Stránkování funguje na technologii AJAX, překreslují se tak vždy pouze výsledky a stránkování, nikoliv celá obrazovka. Zároveň s příslušným odsazením výsledků v databázi se předávají i hodnoty filtrů, proto lze stránkovat i v aktuálně vyfiltrovaných záznamech.

Obrázek 7.9: Ukázka stránkování - začátek výsledků.

Obrázek 7.10: Ukázka stránkování - posouvání zpět a vpřed ve výsledcích.



Obrázek 7.11: Ukázka stránkování - posun oběma směry se zkratkami na začátek a konec.



Obrázek 7.12: Ukázka stránkování - poslední strana.

Kapitola 8

Testování, zhodnocení a možná rozšíření aplikace

Při testování jsme se zaměřili na výsledky vycházející z úkolů hledání. K testům byly použity různé scénáře s rozličnými klíčovými slovy, časy hledání i použitými zdroji. Výsledky dělíme do dvou hlavních kategorií podle doby hledání, aby bylo možné sledovat vliv různých kombinací parametrů na výsledky hledání. U každého z testovacích příkladů představujeme jeho scénář, aby bylo zřejmé jaký klíčový výraz a v jakém kontextu byl hledán. Následuje počet výsledků v testovaném časovém horizontu a rovněž zhodnocení výsledků. V některých případech přidáváme i doporučení, jak hledání vylepšit.

8.1 Krátkodobé výsledky

Následuje výčet provedených krátkodobých testovacích simulací:

1. Hledání klíčového slova „zeman“ v sociální síti Twitter jako hledací dotaz (*query*). Doba hledání 10 hodin.
   Výsledky: Nalezeno 181 výskytů klíčového slova zeman. Úmyslem bylo nalézt zmínky o prezidentu České republiky. V tomto kontextu bylo 16 výsledků relevantních.
   Zhodnocení: Výsledky byly z velké části nerelevantní, ale to se dalo očekávat vzhledem k zadanému klíčovému slovu a zdroji hledání. Po vyfiltrování se ukázalo, že kdyby namísto velmi obecného výrazu „zeman“ bylo zadáno například „prezident zeman“ nebo „miloš zeman“ , tak by se relevance výsledků drasticky zvýšila. Při uvedených pokusech byla úspěšnost více jak 15 z 20 záznamů relevantních. Druhým faktorem ovlivňujícím relevanci byl samotný zdroj, tedy sociální síť Twitter. Některé výsledky mohou působit jako relevantní, ale v řadách uživatelů existuje poměrně velké procento recesistů, kteří používají jména slavných osobností, jako je například Miloš Zeman. Když se však zaměříme na příspěvky jimi napsané, tak odhalíme, že se nejedná o hledanou osobu.

2. Hledání klíčového výrazu „VUT FIT“ v databázi sítě Picasa jako vyhledávací dotaz (*query*). Doba hledání 6 hodin.
   Výsledky: Nalezeno 172 výskytů klíčového výrazu. Cílem bylo nalézt fotografie areálu VUT FIT v Brně. V tomto kontextu kolem 100 výsledků relevantních.
   Zhodnocení: Většina výsledků hledání byla relevantní, i přes použití zkratky fakulty, která mohla být nositelem mnohoznačnosti. Mezi nerelevantní výsledky patřily

fotografie z akcí, například fakultního plesu, nebo jiných aktivit, jež se pořádali v souvislosti se zmiňovanou fakultou. Obvykle však byly takové výsledky spjaty s konkrétním autorem fotografií. Díky tomu nebylo těžké výsledky vyfiltrovat a nerelevantní výsledky smazat. Zároveň se ukázala nečekaná vlastnost výsledků z Picasa v tom, že jich bylo v jednom prohledávaném RSS souboru nezvykle velké množství, řádově stovky. Tento fakt vedl ke zpomalení získávání výsledků pomocí indexeru, především když se provádělo více hledání v síti Picasa najednou.

3. Hledání klíčového výrazu „NVIDIA TITAN“ v sociální síti YouTube jako vyhledávací dotaz (*query*). Doba hledání 6 hodin.
   Výsledky: Nalezeno 27 výskytů klíčového výrazu. Cílem bylo nalézt videa s ukázkami výkonu nejmodernější grafické karty dneška, NVIDIA GeForce GTX TITAN. V tomto kontextu bylo 25 nalezených výsledků relevantních.
   Zhodnocení: Nerelevantní výsledky se vyskytly pouze 2, a to video z takzvaného „unboxingu“ zmíněného produktu a trailer na hru, která dokáže nejlépe využít její výkon. Ostatní výsledky obsahovaly očekávaný obsah, tedy záběry z grafických aplikací a her při využití daného hardwaru.

4. Hledání klíčového výrazu „NVIDIA GTX TITAN“ v sociální síti Twitter jako vyhledávací dotaz (*query*). Doba hledání 6 hodin.
   Výsledky: Nalezeno 18 výskytů klíčového výrazu. Cílem bylo najít komentáře k nejmodernější grafické kartě NVIDIA GeForce GTX TITAN.
   Zhodnocení: Abychom zvýšili relevanci hledaného výrazu, bylo oproti minulému hledání přidáno slovo GTX, pro ještě přesnější zaměření dotazu. Bohužel relevantních nálezů bylo pouze 7, tedy méně než polovina, a to především proto, že velká část výsledků hledání byla v jiném než anglickém nebo českém jazyce. Zbytek obsahoval velmi nízkou informační hodnotu, proto byl zařazen taktéž mezi nerelevantní. Relevantní výsledky obsahovali výkonové grafy pro danou grafickou kartu, srovnání, cenu, tedy hledané a užitečné informace.

5. Hledání klíčového výrazu „Dark Knight“ v sociální síti YouTube jako vyhledávací dotaz (*query*). Doba hledání 6 hodin.
   Výsledky: Nalezeno 28 výskytů klíčového výrazu. Cílem bylo nalézt videa týkající se nové série filmů Batman.
   Zhodnocení: Relevantní byly téměř všechny nalezené výsledky. Nalezený obsah zahrnoval jak trailery, filmové záběry, momenty z natáčení, bonusové materiály, soundtracky a další hledaný materiál, a to i přes fakt, že hledaný výraz mohl působit mnohoznačně. Aktuálnost hledaného obsahu však způsobila, že výsledky obsahovali očekávané informace.

6. Hledání klíčového výrazu „T-34“ v sociální síti Flickr jako vyhledávací dotaz (*query*). Doba hledání 6 hodin.
   Výsledky: Nalezeno 19 výskytů klíčového výrazu. Cílem bylo nalézt fotografie sovětského tanku z druhé světové války s označením T-34.
   Zhodnocení: 8 výsledků bylo relevantních, zbylých 11 nebylo. Příčina se skrývala v označení tanku, které je shodné s označením letadla, které se nacházelo na zbytku fotografií. Výrazné zlepšení relevance by přineslo přidání klíčového slova tank, avšak s tak malým počtem výsledků by se to zřejmě v krátkodobém měřítku nevyplatilo. Nejlepší by bylo zvolit jinou zdrojovou sociální síť, například Picasa nebo YouTube.

7. Hledání klíčového výrazu „Eiffel“ v databázi sítě Picasa jako vyhledávací dotaz (*query*). Doba hledání 6 hodin.
   Výsledky: Nalezeno 424 výskytů klíčového výrazu. Cílem bylo nalézt obrázky a fotografie s Eiffelovou věží.
   Zhodnocení: Relevantní byla valná většina výsledků a hledání tak přineslo velkou zásobu fotografií Eiffelovi věže v různých denních dobách a ročních obdobích. Dotaz byl zřejmě dostatečně konkrétní a zároveň byla zvolena správná zdrojová síť. Již od počátku velký počet výsledků způsobuje výběr zdroje Picasa, který na druhou stranu značně vytěžuje výpočetní nároky serveru.

8. Hledání klíčového výrazu „semolina“ v databázi sítě Instagram jako vyhledávací dotaz (*query*). Doba hledání 6 hodin.
   Výsledky: Nalezeno 16 výskytů klíčového výrazu. Cílem bylo nalézt jídla a recepty z krupice.
   Zhodnocení: Všechny nalezené výsledky byly relevantní ve směru různých pokrmů z krupice. Receptů se vyskytlo sice malé množství, ale daly by se podle názvů pokrmů snadno dohledat klasickými internetovými vyhledávači. Hledání tak považujeme za úspěch, protože si jídlo lze vybrat podle lákavého vzhledu.

9. Hledání klíčového výrazu „DPRK“ v sociální síti Twitter jako vyhledávací dotaz (*query*). Doba hledání 6 hodin.
   Výsledky: Nalezeno 140 výskytů klíčového výrazu. Cílem bylo najít články a postřehy o Severní Koreji.
   Zhodnocení: Hledaný výraz je anglická obdoba českého KLDR (označení pro Severní Koreu jako zemi). Zhruba 40 výsledků bylo v cizí řeči (především Korejština), proto je vyřazujeme jako neplatné. Zbytek se týkal hledaného tématu především díky jednoznačnosti hledané zkratky. Nechtěné výsledky bylo možné snadno vyfiltrovat podle pole autora.

10. Hledání klíčového výrazu „KLDR“ v sociální síti Twitter jako vyhledávací dotaz (*query*). Doba hledání 6 hodin.
    Výsledky: Nalezeno 16 výskytů klíčového výrazu. Cílem bylo najít články a postřehy o Severní Koreji v češtině.
    Zhodnocení: 15 výsledků bylo relevantních. Příspěvky byly všechny v češtině a týkali se hledaného tématu. Menší počet výsledků je způsoben především omezením na češtinu.

11. Hledání klíčového výrazu „KLDR“ ve 12 zvolených českých zpravodajských denících. (*query*). Doba hledání 6 hodin.
    Výsledky: Nalezeny 4 výskyty klíčového výrazu. Cílem bylo najít články a postřehy o Severní Koreji v češtině ze zvolených médií.
    Zhodnocení: Všechny nalezené výsledky byly relevantní, což se dalo očekávat vzhledem ke zvoleným zdrojům. Malé množství výsledků je přímo ovlivněno světovým děním v čase hledání, protože zpravodajské deníky obvykle zachytávají pouze aktuální dění. Z těchto zdrojů proto doporučujeme nechávat hledání po delší časový úsek, pokud nám jde o zachycení nějakého dlouhodobějšího vývoje situace.

8.2 Výsledky trvalejšího běhu

1. Hledání klíčového slova „review“ ve zdroji ze sociální sítě YouTube. Konkrétně se jednalo o recenzenta a komentátora počítačových her, uživatele pod přezdívkou AngryJoe. Doba hledání byla 24 dní.
   Výsledky: Nalezeno 12 výsledků, z toho všech 12 relevantních.
   Zhodnocení: Zde je vidět odlišné využití aplikace, tedy hledání konkrétní informace na jednom konkrétním zdroji. Jelikož bylo ještě před hledáním známo, že zdrojový uživatel vždy striktně pojmenovává vytvářené recenze klíčovým slovem „review“ , bylo tak dosaženo maximální přesnosti. Všechny od doby vytvoření nalezené články kompletně pokrývají oblast hledání. Díky delším intervalům mezi vytvářenými články bylo možné průběžně monitorovat, zda se od poslední návštěvy aplikace neobjevily některé nové, a tak se suplovala funkce odebírání RSS novinek z konkrétního zdroje. Kdyby však existovalo více zdrojů, od kterých by chtěl uživatel odebírat všechny články s recenzemi, nebyl by problém kdykoliv v průběhu požadované zdroje přidat přes editací existujícího hledání.

2. Hledání klíčového výrazu „VUT FIT“ v databázi sítě Picasa jako vyhledávací dotaz (*query*). Doba hledání 3 dny.
   Výsledky: Nalezeno 191 výskytů klíčového výrazu. Cílem bylo nalézt fotografie areálu VUT FIT v Brně. V tomto kontextu bylo asi 110 výsledků relevantních.
   Zhodnocení: Výsledky hledání se od krátkodobé varianty téměř nezměnily. Mezi nové relevantní výsledky patří například snímky nově budovaného křídla školy, tedy poměrně aktuální informace. Pro sledování pouze nových dat na serveru Picasa proto doporučujeme promazat výsledky z první vlny hledání, která je oproti ostatním sítím nestandardně obsáhlá a poté lépe vyniknou nově přidané záznamy v sociální síti.

3. Hledání klíčového výrazu „NVIDIA TITAN“ v sociální síti YouTube jako vyhledávací dotaz (*query*). Doba hledání 3 dny.
   Výsledky: Nalezeno 56 výskytů klíčového výrazu. Cílem bylo nalézt videa s ukázkami výkonu nejmodernější grafické karty dneška, NVIDIA GeForce GTX TITAN. V tomto kontextu bylo 50 výsledků relevantních.
   Zhodnocení: Ve srovnání s krátkodobým hledáním se počet výsledků zdvojnásobil. Opět je valná většina výsledků relevantních. Nové nerelevantní články se týkaly chystaného herního zařízení NVIDIA Shield, které kromě značky NVIDIA nesdílí žádnou jinou podobnost s hledanou grafickou kartou. Tyto výsledky šlo snadno vyfiltrovat a smazat. YouTube tedy přináší pravidelnou dávku příspěvků s novinkami na hledané téma.

4. Hledání klíčového výrazu „NVIDIA GTX TITAN“ v sociální síti Twitter jako vyhledávací dotaz (*query*). Doba hledání 3 dny.
   Výsledky: Nalezeno 75 výskytů klíčového výrazu. Cílem bylo najít komentáře k nejmodernější grafické kartě NVIDIA GeForce GTX TITAN. V tomto kontextu bylo 45 výsledků relevantních.
   Zhodnocení: V tomto případě hledání vidíme velké zlepšení relevantnosti oproti krátkodobému scénáři hledání stejného klíčového výrazu. Mezi nově získané relevantní výsledky patří recenze, srovnání a další informace týkající se hledané grafické karty, stejně jako často komentovaný úspěch prodeje. Právě tento fakt ukazuje, že sociální síť Twitter dokáže při dlouhodobějším hledání jasně upozorňovat na důležité aktu-

ální dění, kdy jedna oficiální zpráva přichází přeposílána z velkého množství zdrojů. Mezi nerelevantní články se řadili většinou příspěvky psané v cizích jazycích, což je způsobeno samotným hledáním, které nespecifikuje jazykovou oblast. Tento fakt dává námět k zamyšlení nad možným rozšířením aplikace ve směru ověřování zdrojového jazyka prohledávaného zdroje, alespoň v rámci angličtiny nebo češtiny.

5. Hledání klíčového výrazu „Dark Knight“ v sociální síti YouTube jako vyhledávací dotaz (*query*). Doba hledání 3 dny.
   Výsledky: Nalezeno 52 výskytů klíčového výrazu. Cílem bylo nalézt videa týkající se nové série filmů Batman. V tomto kontextu bylo 45 výsledků relevantních.
   Zhodnocení: Relevantní byly opět téměř všechny nalezené výsledky. Celkový výsledek tak odpovídá krátkodobému hledání, jenom s vyšším počtem výsledků, které se někdy začaly opakovat (např. trailer k filmu). Nerelevantní výsledky se obvykle týkaly přidružených projektů a filmů, kde se taktéž vyskytuje Batman jako jedna z postav.

6. Hledání klíčového výrazu „T-34“ v sociální síti Flickr jako vyhledávací dotaz (*query*). Doba hledání 3 dny.
   Výsledky: Nalezeno 22 výskytů klíčového výrazu. Cílem bylo nalézt fotografie sovětského tanku z druhé světové války s označením T-34. V tomto kontextu bylo 9 výsledků relevantních.
   Zhodnocení: Ve srovnání s krátkodobým během hledání přibyly 3 výsledky, z nichž byl pouze 1 relevantní a obsahoval fotografii hledaného tanku. Pro lepší výsledky bychom doporučili změnu klíčového výrazu na „T-34-85“ , což je modifikace téhož tanku, která však zamezuje mnohoznačnosti hledaného výrazu (20 z 20 příspěvků relevantních). Frekvence nových výsledků je však velmi nízká, což se dá očekávat vzhledem k historickému tématu hledání. Možná by se vyplatila i změna zdrojové sociální sítě, protože reference na hledaný tank lze s jistotou nalézt v oblasti počítačových her, kam Flicker nezasahuje.

7. Hledání klíčového výrazu „Eiffel“ v databázi sítě Picasa jako vyhledávací dotaz (*query*). Doba hledání 3 dny.
   Výsledky: Nalezeno 509 výskytů klíčového výrazu. Cílem bylo nalézt obrázky a fotografie s Eiffelovou věží.
   Zhodnocení: Relevantní byla opět valná většina výsledků a hledání přineslo velkou zásobu fotografií Eiffelovi věže v různých denních dobách a ročních obdobích. Na rozdíl od textově orientovaného hledání počítáme mezi relevantní i příspěvky v různých jazycích, protože nám jde v první řadě o fotografie, nikoliv o přiložený text. Počet nových příspěvků oproti krátkodobému hledání citelně vzrostl, což zřejmě způsobuje popularita hledané objektu.

8. Hledání klíčového výrazu „semolina“ v databázi sítě Instagram jako vyhledávací dotaz (*query*). Doba hledání 3 dny.
   Výsledky: Nalezeno 34 výskytů klíčového výrazu. Cílem bylo nalézt jídla a recepty z krupice. V tomto kontextu bylo asi 25 výsledků relevantních.
   Zhodnocení: Nerelevantní výsledky buď nesouvisely s hledaným tématem nebo neposkytovalx komentář k jídlu v anglickém či českém jazyce.

9. Hledání klíčového výrazu „DPRK“ v sociální síti Twitter jako vyhledávací dotaz (*query*). Doba hledání 3 dny.

Výsledky: Nalezeno 1075 výskytů klíčového výrazu. Cílem bylo najít články a postřehy o Severní Koreji. Zhruba dvě třetiny výsledků bylo relevantních.
Zhodnocení: Oproti krátkodobému hledání počet výsledků velmi výrazně vzrostl, což způsobuje především výběr zdrojové sociální sítě. Velké množství článků komentuje konkrétní dění, například odpal 3 severokorejských raket. Taková informace se pak lavinově šíří celou sítí, avšak na druhou stranu lze najít i slušné množství názorů a komentářů k takovému dění, proto je počítáme mezi relevantní výsledky. Nerelevantní výsledky byly obvykle v cizí řeči (japonština, korejština, apod.).

10. Hledání klíčového výrazu „KLDR“ v sociální síti Twitter jako vyhledávací dotaz (*query*). Doba hledání 3 dny.
Výsledky: Nalezeno 32 výskytů klíčového výrazu. Cílem bylo najít články a postřehy o Severní Koreji v češtině. Zhruba 23 výsledků bylo v daném kontextu relevantních.
Zhodnocení: Na rozdíl od krátkodobého hledání se vyskytlo poměrně hodně výsledků v angličtině, které zpravidla nebyly relevantní. Ve srovnání s anglickým hledáním (klíčový výraz DPRK) je vidět velký nepoměr ve výsledcích, což značí jasnou jazykovou orientaci sociální sítě Twitter na angličtinu. Často lze spatřit i české autory, kteří přispívají v angličtině. Doporučujeme proto tvořit dotazy hledání v této sociální síti anglicky.

11. Hledání klíčového výrazu „KLDR“ ve 12 zvolených českých zpravodajských denících. (*query*). Doba hledání 3 dny.
Výsledky: Nalezeno 5 výskytů klíčového výrazu. Cílem bylo najít články a postřehy o Severní Koreji v češtině ze zvolených médií. Všechny výsledky byly relevantní.
Zhodnocení: Malý počet výsledků je způsoben faktem, že české zpravodajské deníky častěji využívají označení Severní Korea než zkratku KLDR. Doporučujeme proto u podobných případů používat několik paralelních hledání s různými klíčovými slovy pro pokrytí všech relevantních příspěvků.

8.3 Zhodnocení aplikace a možná rozšíření

Podařilo se nám vytvořit aplikaci s následující charakteristikou:

- Veřejně dostupná a použitelná webová aplikace.
- Automaticky zpracovává a agreguje velké množství zdrojů z internetu do jednotného toku, v kterém lze jednotně vyhledávat klíčová slova nebo fráze, zobrazovat jejich detailní informace nebo přistupovat k jejich zdrojům.
- Aktivně využívá jak klasické RSS zdroje, tak jejich varianty dostupné ze sociálních sítí.
- Obsahuje možnost hodnocení výsledků s přímým vlivem na hodnocení a řazení datových zdrojů, za účelem zvýšení relevance hledaných informací.
- Sbírá statistická data, využitelná pro následná vylepšení vzhledem k použitelnosti aplikace.
- Ukládá hodnocení a aktivity související s datovými zdroji, především z důvodů plánovaných rozšíření směrem ke komplexnímu hodnocení datových zdrojů na internetu.

8.3.1 Stručné shrnutí

Typ aplikace zvolený a následně implementovaný je webová aplikace, především kvůli možnostem interoperability mezi různými typy zařízení, stejně jako jednoduchému sdílení nejnovějších aktualizací a obsahu mezi jednotlivými uživateli. Při implementaci byl použit PHP framework Nette a návrhový vzor MVP pro lepší čitelnost a budoucí rozšiřitelnost aplikace. Hostitelský server byl umístěn na zvláštní stroj sestavený pro účely vývoje a testování aplikace, abychom se vyhnuli omezení ze strany hostingu třetí strany. Po zvážení četných výhod byl za výchozí datový zdroj zvolen formát RSS. Jeho hlavní výhodu představuje možnost integrace datových kanálů z různých zdrojů včetně sociálních sítí. Prohledávání zdrojových URL adres RSS dokumentů probíhá periodicky pomocí zvláštního PHP skriptu, tzv. indexeru. Děje se tak automaticky na pozadí aplikace díky spouštění přes Správce úloh operačního systému Windows. Uživatelské rozhraní je jednoduše a funkčně zařízeno s využitím HTML, CSS a JavaScript. Při jeho implementaci jsme se soustředili především na ovladatelnost a použitelnost. Datové zdroje lze snadno procházet, vytvářet nebo mazat. Při vytváření úkolu hledání se přehledně zobrazuje nabídka zdrojů, v kterých lze hledání volit, včetně varianty hledání ve větším množství zdrojů najednou. Výsledky hledání lze snadno zobrazovat, listovat v nich a zadávat jim hodnocení. Na základě hodnocení se řadí sestupně jak výsledky tak zdroje, v kterých byly nalezeny. Sbírané statistické údaje lze využít k budoucím rozšířením směrem ke komplexním hodnocením datových zdrojů na internetu.

8.3.2 Možná rozšíření

Jak již bylo dříve naznačeno, současný stav aplikace, aniž bychom chtěli jeho roli nějak podceňovat, představuje základ, který se bude pravděpodobně upravovat nebo integrovat do pokročilejšího systému. Současné zaměření se soustředí hledání informací z velkého množství zdrojů na internetu. Kromě toho se nám představují 2 hlavní směry, kudy se mohou rozšíření ubírat.

1. Hlubší dolování dat ze sociálních sítí.
2. Komplexní hodnocení zdrojů dat.

V současné době je aplikace schopna sledovat pouze omezené množství sociálních sítí dostupných na internetu, přičemž jsme omezeni pouze na data ve formátu RSS. Možnosti získávání datových kanálů se mezi jednotlivými sociálními sítěmi velice liší a někde chybí úplně. Jako příklad uveďme sociální síť Facebook, kde lze vytvořit RSS kanál z aktivit uživatele pouze v případě, že patříte do skupiny přátel daného uživatele. Pro ověření identity se pak využívá zvláštního tokenu, který většinou nelze jiným než uvedeným způsobem získat. Tento fakt omezuje naší aplikaci přístup k dolování dat z Facebooku. Řešení však mohou přinést doplňkové aplikace, v případě Facebooku jsou to *OpenGraph* nebo *Graph API*, které dovolují hledat v cílových databázích a v některých případech dokonce umožňují i získaní potřebných tokenů, bez kterých nelze danou operaci provést. Podobných aplikací nebo rozšíření by se k příslušným sociálním sítím zajisté dalo najít více a jejich ovládnutí by mohlo pomoci při rozšiřování portfolia sociálních sítí použitelných pro účely datových zdrojů. Nicméně bylo by nutné prověřit tyto možnosti pro každou sociální síť individuálně, poté najít vhodný způsob implementace a postup, jakým předávat ze strany uživatele parametry hledání. Rozhodně se nejedná o triviální směr vývoje, na druhou stranu se dolování dat ze sociálních sítí jeví jako lukrativní oblast, jak z hlediska zájmu uživatelů, tak z obsahové a statistické hodnoty takto získaných dat, včetně jejich případného ekonomického

potenciálu. Výsledkem by pak byl komplexní nástroj pro dolování dat a statistických informací ze sociálních sítí, kde by spíše než konkrétní hledání zmínek o hledaném výrazu, byly hlavním výstupem přidružené statistické informace o počtech a frekvencích výsledků, autorech, jazycích, a dalších doprovodných informacích souvisejících s hledáním. Ty by poté bylo možné zpracovat pomocí vhodných statistických metod a přiřadit jim konkrétní informační hodnotu, například z oblasti trendů dění ve vybraných zdrojích.

Druhou větev představuje cesta komplexního hodnocení zdrojů na internetu. V tomto případě by se vynaložila značná energie na rozšíření možností hodnocení výsledků, například zadávání konkrétních důvodů pro dané hodnocení či smazání záznamu, hodnocení uživatelských reakcí spojených s výsledkem, jako je počet kliknutí na odkaz vedoucí ke zdrojovému článku, následné činnosti, atd. Bylo by zapotřebí rozšířit zázemí spojené s hodnocením, aby bylo odlišitelné hodnocení konkrétního uživatele a jeho preferencí od globálního ratingu. Hlavním cílem by pak bylo vyloučit zdroje, jež poskytují nerelevantní nebo nesprávná data, a naopak ocenit vysokým hodnocením zdroje poskytující ověřené, reálné a přesné informace. S tím se pojí požadavek na potřebu velkého množství uživatelů, aby mohli být výsledky hodnocení považovány za směrodatné. Ideální by v takovém případě byla možnost sdílení úkolů hledání a jejich výsledků mezi jednotlivými uživateli s možnostmi vzájemného hodnocení. V praxi by pak celá aplikace připomínala vlastní druh sociální sítě, kde by spolu mohli uživatelé vzájemně interagovat, komunikovat, vytvářet skupiny a především vytvářet velká množství hodnocení, kterým by pak mohla být přidělena jasná vypovídající hodnota.

Kapitola 9

Závěr

Ústředním cílem práce bylo navrhnout a implementovat aplikaci, která bude s využitím automatizovaného zpracování a agregace dat z webového prostředí umožňovat hledání a přehledné zobrazovaní relevantních informací vzhledem ke zvoleným klíčovým slovům. Pro tyto účely jsme prostudovali nejčastější datové formáty v prostředí internetu a vybrali z nich ten nejvhodnější, kterým se stal formát RSS. Pro tento formát byly nalezeny využitelné datové kanály jak z oblasti webových portálů, tak z různých sociálních sítích. Následně byl vytvořen návrh webové aplikace s možnostmi indexace velkého množství RSS zdrojů na základě zadaných klíčových slov. Zaměřili jsme se především na agregaci různých datových kanálů do jednotného toku umožňujícího unifikovaný přístup ke hledání.

Od návrhu jsme se přenesli k implementaci aplikace v programovacím jazyce PHP s využitím frameworku Nette a návrhového vzoru MVP (7.3.1). Aplikace je schopná spravovat různé datové zdroje, včetně zdrojů ze sociálních sítí, a hledat v nich za pomoci vytvořeného indexačního skriptu, který je periodicky spouštěn na pozadí webového serveru. Hledání tak neprobíhá okamžitě, ale po delší časový úsek, kdy jsou uživatelům předkládány průběžné výsledky hledání. Volba zdrojů určených k prohledávání je plně v režii uživatele, včetně možnosti přidávání vlastních definovaných zdrojů, čímž se značně zvyšuje relevance výsledků. Zdroje se mezi uživateli sdílí, aby byla poskytnuta co nejširší datová základna pro hledání. Zároveň je dostupná možnost hodnocení výsledků a propagace této informace ke zdroji, odkud byl výsledek získán. Hodnocení tak má hodnotu jak pro uživatele, jimž může pomoci při výběru kvalitních datových zdrojů, tak pro dlouhodobější vývoj aplikace a plánovaná rozšíření.

Aplikace je v současnosti spuštěna na vlastním webovém serveru, kde se provádělo testování, které přineslo uspokojivé výsledky. Možnosti využití aplikace představují značný potenciál, ať již jde o sledování novinek z oblíbených serverů, hledání konkrétní informace napříč spektrem datových zdrojů, sledování aktuálních trendů ve vybraných zdrojích nebo dolování informací ze sociálních sítí. Zároveň je důležité zdůraznit, že se jedná o první fázi vývoje, protože plánujeme aplikaci dále rozšiřovat směrem k možnostem komplexnějšího hodnocení datových zdrojů na internetu.

Závěrem bych dodal, že jedná o velice zajímavé téma, ve kterém bych chtěl v budoucnu nadále pokračovat. Z osobního hlediska mi přinesla práce spoustu praktických zkušeností spojených s vývojem webové aplikace, stejně jako znalosti v oblasti dolování dat z internetu. Zároveň jsem rád vytvářel aplikaci, jež jeví potenciál pro vylepšení uživatelského přístupu k získávání informací na internetu.

Literatura

[1] Agarwal, A.: RSS Feeds Directory for Facebook, YouTube, Pinterest and More. http://www.labnol.org/internet/rss-feeds-directory/21242/, 2013.

[2] Bing, L.: *Web Data Mining*. Springer, 2011, ISBN 978-3-642-19459-7.

[3] Bonchi, F.; Castillo, C.; Gionis, A.; aj.: Social Network Analysis and Mining for Business Applications. *ACM Trans. Intell. Syst. Technol.*, 2011, ISSN 2157-6904.

[4] Cover, R.: Atom Publishing Format and Protocol. http://xml.coverpages.org/atom.html, 2005.

[5] Cover, R.: RDF Rich Site Summary (RSS). http://xml.coverpages.org/rss.html, 2007.

[6] Domingos, P.: Mining social networks for viral marketing. http://homes.cs.washington.edu/ pedrod/papers/iis04.pdf, 2005.

[7] Group, P. D.: PHP: SimpleXML. http://php.net/manual/en/book.simplexml.php, 2013.

[8] Kishore S. Swaminathan: What price information? 2011 [cit. 2012-10-16].

[9] Šlapák, O.: Data, informace, znalosti. *Electronic journal for philosophy*, 2003, ISSN 1211-0442.

[10] Laurent, W.: The Realities of Social Media Data Mining. 2011-03-14 [cit. 2012-10-16].

[11] Nayak; Richi: *XML Data Mining: Process and Applications*. Idea Group Inc. / IGI Global, 2008.

[12] Netcraft: April 2013 Web Server Survey. http://news.netcraft.com/archives/2013/04/02/april-2013-web-server-survey.html, 2013.

[13] Scime, A.: *Web mining*. Idea Group Publishing, 2005, ISBN 15-914-0414-2.

[14] Tiwari; Rajnishand; Buse; aj.: From Electronic to Mobile Commerce: Opportunities Through Technology Convergence for Business Services. http://www.global-innovation.net/publications/PDF/APTM2006.pdf, 2006.

Příloha A

Ukázky zdrojového kódu

```
<?php
/*
 * Summary:     Get results from database with following parameters
 * Parameters:  $user_id - id of current user
 *              $limit - count of results
 *              $offset - result offset
 *              $exp - expression filtering results
 *              $title - title filtering results
 *              $author - author filtering results
 */
  public function findUserResults($user_id, $limit, $offset,
          $exp, $title, $author)
  {
    return $this->database->query("SELECT R.*, T.exp as exp
      FROM results R, tasks T
      WHERE R.task_id = T.id AND T.user_id = ? AND R.ignore = 0 AND
        T.exp LIKE ('%$exp%') AND R.title LIKE ('%$title%')
        AND R.author LIKE ('%$author%')
      ORDER BY R.time DESC LIMIT ? OFFSET ?", $user_id, $limit, $offset);
  }
>
```