Filozofická fakulta Univerzity Karlovy v Praze

Ústav informačních studií a knihovnictví

Studia nových médií

Diplomová práce

Bc. Martin Kubelka

Datamining sociálních sítí

Datamining of Social Networks

Praha 2012

Vedoucí práce: Mgr. Josef Šlerka

Poděkování

Na tomto místě bych rád poděkoval zejména mému vedoucímu práce Josefu Šlerkovi, jak za jeho odborné vedení a cenné připomínky k této práci, tak za jedinečnou možnost se s ním podílet na výzkumech a analýzách týkajících se samotného dataminingu v oblasti sociálních sítí.

Dále bych samozřejmě rád poděkoval své rodině a nejbližším, kteří mi připravili vynikající zázemí a podpůrný servis, bez kterého bych nebyl schopen tuto práci vůbec realizovat.

Čestné prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 1. 1. 2012

…….

Abstrakt

Práce se zabývá využitím dataminingových metod a jejich aplikací na data získaná z prostředí sociálních sítí, potažmo sociálních médií. Představuje základní principy, na kterých tato sociální média staví. Hlavní důraz je pak kladen především na vysoký informační potenciál využití dat ze sociálních sítí. To je v této práci demonstrováno na několika vybraných dataminingových metodách, zejména pak Social Network Analysis a Sentiment Analysis. Další přístupy a možnosti využití těchto dat jsou zmíněny v kapitole o Social Media Monitoringu. Tyto kapitoly jsou též doplněny několika praktickými ukázkami a vlastními výzkumy. V poslední kapitole jsou uvedeny vize a hrozby, které může datamining sociálních sítí přinést do budoucna.

Klíčová slova

Vytěžování dat, dolování dat, sociální sítě, sociální média, analýza sociálních sítí, analýza sentimentu, monitorování sociálních médií

Abstract

This paper is about application of various data mining methods in social networks and social media area. It reveals basic principles of social media with the aim to high information potential of usage of the data from social networks. This is demonstrated on selected data mining methods, especially Social Network Analysis and Sentiment Analysis. Other opportunities of using social media data are shown in chapter about Social Media Monitoring tools. All these chapters are supplemented by practical examples and particular researches. Last chapter reveals visions and threats, which can bring data mining in the future.

Keywords

Data mining, social networks, social media, social network analysis, sentiment analysis, social media monitoring

Bibliografický záznam

KUBELKA, Martin. *Datamining sociálních sítí*. Praha, 2012. Diplomová práce. Univerzita Karlova v Praze, Filozofická fakulta, Ústav informačních studií a knihovnictví, Studia nových médií. Vedoucí práce Mgr. Josef Šlerka.

Obsah

1 Základní teze

1.1 Co jsou to vůbec ta data?

Pro základní pochopení celého problematiky dataminingu si musíme položit zásadní otázku, co jsou to data a vymezit si tak pojem, který nás bude provázet celou touto prací. Pro snadné pochopení si lze vypůjčit velice srozumitelnou a použitelnou definici z Wikipedie.

Data je výraz pro údaje, používané pro popis nějakého jevu nebo vlastnosti pozorovaného objektu. Data se získávají měřením nebo pozorováním. Lze je dělit na data spojitá a data atributivní. Data spojitá se přitom vztahují k nějaké spojité stupnici, zatímco data atributivní nikoliv.

Data jsou:

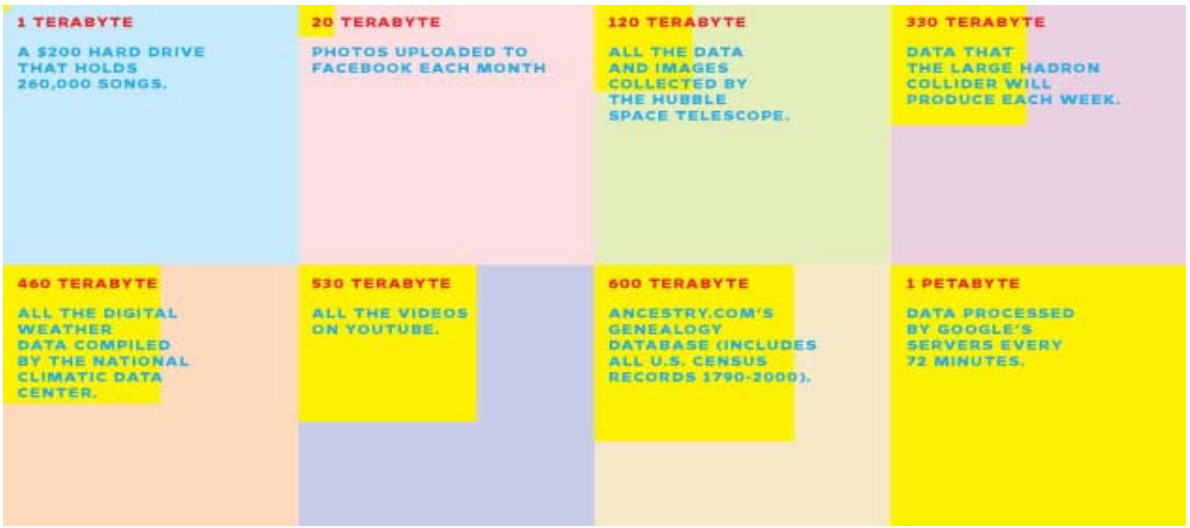
- vyjádření skutečností formálním způsobem tak, aby je bylo možno přenášet nebo zpracovat (např. počítačem)
- číselné nebo jiné symbolicky vyjádřené (reprezentované) údaje a hodnoty nějakých entit nebo událostí
- jakékoliv fyzicky (materiálně) zaznamenané znalosti (vědomosti), poznatky, zkušenosti nebo výsledky pozorování procesů, projevů, činností a prvků reálného světa (reality)
- surovina, z níž se tvoří informace

Data se zejména v humanitních vědách dále dělí na[1]

- tvrdá data, jasně definovaná a obvykle zatížená menší chybou (např. údaje o počtu obyvatelstva).
- měkká data, vyjadřující názory a postoje lidí (např. údaje o oblibě prezidenta republiky z dotazníkového šetření)[2].

[1] Tohoto rozdělení využijeme například v kapitole Social Media Monitoringu v části věnované korelaci měkkých dat ze sociálních sítí a tvrdých dat z exaktních statistik. Této problematice se bude věnovat i jeden z výzkumů, který tato práce obsahuje.
[2] Data. In *Wikipedia : the free encyclopedia* [online].

Žijeme v době, kdy více jak dvě miliardy uživatelů internetu[3] po celém světě vyprodukují každou vteřinou do nedávna nemyslitelné objemy dat. O tomto fenoménu se zmiňuje teoretik Nových médií Lev Manovich ve své studii „Trending: The Promises and the Challenges of Big Social Data“ [4], který vychází ve svých úvahách například z online časopisu Wired[5], který v roce 2008 otevřel novou sekci „The Petabyte Age“[6] („Věk petabytů“) nebo z časopisu The Economist, který v roce 2010 vydal zprávu „Data, data everywhere“ („Data, kam se podíváš“)[7], kde je současný stav nazýván doslova „průmyslovou revolucí dat“.[8]



Obrázek 1 - znázornění velikosti jednoho petabytu, zdroj: www.wired.com (2008)

Takovýto nárůst objemů dat mění celkově i paradigma současné vědy. V průběhu dějin se paradigma vyvíjelo po ose 4 základních paradigmat, která po sobě následovala v následujícím pořadí.

[3] *Internet World Stats : Usage and population statistics* [online].
[4] MANOVICH, Lev. Trending : The Promises and the Challenges of Big Social Data. [online]
[5] The Petabyte Age: Because More Isn't Just More — More Is Different. *Wired.com* [online]
[6] Velikost jednoho petabytu je 10^{15} bytu, čili tisíc terabytu (běžná velikost současného HDD). Pro zajímavost takovýto objem dat zpracují servery Googlu zhruba každých 60 minut (v roce 2008 to bylo zhruba každých 72 minut).
[7] Data, data everywhere. *The Economist* [online].
[8] HAN, Jiawei, Micheline KAMBER a Jian PEI. *Data mining*: *concepts and techniques* [online].

1. **Empirické paradigma** staré více jak tisíc let bylo založené na **pozorování přírody** a jejích jevů.
2. **Teoretické paradigma** používané ještě před několika stovkami let využívá **generalizace problémů** a **vytváření modelů**.
3. **Komputační paradigma** používané před několika desetiletími funguje na principu **simulování komplexních systémů.**
4. **Datové paradigma** je vědeckým paradigmatem současnosti a je založené na **analýze dat**[9].

Kompletní „digitální vesmír“ byl v roce 2006 vyčíslen na 0,18 zettabytů a odhadovaná velikost ke konci roku 2011 je 1.8 zettabytu. Pro představu velikost jednoho zettabytu je 10^{21} bytu, čili tisíc exabytů, jeden milion petabytů, jedna miliarda terabytů. Zdroje těchto dat mohou být různých druhů. Pro ukázku pár příkladů

- New Yorská burza vygeneruje každým dnem zhruba jeden terabyte dat.
- Databáze Facebooku obsahuje zhruba 10 miliard fotek zabírajících více než jeden petabyte na datových úložištích.
- Internetový archiv zálohuje téměř 2 petabyty dat a jeho velikost roste průměrnou rychlostí 20 terabytů za měsíc.
- Velký hadronový urychlovač LHC (Large Hadron Collider) ve švýcarské Ženevě vyprodukuje každý rok 15 petabytů dat.[10]

Celý tento fenomén můžeme shrnout pod termínem „Big Data“ („Velká data“). Tento termín můžeme definovat jako stav, kdy je objem dat natolik veliký, že není možné tato data běžně používanými softwarovými nástroji zachytit, pořádat nebo zpracovávat v únosném časovém horizontu. Velmi podstatná část takovýchto dat pochází nebo se sdílí pomocí sociálních sítí. Díky nim máme dnes přístup k do nedávna nepředstavitelným

[9] The fourth paradigm: data-intensive scientific discovery. Redmond, Wash: Microsoft Research, 2009.
[10] Hadoop: the definitive guide., 2010.

objemům dat obsahujícím texty, digitální fotografie, videa, tabulky, data ze senzorů mobilních zařízení (např. geolokační data) apod.[11] Všechna tato data mají vztah ke konkrétním uživatelům.

Když odhlédneme od samostatných dat (obsahu) a budeme se soustředit na vzájemné sociální vazby v síti, tak zjistíme, že v současnosti miliardy lidí na internetu každou vteřinou kliknutím myši vytvářejí biliony sociálních spojení pomocí veškerých sociálních médií. Každý komentář, email, zpráva instant messagingu, post na blogu či mikroblogu, vzájemné tagování (označování) a odkazování, ba dokonce samotný pohyb na webových stránkách, vytváří jedinečnou vazbu mezi jednotlivými uzly sítě, jedinečnou digitální stopu.[12] Tato data jsou ve valné většině případů volně přístupná a my jich můžeme použít v náš prospěch a vytvořit tak znalosti z dat (data -> knowledge).

Tyto nové možnosti také zcela mění přístup k sociologii. Dříve bylo možné sledovat a vyhodnocovat „deep data" („hluboká data") o malé skupině lidí, ale v současnosti, díky moderním technologiím, je možné získávat „surface data" („povrchová data") od velikého množství uživatelů. Poprvé v historii máme podle Lva Manoviche možnost sledovat představy, názory, myšlenky a pocity stovek milionů lidí. Můžeme vidět obrázky a videa, která vytváří a jak si je navzájem komentují, monitorovat vzájemné konverzace uživatelů, číst jejich blogy a tweety, poslouchat muziku, která se jim líbí z jejich doporučených seznamů a v poslední době i sledovat jejich fyzickou polohu pomocí senzorických dat z jejich mobilních zařízení. Toto všechno je možné jen kvůli tomu, že tito uživatelé se svobodně rozhodli tyto informace sdílet veřejně s kýmkoli, kdo má připojení k internetu. Využití takovýchto „kolektivních vědomostí" umožňuje sledování jednotlivých myšlenek, jak a kde vznikají a jak se prolínají a navzájem ovlivňují. Ve věku petabytů[13] a éře velkých dat totiž více není jen více, ale více je i jinak. [14] O tomto konceptu kolektivní inteligence hovoří například i Josef Šlerka v jeho již přes rok staré prezentaci „Od pavučiny k mraveništi, aneb kolektivní inteligence za času internetu"[15], kde se zabývá právě myšlenkou využití internetu a především služeb Webu 2.0, potažmo sociálních médií, jako

[11] Lev Manovich ve své výše uvedené studii předpokládá, že objem fotografií denně sdílených na Facebooku je větší než veškeré digitalizované dějinné sbírky umění ve všech světových muzeích dohromady.
[12] ČERNÝ, Michal. Znáte své digitální stopy?. *Lupa.cz* [online].
[13] The Petabyte Age: Because More Isn't Just More — More Is Different. *Wired.com* [online].
[14] MANOVICH, Lev. Trending : The Promises and the Challenges of Big Social Data. [online].
[15] ŠLERKA, Josef. *Od pavučiny k mraveništi : aneb kolektivní inteligence za času internetu* [online].

jakési kolektivní paměti, které se člověk musí umět pouze „dobře zeptat“ a dostane se mu velmi cenné odpovědi. Vychází tak z původní ideje Piera Lévyho[16], který jako první hovoří o kolektivní inteligenci, jejímž základním principem je předpoklad, že skupina může být inteligentnější nežli její jednotliví členové. Za nejběžnější příklad se uvádí sociálně žijící hmyz, jako mravenci, termiti, nebo včely. Tento hmyz je sám o sobě velice hloupý, ale jako skupina se dokáže chovat až překvapivě inteligentně. Kolektivní inteligence „skupinového mozku“ vychází z interakcí mezi miliony jednotlivců. Z našeho pohledu toto můžeme brát též jako interakci mezi miliony různorodých komponentů. Například mezi miliony uživatelů sociálních sítí ve spojení s miliony počítačů (strojů). Touto vzájemnou symbiózou tak vyvstane inteligence „vyšší úrovně“[17].

K opravdu komplexním a vyčerpávajícím datům ze sociálních sítí mají přístup pouze společnosti vlastnící tyto služby. Pouze antropolog pracující pro Facebook nebo sociolog pracující v Googlu může využít kompletní data patřící těmto společnostem. Ostatní jsou odkázáni na datamining dat pomocí API[18] těchto služeb, avšak nikdy nebudou schopni získat data kompletní. Důvodem je zabezpečení osobních dat a nastavení soukromí jednotlivých uživatelů, případně kompletní bezpečnostní politiky samotných služeb[19]. Ale i tak se dostává výzkumníkům v této oblasti do rukou mocný nástroj.

Veškerá tato data mají obrovský potenciál pro dataminingové metody a následnou analýzu a my dnes již máme dostatečné možnosti (hardware a software), jak tyto obrovské objemy dat zpracovávat, ukládat, pořádat a vyhodnocovat.[20] Ať již z vědně sociologického hlediska, tak například z ekonomického, marketingového nebo bezpečnostního hlediska má cenu tyto možnosti využívat. Tato práce si klade za úkol poukázat na tyto možnosti a pomocí vlastního výzkumu a praktických příkladů co nejlépe tyto možnosti představit čtenáři.

[16] LÉVY, Pierre. *Collective intelligence*: *mankind's emerging world in cyberspace* [online].
[17] Collective intelligence: creating a prosperous world at peace. Editor Mark Tovey., 2008.
[18] Application Programming Interface – rozhraní pro programování aplikací umožňující využívat přímo data dané služby.
[19] Případně čistě ekonomických zájmů vlastníků těchto služeb.
[20] HANSEN, Derek; SHNEIDERMAN, Ben; SMITH, Marc . *Analyzing Social Media Networks with NodeXL : Insights from a Connected World*.

1.2 Cíle této práce

Cílem této práce je popsat sociální sítě vhodné k získávání užitečných dat a poukázat na potenciál využití takto získaných dat (kapitola sociální sítě). V další části práce představit základní principy dataminingu a na vybraných metodách demonstrovat jeho využitelnost (kapitola datamining). K této demonstraci budou použity zejména nástroje, metriky a případně softwary pro Social Network Analysis, Sentiment Analysis a Social Media Monitoring. V poslední části této práce si autor klade za úkol nastínit možnosti, které tato technologie přinese do budoucna a upozornit na hrozby, které kvůli ní mohou nastat (kapitola vize a hrozby).

2 Sociální sítě

2.1 Proč právě sociální sítě?

Proč by nás jako zdroj cenných informací a dat s potenciálem pro dataminingové techniky měly zajímat zrovna elektronické sociální sítě? Pro odpověď na tuto otázku budeme muset zabrousit trochu do historie. Všechny změny se neodehrály ze dne na den, ale z hlediska inovací v této oblasti je jen málo oborů, kde dochází k rychlejším změnám a vývoji. Uvědomme si, že ještě před 5-6 lety v České republice o nějakých sociálních sítích neměl nikdo z široké veřejnosti ani „šajna“. A služby jako Facebook nebo Twitter byly pouze tématem vzrušených debat několika podivínů z vysokoškolského prostředí, hlavně pak těch, kteří měli možnost vycestovat do zahraničí, případně získat nějaké zahraniční přátele/kontakty. Ale stejně jako tomu bylo se sociálními sítěmi, tak tomu bylo i s ostatními novinkami, které do Čech vždy dorazily s obligátním zpožděním. Na následujících řádcích si definujeme to, jak jsme historicky dospěli k aktuálnímu stavu a proč jsou a budou data ze sociálních sítí tak nesmírně hodnotný informační zdroj. Pro lepší pochopení celé problematiky se proto přesuňme ještě o pár let nazpět.

2.2 Změna komunikačního modelu

Prapůvod rychlých změn a inovací současnosti musíme hledat v dávné i nedávné minulosti. Veškerý vývoj vychází z technologického pokroku v oblasti komunikace informací a v závislosti na něm ze změn v rámci komunikačního modelu společnosti, jak uvádí kupříkladu ve své knize „*The Network Society : Social Aspects of New Media*“ Jan van Dijk[21]. Tyto změny se odehrály v relativně malém časovém úseku. A to zhruba v rozmezí 19. a 20. století[22]. Během technické komunikační revoluce se odehrálo mnoho změn, které zásadně změnily přístup k tomu, jak jsou budovány vzájemné komunikační vazby, jaký obsah se přenáší a jakým způsobem. Když půjdeme ale od prvopočátku, tak mezi hlavní komunikační milníky musíme zahrnout na prvním místě vynález knihtisku ve 14. století. Do té doby bylo sice možné šířit tištěné informace, ale pouze ve velmi malém množství. Další metou bylo rozšíření tisku, které se již dalo považovat za komunikační model one-to-many (jeden k mnoha). V druhé polovině 19. století pak následovala druhá

[21] VAN DIJK, Jan. *The Network Society : Social Aspects of New Media* [online].
[22] Ve srovnání s dobou, kdy převládal převážně komunikační model one-to-one.

vlna komunikační revoluce spočívající ve vynálezu přenosu dat na dlouhé vzdálenosti ať už pomocí drátů, tak volně vzduchem a zároveň nové možnosti analogového uchování dat (fotografie, film, gramofonové desky, magnetofonové pásky). Po II. světové válce pak hovoříme o začátku digitálních médií/digitálního přenosu.

V relativně krátké době se lidstvo dostalo od telegrafu, přes telefon, telex a rádio až k televizi. V období posledních 20 let pak vývoj zrychlil ještě více a dovolil nám do nedávna až nevídané možnosti související s přechodem k síťovému komunikačnímu modelu many-to-many. Ve spojitosti s tímto modelem máme na mysli samozřejmě rozvoj počítačových sítí a Internetu. Internet dnes ale už není jen samotným médiem, ale začal v sobě remediovat ostatní média, brát si jejich vlastnosti a specifika a postupem času je velmi pravděpodobné, že je i zcela pohltí. Otázkou je jaké médium pohltí internet? Budou to sociální sítě?

Vycházejíc ze současného vývoje v této oblasti není tato myšlenka vůbec nepravděpodobná. A pakliže to nebudou sociální sítě takové, jak si je představujeme dnes, tak to může být nějaká jejich budoucí/novější/vyspělejší obdoba. V závislosti na využívaných technologiích se nejpravděpodobněji Internet a ostatní technologie bezprostředně s ním spjaté budou vyvíjet současným směrem, dokud nepřijde další technologická revoluce přinášející objev takového rozsahu, že celý segment posune úplně někam jinam. Například Jan van Dijk takovýto zvrat ale neočekává dříve než v roce 2040[23], čili z tohoto hlediska se nemusíme bát do tohoto modelu investovat náš čas.

2.3 Digital Natives a jejich chování na síti

Jak bylo uvedeno v úvodu, žijeme v éře Velkých dat (Big data), v mraveništi zásobeném miliardou pracovitých „informačních dělníků“, propojeni všichni vzájemně dle komunikačního modelu síťové společnosti (Network society). Je paradoxem, že i když jsme ještě nikdy v historii nebyli svobodnější, tak zároveň jsme ještě nikdy nebyli propojenější a vzájemně na sobě závislejší[24].

[23] VAN DIJK, Jan. *The Network Society : Social Aspects of New Media* [online].
[24] Tamtéž.

Stále více se naše životy odehrávají na elektronické síti, každým dnem zanecháváme své nesmazatelné digitální stopy. Takovýmto způsobem každý jedinec žije svůj vlastní život na síti, do které vtiskává část sebe, svého chování, svého vědomí, svého vědění. Tyto odrazy uživatelského chování jsou v digitální podobě „navždy" zaznamenány a my máme více či méně „plně" k dispozici záznamy o chování zhruba miliardy lidí.

Všechno co uživatelé na elektronických sociálních sítích dělají, by se dalo definovat jako komplement jejich přirozeného reálného (chápejme nikoli virtuálního) jednání. Sociální sítě jim pouze pomáhají překonat prostorové a zejména časové limity, kterými by byli v analogovém světě svázáni. I když doba strávená ve virtuálním světě stále narůstá, tak stále nemůžeme počítat s tím, že by se v dohledné době naše životy přesunuly pouze na sociální sítě. To je myšlenka podobně nepravděpodobná, jako byly vize ohledně zániku tištěných knih, rušení knihoven apod. Stejně tak jako tištěné knihy, knihovny a ostatní součásti lidského života, kterým byl „věštěn" rychlý zánik, tak i fyzické sociální interakce nezaniknou, jen možná budou plnit trochu jinou úlohu nežli dnes. Jako příklad si uveďme firemní poradu, která může být vedena online z domovů jednotlivých účastníků bez nutnosti fyzické přítomnosti na jednom místě. Oproti tomu do hospody na pivo s přáteli si každý radši zaskočí osobně. Stejně tak jako studijní texty, které má dnes prakticky každý v elektronické podobě v počítači, tabletu, mobilním telefonu či elektronické čtečce. Svou oblíbenou detektivku si ale stále rád přečte v papírové podobě s pocitem, že „něco drží v ruce" a vůní papíru.

Tento fakt nesporně ovlivňuje i samotné chování lidí. Uživatelé, kteří měli možnost vnímat různá stádia vývoje ať už samotných počítačů, tak virtuálního prostředí, se chovají na sociální síti zcela rozdílně oproti tzv. Digital Natives[25](lidé již narození do digitálního věku). Dnes je diskutován zejména přístup těchto rozdílných skupin k otázce soukromí. Virtuální součást každého jedince je pro Digital Natives zcela přirozená věc[26]. Musíme tedy chápat, že i přístup ke sdílení informací o sobě berou tito jedinci zcela rozdílně/otevřeněji. Někdy je tento pojem házen do jednoho pytle s rizikovým chováním na sociálních sítích apod. Uvažme ale, že nadměrné sdílení soukromých personálních informací nemusí být u této skupiny lidí vůbec známkou neznalosti rizik vyplývajících ze

[25] Digital native. In *Wikipedia : the free encyclopedia* [online].
[26] WINDISCH, Eva ; MEDMAN, Niclas . Understanding the digital natives. *Ericsson Business Review* [online].

zneužití těchto informací. Takovýto přístup může být zcela legitimně brán jako jejich celkový postoj k chápání soukromí „za času sociální sítě“. Tito uživatelé jsou si od svého prvopočátku vědomi, co o sobě sdělují a komu.

Z toho hlediska nesmíme brát informace o uživatelích získané ze sociálních sítí za 100% pravdivé a věrohodné, protože zároveň s uvědoměním si ztráty soukromí dochází i k manipulaci s pravdivostí informace[27]. Ne každý uživatel chválící na internetu svého zaměstnavatele tak bude vyjadřovat svůj reálný postoj. S vědomím, že je velice snadné pro zaměstnavatele si tyto zprávy vyhledat, tendenčně změní své vyjadřování „na venek“. I v řadách nadšeně chválících tedy mohou být skrytí sabotéři. Zdaleka ne všichni uživatelé jsou ale takovéhoto jednání schopni, ať povahově, tak především inteligenčně. I nadále se tak pravděpodobně budeme mezi Digital Natives setkávat s případy výpovědí od zaměstnavatelů vyhodivších je z důvodu toho, že na sebe na sociálních sítích prozradili prostřednictvím „fotek ze včerejší kalby“, že to rozhodně s jejich nemocenskou nebude tak „žhavé“.

2.4 Co to je sociální síť a na čem staví?

Sociální síť jako taková je síť spojující navzájem (přímo nebo nepřímo) jednotlivce a skupiny jedinců (uzly) za účelem vzájemné komunikace informací (sociální interakce). V češtině můžeme použít též zástupné výrazy jako společenská síť, komunitní síť, komunita. Historie výzkumu v oblasti sociálních sítí sahá až k přelomu 19. a 20. století. Hlavní výzkum v této vědní disciplíně se pak odehrává převážně od poloviny 20. století do současnosti. V počátcích výzkumu v této oblasti se vědci zabývali především sociálními sítěmi z pohledu matematického, antropologického, psychologického a sociologického.

Zcela nového významu nabyl pojem sociální síť s příchodem „nových médií“, zejména Internetu a v poslední době pak s příchodem Webu 2.0.[28] O Webu 2.0 se není

[27] Jak uvádí například Lev Manovich ve své práci (v originále): „Again, this does not mean that we can’t do interesting research by analyzing larger numbers of tweets, Facebook photos, YouTube videos, etc. – we just have to keep in mind that what all this data is not a transparent window into peoples’ imaginations, intentions, motifs, opinions, and ideas. Its more appropriate to think of it as an interface people present to the world“
MANOVICH, Lev. Trending : The Promises and the Challenges of Big Social Data. *Debates in the Digital Humanities* [online].
[28] ZBIEJCZUK, Adam. *Web 2.0 - charakteristiky a služby*. Brno, 2007.

třeba na tomto místě nijak zvláště rozepisovat. Jedná se o stádium vývoje webu, ve kterém přestává být uživatel pouze pasivním příjemcem informace, ale je zároveň začleněn do procesu tvorby informace, její sdílení a předávání dál. Tento uživatelsky garantovaný obsah decentralizuje běžné informační autority, např. klasické zpravodajské servery, online encyklopedie apod. Základními stavebními kameny Webu 2.0 jsou zejména pro tuto práci tak důležité sociální sítě, blogy, mikroblogy, servery sloužící ke sdílení dat, wiki-weby a mashupy (míchanice), které spojují všechny tyto druhy dohromady.[29]

2.5 Druhy, jejich specifika a potenciál pro Datamining

Současná sociální média můžeme rozdělit do mnoha kategorií, z nichž pouze jedna nese přímo název sociální sítě. Základní principy (Web 2.0 viz výše) jako síťovost (vzájemné vztahy mezi uživateli), interaktivnost (reakce, komentáře a zpětná vazba) a uživateli garantovaný obsah (nutnost participace) jsou pro všechna tato média totožná. Pro potřeby této práce proto budeme označovat všechna tato sociální média sociálními sítěmi, protože potenciál pro datamining mají všechny z nich. U každého typu budou vypsána specifika a příklady konkrétních služeb, které do dané kategorie patří.

[29] Web 2.0. In *Wikipedia : the free encyclopedia* [online].

2.5.1 Sociální sítě

Jedná se o služby, které primárně slouží k vzájemnému propojení uživatelů. Staví na interakci a sdílení obsahu (text, fotky, videa, odkazy, poloha) a komunikaci (messaging). Umožňují uživatele clusterovat (třídit) do různých skupin ať podle „přátelství“, fanouškovství nebo fyzického umístění. V současné době se jedná o nejnavštěvovanější elektronické služby na internetu obecně. Předstihly dokonce už i stránky s pornografií, které si dlouhodobě držely 1. místo. Veškeré tyto vlastnosti jsou velice zajímavé z hlediska dataminingu, protože všechny data, která tyto sítě generují, jsou přímo ve vztahu ke konkrétnímu uživateli/skupině uživatel[30]. U známějších služeb budou uvedeny statistiky, které mají demonstrovat jejich růst a tím pádem i čím dál tím větší dataminingový potenciál do budoucna.

Ve světě není v současnosti skloňovanější sociální sítě nežli Facebook[31], dále pak do této kategorie můžeme zařadit jeho obdobu Google+[32] od společnosti Google a od téže společnosti sociální síť Orkut[33], která se ale uchytila prakticky pouze v Brazílii. Na ústupu je nyní sociální síť Myspace[34], která bývala velmi významnou, ale v posledních letech byla převálcována právě těmito službami, zejména pak Facebookem. Dalšími hráči na trhu jsou již více specializované sociální sítě jako např. LinkedIn[35], který slouží jako jakási Human Resources databáze vzájemně propojených profesí/profesionálů. Jako příklad z českého prostředí můžeme uvést např. Lidé.cz[36] nebo líbímseti.cz[37], které jsou dnes díky daleko pokročilejšímu Facebooku také na ústupu. Českou (v dnešní době do zahraničí expandující) sociální sítí s obrovským dataminingovým potenciálem jsou pak amatéři.cz[38], specializující se na sdílení dat (foto a video) s explicitně sexuálním obsahem. Nejedná se o sdílení komerční pornografie, ale o uživatelsky/amatérsky vytvářený obsah. Takovýto obsah by ale z hlediska svého charakteru mohl být velice snadno zneužitelný.

---

[30] Už při samotném založení svého profilu je uživatel směřován k tomu, aby o sobě prozradil co nejvíce. (Pohlaví, věk, místo narození, bydliště, telefonní číslo, osobní stránky, osobní vztahy, rodinné vztahy, sexuální orientace, zájmy, oblíbené knihy, filmy, sporty, životní motto, víra, filosofie apod.) Dále pak je ihned na začátku vybídnut k nahrání své „profilové fotografie“, která na sociální síti reprezentuje jakéhosi avatara.
[31] *Facebook* [online].
[32] *Google+* [online].
[33] *Orkut* [online].
[34] *Myspace* [online].
[35] *LinkedIn* [online].
[36] *Lidé* [online].
[37] *Líbímseti.cz* [online].
[38] *Amatéři.cz* [online].

Facebook

Se svými více jak 800 miliony aktivních uživatelů je Facebook největší sociální síť současnosti. V České republice je to nyní cca 3,5 milionu uživatelů[39]. Jak zaznělo i na letošní konferenci Le Web v Paříži[40], tak z celkového počtu uživatelů ho používá cca 0.5 miliardy každý den. Dále je na statistikách Facebooku jasně patrný veliký růstový potenciál. Během roku 2010 se ke službě připojilo více než 250 milionu uživatelů, což je obrovský nárůst, kterému nemůže konkurovat žádná jiná služba na Internetu. Z celkového počtu uživatelů jich pak více jak 350 milionů přistupuje ke službě z mobilních zařízení. Tento fakt pak demonstruje maximální aktuálnost a operativnost sdílení informací prakticky v reálném čase.[41]

Když zvážíme to, že zároveň letos lidská populace na Zemi dosáhla počtu 7 miliard obyvatel, tak databáze uživatelů Facebooku nabízí jedinečný vzorek cca 11% celkové populace. Vize Facebooku je taková, aby o sobě uživatelé sdíleli co nejvíce a tak nabízí přímou platformu pro sdílení uživatelských fotografií, videí, odkazů, statusů a v poslední době i fyzické pozice (viz Facebook Places v oddělení o geolokačních službách). A to jak ve své klasické, tak v mobilní verzi. Navíc právě ve dnech vzniku této práce uvádí do života nový koncept uživatelského profilu Timeline[42] (Moje historie), který svým pojetím doslova vybízí (až nutí) člověka ke zpětnému doplnění informací a dat obecně i z doby, kdy ještě nebyl uživatelem této služby. Následně nový design stejně tak směřuje uživatele ke stále většímu sdílení fotografií a „zážitků“ ze své historie.

Dále statistiky uvádějí, že každý den uživatelé na Facebook nahrají v průměru 250 milionů fotek. Každých 20 minut napíší zhruba 1,85 milionu statusů a celkově měsíčně na Facebooku stráví neuvěřitelných 9,3 miliardy hodin, což je přes milion let.

To, jak se dá krásně využít takováto komplexní masa dat o velikém množství uživatel v sociologii, dokázali samotní vědci z Facebooku při potvrzení „Small World experimentu“[43].

[39] Czech Republic Facebook Statistics. Socialbakers.com [online].
[40] *Le Web Paris 2011* [online].
[41] Tiskové zprávy. *Facebook.com* [online].
[42] Představujeme profil Moje historie. *Facebook.com* [online].
[43] Viz kapitola Social Network Analysis, sekce o dosahu sítě.

Z pohledu Social Network Analysis nabízela ještě do nedávna síť Facebooku pouze vzájemně rovné nesměrované (undirected) vztahy tzv. přátelství (friendship). Jako reakci na rázný nástup služby Google+, ale umožnila i jednosměrné vztahy tzv. odběry (subscription)[44], kterým se dlouhou dobu bránila. V současné době během konání již výše zmíněné konference Le Web zprovoznil Facebook i odběry u fanouškovských stránek.

Google+

Novinkou poloviny roku 2011 byla sociální síť od společnosti Google. Google se již v minulosti snažil prorazit se svojí sociální sítí Orkut (viz výše). Ta se uchytila prakticky pouze v Brazílii, kde zaujímá roli absolutní jedničky. Síť Google+ do dnešní doby nedosáhla takového úspěchu jako Facebook, ale z hlediska toho jaká firma za ní stojí, má obrovský růstový potenciál. V současnosti spojuje téměř 27 milionu uživatelů, což sice není z daleka tolik, kolik má hlavní konkurent, ale tito uživatelé se rekrutují zejména z prostředí počítačových a informačních specialistů[45]. Obsah vytvářený takovouto specifickou skupinou má z pravidla velice vysokou informační hodnotu.[46]

Z genderového hlediska je tato síť značně nevyvážená oproti Facebooku, kde je poměr mezi pohlavími cca 50:50. U Google+ je to poměr zhruba 75:25 ve prospěch mužů. Z čehož plyne, že tato sociální síť má do budoucna obrovský potenciál růstu ve skupině uživatelek.[47][48]

Síť Google+ proti svému hlavnímu konkurentovi nabídla zcela odlišný koncept vzájemných vztahů mezi uživateli. Stejně jako na Twitteru jsou zde z pohledu Social Network Analysis směrované jednosměrné (directed) vztahy založené na principu tzv. „kruhů“, kdy si uživatel přidává ostatní uživatele do vlastních skupin, kterým má v budoucnu možnost přidělovat přístupová práva k jím tvořenému/sdílenému obsahu[49].

[44] Obdobu „Follow“ na Twitteru.
[45] Hlavními skupinami podle profesí na Google+ jsou zejména inženýři, vývojáři, softwarový inženýři, designéři, programátoři, spisovatelé, fotografové a učitelé.
[46] Infographic: First Google Statistics & Facts. *Digitalbuzzblog.com* [online].
[47] Tamtéž.
[48] Facebook VS. Google [Infographic]. *Singlegrain.com* [online].
[49] Na tento koncept Facebook hbitě zareagoval možností tzv. „odběrů“ (viz oddíl Facebooku).

Obsahově se Google+ neliší od ostatních sociálních sítí, ale její hlavní výhodu spatřuje autor v její plné integraci do samotného Google účtu a spolupráci se všemi ostatními (nesčetnými) „googlovskými“ službami. Zejména pak vynikající integraci do mobilních zařízení s operačním systémem Android. Samotná mobilní aplikace (když jí to dovolíte) zcela automaticky nahrává mobilními zařízeními pořízená data (fotografie, videa) a automaticky k nim ukládá geografická metadata. Následně je uživatel nepřímo vybízen k tomu, aby tento obsah neprodleně sdílel[50]. Je nutno říci, že i mobilní aplikace ostatních sociálních sítí tuto funkcionalitu nabízejí, ale u žádné jiné se nesetkáme s tak silnou tendencí „usurpovat si“ soukromá uživatelská data.

LinkedIn

Z oblasti specializovaných sociálních sítí si na tomto místě zmíníme síť LinkedIn, která se specializuje na „propojování specialistů“ ze všech myslitelných odvětví. Její funkčnost je obdobná jako u ostatních sociálních sítí, unikátní je ale její zaměření na HR[51]. Poskytuje možnost vytvoření profesního profilu, jakéhosi virtuálního CV, které je vystaveno na odiv celému světu, zejména pak firmám a HR agenturám. Vzájemné vazby mezi uživateli jsou řešeny nesměrovanými rovnými vztahy, které lze ještě navíc upřesňovat na základě toho v jaké firmě, nebo na jaké škole se daní uživatelé seznámili.

Hlavním potenciálem této sítě z hlediska dataminingu je v první řadě komplexní databáze lidských zdrojů a v druhé řadě pak nespočet různorodých skupin uživatel, kteří v rámci těchto skupin sdílejí a šíří velice hodnotné informace[52].

[50] Uživatel je vybízen ke sdílení neustálými notifikacemi ohledně právě uploadovaného multimediálního obsahu z jeho mobilního zařízení.
[51] Human Resources – lidské zdroje.
[52] Jako příklad si uveďme skupina zabývající se Social Media Monitoringem http://www.linkedin.com/groups/M%C4%9B%C5%99%C3%ADme-monitorujeme-soci%C3%A1ln%C3%AD-s%C3%ADt%C4%9B-3998566?home=&gid=3998566&trk=anet_ug_hm

2.5.2 Mikroblogy

Mikroblogové služby v sobě nesou z našeho pohledu skoro nejvyšší potenciál. Jsou založeny na principu krátkých zpráv (např. 140 znaků na Twitteru), do kterých je nucen uživatel vměstnat své sdělení. Tato vlastnost je hlavním kladem a někdy také záporem této služby. Informace z mikroblogu jsou vždy značně zhutnělé, často se používá všemožných zkratek a zjednodušení, což někdy může poznamenat srozumitelnost sdělení. Naopak tento fakt velice pozitivně ovlivňuje dynamiku šíření informace. Oproti klasickému blogu se tedy jedná spíše o jakýsi proud (stream) statusů. Z hlediska obsahu jsou touto cestou často šířeny aktuality, novinky, zajímavosti. Nic není staršího nežli hodinu starý tweet. Proto využití dat z mikroblogů je velice dobře použitelné pro monitorování chování a nálad ve společnosti, odhadování trendů apod. Více o této tématice bude v kapitole o Social Media Monitoringu. Stejně tak jako u sociálních sítí, tak mikroblogové služby umožňují sdílet kromě textu též obrázky, videa, odkazy a v neposlední řadě také fyzickou polohu uživatele. Informace o fyzické poloze je zvláště cenná, protože dává možnost filtrovat informace jen z určité oblasti našeho zájmu.

I když počet uživatelů není ještě zdaleka tak vysoký jako u sociálních sítí, tak o to větší je pravděpodobnost růstu v této oblasti. V současné době je jednoznačně nejvýznamnější mikroblogovou službou Twitter[53]. Nepovedeným pokusem konkurovat této službě se v roce 2010 uvedla společnost Google se svou službou Google Buzz[54]. Od svého začátku se ale potýkala s nezájmem uživatelů tuto službu využívat, a tak byla 15. prosince tohoto roku ukončena. Jako další příklad si můžeme uvést ještě třeba Tumblr[55], jehož penetrace ve společnosti není zdaleka tak vysoká jako u jeho konkurenta. Z tohoto důvodu si zde uvedeme specifika mikroblogingu právě na příkladu Twitteru.

[53] *Twitter* [online].
[54] Google Buzz. In *Wikipedia : the free encyclopedia* [online].
[55] *Tumblr* [online].

Twitter

Tato mikroblogingová služba funguje na bázi zasílání krátkých sdělení o 140 znacích zvaných „tweety“ (pípnutí). Do těchto krátkých zpráv je možno též vkládat odkazy, obrázky, videa (odkazy na videa) a to vše lze doplnit o geolokační data, která se vkládají buď díky integrovanému GPS modulu uživatelova mobilního zařízení nebo pomocí lokalizace na základě IP adresy. V případě, že se uživateli líbí nějaký tweet od jiného uživatele, nabízí Twitter možnost tzv. retweetu, čili jakéhosi přeposlání informace svému publiku. Tímto principem je zabezpečeno rychlé šíření těch nejhodnotnějších informací dále sítí a na základě počtu retweetu se dá určovat hodnota informačního sdělení daného tweetu, případně oblíbenost (vliv) jeho autora.

Vztahy mezi uživateli jsou řešeny na bázi směrovaných spojení, tzv. následování (follow). Někdy z tohoto důvodu může tedy na uživatele Twitter mylně působit tak, že píše svoje tweety pro lidi, které následuje, avšak reálně píše pouze pro lidi, kteří následují jeho. Dále pak lze zmiňovat ve tweetech ostatní uživatele za pomoci @ a členit témata pomocí tzv. „hash tagů“ (#). Tato vlastnost také vytváří vztahy mezi uživateli a tématy a je velmi užitečná z hlediska dataminingu při určování témat konkrétního sdělení a vlivnosti uživatel na dané téma. Více se této problematice budeme věnovat v kapitole o Social Network Analysis a Social Media Monitoringu, pro které je Twitter vynikajícím zdrojem dat.

Do dnešního dne získala síť Twitteru cca 300 milionů uživatelů, což jí řadí za Facebook na druhé místo mezi všemi sociálními službami (sítěmi). Každým dnem přibude v průměru dalších 450 tisíc nových účtů[56]. Všichni tito uživatelé pak napíší v průměru 177 milionů tweetů každý den[57].

Otázkou statistik českého Twitteru se zabýval tým Social Media a Marketingu společnosti H1.cz, který koncem roku uveřejnil zprávu „Česko na sociálních sítích“[58]. Tato zpráva vychází z dat, která využívá internetový projekt Klábosení.cz[59] mapující český a

[56] V poslední době vděčí Twitter obrovskému nárůstu počtu uživatelů z hlediska jeho integrace do operačního systému iOS5 mobilních zařízení americké firmy Apple.
[57] 10 Terrific New Twitter Infographics in 2011. *Singlegrain.com* [online].
[58] APPELTAUEROVÁ, Lucie, et al. Česko na sociálních sítích. In H1.cz. [online].
[59] *Klábosení.cz* [online].

slovenský Twitter[60]. V současné době tento server eviduje cca 86 tisíc českých nebo slovenských účtů[61]. Velmi dynamický růst této služby v našich krajinách vystihuje fakt, že zhruba třetina těchto uživatelů se přidala v letošním roce (2011). Průměrný český uživatel pak vyprodukuje cca 13,5 tweetu za měsíc. Na druhou stranu zhruba polovinu obsahu českého a slovenského twitteru generuje 500 nejaktivnějších uživatelů. Tato data opět nastiňují obrovské možnosti, které se nám nově otevírají i v Čechách a na Slovensku. I když je zde Twitterová komunita ještě relativně v plenkách, obrovský nárůst uživatel v posledním roce vykazuje slibný potenciál do budoucna[62].

2.5.3 Blogy

Samotné slovo blog vychází ze sloučení dvou anglických slov Web a Log, které znamenají něco jako webový zápisníček/deníček[63]. Jedná se prakticky o webovou službu, která maximálně zjednodušuje sdílení autorských článků, zamyšlení, odkazů a multimediálního obsahu bez nutnosti vytvářet svou webovou stránku. Oproti mikroblogu není samotný blog tak „pohotový" a aktuální. Jeho obsah ale není nikterak rozsahově omezen, takže informace zde prezentované mohou podat daleko obsáhlejší sdělení. Blog je též v současnosti využíván jako velmi moderní nástroj PR pro komunikaci, jak navenek, tak uvnitř firmy[64]. V tomto případě se ale blog nesmí stát pouze místem zveřejňování firemních tiskových zpráv, jak se zmínil v jednom svém rozhovoru například Adam Javůrek[65].

V dnešní době blogy pokrývají nespočet témat, od ryze osobních deníčků až po zcela konkrétně profesně zaměřené. Z tohoto hlediska tedy obsahují velice různorodá a v případě např. odborných nebo firemních blogů i velice informačně hodnotná data.

[60] Více se tomuto serveru budeme věnovat v kapitole o Social Media Monitoringu.
[61] Tento výběr je volen na základě jazyka nebo geografického umístění uživatele.
[62] I přes to, že je v Čechách Twitter teprve v začátcích, tak svoji sílu dokázal shodou okolností právě v den vzniku této kapitoly v souvislosti se smrtí bývalého českého prezidenta Václava Havla. Tato smutná zpráva nejen že se šířila lavinovou rychlostí po celé síti, ale zároveň se dostala do tzv. Trending Topics – nejvíce zmiňovaných témat na celém Twitteru. Tato twitterová komunikace je též zobrazena v příkladu v kapitole Social Network Analysis.
[63] Blog. In *Wikipedia : the free encyclopedia* [online].
[64] V České republice byla průkopníkem v této oblasti například polská banka mBank se svým firemním blogem, prostřednictvím kterého informuje své zákazníky o změnách, novinkách a nabídkách svých služeb. mBLOG. *mBank.cz* [online].
[65] Adam Javůrek: blog není pro tiskové zprávy. *Itbiz.cz* [online].

Hlavními dataminingovými výzvami v případě blogů jsou právě obsáhlé informace, ať již vytvářené nebo sdílené přímo autorem konkrétního blogu/článku na blogu nebo především komentáře k jednotlivým článkům blogu, které nejlépe vystihují „názory společnosti“ (čtenářské komunity) formou konverzace autora s jeho čtenáři.

Statistiky pro rok 2011 uvádějí, že k polovině tohoto roku bylo založeno zhruba 164 milionů blogerských účtů. Oproti roku 2004, kdy bylo evidováno cca 3 miliony účtů, se jedná o obrovský nárůst. Dvě třetiny blogerů jsou muži a celých 23% všech blogerů se psaní svého blogu věnuje na „plný úvazek“. Co do tématiky jsou nejčastěji zastoupeny blogy obsahující osobní názory (18%), Technologie a internetový marketing (14%), politiku (8%) a business (5%)[66].

Ačkoli se formát nebo design jednotlivých blogových služeb může lišit, tak základní koncept je vždy prakticky stejný. Každý příspěvek obsahuje svůj jedineční nadpis, který bývá zpravidla obsažen i v URL celého článku. Dále pak jméno autora a datum uveřejnění. Prakticky samozřejmostí je dnes i možnost nechávat si posílat nejnovější příspěvky pomocí RSS kanálu[67]. Technologicky pak blogy mohou fungovat např. na bázi open source redakčního systému WordPress (cca 40% blogů) nebo díky online službě Blogger (cca 26% blogů), kterou v roce 2003 koupila společnost Google[68]. Tato služba bude pravděpodobně v nejbližší době integrována do nové sociální sítě Google+ pod názvem Google Blogs[69].

2.5.4 Diskusní fóra

Oproti samotným komentářům u obsahu sdíleného na sociálních sítích nebo příspěvků na blozích jsou fóra samostatnou internetovou službou. Hlavními představiteli jsou především specializované diskusní servery na dané téma nebo oddělení „často kladených dotazů“ (FAQ[70]) na odborných severech, případně firemních serverech. V některých případech bývají též realizovány formou návštěvních knih (guest books).[71]

[66] 2011 Blogging Statistics [infographic]. *Rightmixmarketing.com* [online].
[67] RSS. In *Wikipedia : the free encyclopedia* [online].
[68] Výhodou této služby je opět plná integrace s účtem Google a potažmo pak s ostatními službami, jako např. Picasa, Google analytics apod.
[69] EXCLUSIVE: Google To Retire Blogger & Picasa Brands in Google+ Push . *Mashable.com* [online].
[70] Frequently Asked Questions
[71] Internet forum. In *Wikipedia : the free encyclopedia* [online].

Skupinová diskuse nad tématem je realizována formou hromadného chatu, který ale nemusí na rozdíl od např. konferenční diskuse (v rámci instant messagingu) probíhat v reálném čase. Struktura diskuse se dělí v závislosti na druhy služby (diskusního fóra) na jednoduchou nebo strukturovanou. Jednoduchá struktura se vyznačuje chronologickým řazením příspěvků za sebe podle principu novosti. Z pohledu dataminingu a zejména z hlediska vyhodnocování obsahu je tato struktura obtížnější na analýzu, protože uživatelské reakce řazené chronologicky nemusí[72] reagovat vždy na příspěvek bezprostředně přecházející. Z tohoto hlediska je lepší, když se struktura diskuse větví do stromové struktury na tzv. „vlákna“ (threads), u kterých je daleko snazší určit, který příspěvek reaguje na který. Tento počet vláken může být z důvodu přehlednosti omezen.[73]

Uživatelé na diskusním fóru si mezi sebou budují vzájemné vazby pouze na základě komentářů a stejně tak profilové informace diskutujících nebývají zdaleka tak obsáhlé jako například u sociálních sítí. Hlavní důraz v rámci dataminingu se zde odklání od vazeb mezi uživateli především k vlivu konkrétního uživatele na dané téma (určení „influencerů“). Tento vliv může být vyjádřen například počtem příspěvku k danému vláknu, případně jakousi „karmou“ uživatele, která bývá udělována uživatelům na základě hodnocení jejich odpovědí ostatními uživateli a nebo vychází z celkové aktivity v diskusích[74].

Řízení diskuse může být vedeno administrátorem, který sám hodnotí příspěvky, odpovídá na dotazy a fyzicky se účastní diskuse. Případně vyřazuje prokazatelný spam nebo příspěvky porušující pravidla diskusní služby[75]. Diskuse může být také zcela volná, bez jakéhokoli moderátora a cenzury, řízena pouze skupinovou etikou zúčastněných diskutujících.

Co do obsahu se můžeme setkat obrovskou rozmanitostí. Internetová diskusní fóra se zabývají prakticky jakýmikoli myslitelnými tématy. Ze zahraničních webů si uvedeme

[72] A zpravidla nereagují.
[73] Internet forum. In *Wikipedia : the free encyclopedia* [online].
[74] Více se tomuto tématu budeme věnovat v kapitole o Social Media Monitoringu a marketingovém potenciálu hodnocení „influencerů“ na dané téma v internetových diskusích.
[75] Převážně se jedná o příspěvky urážející ostatní účastníky, případně obsahující vulgarismy apod. Tato funkce může být i automatizována počítačem, za pomoci přednastavených pravidel.

jeden z největších diskusních serverů Yahoo Answers[76]. Z českých pak např. diskusní část patřící pod již dříve zmiňovanou sociální síť Lidé.cz[77] nebo především diskusní fórum na serveru emimino.cz[78]. Toto fórum počtem svých příspěvků a obsahem diskusí v současnosti hlavně předčí celý český a slovenský Twitter!

2.5.5 Sdílení fotografií a videa

Zdrojem obrovského množství dat jsou samozřejmě služby specializované na sdílení uživatelských fotografií a videa. Takováto data jsou velice dobře použitelná ať již sama o sobě, tak jako jedinečné informace mnoho vypovídající o svém autoru[79]. Stejně tak dobře se tato data dají využít v aplikacích na bázi mashupu[80], kde přinášejí zcela novou přidanou hodnotu. Promítnutím fotografií obsahujících metadata s GPS souřadnicemi na mapu získáváme unikátní možnosti nahlédnutí do krajiny, panoramatické pohledy, dokumentace událostí[81] apod. Stejně dobře lze tato data použít v aplikacích rozšířené reality[82] (augmented reality) jako virtuálního průvodce. Použitím v současné době ještě ne zcela spolehlivé technologie rozpoznání obličeje lze tato data velice precizně katalogizovat (tagovat) a získat tak zcela unikátní informace o uživatelích (které ani oni sami o sobě doposud nevěděli)[83].

V dnešní době nejpoužívanějšími službami v kategorii sdílení fotografií jsou Flickr[84] s uploadem cca 3 miliony fotografií za den a celkovým počtem přes 6 miliard fotografií. Tato služba patří pod jinak upadající Yahoo. Podobně oblíbená je i služba Picasa[85] americké společnosti Google, která ale bude pravděpodobně brzy integrována do nové

[76] *Yahoo! Answers* [online].
[77] Forum. *Lide.cz* [online].
[78] *Emimino.cz* [online].
[79] Z praxe bych zde zmínil například studii kladoucí si za cíl lokalizovat zdroje patologického chování mezi dětmi na českém internetu (násilí, šikana, sex, alkohol, drogy apod.), na které jsem se měl možnost podílet. Po hlubší analýze možných zdrojů se primárním zdrojem pro zpracování této analýzy staly servery umožňující sdílení autorských videí, zejména pak server Youtube.com.
[80] Mashup (web application hybrid). In *Wikipedia : the free encyclopedia* [online].
[81] V souvislosti s uvedením nového druhu osobního profilu Timeline nabídl Facebook uživatelům možnost promítnout své fotografie do mapy. Z tohoto hlediska by Facebook mohl velice snadno promítnout fotografie všech uživatel do mapy a získat tak ojedinělou databázi audiovizuálně zmapovaných míst (pomineme-li nastavení soukromí jednotlivých uživatel).
[82] Augmented reality. In *Wikipedia : the free encyclopedia* [online].
[83] Této problematice se budeme podrobněji věnovat v kapitole o vizích a hrozbách.
[84] *Flickr* [online].
[85] *Picasa* [online].

sociální sítě Google+ pod názvem Google Photos[86]. Paradoxně zcela nejpoužívanější službou pro sdílení fotografií se s počtem zhruba 8 milionů nahrání denně stal Facebook. Všechny tyto služby slouží primárně ke sdílení fotografií, ale umožňují nahrávat také videa. Oproti tomu například oblíbená služba Instagram[87] pro mobilní zařízení od firmy Apple se specializuje ryze na sdílení fotografií. Přesto je, se svým počtem 7 milionů uživatelů, kteří nahrají 1,3 milionu fotografií za den a celkovým počtem přes 150 milionů fotografií, hned za těmito „velikány“[88].

Youtube

Mezi službami specializujícími se na sdílení videa má cenu jmenovat pouze Youtube, která množstvím svého obsahu jednoznačně převyšuje všechny své konkurenty. Opět patří pod křídla amerického Googlu a i proto zde můžeme očekávat integraci s ostatními službami této firmy. Je zajímavostí, že samotný Youtube je po své mateřském vyhledavači Google druhým nejpoužívanějším vyhledavačem na světě. Účet na Youtube nově umožňuje propojení jak s účtem Google+, což je více než logické, ale dokonce i s Facebookem, Twitterem apod.

Z praktického hlediska funguje služba tak, že jednotliví uživatelé prostřednictvím svého účtu nahrávají svá videa, která tímto dávají k dispozici ostatním ke zhlédnutí. Z tematicky řazených videí lze následně vytvářet tzv. „kanály“, které ostatní uživatelé mohou odebírat (follow). Videa mohou uživatelé sdílet na sociálních sítích, hodnotit pomocí „líbí/nelíbí“[89]. Každé video má pak své statistiky v závislosti na počtu zhlédnutí dále dělené podle věkových a genderových skupin nejčastějších návštěvníků[90]. Více než 50% videí bylo již ohodnoceno nebo okomentováno někým z uživatelů[91].

Významnost a potenciál pro datamining dat této služby dokládají zcela jistě statistiky. Každou minutou je nahráno cca 48 hodin videí, což tvoří zhruba 8 let videí za

[86] EXCLUSIVE: Google To Retire Blogger & Picasa Brands in Google+ Push . *Mashable.com* [online].
[87] *Instagram* [online].
[88] Instagram Facts & Stats. *Digitalbuzzblog.com* [online].
[89] Takovouto „sociální akci“ provede každý den zhruba 100 milionů uživatelů.
[90] Není náhodou, že například videa mladých chlapců, do půli těla povykujících před svou školou, jsou nejoblíbenější mezi skupinou „muž 45-54“ apod.
[91] Statistiky. *Youtube.com* [online].

den. Uživatelé, jejichž věkové rozmezí je zhruba 18 -54 let si každý den přehrají 3 miliardy videí. Z hlediska zpracování videa a indexace obsahu Youtube každý den zkontroluje cca 100 let videozáznamu. Tento obsah je indexován za pomoci porovnání s více než 6 miliony referenčních souborů, které představují zhruba 300 tisíc hodin materiálu. Z celkového počtu uživatel pak 17 milionů má svůj profil propojený s některou s výše jmenovaných sociálních sítí a přes 12 milionů těchto uživatelů přes tuto síť videa automaticky sdílí. Vysokou integraci do sociálních sítí též dokládá fakt, že každou minutu je napsáno více než 500 tweetů s odkazem na Youtube a uživatelé Facebooku si každý den přehrají videa o délce téměř 150 let[92].

2.5.6 Geolokační služby

Sociální geolokační služby založené na sdílení fyzické pozice s ostatními se dostaly do popředí až v posledních letech, kdy se stal běžnou součástí člověka mobilní telefon s vestavěnou jednotkou GPS[93]. Protože dnes již je penetrace touto technologií v mobilních zařízeních natolik vysoká, tak tyto služby zažívají ohromný růst. Z hlediska dataminingu však geolokační souřadnice mají neuvěřitelný potenciál. Samotná fyzická poloha a její GPS souřadnice jsou něco jako URL daného místa. Ochota uživatelů sdílet tyto souřadnice je motivována herními principy, na kterých se tyto služby zakládají. To vede k tomu, že každou vteřinou jsou k dispozici data mapující reálný pohyb milionů osob. Jako přidanou hodnotu můžeme brát fakt, že většina takovýchto služeb umožňuje uživatelům k daným místům přidávat i další formy cenných dat (texty, obrázky, videa), ba dokonce je k tomuto jednání motivuje. Geolokační služby fungují v zásadě na jednom ze dvou principů. První z nich je nepřetržité real-time sdílení polohy (např. u Google Latitude). Druhým principem jsou služby fungující na bázi tzv. „check-inů“, což jsou jakási ohlášení, že se uživatel zrovna nachází na daném místě (Foursquare, Facebook Places, Google Places).

Službami tohoto typu s největším počtem uživatelů jsou v současnosti geolokační „nástavby“ spadající pod výše zmiňované sociální sítě a to především Facebook Places[94]. A

[92] Statistiky. *Youtube.com* [online].

[93] Méně přesně lze též určovat polohu pomocí polohy vůči GSM stanici BTS v rámci sítě telefonního operátora, nebo např. pomocí některé ze sítí Wi-fi v dosahu mobilního zařízení.

[94] Tato služba se s přechodem na nové profily Facebook Timeline, a koupí samostatné geolokační služby Gowalla, bude více modifikovat do konceptu sdílení „příběhů“ – kde, kdo, kdy s kým a co tam dělali. (samotné heslo FB Places je“ Who, What, When and now Where“)

Google samozřejmě nemůže chybět se svými Google Places. Tyto služby slouží převážně jako podpůrné služby pro celkový koncept sociální sítě. Služba Facebook Places má tedy stejně uživatel jako samotný Facebook a začít ji využívat může kterýkoli uživatel, který si povolí v prohlížeči určování fyzické polohy pomocí IP adresy nebo si zakoupí mobilní zařízení s vestavěnou GPS jednotkou. Samostatnou ryze geolokační službou, která panuje tomuto segmentu je ale bezesporu Foursquare.

Foursquare

Je se svými 15 miliony uživatelů nejoblíbenější samostatná ryze geolokační služba současnosti[95]. Její princip je založen na „checkování" se na místech (venues), která tvoří rozsáhlou databázi. Tato místa mají možnost všichni uživatelé zakládat, vkládat fotografie, psát zajímavé tipy, případně navrhovat jejich úpravy. Možnost editace již založeného místa mají pouze prověření uživatelé tzv. „superuživatelé" (Superusers), kteří navíc v závislosti na svém stupni[96] mohou provádět ještě další druhy editací této databáze míst[97]. Služba v sobě nese též motivační/gamifikační prvky jako například získávání bodů za checkin. Následně je možné se porovnávat s ostatními uživateli podle umístění v tabulce na základě celkového počtu bodů za posledních 7 dní. Případně soutěžit o starostu (mayor) daného místa[98]. V neposlední řadě lze získávat odznáčky (badges) např. za návštěvu určitého počtu specifických kategorií míst[99]. Vztahy mezi uživateli jsou řešeny na principu nesměrovaných vztahů – přátelství. Vaši aktuální polohu mohou vidět pouze vaši přátelé, případně lidé, kteří si zrovna prohlíží místo, na kterém jste checknuti.

Z hlediska dataminingu má tato služba potenciál, ať již opět ve sledování struktur vzájemných přátelství pohledem Social Network Analysis, tak zejména z hlediska sledování reálného pohybu milionů uživatelů. Z takovýchto dat se dá vyvozovat mnohé. Od návštěvnosti akcí všeho druhu (kina, obchodní domy, konference), po sledování rušnosti dopravních uzlů (nádraží, zastávky MHD), až po přímé sledování cestovních návyků jednotlivých uživatelů. Na principu návštěvnosti funguje například projekt využívající real-time data z Foursquaru jménem „Voulez-vous check-in avec moi ce

[95] Blog. *Foursquare.com* [online].
[96] SU 1. stupně a SU 2. Stupně (SU 3 stupně jsou vybíráni na základě přísných pravidel – např. v ČR je zatím pouze jeden)
[97] Slučovat duplicitní venues, měnit jejich kategorie, měnit jejich GPS souřadnice na mapě atd.
[98] Uživatel, který se v poslední době na dané venue checkoval nejvíce (byl zde nejvíce dní).
[99] Například za 5 checkinů v knihovně odznáček „knihomol" apod.

soir?[100]“ Tato služba vám doporučí „nejžhavější“ místa ve městě na základě počtu zrovna checknutých uživatelů[101]. Služba AMEE location footprinter[102] zase vypočítává na základě vašich dat z Foursquare přibližnou ekologickou stopu, kterou zanecháte v podobě vypuštěného CO_2.

Dále pak nedocenitelnou hodnotu má samotná databáze míst pečlivě zkategorizovaných, s přesnými GPS souřadnicemi, adresami, se všemi komentáři, doporučeními, fotografiemi a zejména vlastními statistikami návštěvnosti.

Google Latitude

Oproti kolokačním službám založených na ohlášení polohy pomocí checkinů funguje Google Latitude primárně na principu kontinuálního sdílení polohy za pomoci GPS souřadnic z mobilního zařízení. Přesnost aktuální polohy lze nastavit v rámci zabezpečení soukromí např. pouze na oblast města, případně omezit frekvenci aktualizace z důvodu šetření baterie mobilního zařízení. Je zde ale i možnost nahlásit svojí polohu pouze na způsob checkinu. Sdílení geografické polohy probíhá pouze mezi vzájemně potvrzenými přáteli. Stejně jako u ostatních služeb Google je i tato plně integrována do balíku služeb této společnosti a data zní lze vzájemně používat i v dalších aplikacích[103].

2.5.7 Wiki weby

Princip wiki webů je takový, že každý uživatel (po přihlášení nebo i bez něj) může za pomoci intuitivního značkovacího jazyka editovat dané stránky. Tento jednoduchý princip velice přispívá k aktualitě daných informací, avšak někteří mu vyčítají neověřitelnost informací, což ale není do značné míry oprávněná výčitka, protože většinou u těchto služeb funguje propracovaný systém kontroly obsahu s rozsáhlou databází předchozích verzí příspěvku pro případ návratu k původní verzi. Tento systém je nejvíce oblíben u služeb poskytujících faktografické informace – encyklopedií (Wikipedia[104]) nebo například u učebních skript (v Čechách velice propracovaný projekt Wikiskripta.eu[105] pro studenty lékařských fakult).

[100] Voulez-vous check-in avec moi ce soir? [online].
[101] Zatím funguje jen pro New York, Paříž a San Francisco.
[102] Track your carbon footprint with ALF. *Aboutfoursquare.com* [online].
[103] Google Latitude. Google.com [online].
[104] Wikipedia [online].
[105] Wikiskripta.eu [online].

Wikipedia

Tato otevřená internetová encyklopedie je nejznámějším internetovým wiki projektem současnosti a o její využitelnosti svědčí i hojně používané a velice dobře použitelné citace (zejména definic) v této práci. V dnešní době čítá databáze Wikipedie téměř 26 milionů hesel (článků) ve zhruba 250 různých jazycích. Největší počet hesel je v angličtině cca 3,8 milionu článků. Dalšími jazyky s mnoha hesly jsou zejména španělština, němčina, ruština, polština, portugalština, italština, francouzština a japonština. Čeština se svými více než 216 tisíci hesly patří do 2. skupiny nejčastějších jazyků. Wikipedie funguje za pomoci více než 15 milionů dobrovolníků, z kterých je 1500 administrátorů. Tito uživatelé dohlíží na to, aby všechna hesla byla pravdivá. Dále pak brání chybným nebo úmyslně poškozujícím úpravám databáze[106]. Wikipedie jako taková je krásný příklad ucelené faktografické databáze fungující na principu skupinového sdílení vědomostí[107].

[106] V první ročníku magisterského studia jsme experimentálně zkoušeli poškozovat hesla na Wikipedii v rámci výuky. Interval opravy hesla do původní podoby některým z administrátorů nebyl delší než 10 minut.

[107] Size of wikipedia. In *Wikipedia : the free encyclopedia* [online].

3 Datamining

To že žijeme v éře dat a o jejich různorodosti ve spojitosti s tím v jakých podobách se generují a sdílí v sociálních sítích, jsme se dozvěděli v předešlých kapitolách. V kapitolách následujících se dostáváme ke stěžejní části této práce, která popisuje samotný potenciál dataminingu, jeho základní techniky a vybrané techniky aplikovatelné na data ze sociálních sítí.

Samotné principy vyhledávání specifických vzorů v datech jsou známé již mnoho let. Už v 18. století byl znám Bayesovský teorém, dále pak v 19. století přišly statistické metody regresní analýzy. Obrovský vývoj ale datamining zaznamenal jako většina takovýchto oborů ve chvíli, kdy díky celkovému rozvoji počítačové vědy (informatiky) bylo možné strojově zpracovávat velké objemy dat. Napomohly tomu především objevy v druhé polovině minulého století, jako například neuronové sítě, strojová klastrová analýza, genetické algoritmy, metody rozhodovacích stromů, metody support vector machines[108] atd.

3.1 Co je to datamining?

Do českého jazyka se tento anglický termín překládá jako „dolování dat“ nebo také „vytěžování dat“. To nám ale hned tak nic konkrétního neřekne. Daleko lepší je si pro vysvětlení tohoto pojmu půjčit anglickou definici, kdy se datamining popisuje jako analytická část oblasti získávání znalostí z databází (Knowledge Discovery in Databases - KDD). Klíčové je především slovo „znalosti“, které nám na první pohled říká, že se jedná o už zpracované informace, potažmo informace mající pro nás svou informační hodnotu[109].

Datamining je mladá mezidisciplinární vědecká analytická disciplína spadající pod oblast Computer Science, která se do češtiny překládá jako informatika. Pomocí dataminingových metod, které bývají zpravidla založeny na umělé inteligenci, strojovém učení a statistice se snažíme najít modely/vzory v určitém datovém setu. Cílem dataminingu je pak extrahovat z takovýchto datasetů znalosti v člověku srozumitelné

[108] Tato metoda je podrobněji popsána v kapitole o Sentiment Analysis.
[109] Data mining. In *Wikipedia : the free encyclopedia* [online].

podobě[110]. Často také bývá tento termín chybě používán pro jakékoliv zpracování velkého množství informací nebo bývá generalizován jako část systému pro podporu rozhodování (Decision Support System – DSS), který zahrnuje umělou inteligenci, strojové učení a Business Intelligence. Klíčový pro datamining je právě termín „discovery" tedy objev něčeho nového, v případě dataminingu za pomoci analýzy dat[111].

3.2 Používané techniky

Metody dataminingu se používají zejména k analýze velkých objemů dat, za účelem získání užitečných nových modelů/vzorů (patterns). Konkrétní metody dataminingu pak odhalují specifické výsledky a používají se ke konkrétním účelům. Datamining k tomu používá šest nejzákladnějších metod

- Clusterová analýza (Cluster Analysis) – vyhledává specifické záznamy / skupiny záznamů podobných vlastností (bez pomoci znalosti již známých struktur) za účelem nalezení různých druhů podobností.
- Detekce anomálií (Anomaly/Outlier/Deviation Detection) – detekce neobvyklých druhů záznamů neodpovídajících svými parametry většině (takovéto druhy záznamů mohou mít velkou informační hodnotu).
- Nacházení asociačních pravidel (Association Rule Learning) – objevující zajímavé vazby mezi proměnnými (používáno například u úkolů typu „ten kdo si koupil produkt X si koupil také produkt Y a Z).
- Klasifikační metody - snaží se zobecnit strukturu tak, aby se dala použít na nová data.
- Regresní analýzy – snaží se definovat funkci, která co možná nejlépe popisuje dané modely.
- Sumarizační metody – vytváří lepší reprezentaci daného datasetu za pomoci vizualizace a reportů.

[110] HAN, Jiawei, Micheline KAMBER a Jian PEI. *Data mining*: *concepts and techniques* [online].

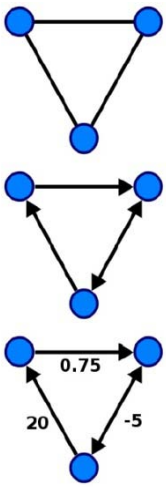
[111] Data mining. In *Wikipedia : the free encyclopedia* [online].

V následující části práce se zaměříme na použití konkrétních specifických dataminingových metod, které jsou použitelné pro data generovaná sociálními sítěmi. Tento seznam není vyčerpávající, ale obsahuje podle mínění autora kvalitní výčet technik a postupů, které demonstrují veliký potenciál dataminingu v oblasti sociálních sítí.

4 Social Network Analysis

Social network analysis[112] byla původně sociologickou technikou, která postupně získala uplatnění také v antropologii, biologii, ekonomii, geografii, komunikační teorii a pro účely této práce především informační vědě. Idea sociální sítě a její analýzy existuje již více než 100 let, ale hlouběji se jí začali vědci zabývat až v druhé polovině 20. století. S nástupem počítačů přestává být tento pojem pouze teoretickou frází a stává se vědním oborem s vlastním analytickým přístupem k tomuto paradigmatu, jasně definovanou teorií a v neposlední řadě analytickými softwary, které hrají v této oblasti v dnešní době klíčovou roli.

Social network analysis představuje zejména mapování, měření a monitorování vztahů a informačních toků mezi jednotlivci, skupinami, organizacemi, počítači, URL a dalšími navzájem propojenými informačními subjekty.[113] Jednotlivé uzly představují tyto konkrétní subjekty nebo jejich skupiny a propojení mezi nimi znázorňuje vzájemné vztahy nebo informační toky. SNA se zabývá jak vizuální, tak matematickou analýzou těchto vzájemných propojení a umožňuje tak lépe pochopit vztahy v síti.



Obrázek 2 - vztahy

Vztahy mezi uzly mohou být znázorněny třemi různými způsoby, v závislosti na tom jaký druh vzájemné vazby mezi uzly panuje.[114]

1. nesměrované (undirected): reprezentující pouze vzájemně symetrické vztahy (například přátelství na Facebooku).

2. směrované (directed): reprezentující navzájem nesymetrické vztahy (například follower na Twitteru).

3. vážené (weighted): reprezentující míru, vzdálenost nebo náklady vzájemného vztahu.

[112] Social network. In *Wikipedia : the free encyclopedia* [online].
[113] Pro potřebu této práce se vymezíme na digitální sociální sítě.
[114] HILBRICH, Robert. *Http://blog.hilbri.ch* [online].

Pro porozumění celé síti musíme v SNA určit polohu jednotlivých uzlů v síti. Tuto polohu můžeme vyjádřit centralitou daného uzlu. Tato centralita vyjadřuje určitou roli daného uzlu v síti. V závislosti na centralitě můžeme určit zhruba tyto základní role uzlů.[115]

a) Konektoři (connectors) – takovýto uzel má vysokou hodnotu degree centrality (vysoký počet vzájemných vazeb).
b) Experti (mavens)
c) Vůdci (leaders)
d) Mosty (bridges) – uzel s vysokou hodnotou betweenness centrality, spojují části sítě, které by se v důsledku jejich výpadku zcela izolovaly.
e) Osamělci (isolates) – uzly v odlehlých částech sítě. Jejich pozice by se mohla zdát dost nevýhodná, avšak toto zdání je pouze relativní, protože se snadno mohou stát velmi důležitými mosty vůči zcela jiné síti.

Samotná centralita uzlu má pak čtyři základní parametry, které můžeme měřit. Patří mezi ně stupeň uzlu nebo také počet hran (Degree centrality), centralita středové blízkosti (Closeness centrality), centralita mezilehlosti (Betweenness centrality) a vážený počet hran (Eigenvector centrality)[116]. Z důvodu relativně obtížně přeložitelné a ne zcela jednotné terminologie, se při popisu těchto veličin uchýlím k anglickým termínům, u kterých nedojde k jakékoli nejasnosti nebo záměně.[117]

Jednotlivé grafy sítí byly zpracovány pomocí softwaru NodeXL, což je nástavba tabulkového procesoru Microsoft Excel.[118]

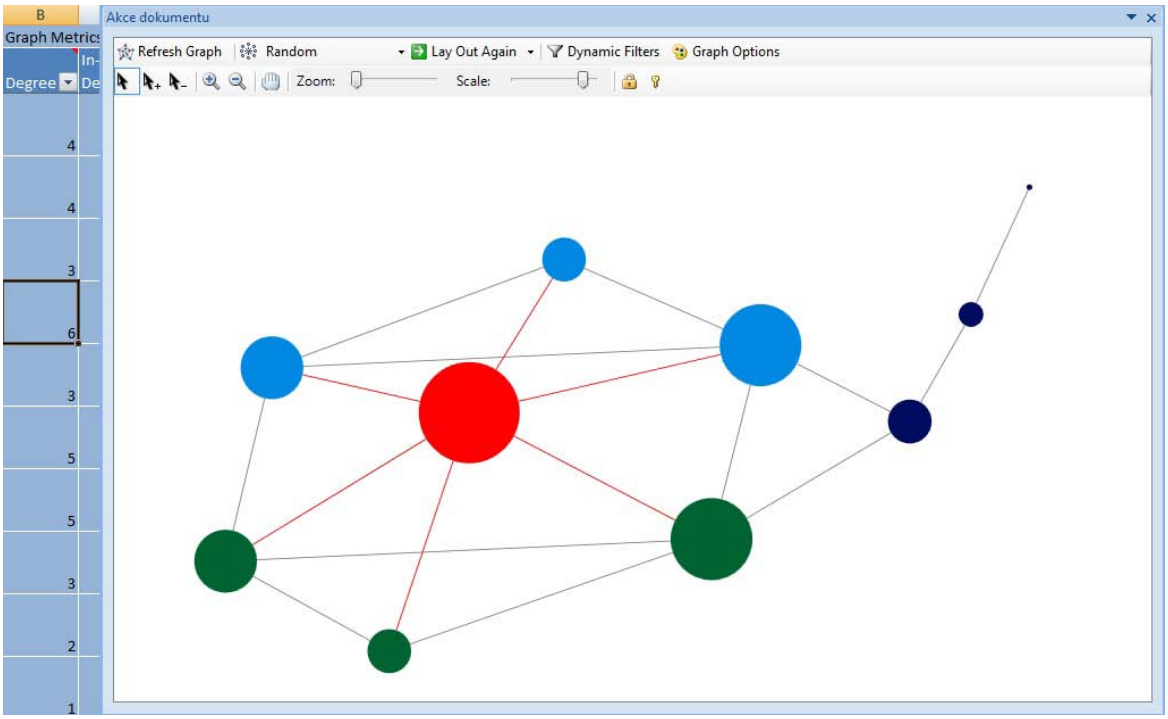
[115] Social Network Analysis: A Brief Introduction. *Orgnet.com* [online].
[116] Také můžeme použít Katz centralitu, která nepatří do základní čtveřice.
[117] Centrality. In *Wikipedia : the free encyclopedia* [online].
[118] *NodeXL : Network Overview, Discovery and Exploration for Excel* [online].

4.1 Degree centrality

Degree centrality se v češtině často překládá jako stupeň uzlu nebo jako počet hran uzlu. Tento parametr vyjadřuje v podstatě počet přímých vazeb, které mezi sebou jednotlivé uzly mají. Také jinak, čím vyšší stupeň uzlu daný uzel má, tím má větší počet přímých vazeb na ostatní uzly. Jestliže se pak jednotlivé uzly nacházejí v orientované síti (jsou známé orientace vazeb), tak můžeme tento parametr ještě rozdělit na Indegree centrality a Outdegree centrality. Indegree centrality vyjadřuje počet vazeb směřujících z okolí k danému uzlu. Outdegree centrality vyjadřuje naopak počet vazeb vedoucích od uzlu směrem k okolním uzlům. Indegree tedy můžeme chápat jako jakousi „oblíbenost" nebo „popularitu" v síti, naproti tomu Outdegree si lze vysvětlit jako „společenskost" daného uzlu.[119] [120]

Uzly v síti s nejvyšším stupněm centrality patří k nejaktivnějším v síti a říká se jim „spojovatelé" (Connector nebo Hub). Jak můžeme vidět na Obrázku 3, tak by se na první pohled dalo říci, že čím větší stupeň centrality (na obrázku znázorněn velikostí uzlu), tím více vazeb s ostatními uzly a tím pádem lepší umístění v síti. Toto rčení není vždy zase až tak pravda, protože si zde při hlubším prozkoumání můžeme povšimnout, že středový uzel spojuje zejména ty uzly, které už vzájemné vazby mezi sebou mají.[121]

[119] Centrality. In *Wikipedia : the free encyclopedia* [online].
[120] Social Network Analysis: A Brief Introduction. *Orgnet.com* [online].
[121] Tamtéž.

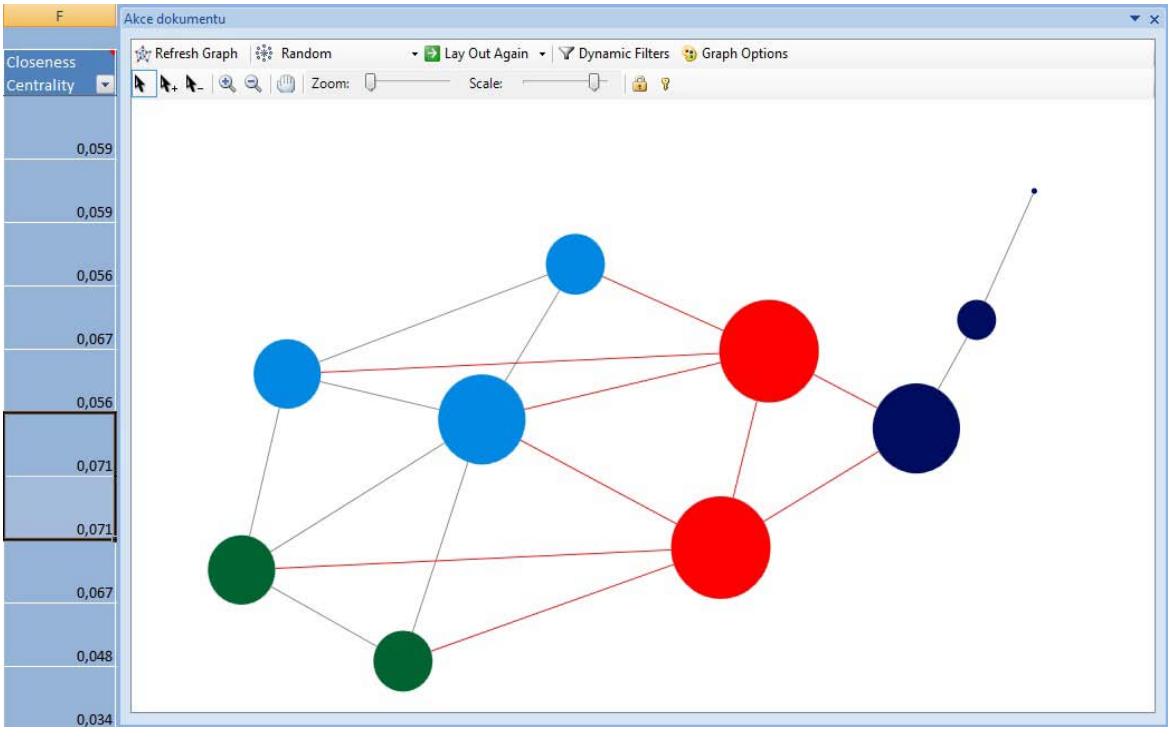


Obrázek 3 - Degree centrality[122]

[122] *NodeXL : Network Overview, Discovery and Exploration for Excel* [online].

4.2 Closeness centrality

Closeness centrality se do češtiny překládá jako středová blízkost. Hodnota konkrétního uzlu se pak získá součtem všech nejkratších vzdáleností ke všem ostatním uzlům v síti. Jde tedy o nejkratší možnou vzdálenost ke všem ostatním uzlům od konkrétního uzlu. Jinými slovy se tedy jedná o jakousi nejlepší/nejvýhodnější polohu v síti, kdy uzel s největší středovou blízkostí má ze všech uzlů sítě ke všem ostatním uzlům sítě nejblíže. Jak je dobře na první pohled vidět na Obrázku 4 zobrazujícím jednotlivé uzly, které svojí velikostí znázorňují svojí středovou blízkost. Uzly s největší středovou blízkostí jsou v tomto případě vzdáleny od všech ostatních uzlů sítě maximálně dvěma kroky. Uzly s takovýmto umístěním mají výbornou možnost co nejlépe monitorovat informační toky v síti.[123] Proudí přes ně veliké množství dat.

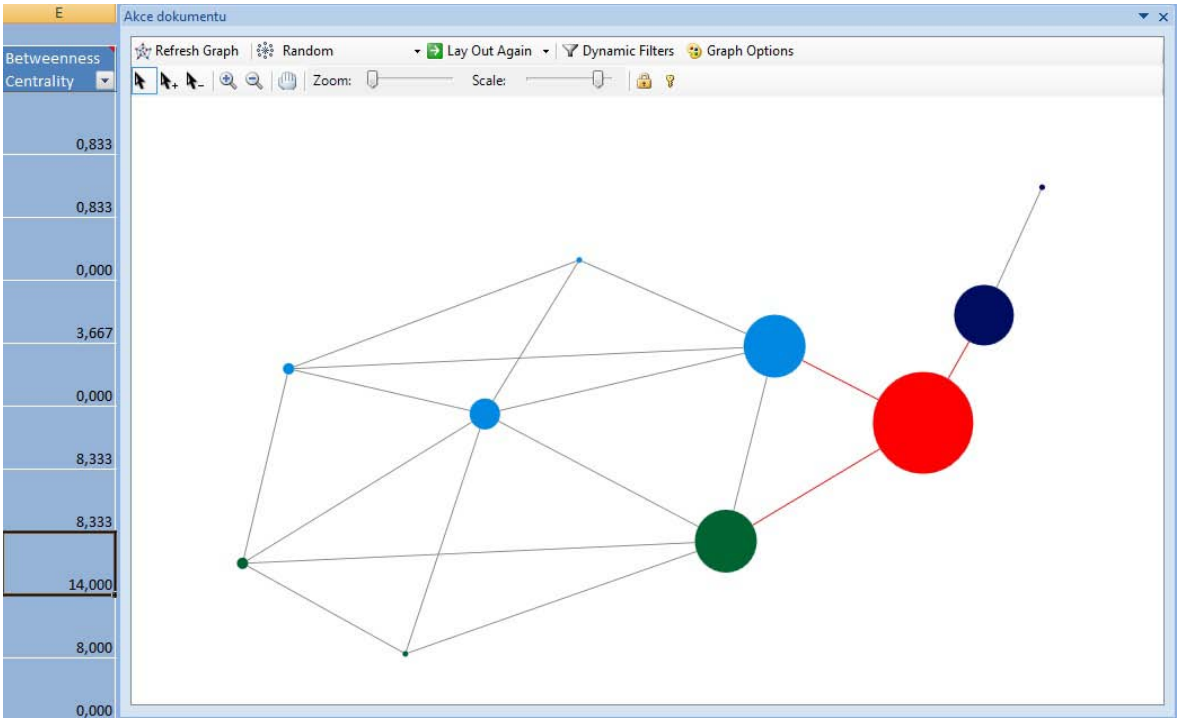


Obrázek 4 - Closeness centrality[124]

[123] Social Network Analysis: A Brief Introduction. *Orgnet.com* [online].
[124] *NodeXL : Network Overview, Discovery and Exploration for Excel* [online].

4.3 Betweenness centrality

Betweenness centrality se do češtiny překládá nejčastěji jako mezilehlost v síti a vyjadřuje, kolikrát stojí konkrétní uzel „v cestě“ ostatním uzlům při výpočtu Closeness centrality, tedy při výpočtu nejkratších tras ke všem uzlům sítě. Jak je patrné z obrázku 4, který znázorňuje velikostí uzlu hodnotu mezilehlosti, tak takovéto uzly zpravidla propojují oddělené části/oblasti sítě a hrají tedy v dané síti velmi důležitou roli „zprostředkovatele“. Říká se jim z tohoto důvodu také „Mosty“. Na druhé straně v případě výpadku tohoto uzlu se rozpadá daná síť na dvě a více částí. Kvůli tomu mají uzly s vysokou mezilehlostí obrovský vliv na informační toky v síti. [125] [126]

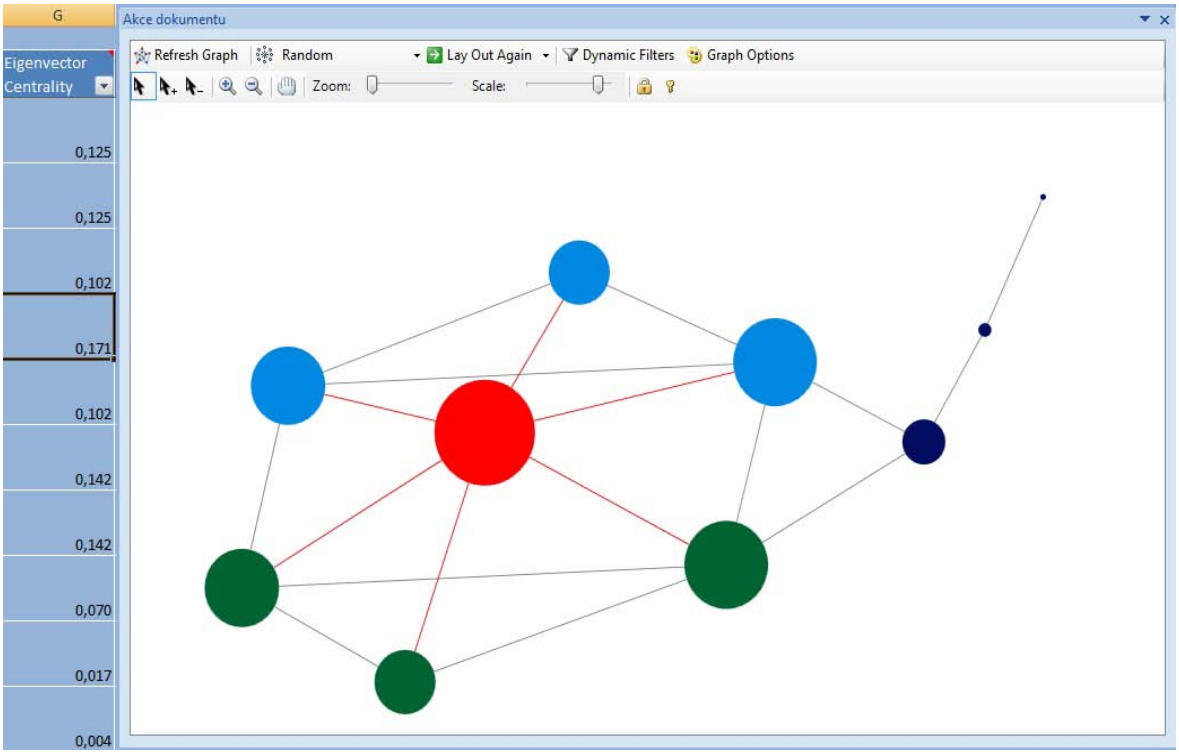


Obrázek 5 - Betweeness centrality[127]

[125] Social Network Analysis: A Brief Introduction. *Orgnet.com* [online].
[126] Centrality. In *Wikipedia : the free encyclopedia* [online].
[127] Social Network Analysis: A Brief Introduction. *Orgnet.com* [online].

4.4 Eigenvector centrality

Je veličina vyjadřující „důležitost“ uzlu v síti. Do češtiny se také někdy překládá jako vážený počet hran uzlu. Velikost tohoto parametru vychází z předpokladu, že spojení s uzly s vyšší důležitostí zvětšují důležitost tohoto uzlu více nežli spojení s uzly důležitosti menší. Na podobném principu funguje např. Google Page rank.[128] Na obrázku můžeme vidět, že v naší vzorkové síti má nejvyšší hodnotu Eigenvector centrality centrální uzel, který má i vysokou hodnotu Degree (je propojen s velkým množstvím dalších uzlů) a zároveň uzly s kterými je propojen, tak mají také relativně velký počet ostatních vazeb.



Obrázek 6 - Eigenvector centrality[129]

[128] Centrality. In *Wikipedia : the free encyclopedia* [online].
[129] Social Network Analysis: A Brief Introduction. *Orgnet.com* [online].

4.5 Katz centrality

Katzova centralita představená Leo Katzem[130] v roce 1956 je obdobou Eigenvector centrality, čili se používá k měření významu uzlu v síti. Tento relativní vliv měří na základě výpočtu, který vychází ze součtu bezprostředních vazeb uzlu (vazby 1. úrovně), ke kterým připočítává i všechny další vazby mezi uzly v síti. Váha těchto vazeb (2. – n. úrovně) je ale penalizována útlumovým faktorem α. Každá cesta mezi dvěma uzly je počítána jako n kroků, které se musí udělat k vzájemnému dosažení. Na základě počtu těchto kroků je určena vazba n. úrovně a hodnota faktoru α, který je roven α^{n-1}. Oproti Eigenvector centrality má výhodu, že se dá použít i v acyklických směrovaných sítích, kde je předchozí metrika většinou nepoužitelná[131].

4.6 Celková centralita sítě

Z dalšího hlediska se nám jedná o celkovou centralitu sítě. Síť, která je vysoce centralizovaná, což znamená, že obsahuje pouze několik centrálních uzlů, které tvoří hlavní spojnici mezi ostatními uzly je velice náchylná na výpadky přenosu informace. Těmto centrálním uzlům se říká HUBy a mají vysoké hodnoty degrese a betweenness centrality. Pakliže jsou tyto centrální uzly odstraněny (vyřazeny z provozu), tak nastává situace, kdy se celá síť rozdělí na mnoho menších, navzájem nepropojených sítí.[132]

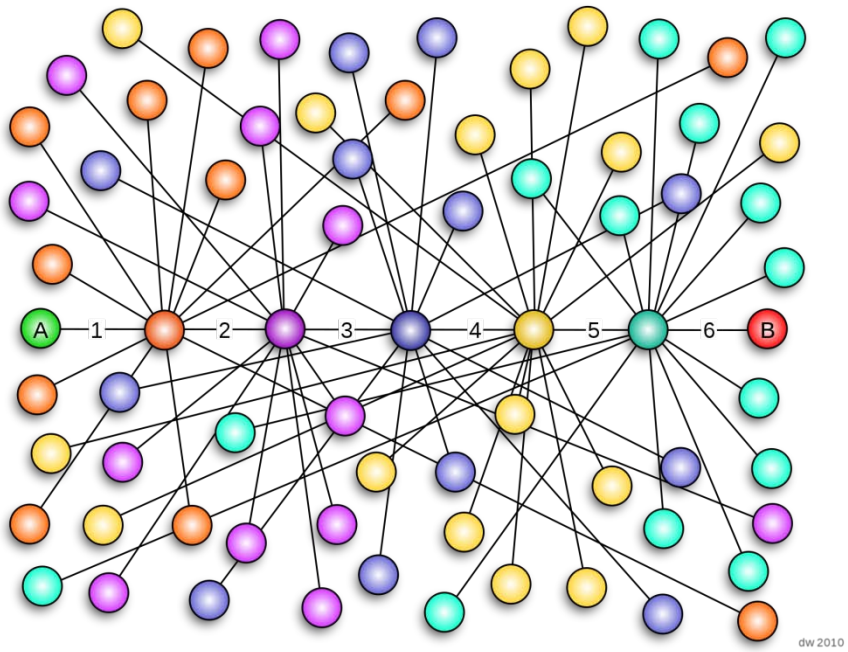
4.7 Dosah sítě a šest stupňů separace

Důležitým měřítkem je také dosah každého uzlu v síti a vzdálenost vzájemných vazeb, který nebývá zpravidla větší než 3 kroky. Tento počet většinou určuje pomyslný horizont, za který není „vidět“. Každý uzel tedy žije v jakémsi svém „malém světě“, za jehož horizontem již nemá dostatečnou možnost šířit své informace, případně svůj vliv. Běžné je, že má uzel dosah maximálně přes 2 kroky, který odpovídá „přátelům přátel“ (Friend of a Friend - FOAFs) Na tomto principu byl, ještě dávno před spuštěním Facebooku[133], spuštěn projekt FOAF-project[134]. Ten si kladl za úkol vytvořit strojově

[130] Americký statistik, který svými objevy velice přispěl Socila Network Analysis.
[131] Katz centrality. In *Wikipedia : the free encyclopedia* [online].
[132] Social Network Analysis: A Brief Introduction. *Orgnet.com* [online].
[133] Facebook byl spuštěn v únoru 2004.

čitelnou podobu osobních stránek uživatelů a zachytit v ní jejich vzájemné vztahy a propojenost mezi aktivitami, které dělají a věcmi, které tvoří. Toto všechno bylo možné díky technologii RDF[135].

Z tohoto pohledu, když by dosah sítě byl větší než 2, maximálně 3 kroky, tak by se mohl prakticky „znát každý s každým“. Podle zákonitostí, které měl dokázat „Experiment malého světa“ (Small World Experiment)[136] Stanleyho Milgrama. Tomuto experimentu se také často říká „Šest stupňů separace“ (Six degrees of separation)[137]. Experiment byl založen na zvědavosti zjistit, jaká je pravděpodobnost, že by se dva zcela náhodně zvolení lidé mohli navzájem „znát“. Z našeho pohledu Social Network Analysis, bychom řekli, že nás zajímá, jakou průměrnou vzdálenost (vyjádřenou počtem kroků) mezi sebou mají dva náhodně vybraní lidé.



Obrázek 7 - graf znázorňující spojení dvou uzlů pomocí max. 6 kroků (zdroj: wikimedia.org)

I když tento pokus byl uskutečněn na konci 60. let minulého století z dnešního měřítka za spartánských podmínek (metoda kontaktování byla realizována běžnou poštou) a pouze na území USA. Pokus dokázal, že průměrný počet kroků mezi dvěma naprosto neznámými lidmi byl v rozmezí 5,5 – 6 kroku. Nedávno se stejnou hypotézu

[134] *The Friend of a Friend (FOAF) project* [online].
[135] *Resource Description Framework (RDF)* [online].
[136] Small world experiment. In *Wikipedia : the free encyclopedia* [online].
[137] Název termínu inspirovaný knihou maďarského autora Frigyese Karinthyho.

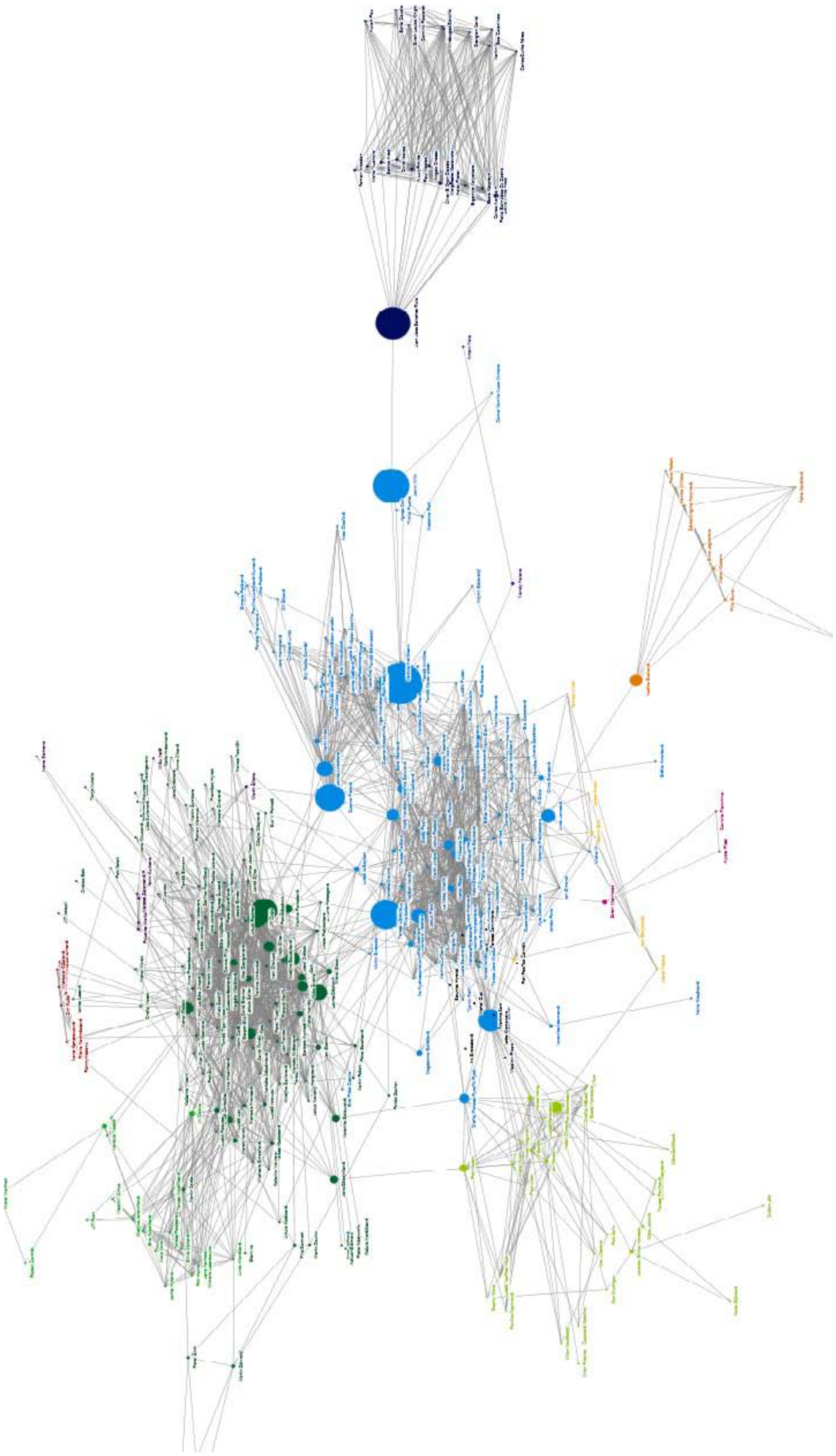
pokusili ověřit vědci ze samotného Facebooku. Zkoumáno bylo cca 700 milionů uživatelů (vzorek 10% celkové populace světa), mezi nimiž bylo cca 69 miliard vzájemných vazeb. I když čísla nelze přímo srovnávat, protože metodiky byly odlišné (Milgram měl k dispozici pouze pohlednice a vzorek 296 dobrovolníků[138]) a vztahy mezi nimi. Oproti vědcům z Facebooku, kteří měli komplexní data o všech svých uživatelích. Výsledek ale vyšel jednoznačně a vypovídá o tom, že vzájemné vazby mezi všemi uživateli se rok od roku zkracují. V roce 2008 to bylo podle dat Facebooku zhruba 5,28 kroku oproti současnosti, kdy vzdálenost tvoří pouze 4,74 kroku.[139] [140] A jejich zkracování bude probíhat i nadále v důsledku rozvoje Facebooku.

4.8 Využití v praxi (příklady NodeXL)

4.8.1 Autorova síť přátel na Facebooku

V praxi nám určení těchto parametrů pomůže určit například hlavní konektory v síti. (viz Obrázek 8 - schéma autorovy sítě přátel na Facebooku.) Na obrázku je pomocí softwaru NodeXL znázorněná autorova síť přátel na Facebooku. Jednotlivé barvy krásně určují, které podsložky sítě (podskupiny uzlů – „přátel“) spolu souvisí nejvíce[141] (mají nejvíce vzájemných vazeb) a velikost jednotlivých nodů je v tomto případě odvozena od betweenness centrality (mezilehlosti), která jednoznačně určuje uzly, které jsou nejdůležitějšími konektory v síti (tj. propojují oddělené části sítě mezi sebou). Z tohoto hlediska v grafu chybí uzel samotného autora, protože je zřejmé, že by byl v centru a v závislosti na mezilehlosti by byl i zároveň největším uzlem v síti, propojujícím veškeré její součásti[142].

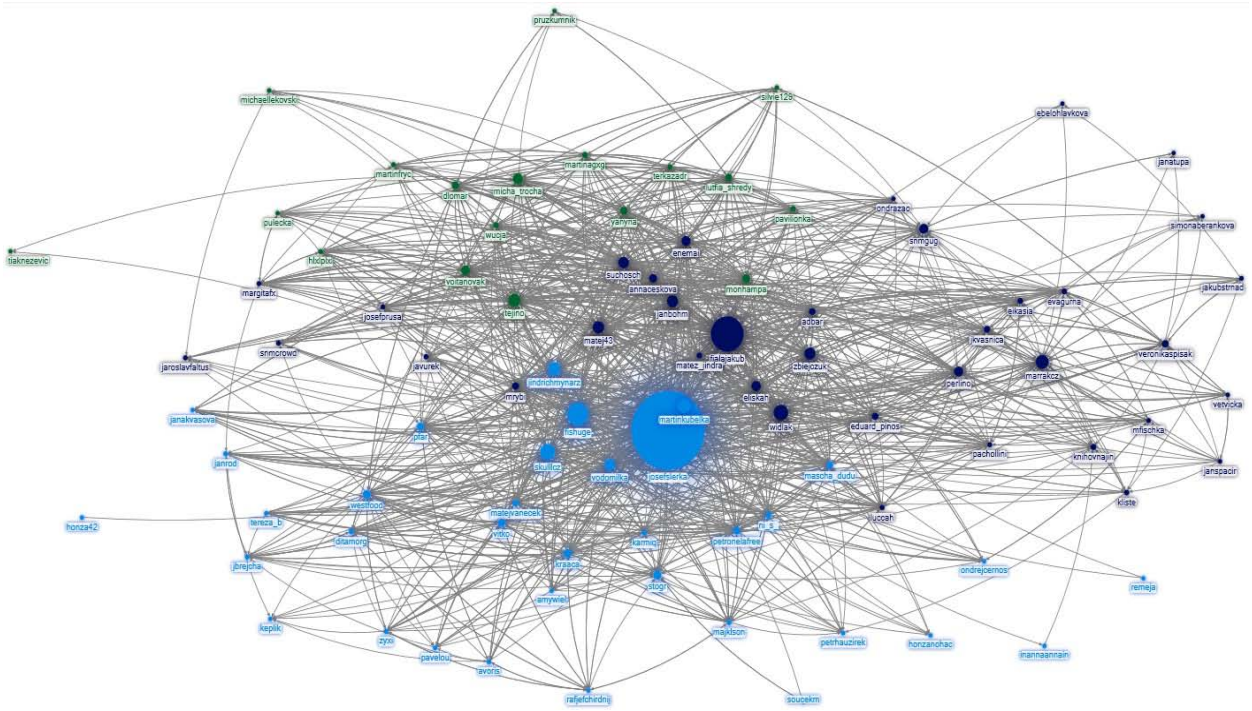
[138] Ze kterých k cíli dorazilo pouze 64 dopisů.
[139] UGANDER, Johan . *The Anatomy of the Facebook Social Graph*. [online].
[140] BACKSTROM, Lars . *Four Degrees of Separation*. [online].
[141] Zelená barva určuje autorovi dlouholeté přátele zejména z Prahy. Světle modrá skupina uzlů se ještě sama odděluje lehce na dvě části zhruba ve své polovině. Tato skupina znázorňuje vazby mezi uzly reprezentující autorovi přátele z vysoké školy. A to jak z bakalářké, tak magisterské části studia (proto dvě části). Tmavě modrá skupina pak reprezentuje přátele poznané na studiích v zahraničí. Ostatní menší skupinky znázorňují skupiny přátel, kteří mají nějaký další „společný jmenovatel“.
[142] Kompletní .xlsx soubor s grafem se nachází v externí příloze 1. Prohlížet ho lze v softwaru MS Excel po doinstalování pluginu NodeXL.
NodeXL : Network Overview, Discovery and Exploration for Excel [online].



Obrázek 8 - schéma autorovy sítě přátel na Facebooku

4.8.2 Graf sítě followerů skupiny @stunome/stunome

Na Obrázku 9 pak vidíme graf vytvořený taktéž pomocí softwaru NodeXL, který znázorňuje propojení uzlů ve skupině Stunome na Twitteru.[143] Vzájemné propojení demonstruje vztahy jednotlivých uzlů mezi sebou a velikost uzlu udává Closeness centrality, která představuje jakousi nejvýhodnější polohu v síti (viz Closeness centrality) V tomto ohledu ve skupině Stunome[144] má nejvýhodnější polohu uživatel Josef Šlerka[145] a Jakub Fiala[146](což vzhledem k jejich funkci není vůbec překvapivé). Tito uživatelé mohou snadno zasahovat do veškerých diskusí, které se vedou v této sociální skupině a zároveň velice snadno monitorovat tyto informační toky[147].



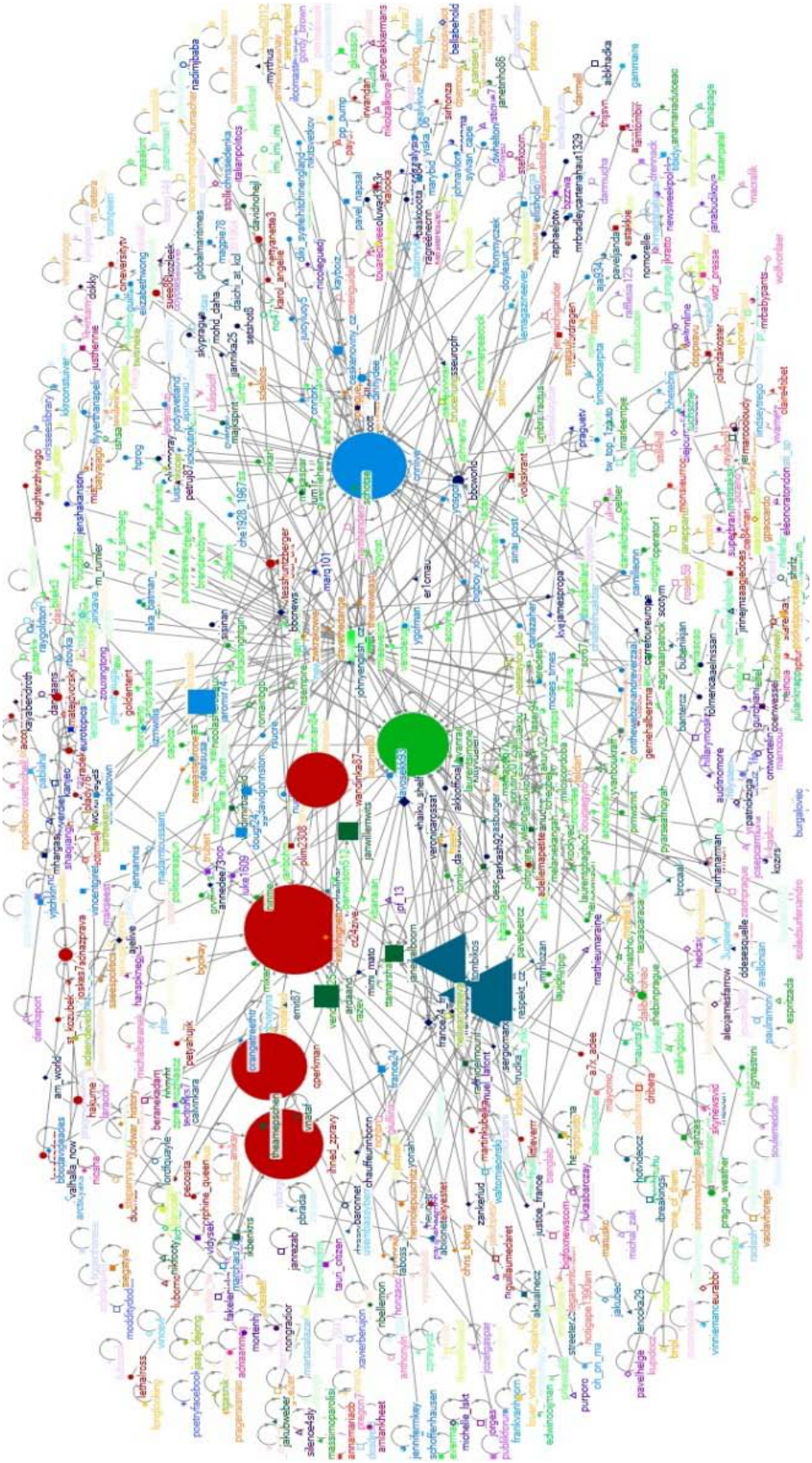
Obrázek 9 - graf znázorňující skupinu @stunome/stunome na Twitteru

[143] @stunome/stunome – seznam uživatelů twitteru, kteří mají něco společného s tímto studijním oborem.
[144] Studia Nových Médií při FF UK.
[145] Vedoucí Studií Nových Médií.
[146] Tajemník Studií Nových Médií.
[147] Kompletní .xlsx soubor s grafem se nachází v externí příloze 2. Prohlížet ho lze v softwaru MS Excel po doinstalování pluginu NodeXL.
NodeXL : Network Overview, Discovery and Exploration for Excel [online].

4.8.3 Graf twitterové komunikace v souvislosti se mrtí Václava Havla

Na následujícím obrázku je znázorněna komunikace na Twitteru v průběhu pohřbu Václava Havla. Velikost uzlů znázorňuje Betweenness centrality a ukazuje nám, že v centru konverzací se umísťovaly především twitterové účty zpravodajských stanic. Za českou republiku to je @ct24_zive, @ihned_zpravy a @respekt_cz. Ze zahraničních serverů jsou to hlavně @cnnlive a @bbcworld. Dále pak můžeme vidět i další konkrétní velmi retweetované a do konverzací zapojené uživatele z celého světa[148].

[148] Kompletní .xlsx soubor s grafem se nachází v externí příloze 3. Prohlížet ho lze v softwaru MS Excel po doinstalování pluginu NodeXL.
NodeXL : Network Overview, Discovery and Exploration for Excel [online].



Obrázek 10 - graf znázorňující komunikaci na Twitteru ve spojitosti s úmrtím Václava Havla.

4.8.4 Graf twitterové komunikace na akci Barcamp

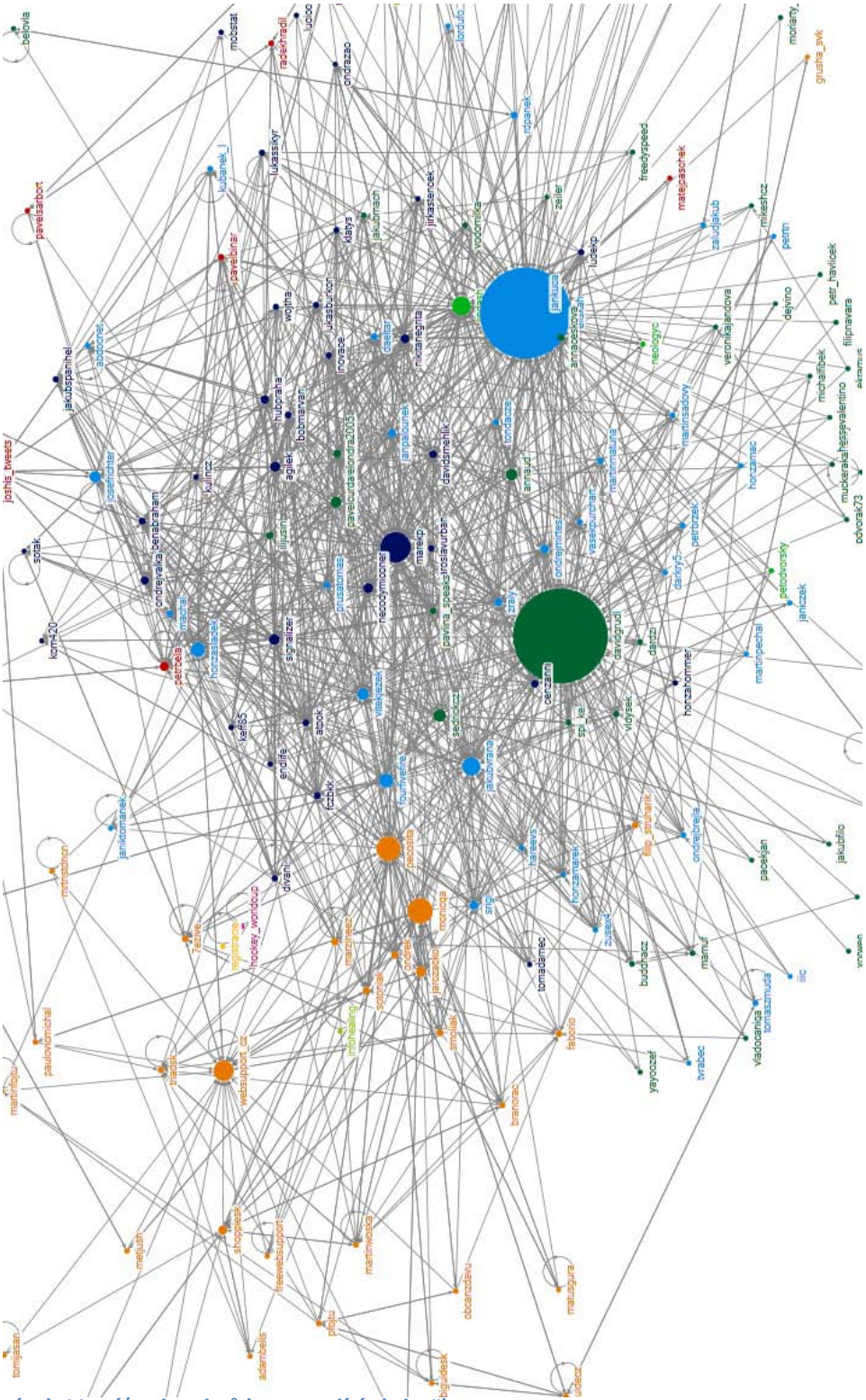
Jako poslední příklad si uvedeme síť vzájemných vztahů twitteristů komentujících akci BarcampPraha2011, která se odehrála 10. 12. 2011 v Praze[149]. Jednalo se o jednodenní konferenci a workshop věnovaný moderním technologiím a sociálním médiím. Opět pomocí SNA softwaru NodeXL je velice snadné určit hlavní spojující články celé skupiny těchto social media specialistů. Pomocí metriky Betweenness centrality jsou opět znázorněny uzly (uživatelé), které tvoří hlavní pojící prvky (mosty) v této specifické komunitě. Jak jasně vyplývá z obrázku níže, tak tuto důležitou roli zde plní zejména dva uživatelé a to David Grudl[150] a Eliška Hutníková[151].[152]

[149] *Barcamppraha.cz* [online].

[150] Český programátor, který se zejména proslavil svými kontroverzními výroky na twitteru, za které si u některých uživatelů vysloužil přezdívku „největší prase českého internetu".
ZANDL, Patrick. Hon na i-prasata a cyniky z diskusního podpalubí. *Lupa.cz* [online].

[151] Social Media Geek Girl pracující v H1.cz.

[152] Kompletní .xlsx soubor s grafem se nachází v externí příloze 4. Prohlížet ho lze v softwaru MS Excel po doinstalování pluginu NodeXL.
NodeXL : Network Overview, Discovery and Exploration for Excel [online].



Obrázek 11 - síť twitteristů komentující akci s #barcampcz

4.9 Alternativní metriky a měření vlivu

Právě s příchodem sociálních sítí začalo být velice aktuální měření vlivu v elektronické sféře. Existuje relativně velké množství služeb, které se pokouší určovat tento vliv za pomoci různých metrik. Tyto metriky většinou fungují za pomoci sofistikovaných algoritmů, které si vývojáři těchto služeb střeží jako oko v hlavě. Pro příklad si můžeme uvést třeba Peer Index[153], Tweet Rank[154] a podobné. Nejpoužívanější a nejdiskutovanější v současnosti je bezesporu služba Klout[155].

[153] Peer Index [online].
[154] Tweet Rank [online].
[155] Klout [online].

4.9.1 Klout

Když upustíme od určování důležitostí uzlu (uživatele) na základě polohy v síti, která je nesporně vypovídající, tak přicházejí do popředí metriky, které sice mohou být označeny za ne zcela průkazné (z hlediska sporné metodologie[156]) ale v praxi za velice dobře použitelné. V současnosti je to zejména projekt Klout[157], který si dává ambice měřit vliv (nebo spíše oblíbenost) na sociálních sítích. Dále pak si dává za úkol určit témata nebo oblasti témat, v kterých je daný uživatel influencerem (vlivnou osobností). To že je takto metrika relativně dobře použitelná v praxi dokazuje její využití například na serveru klaboseni.cz[158].

Klout vyjadřuje vlivnost/oblíbenost[159] uživatele na sociálních sítích pomocí tzv. Klout Score a ještě dalších parametrů blízce s ním souvisejících. Jedná se o True Reach (reálný dosah), Amplification (míra šíření) a Network (parametr sítě).

Klout Score – je hodnota pohybující se od 1 do 100, vyjadřující hodnotu vlivu uživatele na sociální síti/ích. Čím je hodnota vyšší, tím vyšší tento vliv je.

True Reach – je reálná velikost publika, ke kterému se dostane obsah šířený uživatelem. Od těchto uživatelů se očekává potenciální reakce na tento obsah.

Amplification – udává míru „ovlivnění“ ostatních uživatelů, míru/množství reakcí na příspěvek (komentář, sdílení apod.).

Network – udává míru vlivu uživatelů v dosahu True Reach, především jak moc reagují ti „nejvlivnější“. Tento parametr pak udává celkovou vlivnost uživatelovi sítě.

---

[156] Samotný Klout z logických důvodů nikde neuveřejňuje svůj přesný algoritmus, kterým vypočítává jednotlivé hodnoty parametrů Klout Score.

[157] About Klout. *Klout.com* [online].

[158] *Klábosení : O čem se klábosí na českém a slovenském Twitteru.* [online].

[159] Z důvodu toho, že výpovědní hodnota Klout Score není vždy 100% vypovídající, tak bude ve spojitosti s touto metrikou uváděna spíše oblíbenost v sociální síti, nežli skutečná vlivnost. Podrobně se touto problematikou zabýval v prostředí českých sociálních sítí například Daniel Dočekal (internetový publicista) ve svém článku k příležitosti změny algoritmu Kloutu:
DOČEKAL, Daniel. Klout změnil algoritmus a TOP českého Twitteru se otřáslo. *Justit.cz* [online].
A zejména pak ve svém experimentu týkajícího se vylepšení svého Klout Score na základě odpojování některých měřených sociálních sítí.
DOČEKAL, Daniel. Experiment #klout - vylepšení skóre? Odebírejte věci, jak snadné. *Pooh.cz* [online].
Tento experiment však neměl zcela jasný výsledek a spíše dokázal relativní použitelnost Klout Score.

Co klout reálně měří?

Všechny tyto parametry mohou znít docela „tajuplně“, ale v reálu se za nimi skrývá relativně jasný výčet parametrů, které nejvíce ovlivňují výsledné hodnoty Klout Score a jeho ostatních prvků. Klout si bere data získávaná z mnoha druhů sociálních sítí jako například Twitter, Facebook, Google+, LinkedIn, Foursquare, Youtube, Blogger, Wordpress a dalších. V současné době ale nejsou zpracovány metriky ani ne pro polovinu z nich. Do Klout Score a jeho podsložek se tedy započítávají zatím jen tyto parametry[160]:

- Twitter: retweety a zmínky
- Facebook: Komentáře, Zprávy na zdi, „Lajky“ (likes)
- Google+: Komentáře, Sdílení dál, +1
- LinkedIn: Komentáře, „Lajky“ (likes)
- Foursquare: Tipy a splněné tipy (ToDo’s)

Klout se též vůbec nevypořádává s problematikou jedinečnosti šířeného obsahu. Z pohledu této metriky bude mít daleko větší vliv člověk pouze sdílející/přeposílající nějaký obsah, nežli člověk generující jedinečný obsah, avšak v daleko menší míře. Řešení tohoto problému je ale obecně velice obtížné.

Role

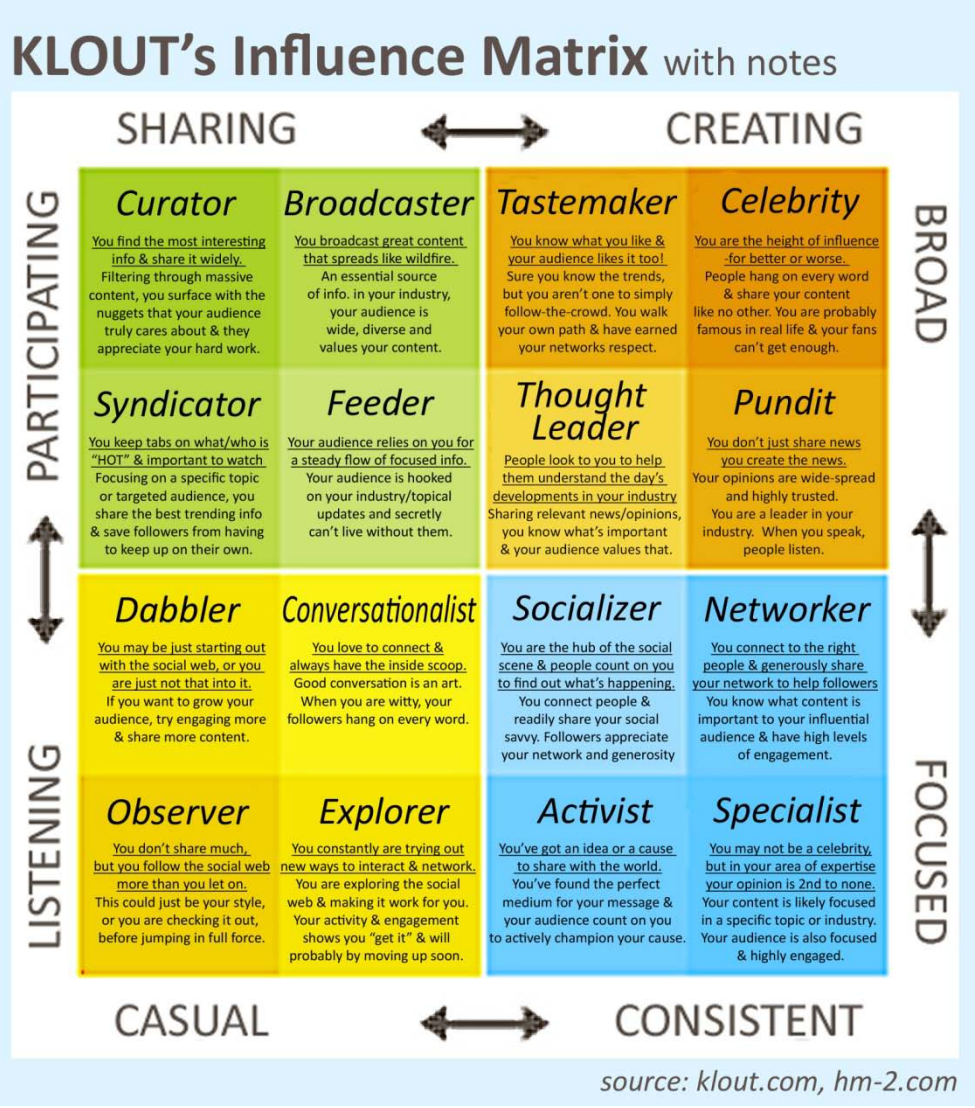
Na základě chování uživatelů v sociálních sítích a hodnot jejich jednotlivých parametrů Klout Score přiděluje Klout pro lepší porozumění uživatelům role (Klout Style). Tato role je odvozena zejména podle:

- míry jejich záběru a zaměření (široký záběr vs. úzce zaměřený – odvozuje se pravděpodobně od rozličnosti a počtu témat, viz níže oddíl témata)
- míry sdílení oproti míře vytváření (nejedná se o jedinečnost obsahu jako takového, ale o jedinečnost konkrétního sdělení, např. tweetu na Twitteru oproti „sdílení“ sdělení – retweetu)
- míry participace (naslouchající vs. živě se účastnící diskuse)
- míry účasti (pravidelná vs. příležitostná)

[160] What does Klout Measure?. *Klout.com* [online].

Musíme mít ale na paměti, že uživatelské role podle Kloutu jsou vytvářeny právě a jedině na základě toho, jakým způsobem uživatelé vytvářejí/šíří/sdílí informace vůči svému okolí. V žádném případě tyto role nemohou reflektovat skutečnou povahu osobnosti člověka. Z metrik Kloutu může i skutečná celebrita, generující naprosto unikátní obsah, obdržet v porovnání s ostatními uživateli zcela podprůměrné statistiky. Důvodem je, že Klout neupřednostňuje (a ani ze svého principu nemůže upřednostňovat) kvalitu nad kvantitou.

Nejlépe jednotlivé role vystihuje následující tabulka znázorňující je vzhledem k jejich poloze v matici v závislosti na výše uvedených parametrech.



Obrázek 12 - role podle Kloutu, zdroj. Klout.com[161]

[161] Klout Doubt?. *Likeable.com* [online].

Témata

V neposlední řadě je velice důležitým parametrem funkčnosti Kloutu určování témat, v kterých jsou uživatelé influencery (vlivnými). Na základě textové analýzy všech sdělení jsou uživateli navrhnuta témata, v kterých je (pravděpodobně) vlivný. Nestačí ale pouze psát sdělení na nějaké téma, ale ostatní uživatelé musí na tato sdělení také dostatečně reagovat. Následně je možné seznam témat editovat, protože tato funkce není zdaleka bezchybná a někdy Klout přidělí uživateli naprosto nesouvisející témata. Jelikož se jedná o textovou analýzu, tak má Klout zejména problém s uživateli, kteří píší v jiném jazyce, nežli v angličtině (pak se člověk někdy dočká vskutku kuriózních výsledků). Pro zpřesnění a utvrzení se ve správnosti výběru témat mohou ostatní uživatelé přidávat body „K“ k tématům ostatních uživatel. Tato funkčnost zde figuruje jako lidská korektura zpřesňující automatickou detekci.[162]

[162] Improved Klout Topics. *Klout.com* [online].

5 Sentiment Analysis

V předchozí kapitole jsme se zabývali tím, jak je důležitá poloha v síti a jakým způsobem měřit vlivnost uživatel. Tato kapitola nám prozradí, jak využívat data ze sociálních sítí ke zjištění toho, co si lidé myslí. Jak již bylo vylíčeno v samotném úvodu práce, v poslední době s nástupem sociálních služeb jsme se dostali do stavu, kdy jsme zavaleni daty ze sociálních sítí, která jsou doslova „názorů plná“ a samotní uživatelé jsou dychtiví „pověsit“ své názory na internet. My máme jedinečnou možnost tyto názory použít v náš prospěch. Správná analýza těchto dat nám dává do rukou cenné informace použitelné například v oblasti marketingu. Znalost toho co si ostatní lidé myslí nám může také velice usnadnit rozhodovací procesy ve firmě/osobním životě apod.[163]

I když dnes již konečně máme možnost pracovat s takovýmto množstvím dat ze sociálních médií, nese to sebou jednoznačný problém, jak toto kvantum článků, příspěvků, komentářů atd. co nejsnadněji a nejlépe automaticky zpracovat. V tuto chvíli přicházejí na řadu metody Sentiment Analysis a Opinion Miningu[164].

Základním úkolem Sentiment Analysis je určení „sentimentu“ daného výroku/zprávy/článku. Zpravidla se jedná o to, jestli je výrok pozitivní, neutrální nebo negativní. Případně může mít tato stupnice ještě více podstupňů. U sofistikovanějšího dělení můžeme ještě rozdělit určení sentimentu pro nositele postoje[165] (holder) a cíl (target), ke kterému je daný postoj zaujat. K tomuto cíli můžeme jít hned několika různými způsoby. Využívají se techniky z oblastí zpracování přirozeného jazyka (Natural Language Processing), počítačové lingvistiky (Computational Linguistic) a textové analýzy (Text Analysis). Konkrétně technicky jde např. o strojové učení (Machine Learning), latentní sémantickou analýzu (Latent Semantic Analysis), SVM (support report machines), Bayesovské filtrování (Bayes Filtering) a další. Tyto techniky můžeme rozdělit do hlavních dvou skupin, což jsou pravděpodobnostní techniky a techniky založené na porovnávání dat s referenčními slovníky. Přesnost tohoto strojového určení se měří srovnáním s tím,

[163] Zejména firmy jsou lačné po informacích, co si o nich jejich zákazníci povídají, jaké názory mají na jejich produkty a služby a v jaké souvislosti jsou tyto produkty zmiňovány. (Více k této problematice bude též v kapitole o Social Media Monitoringu.)

[164] PANG, Bo ; Lillian LEE. Opinion Mining and Sentiment Analysis. Foundations and Trends in Infomration Retrieva [online]. 2008, s. 1-135 [cit. 2011-12-21].

[165] Osoba, která zaujímá daným výrokem nějaký postoj (kladný, neutrální, záporný)

jak by daný výrok hodnotil člověk[166]. Porovnání je vyjádřeno na dvou typech hodnot Precission a Recall[167] [168]. Tyto parametry přesnosti bývají často prubířským kamenem celé problematiky.

5.1 Měření úspěšnosti - Precission a Recall

Hodnotu parametru hodnocení úspěšnosti Precission získáme, pakliže například u určení sentimentu slov program vyhodnotí, že 60 slov je s kladným sentimentem. Z těchto slov ale bylo vyhodnoceno 20 slov chybně, protože kladný sentiment neměly. Hodnota precission tedy bude 40/60 tedy 2/3. V celkové množině slov ale program dalších 80 slov s pozitivním sentimentem vůbec nenašel. Obdobným způsobem pak získáme hodnotu Recall, které se pak rovná 40/120, tedy 1/3. Parametr Precission se tedy rovná počet správně vyhodnocených slov s kladným sentimentem děleno počtem všech, které byly takto vyhodnoceny. Parametr Recall se pak rovná počet správně vyhodnocených s kladným sentimentem děleno počtem všech správně vyhodnocených spolu se všemi ostatními s kladným sentimentem, které takto nebyly vyhodnoceny[169].

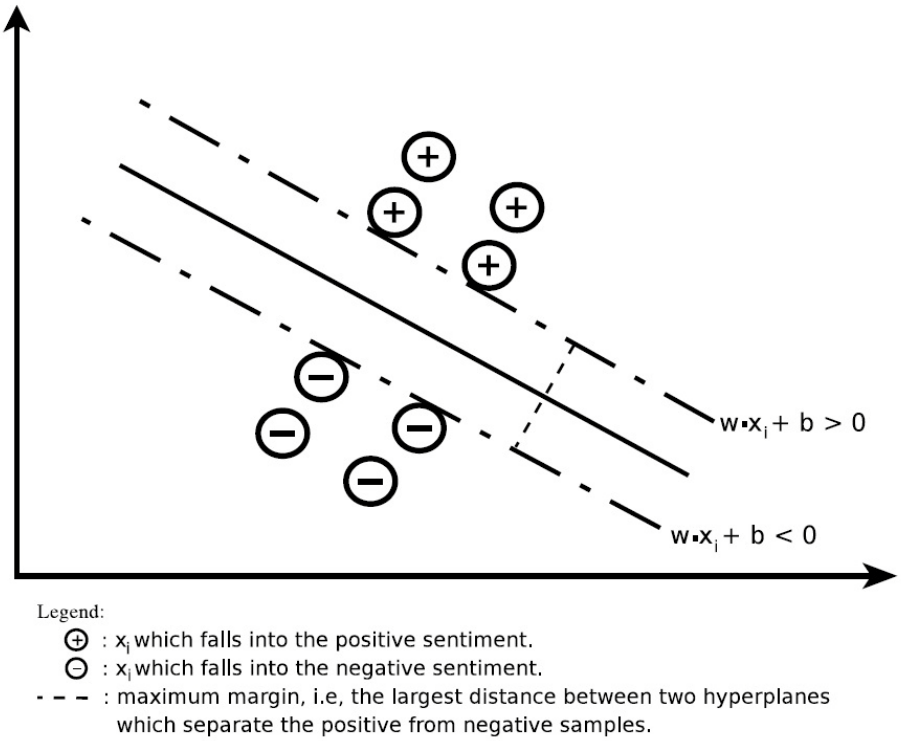
5.2 Techniky strojového učení

Strojové učení je technika založená na umělé inteligenci. Pomocí speciálních algoritmů si je počítač schopen vštípit určitá pravidla, která se naučí na základě empirických dat z okolí. K takovéto výuce se používají „treningová“ data, která slouží počítači k vybudování souboru pravidel. Tato pravidla následně používá k zpracování „ostrých“ dat[170].

[166] Konkrétně tomuto problému se budeme věnovat níže, v části týkající se experimentu zabývajícím se rozličností nazírání na sentiment mezi skupinou lidí.
[167] Sentiment analysis. In *Wikipedia : the free encyclopedia* [online].
[168] PRABOWO, Rudy ; Mike THELWALL. Sentiment Analysis: A Combined Approach [online].
[169] Precision and recall. In *Wikipedia : the free encyclopedia* [online].
[170] Machine learning. In *Wikipedia : the free encyclopedia* [online].

5.3 Support Vector Machine (SVM)

Support Vector Machines je metoda patřící do skupiny metod strojového učení. V klasifikačních úlohách pracuje SVM na principu určení nadroviny, která v prostoru příznaků ideálně rozděluje tréninková data (nadrovina je lineární funkcí). Optimální nadrovina pak vypadá tak, že body leží v opačných poloprostorech a hodnota minima vzdáleností bodů od roviny je co největší. Nejlépe je to vidět na následujícím obrázku, kde je znázorněno, jak po obou stranách roviny je prostor prostý jakýchkoli bodů. Popis nadroviny se provádí za pomoci nejbližších bodů, kterých nebývá mnoho. Tyto body nazýváme podpůrné vektory, od toho je pak odvozeno jméno této metody. Tato metoda rozděluje data do dvou tříd, je tedy binární.[171] [172]



Obrázek 13 - metoda Support Vector Machines zdroj: Sentiment Analysis: Combined Approach[173]

[171] Support vector machines. In *Wikipedia : the free encyclopedia* [online].
[172] PRABOWO, Rudy ; Mike THELWALL. Sentiment Analysis: A Combined Approach [online].
[173] Tamtéž.

5.4 Bayesovské filtrování

Bayesovské filtrování funguje na principu klasifikace pomocí výpočtu pravděpodobnosti, že konkrétní slovo nebo znak bude splňovat určité parametry (bude podmiňovat kladný/záporný sentiment) Tyto pravděpodobnosti se postupně mění na základě předchozí zkušenosti. Když se určité slovo bude často objevovat ve sděleních s určitým sentimentem, tak se zvýší pravděpodobnost, že tento sentiment implikuje. [174]

5.5 Slovníkové metody

U slovníkových metod porovnáváme slova obsažená ve zkoumaném textu se slovy z takzvaných sentimentových polarizačních slovníků. Tyto slovníky obsahují pozitivní, neutrální a negativní slova, které ovlivňují výsledný sentiment celého sdělení. Největším problémem této metody u analýzy textu ze sociálních sítí bývá to, že takovýto text obsahuje velkou spoustu slov a znaků, které jsou tvořeny zkratkami, emotikonami a slangovými nebo jinak specifickými výrazy (lolspeak[175]), které slovník neobsahuje a výsledné určení sentimentu může být tímto značně ovlivněno. U takovýchto slov pak sentiment tato metoda neurčí a z pohledu úspěšnosti tak klesá rapidně hodnota Recall[176].

5.6 Latentní sémantická analýza (LSA)

Je technika zpracování přirozeného jazyka, která se používá k určení vztahů mezi jednotlivými slovy nebo výrazy. Základní předpoklad LSA je, že slova, která jsou spolu příbuzná významem, stojí blízko sebe i v textu. V praxi se tato technika používá například pro řešení problémů synonymie (dva výrazy znamenají totéž) a polysémie (jeden výraz má více významů)[177].

[174] Bayesian spam filtering. In *Wikipedia : the free encyclopedia* [online].
[175] Psaní jednotlivých slov nebo celých sdělení jinými slovy, které stejně nebo podobně znějí. Například v Aj „*i can has ice scream? (I can have ice cream?)* “ apod.
[176] ZHANG, Lei, Riddhiman GHOSH, Mohamed DEKHIL, Meichun HSU a Bing LIU. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. HP laboratories [online].
[177] Latent semantic analysis. In *Wikipedia : the free encyclopedia* [online].

5.7 Hlavní problémy při určování sentimentu

Ať již použijeme jakékoli metody, tak se setkáme s takovými situacemi, v kterých bude automatická počítačová analýza stále velice obtížná. A to zejména

- Ovlivnění celkového sentimentu neexistujícími slovy a nestandardní skladbou věty („Včeeera to bylo zase jako málo huuuuusty“).
- Neznalost celkového kontextu sdělení (Zejména u diskusí, kde nemusí navazovat jednotlivé příspěvky přímo na sebe, může být pro stroj těžké určit, k čemu se reakce vztahuje a k čemu se vztahuje tedy určení sentimentu).
- Ironie („Iveta Bartošová je opravdu chytrá holka“).

5.8 Vlastní výzkum

Vycházíme-li z předpokladu, že použití kombinací několika výše zmíněných přístupů lze dosáhnout úspěšnosti určení sentimentu[178] až 70% (např. kombinace Bayesovských filtrů a SVM[179]), jak si můžeme být jisti, že tato úspěšnost je pravdivá. Nikdo nám totiž nemůže v takto subjektivní oblasti, jako je určování sentimentu říci, jaký výsledek určení je správný. Není zkrátka k dispozici žádný autoritativní dataset z reálných dat, vůči kterému bychom mohli měřit úspěšnost těchto postupů.

Proto jsme s Mgr. Josefem Šlerkou vycházeli z předpokladu, že sami lidé se mezi sebou nebudou schopni shodnout na tom, jaký je sentiment jednotlivých výroků. Tuto hypotézu jsme se pokusili experimentálně ověřit. Naše výsledky následně publikoval Josef Šlerka na internetovém online magazínu Lupa.cz, zabývajícím se informačními technologiemi, v článku „O sentiment analýze bez sentimentu aneb jeden malý experiment“[180].

[178] V rámci tohoto výzkumu berme výraz „sentiment“ jako určení emočního zabarvení (pozitivní/neutrální/negativní).

[179] Popsáno např. v již dříve citovaném článku.
PRABOWO, Rudy ; Mike THELWALL. Sentiment Analysis: A Combined Approach [online].

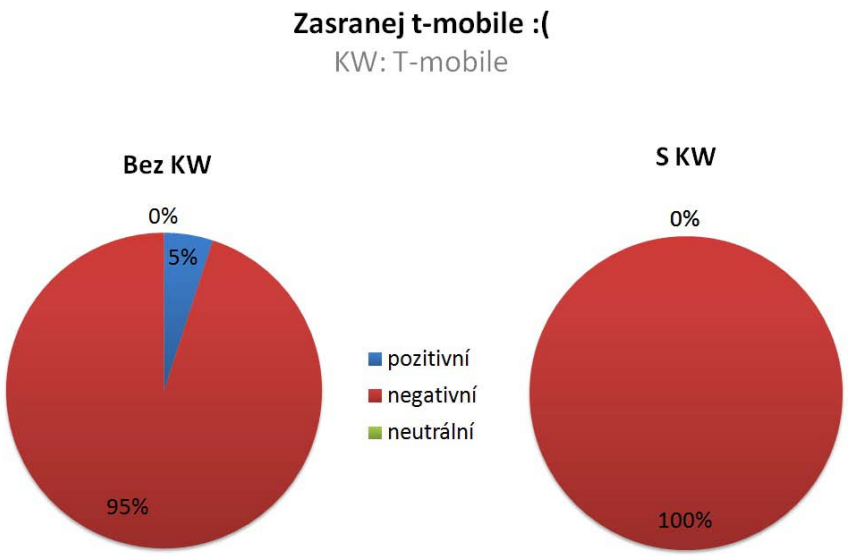
[180] ŠLERKA, Josef. O sentiment analýze bez sentimentu : aneb jeden malý experiment. Lupa.cz [online].

5.8.1 Metodologie

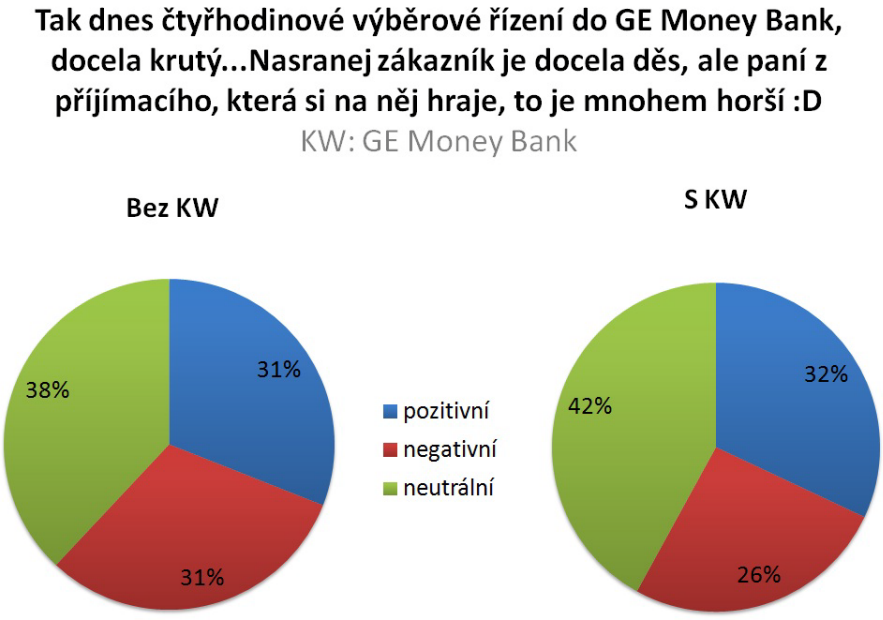
Náhodně jsme vybrali 90 výroků získaných pomocí nástroje pro social media monitoring Ataxo Social Insider, který zpracovává data z českých sociálních sítí (zejména Twitteru a Facebooku). Statusy byly různé délky, různé jazykové úrovně, obsahovaly různá klíčová slova a respondentům byly dány bez jakékoli znalosti kontextu celého vlákna konverzace (tak jak by je obdržel stroj). Tyto statusy jsme rozdělili do 3 balíčků po 30 a pomocí webových formulářů jsme tyto zmínky nechali ohodnotit českou twitterovou komunitou, aby posoudila, jaký sentiment ten který výrok má. Možnosti byli pouze tři – pozitivní / neutrální / negativní. Více datasetů jsme použili ze dvou důvodů. Protože když hodnotí sentiment počítač za pomoci metod strojového učení, jako jsou Bayesovské filtry nebo SVM, tak stroj neví, k jakému slovu se daný sentiment vztahuje (u technik založených na polarizačních slovnících to vědět může). Proto je dobré mít vyhodnoceny datasety více než dva. Dále pak u metod strojového učení hodně záleží na cvičícím materiálu, takže motivací bylo i vytvoření základu pro budoucí cvičící korpus. Rozdílné datasety se zpřístupňovaly respondentům na základě prokliku na jeden za tří formulářů.

Tyto tři formuláře pak byly vyplňovány ve dvou kolech. V prvním kole byly respondentům dány výroky, u kterých nevěděli, ke kterému klíčovému slovu sentiment vztahuje. Ve druhém kole byly tyto výroky doplněny o jednoznačné klíčové slovo. Takže například u výroku: „Nevolejte,nepiste mi na T-Mobile!!! Posral se mi IPhone!“ nebylo v prvním kole jasné, ke kterému slovu se sentiment bude určovat. V kole druhém bylo upřesněno, že emocionální zabarvení se bude vztahovat ke slovu T-mobile, což značně změnilo výsledné tipování respondentů.

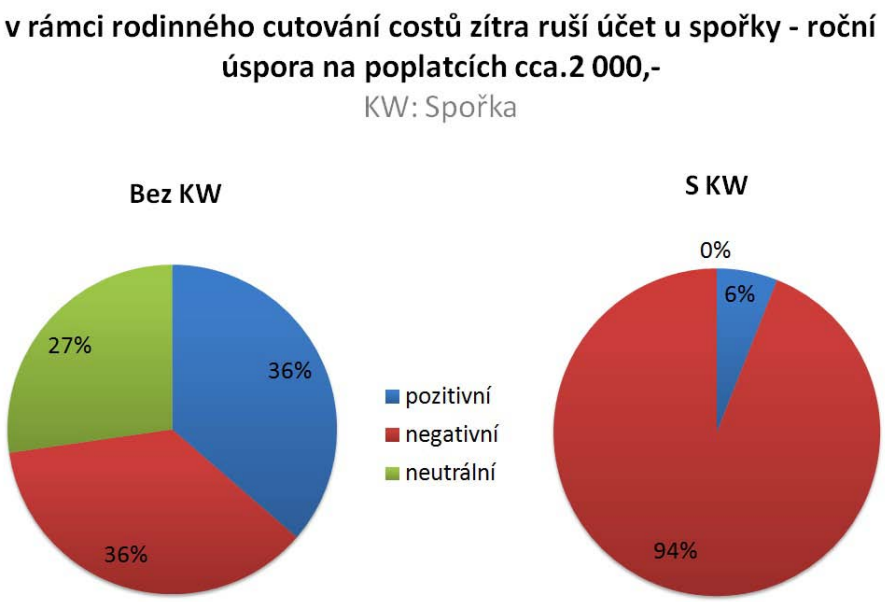
5.8.2 Příklady statusů a jejich určení respondenty



Obrázek 14 - příklad relativně jednoznačného výroku



Obrázek 15 - příklad velice nerozhodného výroku



Obrázek 16 - příklad výroku, jehož hodnocení se po přidání klíčového slova nejvíce změnilo

5.8.3 Výsledky

U každého ze všech 6 formulářů byla získána data minimálně od 30 respondentů[181]. Konečné výsledky prokázaly, že v prvním kole byla shoda nad 70% pouze ve 30 případech z 90, čili 1/3. Přidáním klíčového slova v druhém kole se shoda nad 70% zvedla na 43 případů z 90, což stále není ani polovina. Náš předpoklad se tedy potvrdil. Výsledkem je, že běžní uživatelé se nejsou vůbec schopni shodnout na tom, jaký sentiment zmínka vyjadřuje[182].

[181] U většiny formulářů toto číslo bylo ještě mnohem vyšší, např. u prvního formuláře to bylo přes 100 respondentů.
[182] Kompletní výsledky ke všem grafům ve formátu .pptx (Microsoft PowerPoint) lze najít v externí příloze 5.

5.8.4 Závěry

- Samotní lidé mají mezi sebou velký problém se shodnout na tom jak je který status emocionálně zabarvený (jaký má sentiment).
- Vzájemná shoda narůstá zejména v případech, kdy sdělení obsahuje nějaké slovo z polarizačního slovníku, ještě více pak, když je takové slovo z vulgárního slovníku[183].
- Většina zmínek má bez znalosti kontextu celého komunikačního vlákna diskuse zcela ambivalentní význam.
- Lépe se detekuje extrémní negativita, nežli drobná pozitivita, která často spadá do kategorie neutrální[184].[185]

V současné době použitelnost automatizované technologie Sentiment Analysis ovlivňují tedy významně dva faktory. Zaprvé automatické techniky nejsou ještě na takové úrovni, aby byly vždy zcela spolehlivé. Když už mají některé techniky vysoké hodnoty Precision, tak mohou trpět velice nízkými hodnotami Recall, zvláště pak v případě tak specifických sdělení, které generují sociální sítě. Z tohoto důvodu je nutné ideálně metody kombinovat. A za druhé, když už optimálně vyladíme postup, jakým budeme sentiment určovat, nejlépe specificky pro každý druh informačního média, tak musíme brát v potaz, že není žádná univerzální autorita, která by potvrdila, že naše výsledky jsou zcela správné, protože ani samotní lidé se mezi sebou neshodnou.

Jak píše ve své eseji o určování sentimentu například Lev Manovich, tak ideální postup je využít lidskou schopnost porozumět a správně interpretovat (s kterou mají počítače stále ještě problémy) se schopností počítačů analyzovat obrovské množství dat najednou, za pomoci algoritmů, které pro něj utvoříme[186]. Sám počítač ještě dnes není schopen pochopit kontext celého sdělení a to jsme zatím nebrali v potaz například sdělení s obrázkem, nebo dokonce videem. Výsledný závěr tedy je, že nám počítač může usnadnit mnoho práce, ale ve výsledku je dnes stále ještě potřeba konečná lidská analýza.

[183] Viz příkladový obrázek relativně jednoznačného výroku.
[184] Viz opět případ relativně jednoznačného výroku
[185] ŠLERKA, Josef. O sentiment analýze bez sentimentu : aneb jeden malý experiment. Lupa.cz [online].
[186] MANOVICH, Lev. Trending : The Promises and the Challenges of Big Social Data. [online].

6 Social Media Monitoring

I když oblast Social Media Monitoringu pod sebe zahrnuje a využívá techniky jak Social Network Analysis, tak Sentiment Analysis a Opinion Miningu je tato kapitola začleněna až po těchto praktických kapitolách. V této kapitole se budeme věnovat využití a potenciálu Social Media Monitoringu, který staví na dataminingu sociálních sítí.

6.1 Co můžeme pod Social Media Monitoringem chápat?

Social Media Monitoring, respektive media monitoring, není vůbec nic nového. Funguje naprosto na stejném principu jako výstřižková služba, jejíž historie sahá až do poloviny 19. století. Hlavní motivací bylo již tehdy snížení nákladů za cenu informací a to i v době, kdy bylo nesrovnatelně menší množství možných zdrojů. Jak volně vykládá Josef Šlerka ve své prezentaci věnující se monitoringu sociálních sítí myšlenku Norberta Wienera, tak hodnota informace získaná z jednoho zdroje je rovna energii, kterou bychom museli vynaložit na získání stejné informace z jiného zdroje[187]. Z tohoto pohledu Social Media Monitoring založený na datech ze sociálních sítí šetří náklady spojené s nutností sledovat tato veškerá média jinými způsoby. Z hlediska toho, jaký objem informací je v současnosti v těchto datech obsažen[188], tak jsou dnes specializované služby Social Media Monitoringu nejen obrovskou úsporou zdrojů, ale prakticky jedinou použitelnou možností, jak tyto informace získat.

6.2 Zdroje dat pro Social Media Monitoring

Obecně zdroje dat monitoringu budou v našem případě samozřejmě sociální sítě, čili to mohou být veškeré služby popsané podrobně výše v kapitole o sociálních sítích. Dále to mohou být navíc ještě například zpravodajské servery a RSS kanály. Každý takovýto zdroj má svá specifika a jeho využití bude výhodné v závislosti na tom, na co se budeme ptát – jaká bude naše informační potřeba.

[187] ŠLERKA, Josef. Social Media Monitoring. ATAXO. Slideshare.net [online].
[188] Viz úvodní kapitola této práce.

6.3 Na jakém principu fungují nástroje SMM?

Jak už bylo uvedeno v úvodu celé práce, tak přístup ke kompletním datům mají pouze vlastníci konkrétních služeb. Všichni ostatní se musí spolehnout na alternativní způsoby získávání dat ze sociálních sítí. Jedná se zejména o

- Indexování podobně jako to dělají internetové vyhledávače
- Získávání dat pomocí API pod konkrétním uživatelským účtem
- Kombinace obou předchozích metod

Následná práce se získanými daty funguje na principu klíčových slov (KW), pomocí kterých jsou zadávány dotazy do databáze, z které již vycházejí výstupy v podobě textů[189].

6.4 Co můžeme od Social Media Monitoringu chtít?

Hodnota informace je dále závislá na tom, jaká je naše informační potřeba. Od toho se dále odvíjí to, co budeme od Social Media Monitoringu očekávat. Potenciál využitelnosti Social Media Monitoringu je tak široký, jak moc jsou široká témata, která se v sociálních médiích objevují. Například nám může jít o

- Komunikaci, vytváření a upevňování vztahu s klienty
- Získávání zpětné vazby na kampaně v prostoru sociálních sítí
- Vytváření pozitivní vazby ke značce/produktu
- Identifikace problémů
- Redukce negativních reakcí
- Sledování konkurence
- Identifikace vlivných uživatelů - Influencerů
- Analýza trendů
- Prediktivní analýzy vývoje
- Analýza bezpečnostních rizik (bezpečnostní agentury)

[189] ŠLERKA, Josef. Social Media Monitoring. ATAXO. Slideshare.net [online].

6.4.1 Online marketing

V komerční sféře se dnes klade obrovský důraz na online marketing. Miliony ať už stávajících nebo potencionálních klientů jsou online na sociálních sítích a využití tohoto kanálu ke komunikaci a k jejich oslovení prostřednictvím tohoto kanálu je tématem mnoha teoretických i praktických příruček a objektem zkoumání nejedné diplomové či bakalářské práce. Výstupy Social Media Monitoringu jsou pak velice cenným podpůrným prostředkem našeho marketingového snažení.

6.4.2 Komunikace s klienty a řešení problémů

Aby mohla komunikace s klienty na sociálních sítích probíhat co nejlépe, tak je potřeba sledovat, co se o naší firmě/značce/produktech píše. Pro firmu není nic horšího, než když není schopna zachytit ohniska „špatných nálad“ a včas nereagovat na tyto stížnosti. Stále častěji se stává, že uživatelé očekávají a berou za samozřejmé, že když si někde na sociální síti postěžují na nějakou službu/produkt, tak by se jim měla inkriminovaná firma ozvat a jejich problém řešit[190]. K tomuto účelu lze (dokonce je nutno) použít Social Media Monitoring. Včasné zachycení ohnisek diskusí týkajících se předmětů našeho zájmu je zcela klíčové a napomáhá k co nejrychlejší identifikaci a lokalizaci problému. Jeho následným řešením a informováním zákazníků ohledně průběhu tohoto řešení pak minimalizujeme další negativní reakce.

Samotné odhalení negativních sdělení lze částečně automatizovat za pomoci Sentiment Analysis, kterou jsme si popsali v předešlé kapitole. Dále je možné si, díky sentiment analýze, nechat generovat alerty v případě, že se někde v sítí začíná zvedat „vlna“ negativity. Pro tento případ ale musíme mýt jasně definováno, co považujeme za „vlnu“ a co je pro nás negativní. Respektive kolik negativních reakcí je pro nás již vlnou hodnou braní na zřetel a jak moc negativní musí být, abychom je k této vlně započítávali.

Jak je výše uvedeno ve výčtu, tak na stejném principu můžeme monitorovat ohlasy na naše kampaně. Tyto kampaně nemusí být pouze z prostředí sociálních sítí, avšak

[190] V této oblasti je v České republice průkopníkem například firma Telefonica O2, která má postaven celý Social Media tým, který monitoruje veškeré zmínky o O2 a následně na ně prostřednictvím sociálních sítí reaguje a se stěžovateli problémy řeší.

pomocí sociálních sítí je lze velice dobře vyhodnocovat, protože mnoho uživatelů se zde o takovýchto tématech baví.

6.4.3 Sledování konkurence

Další možností je sledování chování naší konkurence. Tímto způsobem se můžeme nechat inspirovat úspěšnými konkurenčními marketingovými tahy, případně se naopak poučit z cizích chyb.

6.4.4 Identifikace Influencerů

Další možností je na základě monitoringu sociálních sítí identifikovat uživatele vlivné v nějaké oblasti. Vlivnost ve smyslu ovlivnění ostatních uživatelů můžeme měřit za pomoci metrik Social Network Analysis, případně si pomoci ještě metrikami, které přináší například Klout. Takto je možné měřit vlivnost uživatel na dané téma na základě počtu příspěvků, komentářů nebo reakcí na určitá témata v sociálních sítích a počtu zpětných reakcí na tyto příspěvky. Posouzení míry erudovanosti odpovědí daného uživatele je sice na osobě hodnotitele. Ale prostým výpočtem toho, kteří uživatelé reagují na konkrétní téma nejvíce (a vyvolávají u ostatních uživatel potřebu na jejich informace zpětně odpovídat) lze relativně snadno získat základní hodnoty míry vlivu i přehlednou síť mapující tuto komunikaci[191]. Na základě těchto výsledků si firmy mohou např. vytipovat uživatele, kteří pracují v jejich prospěch a „hýčkat si je“. Případně tento systém může fungovat stejně dobře jako nástroj HR, pro nábor nových zaměstnanců.

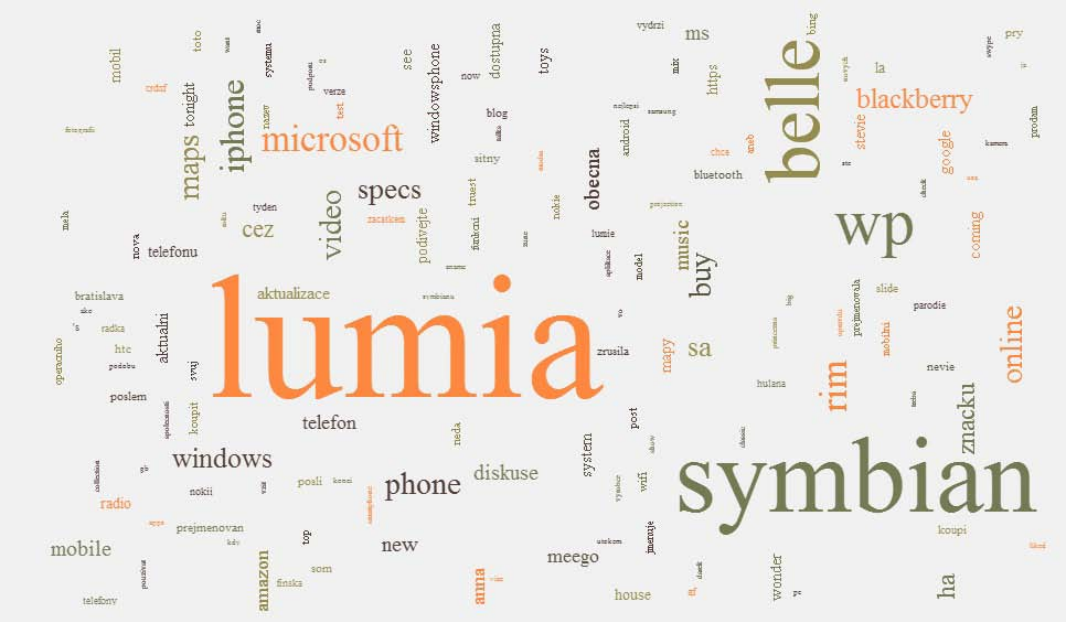
6.4.5 Identifikace témat a analýza trendů

Důležitou součástí monitoringu sociálních sítí je identifikace témat a na jejím základě analýza trendů spojených s určitým tématem. V tomto případě si můžeme pomoci například technikou word cloudu[192] (někdy také tag cloudu). Jedná se o grafické znázornění frekvence výskytů určitých termínů (slov) v textu, pomocí „mraku“ slov, jejichž velikost vyjadřuje právě hodnotou četnosti ve zkoumaném textu. Vynechány pak bývají zpravidla slovní druhy jako předložky, spojky a běžná slova v daném jazyce. V našem

[191] Viz kapitola o Social Media Analysis, zejména pak poslední příklad analýzy komunikace pomocí softwaru NodeXL.
[192] Tag cloud. In *Wikipedia : the free encyclopedia* [online]

případě tak zkoumaný text může být například soubor zmínek a komentářů ze sociálních sítí filtrovaných podle určeného klíčového slova.

Následující obrázek zobrazuje word cloud nad klíčovým slovem #Nokia v posledních 300 tweetech zaznamenaných serverem klábosení.cz, který mapuje český a slovenský twitter. Na první pohled je patrné, že ve spojitosti s touto upadající finskou značkou se nejčastěji mluví o „lumia“, což je název vlajkové lodi tohoto výrobce, který je právě uváděn na trh. Dále pak je zmiňován Symbian, operační systém, jehož implementace v mobilních telefonech byla nedávno ukončena a o této novince se dosti diskutuje na sociálních sítích. Podobným způsobem bychom mohli rozklíčovat i další slova z word cloudu a získat tak relativně rychle dobrý přehled o tématech, která jsou s Nokií spojována.



Obrázek 17 - Word Cloud nad klíčovým slovem #Nokia zdroj: klaboseni.cz

Dalším stupněm takovéto analýzy může být porovnání vzájemných korelací termínů mezi sebou za účelem zjištění, která témata jsou spolu nejčastěji spojována. Za použití korpusu obsahujícího 325 milionů statusů z Facebooku nabízí tuto možnost Mechanická Adlina[193], sestra Mechanické Popelky[194], která službu poskytuje nad statistickými daty z vyhledávání na Googlu.

[193] Mechanická Adlina [online].

Následující obrázek popisuje vzájemné vazby mezi nemocemi a léky. Čím je číslo menší, tím je větší pravděpodobnost, že tato slova budou stát v textu (statusu z FB) spolu. Na první pohled je patrné, že například Paralen a Coldrex je nejvíce spojován s teplotou a angínou. Dále pak s angínou jsou spojována nejčastěji antibiotika.

	ibalgin	paralen	coldrex	antibiotika
chřipka	INF	0.42050202	INF	0.41078834
angína	INF	0.33825966	0.29422665	0.34611935
kašel	0.47334818	0.49606154	0.43148665	0.54456261
teplota	0.43326569	0.38316022	0.39494894	0.43990975
bolest	INF	0.58319345	INF	0.56415298

Obdobně jako na našem případu, který byl schválně vybrán jednoznačně pro demonstraci správnosti výsledku, můžeme poměřovat i spoustu dalších zdánlivě nesouvisejících termínů a získat tak někdy až překvapivé výsledky. Nesmíme ale zapomínat, že nejdůležitější je výsledná interpretace výsledků, která nemusí být zdaleka vždy tak jednoduchá a jednoznačná jako na našem příkladu s nemocemi.

[194] The Mechanical Cinderella [online].

6.5 Vlastní výzkum - Prediktivní analýza a odhad budoucího vývoje

V závěru této kapitoly se dostáváme k tomu nejzajímavějšímu, což je možnost predikce budoucího stavu a srovnání „měkkých dat“ z dataminingu sociálních sítí s „tvrdými“ statistickými daty. Za tímto účelem jsme se rozhodli učinit s Josefem Šlerkou výzkum zakládající se na srovnání reálné návštěvnosti filmů v roce 2010 a 2011 s tím, jak se o nich mluvilo na sociálních sítích. Data pro tuto analýzu byla čerpána ze softwaru pro Social Media Monitoring Ataxo Social Insider[195]. Hlavním cílem tohoto výzkumu bylo potvrdit (či vyvrátit), zdali je možné v současnosti použít data z českého prostředí sociálních sítí k predikci návštěvnické úspěšnosti vybraných filmů. Případně zjistit za jakých podmínek je to možné.

6.5.1 Metodologie

Filmy byly záměrně vybrány takové, u kterých se dal očekávat zvýšený zájem publika. To z toho důvodu, že zejména český Twitter[196] a potažmo i Facebook nedosahuje ještě takové velkosti. Nebyl by tedy dostatečný počet zmínek na sociálních sítích. Jako filmy s premiérou v roce 2010 byly vybrány zahraniční filmy Inception (Počátek)[197] a Alenka v říši divů[198]. Pro srovnání jsem podrobil výzkumu ještě český film Alois Nebel, který měl premiéru koncem roku 2011. Tento film neměl potenciál stát se kasovním trhákem, ale díky své revoluční technologii zpracování vzbuzoval mezi diváky také dostatečný rozruch[199].

Samotná praktická část spočívala v porovnání tvrdých statistických dat[200] návštěvnosti po jednotlivých týdnech, v období kdy se drželi v TOP 20 nejnavštěvovanějších filmů v Čechách (dle dat Unie filmových distributorů), s počtem zmínek v těchto týdnech na sociálních sítích.

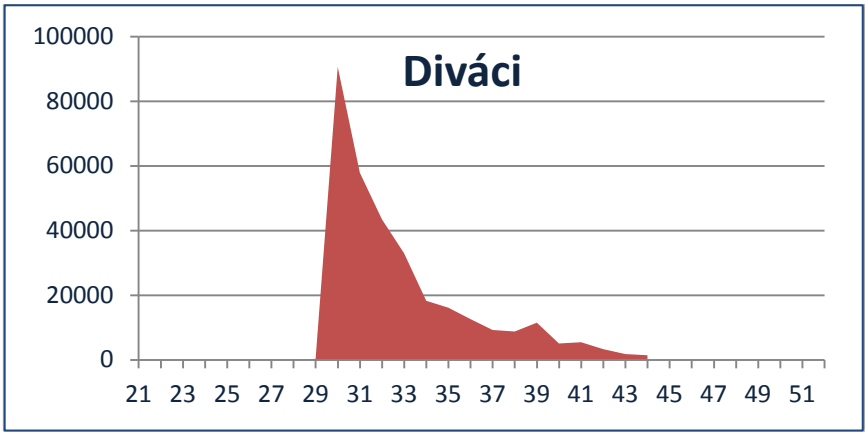
Čísla v tabulkách a grafech uvádějí vždy číslo týdne v měsíci a k němu odpovídající počet diváků, sumu tržeb v českých korunách a počet zmínek na sociálních sítích získaný pomocí softwaru ASI.

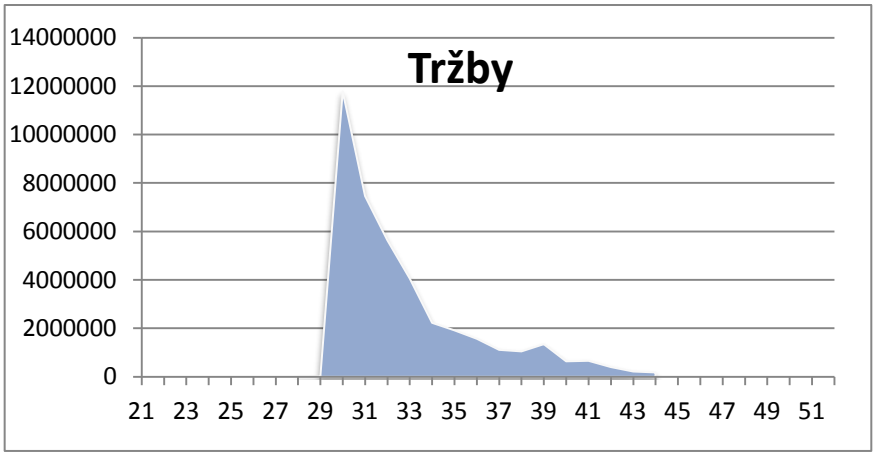
---

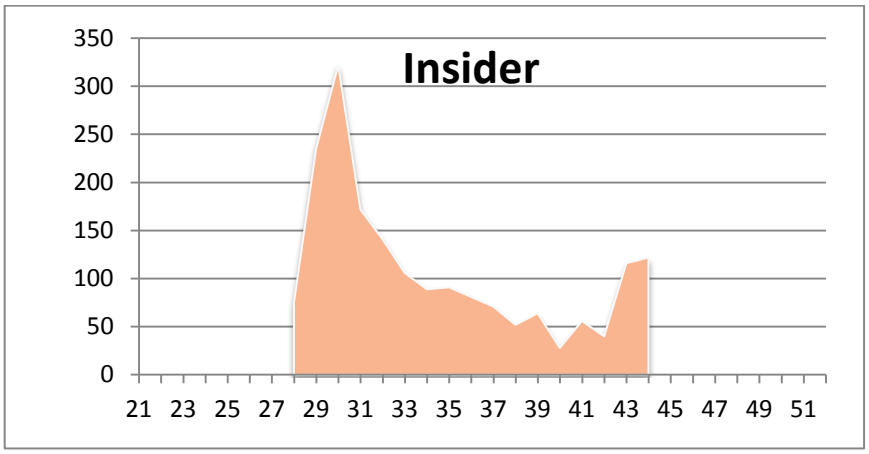
[195] *Ataxo Social Insider* [online].
[196] Viz statistiky českého Twitteru v kapitole o sociálních sítích.
[197] Počet zmínek se pohyboval v desítkách až stovkách.
[198] Počet zmínek spíše pouze v desítkách.
[199] Počet zmínek spíše pouze v desítkách.
[200] TOP 20 filmů. *Unie Filmových Distributorů* [online].

6.5.2 Výsledky pro film Inception (Počátek)

21					
22					
23					
24					
25					
26				Korelace -1 týden	0,85855
27					
28	0	0	75		
29	317	15446	235		
30	90615	11791429	322		
31	57882	7469105	172		
32	43452	5639374	141		
33	32970	4063153	106		
34	18279	2231489	89		
35	16136	1920623	91		
36	12611	1574412	81		
37	9242	1115302	71		
38	8771	1051376	52		
39	11500	1349383	64		
40	5067	641665	28		
41	5455	662671	56		
42	3315	404818	40		
43	1794	220771	116		
44	1424	182726	122		
45					
46					
47					
48					
49					
50					
51					
52					



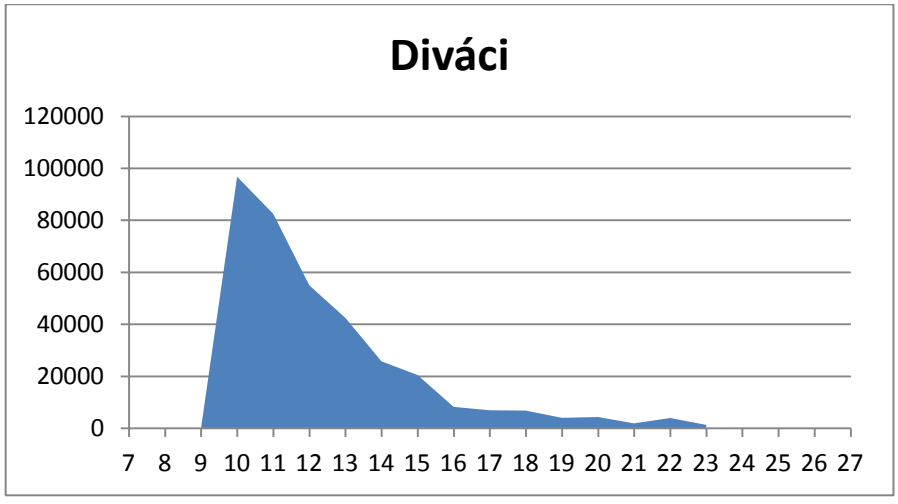


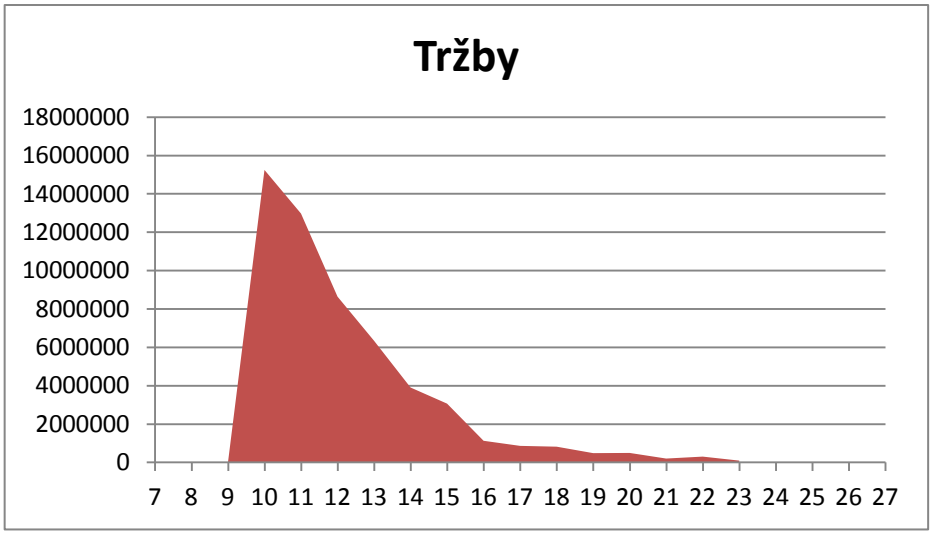


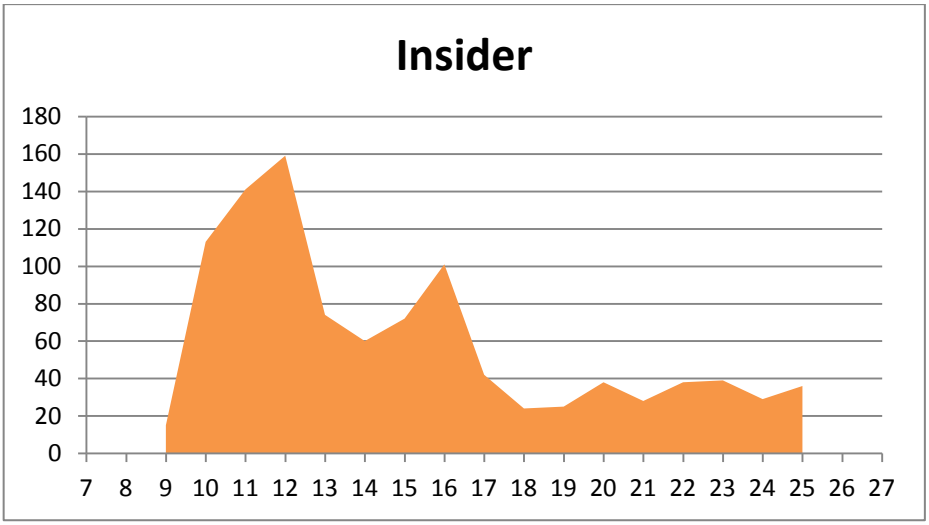
Jak znázorňují předešlé tabulky a obrázky, tak u snímku Inception jednotlivé grafy vykazují značnou podobnost. Jejich vzájemná korelace dosahuje hodnoty 0,85855. O to zajímavější je, že této korelace nabývá tento film s posunem -1 týden na straně „měkkých dat“ ze sociálních sítí. Čili „šeptanda“ na sociálních sítích předcházela reálný vývoj počtu diváků a výše tržeb. U tohoto filmu tedy pomocí zmínek na sociálních sítích bylo možné předpovídat budoucí vývoj.

6.5.3 Výsledky pro film Alenka v říši divů

1					
2					
3					
4					
5					
6				Korelace	0,82113949
7					
8					
9	143	0	15		
10	96849	15253732	113		
11	82573	12981258	141		
12	55060	8651490	159		
13	42530	6351682	74		
14	25822	3914542	60		
15	20568	3074268	72		
16	8302	1137705	101		
17	7006	872691	42		
18	6884	831058	24		
19	4096	494406	25		
20	4405	504904	38		
21	1960	211870	28		
22	4033	313396	38		
23	1372	105247	39		
24			29		
25	1565	104977	36		
26					
27					
28					
29					
30					
31					
32					
33					
34					
35					
36					
37					



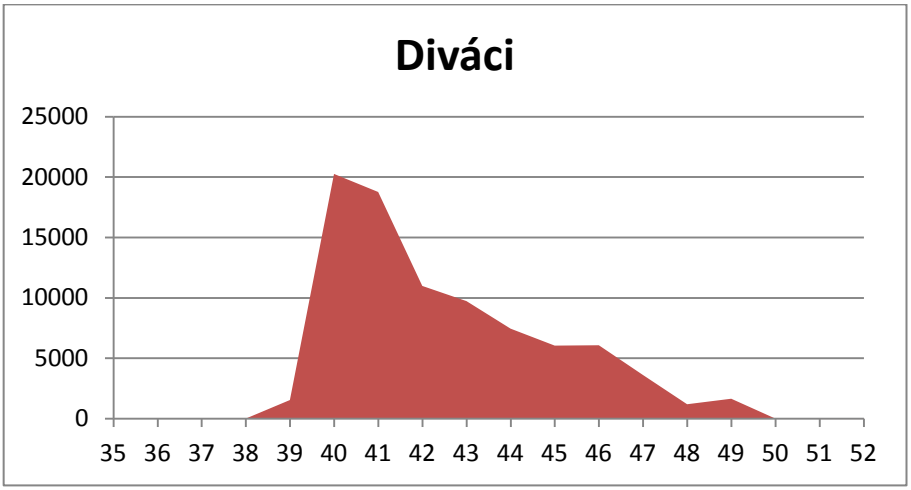


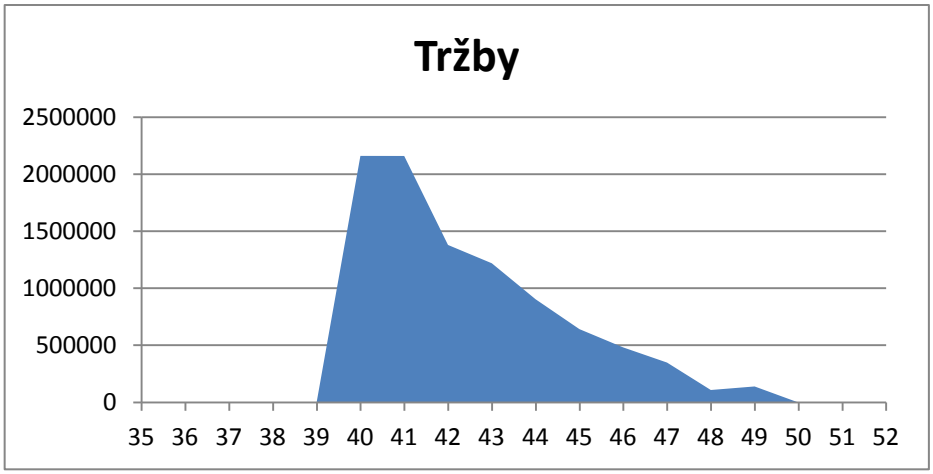


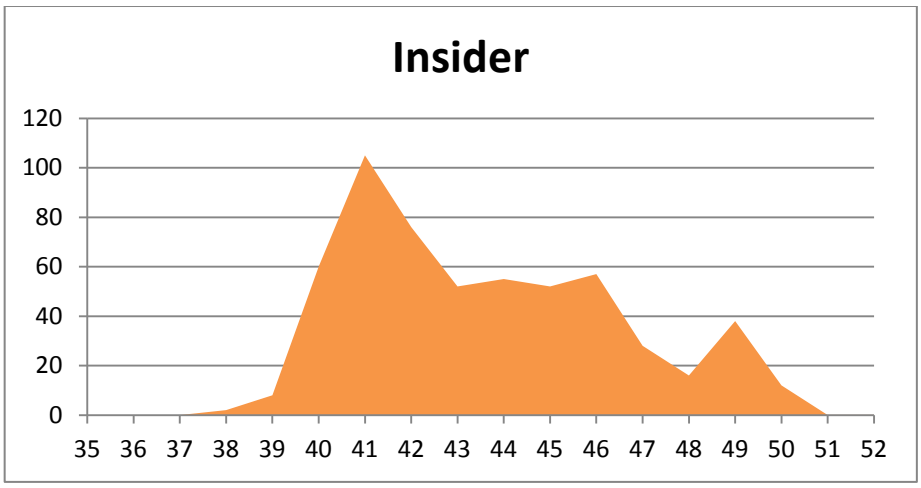
U filmu Alenka v říši divů nepředchází zmínky na sociálních sítích statistické údaje o návštěvnosti o týden, jako tomu bylo u filmu Počátek, ale korelace mezi „měkkými“ a „tvrdými“ daty byla též hodně vysoká. Konkrétně je rovna 0,8211. Je tedy zřejmé, že ve velké míře konverzace na sociálních sítích korespondují s reálnou návštěvností tohoto filmu.

6.5.4 Výsledky pro film Alois Nebel

31					
32					
33					
34					
35				Korelace	0,808445
36					
37					
38			2		
39	1 535	0	8		
40	20 255	2 161 375	60		
41	18 761	2 160 655	105		
42	10 982	1 380 737	76		
43	9 728	1 220 206	52		
44	7 446	904 710	55		
45	6 041	643 021	52		
46	6 071	482 721	57		
47	3 608	349 986	28		
48	1 187	109 452	16		
49	1 643	140 569	38		
50			12		
51					
52					







U filmu Alois Nebel nebyly výsledky také tak přesvědčivé jako u Počátku, ale i tak vzájemná korelace mezi počtem diváků a zmínkami na sociálních sítích dosáhla hodnoty 0,8084. U tohoto filmu též jako u Alenky nedošlo k tomu, že by se o něm mluvilo s týdenním předstihem, avšak míra „šuškandy“ přímo koresponduje s vývojem návštěvnosti a tržeb tohoto filmu. Tento výsledek mohl být ovlivněn např. tím, že pro výzkum nebyla k dispozici podrobnější data (např. denní přehledy).

6.5.5 Závěry

Z těchto dat jsem vyvodil několik závěrů, které by měly demonstrovat potenciál využití takovýchto analýz budoucího vývoje na základě dat ze sociálních sítí (zejména v budoucnu).

- U filmů u kterých se dá z jakéhokoli důvodu[201] očekávat vyšší zájem veřejnosti, tak lze již dnes predikovat křivku návštěvnosti a potažmo pak úspěch a tržby. Jako tomu bylo například u námi zkoumaného snímku Počátek[202] (Inception), který má v současnosti hodnocení 89% ze 100%. To svědčí o jeho veliké oblíbenosti u diváků.
- Filmy s nižší oblíbeností u publika (nežli u předcházející skupiny snímků) ukazují, že data ze sociálních sítí jsou zatím jen velmi věrnou kopií skutečných statistických dat návštěvnosti (tržeb). Shodou náhod jak film Alois Nebel[203], tak Alenka v říši divů[204] mají totožné divácké hodnocení 70%. To je velice dobré hodnocení, ale neřadí je ještě do kategorie nejoblíbenějších filmů.
- U málo známých filmů v současnosti nelze použít data ze sociálních sítí pro prediktivní analýzu, protože jich není dostatečné množství pro vyvození závěrů.

6.5.6 Možná současná omezení

- V současné době je většina omezení, která se nám staví do cesty při snaze predikovat budoucí vývoj z dat na českých a slovenských sociálních sítích zejména kvantitativního charakteru. Z tohoto důvodu prozatím není možné tyto analýzy uplatnit například i na méně známé filmy.
- Přesnosti predikce by též přispělo, kdybychom měli k dispozici přesnější data. Např. na našem příkladu s návštěvností filmů byly k dispozici pouze týdenní přehledy.
- Celkově brání vyšším možnostem srovnávání „měkkých“ a „tvrdých“ dat nedostatek volně dostupných dostatečně přesných statistických informací.

[201] Například hlediska mediálních kampaní, oblíbenosti hlavních aktérů a realizačního týmu, využití nových technologií apod.
[202] Počátek. *Česko-slovenská filmová databáze* [online].
[203] Alois Nebel. *Česko-slovenská filmová databáze* [online].
[204] Alenka v říši divů. *Česko-slovenská filmová databáze* [online].

7 Vize a hrozby

7.1 Střízlivý pohled

V této závěrečné kapitole by autor rád nabídl pár soukromých pohledů na vývoj v oblasti dataminingu sociálních sítí a sociálních médií obecně. Kapitola je sice rozdělena na podkapitoly vize a hrozby, z níž vize by měly být z oblasti pozitivních (techno-optimistických) pohledů na věc. Hrozby by pak měly vyjadřovat, jaká rizika může mít používání/zneužívání těchto technologií. Pohled na tuto problematiku může být rozdílný případ od případu a hlavně názor od názoru. Některé věci, které vidí autor pozitivně, tak může čtenář shledávat hrozivými a naopak.

Vycházejíc především ze statistických dat obsažených v kapitole o sociálních sítích a z ambiciózních plánů majitelů většiny v této práci uvedených sociálních sítí/služeb/médií musíme konstatovat, že očekávání dalšího masivního nárůstu v této oblasti je více než reálné. I když například 1/10 populace v případě Facebooku je mimořádně veliké číslo, tak stále je to pouze 1/10. Zejména u některých skupin uživatel lze očekávat veliký nárůst. Autor si tedy v tomto kontextu dovolí předpovídat

- Stále větší růst sociálních médií (jak z pohledu růstu počtu uživatelů, tak z pohledu množství generovaných dat těmito uživateli)
- Rostoucí ochotu uživatelů sociálních médií o sobě sdílet více dat (i ryze osobního charakteru)
- Přesun stále většího množství uživatelů „online“ (z pohledu toho, že jejich identita se pro ně nestane pouze nástrojem komunikace, ale i komplementem jejich reálného já)
- Snížení Digital Divide[205] (ať už z důvodu levnějších technologií, tak z důvodu „nutnosti“ uživatelů být „na síti“[206] a nestranit se tak od budoucí většiny)

[205] Digital Divide. In *Wikipedia : the free encyclopedia* [online].
[206] Vycházejíc z objemu diskusí na serverech zaměřených na maminky na mateřské dovolené (emimino.cz apod.), dejme možnost divačkám pořadu „Sama doma“ psát své dotazy na Twitter a rázem se rozroste velikost české Twitterové scény o pěkných pár tisíc nových uživatelů/uživatelek.

Všechny tyto předpoklady mají za následek enormní nárůst uživateli generovaných dat. To sebou nese právě již zmíněný obrovský potenciál v případě správného využití a stejně ruku v ruce nesmírná rizika zneužití. Pro samotný datamining nesou výše zmíněné premisy zhruba takovýto předpoklad vývoje[207]

- Analýza dat dataminingovými metodami bude stále více používaná zejména v oblasti businessu, e-komerci a e-marketingu.
- Dataminingové nástroje budou v budoucnu nejen schopné se vypořádat s masivními a stále rostoucími objemy dat, ale zároveň budou umět pracovat s těmito daty interaktivně.
- Dataminingové nástroje budou více a více integrovány do vyhledávačů, databázových systémů, datových skladů a systémů na bázi Cloud computingu.
- Do hlavní oblasti zájmu se dostane též datamining „časoprostorových" dat na základě masivního rozšíření mobilních zařízení s GPS senzory a bezdrátovým vysokorychlostním připojím k síti.
- Pokročilé dynamické distribuované metody umožní datamining datového toku v reálném čase (takto bude možné zefektivnit boj proti terorizmu, kde jde o každou minutu apod.).
- Nové možnosti nastanou také v oblasti dataminingu multimediálních dat[208].
- V souvislosti s předchozím bodem dojde k nárůstu dataminingu biometrických dat[209].

V následujících dvou podkapitolách bude uveden „odvážnější" pohled autora na perspektivy v této oblasti.

[207] HAN, Jiawei, Micheline KAMBER a Jian PEI. *Data mining*: *concepts and techniques* [online].
[208] Více v následujících dvou kapitolách.
[209] Taktéž.

7.2 Vize

Nové možnosti jsou stále v oblasti propojení dat ze sociálních sítí s „běžnými daty“ ve formě různých mashupů (míchanic). V minulosti bylo oblíbenou technikou tato data promítat například na mapu[210]. Na tomto místě vidí autor možnosti v aplikacích rozšířené reality (augmented reality) na základě dat získaných dataminingem ze sociálních sítí. Společně s použitím technologií na principu rozpoznání osob podle různých biometrických parametrů[211]. Stejně tak jako fyzická poloha je jedinečný identifikační údaj[212], tak lidské biometrické údaje jsou jednoznačným identifikátorem, který každý uživatel „sdílí“, ať chce, či nechce celý svůj život.

Na tomto principu stavěla například letošní letní kampaň FaceLook firmy Coca Cola, která nabídla svým zákazníkům[213] možnost pomocí speciálních zařízení nahrávat fotografie a psát nové statusy na Facebook. Jako login i heslo do systému sloužila pouze uživatelova tvář[214].

Tímto směrem se pravděpodobně bude ubírat i další vývoj v závislosti na přesnosti a spolehlivosti technologií rozpoznání obličeje. Tento předpoklad potvrzuje i samotný Google, který jako novinku do své nejnovější verze operačního systému pro chytrá mobilní zařízení (telefony a tablety) Android 4.0 Ice Cream Sandwich připravil podporu odemknutí klávesnice za pomoci rozpoznání majitelovy tváře. Stejně jako dnes funguje služba rozšířené reality Google Goggles[215], tak v budoucnu bude možné rozpoznávat lidské tváře právě na základě dat (fotografie, profilové fotografie, videa) ze sociálních sítí. Dále budou tyto informace doplněny o veškeré osobní údaje, které o sobě v minulosti daný člověk prozradil na sociálních sítích a to právě díky real-time dataminingu. Takového sofistikované aplikace bude možné provozovat i na průměrných mobilních zařízeních. Dnes aplikace tohoto typu sloužila k naskenování „rozsypaného čaje“ na dovolené v Japonsku a rychlému odhalení jeho významu. Nebo k rozpoznání produktu v kamenném krámě a následné lacinější koupi v internetovém obchodě. V budoucnu bude možné

[210] Služby jako http://trendsmap.com/ zobrazující trendy z Twitteru na mapě, nebo http://twittervision.com/ zobrazující již konkrétní tweety a obrázky z Twitteru.
[211] Například technologie identifikace člověka na základě rysů jeho tváře (Face recognition), nebo technologie rozpoznávající typický pohyb/chůzi člověka apod.
[212] Jakési URL místa.
[213] Především z řad teenagerů.
[214] FaceLook: Coca-Cola’s Facial Recognition App. *Digitalbuzzblog.com* [online].
[215] Goggles. *Google.com* [online].

takovéto aplikace využít například v prostředcích MHD k snadné identifikaci svobodných slečen, případně těch, které právě napsaly do svého statusu, že jejich přítel je idiot.

Takovéto představy z pohledu autora nejsou vůbec nereálné. Trochu pokročilejší verzi této technologie nabídli například autoři nového „akčního trháku“ Mission Impossible – Ghost Protocol[216], kde hlavní hrdina využívá kontaktní čočky, které disponují obdobnou funkcí. Když budeme pokračovat v takovýchto úvahách dále, tak dojdeme k závěru, proč jednou tuto technologii neaplikovat v podobě implantátu rovnou do mozku?

7.3 Hrozby

Stejně tak jako hlavní pozitiva jsou v propojení reálných dat a s daty ze sociálních sítí, tak na stejném místě jsou i rizika této technologie. Propojením výše zmíněných biometrických identifikačních dat s daty ze sociálních sítí dostáváme jedineční identifikátor s komplexními osobními údaji. Dále se v dnešní době hlásí o slovo technologie automatického tagování[217] (označování) fungující též na principu rozpoznávání tváře. Díky této technice se brzy dočkáme chvíle, kdy kdokoli nás kdekoli vyfotí a bude tyto fotografie sdílet na sociální síti, tak budeme jednoznačně identifikování a „odhaleni“ veřejnosti[218].

Za použití Surveillance[219] systémů na sledování osob pak můžeme získat kompletní přehled o činnosti konkrétních osob. Už dnes jsme sledování prakticky na každém rohu a při použití takovéto technologie bychom byli navíc jednoznačně identifikovatelní. Dále pak by sledování nebylo prováděno pouze za pomoci již nainstalovaných kamerových systémů, ale data by bylo možné získávat ze zařízení jednotlivých uživatelů. Takováto technologie by byla velice snadno zneužitelná všudypřítomnému „fízlování“ občanů a potažmo tak k následné deformaci obecné morálky.

---

[216] Mission Impossible - Ghost Protocol. *Csfd.cz* [online].

[217] Tyto technologie v současnosti již používá jak Facebook, tak Google+. V případě Google+ je alespoň možné tuto službu vypnout. V nynější podobě ale tato služba funguje pouze jako podpůrná služba, kdy uživatelé dostávají návrhy, zdali se na dané fotografii nenachází některý z jejich přátel.

[218] Jen málo kdo by chtěl, aby jeho manželka/přítelkyně viděla fotografii zcela cizích lidí z restaurace, na jejímž pozadí se její manžel tulí ke sličné servírce.

[219] Surveillance. In *Wikipedia : the free encyclopedia* [online].

Už samotné vědomí, že je člověk monitorován (jinými slovy, že ho někdo sleduje) mění jeho chování v rozličných situacích. Tento fakt následně deformuje morálku osob žijících v takovémto prostředí. Člověk se přestává chovat podle svého přesvědčení, morálních hodnot a charakteru, ale je deformován všudypřítomným strachem, že ho někdo uvidí[220]. Nedodržuje pak zákony z důvodu, že to je správné, ale z důvodu, že může být kdykoli snadno odhalen, usvědčen a potrestán[221]. Z tohoto negativního pohledu na věc se stávají naše biometrické a osobní údaje cennou informací a měli bychom si je v dnešní rozvolněné době sdílení co nejvíce chránit.

Závěrem musíme zmínit, že veškerá práce s osobními údaji (získávání, ukládání, nakládání s nimi) podléhá právním regulacím, zejména pak zákonu na ochranu osobních údajů. Čili prozatím jsme proti takovýmto scénářům „relativně" chráněni. Uplatnění těchto technologií, ať už v pozitivním nebo negativním slova smyslu nastíněném na řádcích výše, je tedy spíše politickou nežli technologickou otázkou. Analýza těchto právních hledisek je ale tak rozsáhlé téma hodné samostatné diplomové práce.

[220] HALE, Benjamin. Identity Crisis : Face Recognition Technology and Freedom of the Will. *Ethics Place and Environment* [online].
[221] KUBELKA, Martin. *Ochrana soukromí v informační společnosti*: *problematika kamerových sledovacích systémů.*

8 Zdroje

10 Terrific New Twitter Infographics in 2011. *Singlegrain.com* [online]. [cit. 2012-01-02]. Dostupné z: http://blog.hubspot.com/blog/tabid/6307/bid/19023/10-Terrific-New-Twitter-Infographics-in-2011.aspx

2011 Blogging Statistics [infographic]. *Rightmixmarketing.com* [online]. [cit. 2012-01-02]. Dostupné z: http://www.rightmixmarketing.com/right-mix-blog/blogging-statistics

About Klout. *Klout.com* [online]. 2011 [cit. 2012-01-02]. Dostupné z: http://www.klout.com/corp/about

Adam Javůrek: blog není pro tiskové zprávy. *Itbiz.cz* [online]. [cit. 2012-01-02]. Dostupné z: http://www.itbiz.cz/rozhovor-adam-javurek

Alenka v říši divů. *Česko-slovenská filmová databáze* [online]. 2011 [cit. 2011-12-29]. Dostupné z: http://www.csfd.cz/film/235922-alenka-v-risi-divu/

Alois Nebel. *Česko-slovenská filmová databáze* [online]. 2011 [cit. 2011-12-29]. Dostupné z: http://www.csfd.cz/film/242734-alois-nebel/

Amatéři.cz [online]. 2011 [cit. 2011-12-15]. Dostupné z WWW: <http://www.amateri.cz/>.

APPELTAUEROVÁ, Lucie, et al. Česko na sociálních sítích. In H1.cz. [online]. Praha, 8. 11. 2011 [cit. 2011-12-18]. Dostupné po registraci z WWW: <http://www.h1.cz/files/h1cz-cesko-socialni-site-2011.pdf>.

Ataxo Social Insider [online]. 2011 [cit. 2011-12-29]. Dostupné z: http://ataxosocialinsider.cz/

Augmented reality. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 15 September 2002, last modified on 19 December 2011 [cit. 2011-12-20]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Augmented_reality>.

Barcamppraha.cz [online]. 2011 [cit. 2011-12-12]. Dostupné z WWW: <http://barcamppraha.cz/>.

BACKSTROM, Lars . *Four Degrees of Separation*. Palo Alto, 2011. 13 s. Oborová práce. Facebook, Palo Alto, CA, USA. Dostupné z WWW: <http://arxiv.org/PS_cache/arxiv/pdf/1111/1111.4570v2.pdf>.

Bayesian spam filtering. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 9 March 2004, last modified on 17 December 2011 [cit. 2011-12-22]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Bayesian_spam_filtering>.

Blog. *Foursquare.com* [online]. [cit. 2012-01-02]. Dostupné z: http://blog.foursquare.com/

Blog. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 14. 2. 2005, last modified on 19. 12. 2011 [cit. 2011-12-19]. Dostupné z WWW: <http://cs.wikipedia.org/wiki/Blog>.

Centrality. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 3 February 2005, last modified on 20 November 2011 [cit. 2011-11-24]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Centrality>.

ČERNÝ, Michal. Znáte své digitální stopy?. *Lupa.cz* [online]. 2007, 1, [cit. 2011-12-01]. Dostupný z WWW: <http://www.lupa.cz/clanky/znate-sve-digitalni-stopy/>.

Czech Republic Facebook Statistics. Socialbakers.com [online]. [cit. 2011-12-27]. Dostupné z: http://www.socialbakers.com/facebook-statistics/czech-republic

Data, data everywhere. *The Economist* [online]. Feb 25th 2010, 1, [cit. 2011-12-01]. Dostupný z WWW: <http://www.economist.com/node/15557443>.

Data. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 8. 8. 2005, last modified on 5. 12. 2011 [cit. 2011-12-25]. Dostupné z WWW: <http://cs.wikipedia.org/wiki/Data>.

Data mining. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 28 February 2002, last modified on 28 December 2011 [cit. 2011-12-30]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Data_mining>.

Digital Divide. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 16 September 2004, last modified on 30 November 2007 [cit. 2012-01-03]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Digital_Divide>.

Digital native. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 22 May 2007, last modified on 1 December 2011 [cit. 2011-12-13]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Digital_native>.

DOČEKAL, Daniel. Experiment #klout - vylepšení skóre? Odebírejte věci, jak snadné. *Pooh.cz* [online]. 27/10/2011 [cit. 2012-01-02]. Dostupné z: http://www.pooh.cz/pooh/a.asp?a=2017462

DOČEKAL, Daniel. Klout změnil algoritmus a TOP českého Twitteru se otřáslo. *Justit.cz* [online]. [cit. 2012-01-02]. Dostupné z: http://www.justit.cz/wordpress/2011/10/27/klout-zmenil-algoritmus-a-top-ceskeho-twitteru-se-otraslo/

Emimino.cz [online]. 2011 [cit. 2011-12-19]. Diskusní fórum. Dostupné z WWW: <http://www.emimino.cz/diskuse/>.

EXCLUSIVE: Google To Retire Blogger & Picasa Brands in Google+ Push . *Mashable.com* [online]. [cit. 2012-01-02]. Dostupné z: http://mashable.com/2011/07/05/google-blogger-picasa-rebranding/

Facebook [online]. 2011 [cit. 2011-12-15]. Dostupné z WWW: <https://www.facebook.com/>.

Facebook VS. Google [Infographic]. *Singlegrain.com* [online]. [cit. 2012-01-02]. Dostupné z: http://www.singlegrain.com/blog/facebook-vs-google-plus/

FaceLook: Coca-Cola's Facial Recognition App. *Digitalbuzzblog.com* [online]. Aug 2, 2011 [cit. 2012-01-03]. Dostupné z: http://www.digitalbuzzblog.com/facelook-coca-colas-facial-recognition-app/

Flickr [online]. 2011 [cit. 2011-12-20]. Dostupné z WWW: <http://www.flickr.com/>.

Forum. *Lide.cz* [online]. [cit. 2012-01-02]. Dostupné z: http://forum.lide.cz/

Goggles. *Google.com* [online]. [cit. 2012-01-03]. Dostupné z: http://www.google.com/mobile/goggles/#text

Google+ [online]. 2011 [cit. 2011-12-15]. Dostupné z WWW: <https://plus.google.com/>.

Google Buzz. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 9 February 2010, last modified on 16 December 2011 [cit. 2011-12-17]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Google_Buzz>.

Google Latitude. Google.com [online]. [cit. 2011-12-27]. Dostupné z: https://www.google.com/latitude

Hadoop: the definitive guide. 2nd ed. Farnham: O'Reilly, 2010. ISBN 978-144-9389-734.

HALE, Benjamin. Identity Crisis : Face Recognition Technology and Freedom of the Will. *Ethics Place and Environment* [online]. June 2005, Vol. 8, No. 2, [cit. 2011-01-03]. Dostupný z WWW: <http://cstpr.colorado.edu/admin/publication_files/resource-2604-2005.58.pdf>.

HAN, Jiawei, Micheline KAMBER a Jian PEI. *Data mining*: *concepts and techniques* [online]. 3rd ed. Waltham: Morgan Kaufmann, c2012, 703 s. [cit. 2011-12-30]. Morgan Kaufmann series in data management systems. ISBN 978-012-3814-791.

HANSEN, Derek; SHNEIDERMAN, Ben; SMITH, Marc . *Analyzing Social Media Networks with NodeXL : Insights from a Connected World*. Amsterdam : Elsevier, 2011. 304 s. ISBN 978-0-12-382229-1.

HILBRICH, Robert. *Http://blog.hilbri.ch* [online]. November 27, 2007 [cit. 2011-12-10]. Social Network Analysis using Graph Metrics of Web-based Social Networks. Dostupné z WWW: <http://blog.hilbri.ch/wp-content/uploads/2008/04/sna-presentation.pdf>.

Improved Klout Topics. *Klout.com* [online]. December 6th, 2011 [cit. 2012-01-02]. Dostupné z: http://corp.klout.com/blog/2011/12/improved-klout-topics/

Infographic: First Google Statistics & Facts. *Digitalbuzzblog.com* [online]. [cit. 2012-01-02]. Dostupné z: http://www.digitalbuzzblog.com/infographic-first-google-plus-statistics-and-facts/

Instagram [online]. 2011 [cit. 2011-12-20]. Dostupné z WWW: <http://instagram.com/>.

Instagram Facts & Stats. *Digitalbuzzblog.com* [online]. [cit. 2012-01-02]. Dostupné z: http://www.digitalbuzzblog.com/infographic-instagram-facts/

Internet forum. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 12 August 2003, last modified on 17 December 2011 [cit. 2011-12-19]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Internet_forum>.

Internet World Stats : Usage and population statistics [online]. 2011 [cit. 2011-12-01]. Dostupné z WWW: <http://www.internetworldstats.com/stats.htm>.

Katz centrality. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 26 April 2011, last modified on 19 November 2011 [cit. 2011-12-15]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Katz_centrality>.

Klábosení : O čem se klábosí na českém a slovenském Twitteru. [online]. 2010 [cit. 2011-12-11]. Dostupné z WWW: <http://www.klaboseni.cz/index.php>.

Klout [online]. 2011 [cit. 2011-12-27]. Dostupné z: http://www.klout.com/

Klout Doubt?. *Likeable.com* [online]. 20. JUN, 2011 [cit. 2012-01-02]. Dostupné z: http://www.likeable.com/2011/06/kloutdoubt/

KUBELKA, Martin. *Ochrana soukromí v informační společnosti*: *problematika kamerových sledovacích systémů.* Praha, 2011. Seminární práce. Univerzita Karlova, Ústav informačních studií a knihovnictví, Studia nových médií.

Latent semantic analysis. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 29 May 2004 , last modified on 15 December 2011 [cit. 2011-12-22]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Latent_semantic_analysis>.

Le Web Paris 2011 [online]. 2011 [cit. 2011-12-15]. Dostupné z WWW: <http://leweb.net/>.

LÉVY, Pierre. *Collective intelligence*: *mankind's emerging world in cyberspace* [online]. New York: Plenum Trade, c1997, 277 s. [cit. 2012-01-02]. ISBN 03-064-5635-4.

Líbímseti.cz [online]. 2011 [cit. 2011-12-15]. Dostupné z WWW: <http://libimseti.cz/>.

Lidé [online]. 2011 [cit. 2011-12-15]. Dostupné z WWW: <http://www.lide.cz/>.

LinkedIn [online]. 2011 [cit. 2011-12-15]. Dostupné z WWW: <www.linkedin.com/>.

Machine learning. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 25 May 2003, last modified on 12 December 2011 [cit. 2011-12-22]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Machine_learning>.

MANOVICH, Lev. Trending : The Promises and the Challenges of Big Social Data. *Debates in the Digital Humanities* [online]. 2011, 1, [cit. 2011-12-01]. Dostupný z WWW: <http://www.manovich.net/DOCS/Manovich_trending_paper.pdf>.

Mashup (web application hybrid). In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 19 September 2005 , last modified on 19 December 2011 [cit. 2011-12-20]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid)>.

mBLOG. *mBank.cz* [online]. [cit. 2012-01-02]. Dostupné z: http://www.mbank.cz/blog/

Mechanická Adlina [online]. 2011 [cit. 2011-12-29]. Dostupné z: http://www.mechanicalcinderella.com/adlina.php

Mission Impossible - Ghost Protocol. *Csfd.cz* [online]. [cit. 2012-01-03]. Dostupné z: http://www.csfd.cz/film/245639-mission-impossible-ghost-protocol/

Myspace [online]. 2011 [cit. 2011-12-15]. Dostupné z WWW: <www.myspace.com/>.

NodeXL : Network Overview, Discovery and Exploration for Excel [online]. Microsoft, © 2006-2011, 2011.10.12 [cit. 2011-11-24]. Dostupné z WWW: <http://nodexl.codeplex.com/>.

Orkut [online]. 2011 [cit. 2011-12-15]. Dostupné z WWW: <http://www.orkut.com/>.

Peer Index [online]. 2011 [cit. 2011-12-27]. Dostupné z: http://www.peerindex.com/

Picasa [online]. 2011 [cit. 2011-12-20]. Dostupné z WWW: <http://picasa.google.com/>.

Počátek. *Česko-slovenská filmová databáze* [online]. 2011 [cit. 2011-12-29]. Dostupné z: http://www.csfd.cz/film/254156-pocatek/

PRABOWO, Rudy ; Mike THELWALL. Sentiment Analysis: A Combined Approach [online]. Wolverhampton, UK, 2011 [cit. 2011-12-22]. Dostupné z: http://www.cyberemotions.eu/rudy-sentiment-preprint.pdf. School of Computing and Information Technology University of Wolverhampton.

Precision and recall. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 21 November 2007, last modified on 12 December 2011 [cit. 2011-12-24]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Precision_and_recall>.

Představujeme profil Moje historie. *Facebook.com* [online]. [cit. 2012-01-02]. Dostupné z: https://www.facebook.com/about/timeline

Resource Description Framework (RDF) [online]. 1999-02-22 [cit. 2011-12-10]. Dostupné z WWW: <http://www.w3.org/RDF/>.

RSS. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 7. 10. 2005, last modified on 26. 10. 2011 [cit. 2011-12-19]. Dostupné z WWW: <http://cs.wikipedia.org/wiki/RSS>.

Sentiment analysis. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 13 August 2006, last modified on 17 December 2011 [cit. 2011-12-22]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Sentiment_analysis>.

Size of wikipedia. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 13 December 2004 , last modified on 27 December 2011 [cit. 2011-12-27]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Size_of_wikipedia>.

Small world experiment. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 5 May 2004, last modified on 23 November 2011 [cit. 2011-12-15]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Small_world_experiment>.

Social network. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 23 September 2003, last modified on 21 November 2011 [cit. 2011-11-23]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Social_network>.

Social Network Analysis: A Brief Introduction. *Orgnet.com* [online]. [cit. 2012-01-02]. Dostupné z: http://www.orgnet.com/sna.html

Statistiky. *Youtube.com* [online]. [cit. 2012-01-02]. Dostupné z: http://www.youtube.com/t/press_statistics

Support vector machines. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 7 October 2004, last modified on 7 October 2004 [cit. 2011-12-22]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Support_vector_machines>.

Surveillance. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 2008-01-10, last modified on 2011-06-08 [cit. 2011-01-01]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Surveillance>.

ŠLERKA, Josef. Od pavučiny k mraveništi : aneb kolektivní intelignece za času internetu [online]. 8.12.2009 [cit. 2011-12-12]. Slideshare.net/josefslerka. Dostupné z WWW: <http://www.slideshare.net/josefslerka/od-st-k-mraveniti>.

ŠLERKA, Josef. O sentiment analýze bez sentimentu : aneb jeden malý experiment. Lupa.cz [online]. 2011[cit. 2011-12-24]. Dostupné z: http://www.lupa.cz/clanky/o-sentiment-analyze-bez-sentimentu-aneb-jeden-maly-experiment/

ŠLERKA, Josef. Social Media Monitoring. ATAXO. Slideshare.net [online]. [cit. 2011-12-27]. Dostupné z: http://www.slideshare.net/josefslerka/social-media-monitoring-9608932

Tag cloud. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 7 June 2005, last modified on 28 November 2011 [cit. 2011-12-29]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Tag_cloud>.

The fourth paradigm: data-intensive scientific discovery. Redmond, Wash: Microsoft Research, 2009. ISBN 978-098-2544-204.

The Friend of a Friend (FOAF) project [online]. 2000 [cit. 2011-12-10]. Dostupné z WWW: <http://www.foaf-project.org>.

The Mechanical Cinderella [online]. 2011 [cit. 2011-12-29]. Dostupné z: http://www.mechanicalcinderella.com/index.php

The Petabyte Age: Because More Isn't Just More — More Is Different. *Wired.com* [online]. [cit. 2012-01-02]. Dostupné z: http://www.wired.com/science/discoveries/magazine/16-07/pb_intro

Tiskové zprávy. *Facebook.com* [online]. [cit. 2012-01-02]. Dostupné z: https://www.facebook.com/press/releases.php

TOP 20 filmů. *Unie Filmových Distributorů* [online]. 2011 [cit. 2011-12-29]. Dostupné z: http://www.ufd.cz/top-20-filmu

Track your carbon footprint with ALF. *Aboutfoursquare.com* [online]. [cit. 2012-01-02]. Dostupné z: http://aboutfoursquare.com/amee-location-footprinter/

Tumblr [online]. 2011 [cit. 2011-12-17]. Dostupné z WWW: <https://www.tumblr.com/>.

Twitter [online]. 2011 [cit. 2011-12-17]. Dostupné z WWW: <https://twitter.com>.

Tweet Rank [online]. 2011 [cit. 2011-12-27]. Dostupné z: http://www.retweetrank.com/

UGANDER, Johan . *The Anatomy of the Facebook Social Graph*. Palo Alto, CA, USA, 2011. 17 s. Oborová práce. Cornell University, Ithaca, NY, USA. Dostupné z WWW: <http://arxiv.org/PS_cache/arxiv/pdf/1111/1111.4503v1.pdf>.

VAN DIJK, Jan. *The Network Society : Social Aspects of New Media* [online]. 2nd. London : SAGE, 2006 [cit. 2011-12-12]. Dostupné z WWW: <1-4129-0867-1>.

Voulez-vous check-in avec moi ce soir? [online]. [cit. 2011-12-26]. Dostupné z: http://www.checkinavecmoi.com/

Web 2.0. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 28 February 2005 , last modified on 30 November 2011 [cit. 2011-11-30]. Dostupné z WWW: <http://en.wikipedia.org/wiki/Web_2.0>.

What does Klout Measure?. *Klout.com* [online]. [cit. 2012-01-02]. Dostupné z: http://corp.klout.com/blog/2011/12/what-does-klout-measure

Wikipedia [online]. [cit. 2011-12-27]. Dostupné z: http://www.wikipedia.org/

Wikiskripta.eu [online]. 2011 [cit. 2011-12-27]. Dostupné z: http://www.wikiskripta.eu/index.php/Home

WINDISCH, Eva ; MEDMAN, Niclas . Understanding the digital natives. *Ericsson Business Review* [online]. 2008, 1, [cit. 2011-12-13]. s. 36-39. Dostupný z WWW: <http://www.ericsson.com/ericsson/corpinfo/publications/ericsson_business_review/pdf/108/understanding_digital_natives.pdf>.

Yahoo! Answers [online]. 2011 [cit. 2011-12-19]. Dostupné z WWW: <http://answers.yahoo.com/>.

ZANDL, Patrick. Hon na i-prasata a cyniky z diskusního podpalubí. *Lupa.cz* [online]. 23. 3. 2011 [cit. 2012-01-02]. Dostupné z: http://www.lupa.cz/clanky/hon-na-i-prasata-a-cyniky-z-diskusniho-podpalubi/

ZBIEJCZUK, Adam. *Web 2.0 - charakteristiky a služby*. Brno, 2007. 71 s. Diplomová práce. Masarykova Univerzita v Brně, Fakulta sociálních studií, Katedra mediálních studií a ţurnalistiky. Dostupné z WWW: <http://www.zbiejczuk.com/adam/zbiejczuk_web20.pdf>.

ZHANG, Lei; Riddhiman, GHOSH; Mohamed, DEKHIL; Meichun, HSU; Bing LIU. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. HP laboratories [online]. 2011[cit. 2011-12-24]. Dostupné z: http://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf