Univerzita Pardubice
Fakulta ekonomicko-správní

Využití metody detektivního vytěžování databází

Petr Kališ

Bakalářská práce

2011

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: **2010/2011**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Petr KALIŠ**
Osobní číslo:	**E07056**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Informační a bezpečnostní systémy**
Název tématu:	**Využití metody detektivního vytěžování databází**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

Získávání informací z otevřených zdrojů
Vytěžování dokumentů
Využití vytěžování dokumentů v procesu detektivního rozpracování

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

BRABEC, František, et al. Soukromé detektivní služby . [s.l.] : Eurounion, 1995. 263 s. ISBN 80-85858-16-9.
BRABEC, František. Technologie detektivní činnosti. Zlín : Univerzita Tomáše Bati ve Zlíně, 2009. ISBN 978-807318-780-4.
KAMENÍK, Jiří, BRABEC, František. Komerční bezpečnost. Praha : Aspi, 2007. ISBN 978-80-7357-309-6.

Vedoucí bakalářské práce: **doc. Ing. Pavel Petr, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **4. října 2010**

Termín odevzdání bakalářské práce: **6. května 2011**

L.S.

doc. Ing. Renáta Myšková, Ph.D.
děkanka

doc. Ing. Jiří Křupka, Ph.D.
vedoucí ústavu

V Pardubicích dne 4. října 2010

Prohlašuji:
Tuto práci jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 5. 8. 2011

Petr Kališ

PODĚKOVÁNÍ

Rád bych touto cestou poděkoval doc. Ing. Pavlu Petrovi, PhD. za poskytnutou odbornou pomoc a vedení při tvorbě bakalářské práce.

SOUHRN

Bakalářská práce se zabývá problematikou vytěžování databází a dokumentů se zaměřením na vytěžování otevřených zdrojů. Práce nejprve mapuje typy otevřených zdrojů informací a možné způsoby jejich využití v detektivní a zpravodajské práci. Zaměřuje se na vytěžování elektronických dokumentů a ukazuje některé metody a přístupy k jejich zpracování. Dále se práce zabývá vyhledáváním informací na internetu a sociálních sítích, jakožto největším otevřeném zdroji informací, a na závěr ukazuje možnosti zpravodajských technologií při zpracovávání informací.

KLÍČOVÁ SLOVA

otevřené zdroje, detektivní vytěžování databází a dokumentů, vyhledávání informací na internetu, vytěžování sociálních sítí, zpravodajské technologie

TITLE

Detective usage of database mining methods

ABSTRACT

This thesis deals with mining of databases and documents focusing on the exploitation of open sources. The thesis first maps the types of open sources of information and possible ways of their use in detective and intelligence work. It focuses on the exploitation of electronic documents and shows some methods and approaches to its treatment. Furthermore, the work deals with searching for information on the Internet and social networks as the largest open source of information, and finally shows the possibilities of intelligence technologies in processing information.

KEYWORDS

open sources, detective mining of databases and documents, searching for information on the Internet, mining social networks, intelligence technologies

OBSAH

Úvod

Po pádu mnoha totalitních režimů a diktatur ve světě a s tím souvisejícím zrušením cenzur došlo k výraznému nárůstu publikování informací. S následnou elektronizací a nástupem celosvětové sítě internet přišel informační boom. Jak roste počet uživatelů internetu, roste i počet informací publikovaných prostřednictvím této sítě, zároveň ale klesá jejich hodnota. Lidé jsou zahlcováni informacemi v televizi, tištěných médiích a na internetu především, a vybrat z nich ty relevantní a pravdivé je stále těžší.

Získávání relevantních informací se stává pro všechny instituce životně důležité, vzhledem k stále sílícímu konkurenčnímu prostředí. Informace jsou vysoce ceněnou komoditou a jejich nárůst umožnil vznik mnoha nových oborů, zabývajících se vyhledáváním, analýzou a zpracováním informací a jejich přeměnou na využitelné znalosti. Tyto nové postupy a metody se uplatňují i v oblasti detektivní a zpravodajské práce. Tato fakta mě přesvědčila, abych si vybral za téma své bakalářské práce Využití metody detektivního vytěžování databází.

Cílem této práce je uvést do problematiky vytěžování dokumentů, databází a otevřených zdrojů obecně se zaměřením na využití v detektivní a zpravodajské práci, zejména v procesu detektivního rozpracování. Cílem práce je dále seznámit se s možnostmi vyhledávání informací na internetu a sociálních sítích coby nově se rozvíjejícím fenoménu, zmíněny budou některé nástroje umožňující efektivní sběr informací z těchto zdrojů. V závěru práce budou popsány některé postupy a technologie zpravodajské činnosti, umožňující vyhledávání a analýzu informací.

1 Vytěžování otevřených zdrojů

Informace znamenají v dnešní době značnou konkurenční výhodu ve všech oblastech a oborech lidské činnosti. Informace jsou dnes nejen pasivně přijímány, ale i záměrně vytvářeny a využívány k oslabení, nebo zničení nepřítele – například konkurenční firmy. Informace nestačí jen získat, daleko důležitější je informace interpretovat a přeměnit na znalosti, které lze účelně využít, což je předmětem detektivní práce a zpravodajství. Detektivní činnost je realizována prostřednictvím forem detektivní činnosti, které pomáhají detektivovi organizovat práci, stanovit postup vyšetřování a správně zvolit metody a prostředky, které bude využívat. Jednou z nejdůležitějších forem detektivní činnosti je detektivní rozpracování.

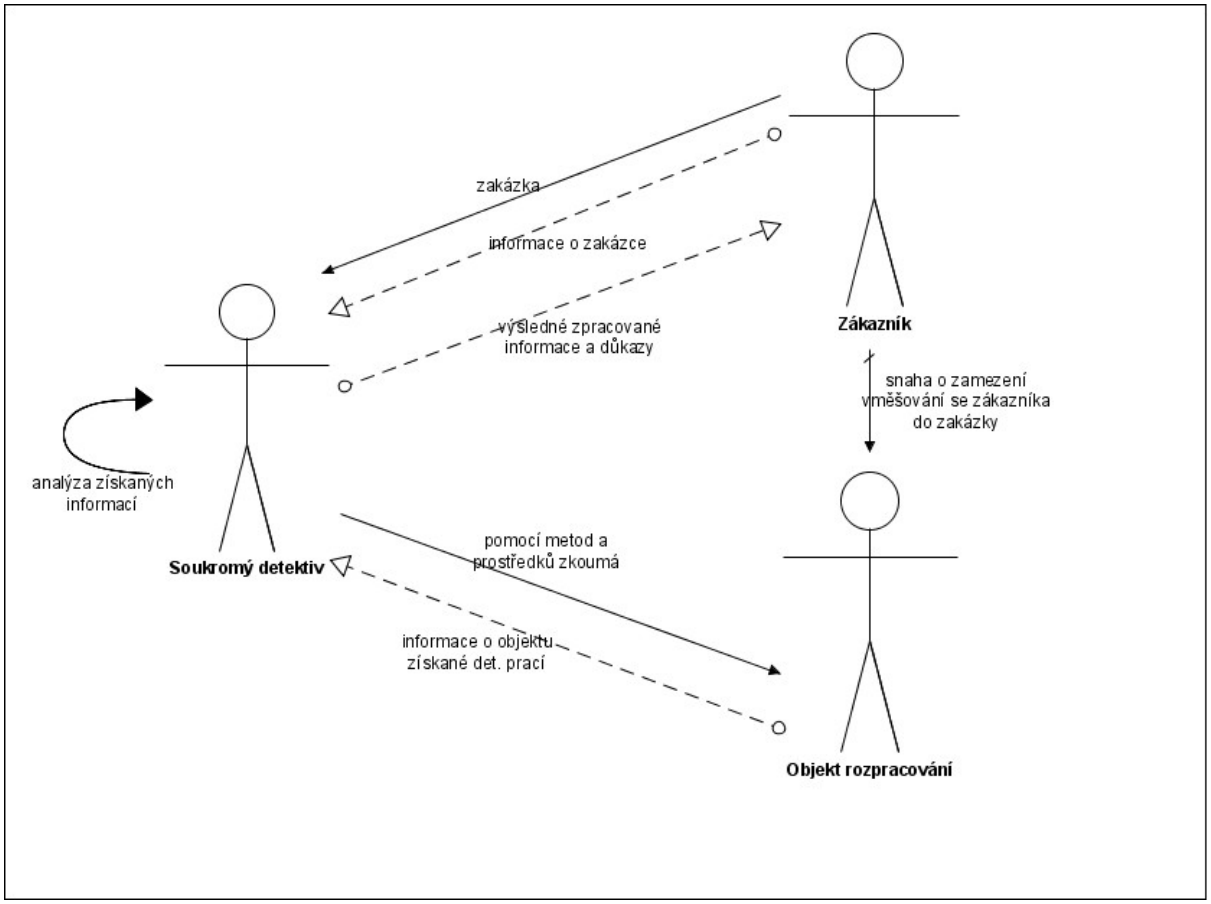
1.1 Detektivní rozpracování

Detektivní rozpracování je vedle detektivního pátrání, detektivní prověrky, detektivní ochrany a detektivního zpravodajství jednou z forem detektivní činnosti. Detektivní rozpracování směřuje zejména k zajištění důkazů a informací o důkazech o protiprávním jednání osob. Jedná se o odborný postup uplatňování kriminalistických a detektivních postupů a metod tak, aby vždy byla zachována průkaznost a ověřitelnost zachycených informací. Z detektivních metod jsou v průběhu detektivního rozpracování uplatňovány především [3]:

- metoda detektivního vytěžování,
- metoda detektivního pozorování,
- metoda detektivního vyhodnocování dokumentů,
- metoda detektivního vedení a vytěžování databází,
- metoda detektivní legendy,
- metoda detektivní obhlídky místa.

Detektivní rozpracování lze charakterizovat jako nejsložitější a nejkomplexnější formu detektivní činnosti. Schematické zobrazení průběhu detektivního rozpracování je znázorněno na obrázku 1, kde plné čáry znázorňují činnosti, přerušované čáry informační toky. Detektivní rozpracování můžeme rozdělit na 4 fáze [3]:

1. **Fáze formulace problému**, kdy je zakázka přijata a jsou shromažďovány informace potřebné pro její splnění. Již v této fázi je potřeba si uvědomit, zda půjde o práci individuální, či týmovou. Jestli ji zvládne jednotlivec, více pracovníků, či dokonce více detektivních agentur.
2. Ve **Fázi detektivního rozpracování** dochází k analýze sebraných vstupních informací, vytyčení detektivních verzí a zpracování plánu operativního rozpracování. Je potřeba rozhodnout o počtu pracovníků a vazeb na ostatní agentury, s ohledem na splnění úkolu a zachování diskrétnosti. Dochází k prolínání do fáze třetí a ze třetí zpět, je potřeba neustále operativně plánovat a analyzovat nové informace a vylučovat nepotřebné detektivní verze.
3. **Fáze realizace detektivního rozpracování**, kde je uplatňována vlastní detektivní činnost s využitím metod detektivní činnosti. Tato fáze je představována na sebe navazujícími a prolínajícími se cykly. V této fázi je nutné sledovat zákazníka, aby z důvodu například netrpělivosti nenarušil vyšetřování a také je potřeba dohlížet a koordinovat detektivy a nasazené týmy či agentury.
4. Poslední je **Fáze uzavření případu,** ve které jsou vyhodnocovány výsledky rozpracování z hlediska jejich použitelnosti, zajišťování důkazů a zpracování závěrečné zprávy pro klienta.



Obrázek 1 Schéma průběhu detektivního rozpracování [vlastní]

Významným dokumentem v detektivním rozpracování je Plán detektivního rozpracování, který je potřeba chápat jako cyklus postupů a opatření, probíhajících po celou dobu detektivního rozpracování. Plán, jakožto písemný materiál, by měl obsahovat [3]:

- zhodnocení situace,
- vytýčení jednotlivých kroků,
- stanovení osobní odpovědnosti za plnění a termínů plnění,
- záznam o výsledku plnění.

Nejdůležitějšími metodami detektivní a zpravodajské práce se v dnešní době stávají metody vytěžování databází, dokumentů a otevřených informačních zdrojů obecně. S nástupem digitalizace společnosti a s rozmachem Internetu se stávají tyto metody nejrychlejší cestou k získání informací. Právě vytěžováním otevřených zdrojů lze dnes získat největší množství informací za relativně nejnižší cenu. Tím se tyto dříve spíše podpůrné metody detektivní práce dostávají do popředí a jsou značně využívány v procesu detektivního rozpracování.

Stále je ovšem nutné využívat komplexně i ostatní metody a jejich pomocí doplňovat chybějící informace, zvláště pak při zpracovávání informací a jejich přeměně na znalosti.

Při vytěžování otevřených zdrojů vzniká řada produktů. Jedná se především o [4]:

- **tiskové monitoringy**,
- **monitoringy z výstav a veletrhů**,
- **tematické rešerše**, které slouží k zmapování informací okolo nějakého tématu, osoby, organizace jevu atd.,
- **přehledové rešerše**, jejichž předmětem je odkrývání souvislostí ve velikém množství informací, shlukování těchto informací a sledování trendů,
- **kontextové rešerše**, při jejichž zpracování je kladen důraz zejména na to, jak budou informace dále využity,
- **investigativní analýzy,** představující pochopení významu informací a jejich přeměna na znalosti o řešeném problému. Může se jednat o situační analýzy, vývojové analýzy, či kontextové analýzy.

1.2 Vytěžování databází a dokumentů

Vytěžování databází jako takových souvisí s nástupem počítačů a digitalizací. Předchůdcem databází byly papírové kartotéky, které byly využívány pro třídění informací. Pro každé kriterium třídění bylo ovšem třeba budovat další kartotéku. Papírové kartotéky jsou dnes již zastaralý systém a jsou nahrazovány digitálními kartotékami a databázemi, které umožňují mnoho pohledů do své struktury, bez nutnosti jakkoli měnit své uspořádání. Pod pojmem databáze si dnes ovšem nelze představit pouze nástupce papírových kartoték. Databáze se staly multimediálními, interaktivními a tím i rychle se měnícími v čase a jejich vytěžování je obtížnější.

Databáze obsahují data již nějak systematicky uspořádaná. Jsou dnes vedeny v podstatě všemi organizacemi. Může se jednat o profesní databáze, sdružující informace o určitém oboru, nebo rejstříky veřejných organizací, evidující informace o určitém ekonomickém prostoru, nebo databáze médií a tisku, informující o aktuálním dění. Databáze vedou knihovny a univerzity, evidující výzkum a informace o určitém tématu, stejně jako komerční organizace informující o svých produktech [4].

Cenné informace je důležité uchovávat pro možné budoucí využití. Proto je i pro každou detektivní a zpravodajskou kancelář vhodné budovat účelové poznatkové evidence, registrace a databáze, které zvyšují efektivnost budoucí práce.

Databáze mohou být tvořeny samotnou osobou detektiva, pro jeho operativní účely, nebo na úrovni agentury, pro využití všemi pověřenými zaměstnanci, nebo na úrovni sdružení více agentur, kde jsou poznatky nejlépe a nejefektivněji předávány a shromažďovány. Dalším typem databáze vedené agenturou jsou databáze povinného charakteru, vyplývajícího ze zákona pro fungování soukromé činnosti. Může se jednat o spisovou evidenci, evidenci zákroků, evidenci zakázek a další [3].

Relační či fulltextové databáze jsou důležité pro další zpracovávání informací a jejich přeměnu na znalosti. Lze v nich jednoduše vyhledávat, třídit a porovnávat informace a sledovat společný kontext. Bez správně utříděných a strukturovaných informací se dnes neobejde žádná společnost. Veliké množství informací je ovšem stále obsaženo v nestrukturovaných otevřených informačních zdrojích.

S vytěžováním otevřených zdrojů souvisí i vytěžování a vyhodnocování dokumentů. Dokumenty jsou významným nositelem informací a jejich vytěžování a vyhodnocování je důležitou součástí detektivní a zpravodajské práce. Zejména písemné dokumenty mohou lépe než kterýkoli jiný informační zdroj sloužit jako důkazní prostředek.

Dokumenty jsou v detektivní práci vyhodnocovány podle [3]:

- **obsahu,** tedy zda byly dokumenty získány cestou oficiální nebo neoficiální, jestli mají vztah k nějaké osobě, události nebo stavu, a jestli se jedná o dokumenty ještě v oběhu, dokumenty uložené nebo dokumenty již archivované,
- **formy,** tedy jestli se jedná o dopisy a podobné písemnosti, doklady prokazující totožnost nebo jiný stav, technické dokumentace a protokoly nebo ostatní spisový materiál,
- **pravosti,** jedná-li se o dokumenty pravé nebo nepravé, což může detektiv pouze odhadovat, ale konečné slovo má až soudní znalec, jehož závěr je platný i u soudního jednání,
- **informační hodnoty,** tedy z hlediska objemu informací a jejich objektivitě,
- **důkazní hodnoty,** zda bude dokument použit jako důkazní prostředek při soudním jednání nebo správním řízení či nikoliv. Podle toho je nutno k dokumentu přistupovat

a patřičně s ním nakládat, aby nedošlo k znehodnocení důkazní hodnoty, zejména prokázat jeho původ.

Vytěžování tištěných dokumentů je velmi zdlouhavé a náročné. Kvůli úspoře času a prostředků se v dnešní době téměř všechny tištěné dokumenty převádějí do elektronické podoby a následně se zpracovávají elektronicky. Převod písemností do digitální podoby probíhá skenováním a následným převodem obrazové informace na textovou technologií OCR/ICR. Takto zpracované dokumenty lze pak již automatizovanou formou dále zpracovávat, indexovat, strukturovat a ukládat do databází. Nutno ovšem podotknout, že naprostá většina dokumentů dnes vzniká již v elektronické formě. Elektronické dokumenty mohou být použity jako vstup indexovacím a vyhledávacím metodám a metodám Text miningu právě k automatizovanému zpracování. Výhody digitalizace jsou zřejmé, především se jedná o úsporu místa při archivaci, vyšší zabezpečení, lepší dostupnost, kvalitnější prohledávání a analýzu, ale hlavně nižší náklady oproti manuálnímu zpracovávání.

1.3 Indexace dokumentů

Indexace dokumentů je proces, při kterém dochází k obsahové analýze dokumentů. Neboli dochází k vyjádření obsahu dokumentu pomocí termínů určitého selekčního jazyka. Tím je pak usnadněno jeho obsahové zařazení a následné vyhledávání. Postupy indexace nejprve určí předmět dokumentu a následně tento předmět vyjádří pojmy selekčního jazyka. Podle použitých postupů lze indexaci rozdělit na [5]:

- **automatickou indexaci**, kde proces indexace probíhá pomocí automatických postupů ve všech fázích,
- **intelektuální indexaci**, při které všechny fáze probíhají pomocí intelektuálních postupů ve všech fázích indexace,
- **poloautomatickou indexaci**, při které dochází ke kombinaci postupů automatických s intelektuálními.

Automatická indexace je plně realizována pomocí automatických postupů, tedy za využití informačních technologií a počítačů. Souvisí tedy právě s vývojem informačních technologií. S rozvojem internetu a stoupajícím množstvím publikovaných dokumentů je právě automatický proces indexace již nezbytný a nenahraditelný. Automatická indexace dokáže

objektivně označit obsah dokumentu pomocí příslušných pojmů selekčního jazyka. Je to metoda rychlá, levná, relativně spolehlivá a efektivní. Ke všem dokumentům přistupuje bez subjektivního vjemu [5].

Do procesu automatické indexace vstupuje úplný text dokumentu, který je následně zpracováván procesy obsahové analýzy. Dochází při nich k výběru slov, která mohou mít význam pro vystižení obsahu dokumentu. Výběr slov se děje například porovnáváním s deskriptory nebo znaky tezauru, nebo podle umístění ve větě a na stránce, nebo podle počtu výskytů v textu, atd. Vybraná významová slova jsou převáděna na svůj základní gramatický tvar. Následuje výběr indexních termínů, které odpovídají vybraným významovým slovům. Jedná se tedy vlastně o jakýsi překlad z jazyka přirozeného do zvoleného jazyka selekčního [7].

Automatická indexace má ovšem i své nevýhody a mezery. Počítač totiž zatím nemůže zpracovávanému textu porozumět tolik jako člověk. Některá slova nebo fráze mají v odlišném kontextu odlišný význam. Je nutno vzít v úvahu použití synonym, homonym, výrazů nerelevantních pro obsah dokumentu, odkazů na jiný obsah, či tvarosloví. Na to musí být brán zřetel a tomu se musí přizpůsobovat algoritmy automatické indexace. Aby automatická indexace probíhala co nejrychleji, dochází k určitým zjednodušením a zkratkám, které také snižují kvalitu výsledného indexování. Například dochází k vynechání takzvaných stop slov, čili spojek, předložek a podobných větných členů. Tím také může dojít ke změně významu některých slov a vět a ztížení správné indexace [5].

Všem výše uvedeným problémům se dokáže vyhnout indexace intelektuální. Ta je ve svém výsledku kvalitnější než automatická. Člověk dokáže zpracovávanému textu porozumět bez větších problémů a správně ho indexovat. Ovšem značně záleží na osobě indexátora, jeho znalostech, zkušenostech a schopnostech. Tím se stává intelektuální indexace značně subjektivní záležitostí, nehledě na to, že člověk nedokáže zpracovat tak obrovské množství dokumentů jako stroj. Intelektuální indexace je pomalá, náročná a nákladná [5].

Pokud bychom porovnávali míru shody indexace dvou dokumentů, vítězem by byla indexace automatická, která využívá stále stejné postupy a přistupuje ke všem dokumentům objektivně. Není ovlivňována únavou, nepozorností a dalšími faktory. Intelektuální indexace oproti tomu dokáže lépe pochopit a vystihnout obsah zpracovávaného dokumentu. Automatická i intelektuální indexace mají obě své výhody a nevýhody, ale směřují ke stejnému cíli. Jako

nejúčinnější se jeví kombinace těchto metod, tedy indexace poloautomatická, při které lze nevýhody obou metod snížit.

S rostoucími počty vytěžovaných dokumentů stoupá i snaha o zautomatizování extrakce informací a znalostí. Pokročilými metodami, které dokáží nejen identifikovat obsah dokumentů, ale i z dokumentů vytěžit užitečné informace a znalosti, jsou metody text miningu.

1.4 Text mining

Text mining, neboli „dolování znalostí z textu“, se nazývá soubor technik pro extrakci informací z nestrukturovaných dokumentů. Text mining se snaží dokumenty automaticky filtrovat na vstupu, shlukovat a klasifikovat a vyvozovat z nich informace a znalosti. Dovede identifikovat hledané informace a odhalit v textu skryté trendy a vztahy s těmito informacemi související [17].

Vstupem pro text mining mohou být libovolné elektronické textové dokumenty. Může se jednat o elektronickou poštu, blogy, články, reporty a zprávy a další internetové dokumenty. S nestrukturovaností vstupních dokumentů souvisí i mnoho problémů, s kterými se text mining musí potýkat. Jedná se o obrovské soubory dat, jejichž obsah se často mění, jsou psány v různých jazycích, mohou se v nich vyskytovat chyby, dvojsmysly. Jsou uloženy v různých formátech s různým kódováním. Jsou psány na rozdílné úrovni jazyka i znalostí. Na to vše musí být brán zřetel. Proto je kladen veliký důraz na předzpracování dokumentů a jejich normalizaci do tvaru, v kterém jsou nadále zpracovávány a kde jsou z nich získávány znalosti. Předzpracování a získávání znalostí jsou dvě základní části text miningu a text miningové nástroje můžeme tedy rozdělit na filtry a analyzátory [19].

V části předzpracování je z dokumentu extrahován čistý text, který je normalizován, tzn, převeden na stejný font, velikost písma a zvýraznění. Grafické nebo jiné informace jsou ignorovány, pracuje se pouze s textem. Struktura textu se pečlivě zachovává, protože může být jediným vodítkem k vytvoření struktury textu a určení významu termů. *[17]„Termy představují základní objekty, s nimiž se provádí další zpracování. Pojem term nemá v text miningu pevně stanovený význam. Jedná se o základní prvek, s nímž probíhá zpracování, ale jeho tvar se může lišit podle metody, která ho využívá např. věta při sumarizaci textu nebo jednotlivá slova (souslovi) při extrakci informací.“* Po identifikaci a extrakci termů dochází v části získávání znalostí k analýzám termů. Jsou provedeny operace vyhledávání vědomostí a

rozhodovací procesy, které vedou k požadovaným výsledkům. Závěry bývají zobrazeny pomocí vizualizačních nástrojů [17].

Text mining využívá celou řadu metod a postupů, které navzájem kombinuje při dolování znalostí v textu. Zejména se jedná o zpracování přirozeného jazyka (Natural Language Processing neboli NLP). NLP neboli počítačová lingvistika se snaží vyřešit jeden s nejstarších a nejsložitějších problémů v oblasti počítačové vědy a umělé inteligence, a to je porozumění přirozenému jazyku strojem. Mezi úkoly NLP, které Text Mining využívá, patří také strojový překlad (Machine translation), neboli automatický překlad z jednoho přirozeného jazyka do jiného, získávání informací (Information retrieval), mezi které patří služby vyhledávačů, o kterých bude zmínka, Extrakce informací (Information extraction), a další.

Mezi aplikace dolování v textech patří [19]:

- **kategorizace textů**, která je jedním z úkolů text miningu, při kterém jsou dokumenty automaticky zařazovány do předdefinovaných množin kategorií. Kategorie mohou být rozděleny podle tématu, autora, klíčových slov, názvu, oboru atd.,
- **extrakce informací**, jež je jedním z úkolů NLP, se snaží automaticky nacházet strukturované, nebo semistrukturované informace v nestrukturovaných textech podle jejich obsahu. Tato činnost zpravidla předzpracovává dokumenty a její výstup poskytuje informace dalším metodám na jejich vstupu,
- **shlukování textů**, což je automatický proces, který se snaží dokumenty shlukovat do skupin podle jejich vnitřní obsahové podobnosti. Vzájemná podobnost dokumentů se maximalizuje, podobnost skupin mezi sebou se minimalizuje. K identifikaci námětů a témat pro zařazení dokumentů do jednotlivých skupin se používají termy, které jsou pro danou skupinu typické,
- **automatická identifikace jazyka dokumentu**, která probíhá na základě porovnávání frekvencí charakteristických dvojic a trojic písmen s předem vytvořenými tabulkami konkrétních jazyků, čímž jazyk dokumentu s určitou přesností identifikuje. Přesnost závisí na délce dokumentu,
- **A další** aplikace text miningu, jako automatické rozdělení dokumentu, určení autora dokumentu, již zmíněný automatický překlad, identifikace kopírování dokumentu atd.

Text miningové nástroje, spolu s nástroji web miningovými, dokáží v relativně krátkém čase zpracovat veliké množství informací, což je při dnešním zahlcení informacemi důležité snad

ve všech oborech lidské činnosti. Kromě běžných odvětví kde se uplatňují, jako je marketing, věda atd., nachází uplatnění i při odhalování kriminality a prevenci. Může se jednat o detekci hrozeb a bezpečnostních rizik, předcházení teroristickým a kybernetickým útokům, odhalování podvodů v bankovnictví a pojišťovnictví, vše pomocí modelů vytvořených těmito technikami.

2 Typologie informačních zdrojů

Informační zdroj je *[25]"Informační objekt, který obsahuje dostupné informace odpovídající informačním potřebám uživatele"*. Informační zdroje mohou být rozděleny podle mnoha hledisek a třídění podle nich se vzájemně prolíná. Způsob vytěžování informačních zdrojů souvisí především s typem informačního zdroje. Každý typ informačního zdroje vyžaduje jiný přístup k vytěžování informací.

Informační zdroje můžeme rozdělit podle:

- **detektivního a zpravodajského pohledu** na primární a sekundární,
- **dostupnosti** na veřejné, neveřejné a tajné,
- **formy publikování** na tištěné, mikrografické a elektronické,
- **způsobu prezentace informací** na textové, numerické, obrazové, zvukové, audiovizuální a multimediální.

Zdroje z detektivního a zpravodajského pohledu

Zdroje informací z pohledu detektivního a zpravodajského je možné rozdělit na [4]:

- **primární**, které musí být speciálními metodami a prostředky detektivní činnosti vytěžovány z takzvaných nosičů. Nosiči mohou být lidé, firmy, prostředí apod. Mezi metody pak patří detektivní vytěžování, pozorování, legenda, osobní pátrání a další. Primární zdroje informací nejsou tedy zaznamenány, či shromážděny,
- **sekundární**, což jsou pak ty, které již v nějaké formě, ať již tištěné, zvukové, obrazové, či elektronické, zaznamenány jsou. Jsou tedy výstupem detektivní činnosti při vytěžování primárních zdrojů.

Z pohledu detektivního a zpravodajského jsou v podstatě všechny sekundární zdroje zdroji otevřenými, nicméně otevřené zdroje jsou obecně chápány ještě úžeji.

Při detektivní a zpravodajské práci mají otevřené, tedy sekundární zdroje informací, rozhodující vliv. Až 80% poznatků pochází právě z otevřených zdrojů (z toho 10% z vlastních databází), 20% poznatků je výsledkem vytěžování primárních zdrojů informací dalšími metodami detektivní činnosti [4].

Zdroje podle dostupnosti

Pro detektivní a zpravodajskou práci je velmi důležitá také dostupnost informací, respektive legalita vstupu detektiva do určitých evidencí a registrací.

Podle dostupnosti se tedy informační zdroje dále dělí na [14]:

- **tajné**, neboli utajované zdroje informací, se řídí zákonem č. 412/2005 Sb., který *[27]„upravuje zásady pro stanovení informací jako informací utajovaných, podmínky pro přístup k nim a další požadavky na jejich ochranu, zásady pro stanovení citlivých činností a podmínky pro jejich výkon a s tím spojený výkon státní správy“*. Jsou to například systémy zajišťující obranu státu, různé policejní databáze apod.,
- **neveřejné** zdroje obsahují *[25]„Informace dostupné omezenému okruhu uživatelů. Omezení může být pasivní (např. informace, která nebyla dosud publikována) nebo aktivní (k informaci je zamezen přístup neoprávněných uživatelů). Nejčastější techniky ochrany před neoprávněným zveřejněním a přístupem představuje zachovávání mlčenlivosti a omezování přístupu k informacím pouze na oprávněné uživatele.“* Může se jednat např. o firemní informační systémy,
- **veřejné** zdroje informací mají neomezený okruh příjemců a zpřístupňují informace široké veřejnosti.

Do utajovaných evidencí mají přístup pouze povolané osoby a vstup jiných osob je nelegální. Podle stupně utajení evidence se nepovolaná osoba může dopustit trestného činu. Vytěžování těchto informačních zdrojů je v podstatě vymezeno pouze pro státní úřady.

Neveřejné informační zdroje mohou sloužit jako velmi cenný zdroj informací a důkazů v procesu detektivního rozpracování i s ohledem na to, že se jedná o informace využitelné k získání konkurenční výhody. Je potřeba ale vždy zhodnotit, jestli přístupem k těmto

informacím není naplněna skutková podstata trestného činu proti hospodářskému, obchodnímu, či služebnímu tajemství.

Veřejně přístupné informační zdroje jsou přístupné pro každého občana, buď zdarma, nebo po zaplacení poplatku. Zde je situace pro detektiva nejobtížnější, protože informací je mnoho a jsou třeba neúplné. V tomto případě je vhodné například provázat vytěžování tohoto zdroje informací s vytěžováním nějakého zdroje primárního. Třeba navázání kontaktu se zájmovou osobou apod.

Zdroje podle formy publikování

Z hlediska formy, v jaké jsou publikovány, můžeme informační zdroje rozdělit na [1]:

- **tištěné**,
- **mikrografické**,
- **elektronické**.

Tištěná forma

Nejstarší a donedávna nejpoužívanější formou publikování dokumentů je forma tištěná. Existuje v podstatě již od sedmého století po Kristu. Největšího rozmachu dosáhla na konci patnáctého století s vynálezem knihtisku. Z vývojového hlediska jde o již uzavřenou etapu.

Zdroje v tištěné formě je obtížné a poměrně zdlouhavé vytěžovat. I převod do elektronické podoby s sebou nese jisté problémy. Pro kvalitní vyhledávání v tištěných dokumentech je důležitá správná katalogizace, neboli jmenný a věcný popis dokumentů. Pro účely vyhledávání dokumentů jsou tvořeny různé druhy katalogů.

Jedná se například o knihovní a souborné katalogy, které obsahují informace o dokumentech, které mají knihovny nebo archivy ve svém fondu. Dříve byly používány katalogy lístkové, které jsou ale nahrazovány katalogy digitálními, zefektivňujícími vyhledávání. Příkladem může být online katalog Národní knihovny ČR – http://aleph.nkp.cz/cze/nkc , nebo Souborný katalog České republiky – HTTP://ALEPH.NKP.CZ/CZE/SKC. [1]

Pro vyhledávání podle oboru jsou tvořeny bibliografie, resp. bibliografické databáze, obsahující seznamy literatury vypracované na dané téma. Předchůdcem bibliografických

databází byly referátové časopisy. Příkladem bibliografické databáze může být Česká národní bibliografie - HTTP://ALEPH.NKP.CZ/CZE/CNB [1]

Významným zdrojem informací o obsahu odborných a vědeckých periodik jsou tzv. current contents dokumenty, Dnes přístupné v databázové formě. Tyto databáze produkované Institutem pro vědecké informace – Thomson ISI, shrnují obsahy přes 8500 časopisů, 2000 knih a více než 7000 webových adres. [1]

Dalšími zdroji informačních pramenů mohou být ještě citační rejstříky a nakladatelské katalogy. Většina tištěných dokumentů je dnes převáděna do elektronické podoby a nové tištěné dokumenty mívají ve většině případů i formu elektronickou, v které pak není problém hledat za pomoci již zmiňovaných automatizovaných nástrojů. Tištěné dokumenty ovšem stále hrají velkou roli v důkazním řízení, jsou obecně nejlépe přijímány jako důkaz u soudu.

Mikrografická forma

Druhou formou ukládání a publikování dokumentů je forma mikrografická. Jedná se o dokumenty uložené na mikrografických médiích (mikrofilmech, mikrofiších). Tato metoda uchovávání informací byla hojně využívána za druhé světové války a rozmachu se jí dostalo v druhé polovině dvacátého století. Začala být používána jako kompaktnější forma uchovávání textových a grafických informací. Knihovny převedením především tištěných novin a časopisů ušetřily až 95 procent kapacit v porovnání s archivováním celých výtisků. Archivovány tímto způsobem byly i technické výkresy, celé knihy, apod [1].

V dnešní době digitalizace pozbývá tato analogová forma ukládání dokumentů významu. Vytěžování je obtížné, jsou k němu zapotřebí speciální zobrazovací prostředky, zvláštní zacházení atd. Výhodami ve své době byla především úspora místa, nízké náklady na transport, při zachování určitých zásad se jednalo o stálé médium. Nevýhodou bylo špatné zobrazování grafických informací – fotek a kreseb, obtížná tvorba tištěných kopií, již zmíněná nutnost speciálního technického vybavení. Tuto formu uchování informací bychom mohli označit za předchůdce digitalizace, protože poprvé řešila nedostatek úložného prostoru pro stále narůstající množství publikovaných informací.

Elektronická forma

Třetí a nejdůležitější formou publikování dokumentů je forma elektronická. Za elektronickou formu publikování informací lze považovat formu zaznamenanou a prezentovanou pomocí elektronického zařízení. Její nejdůležitější a dnes již nejrozšířenější podmnožinou je forma **digitální**, zaznamenávající informaci jako řetězce bitů. Jedná se v podstatě o nový způsob uchovávání informací, je perspektivní a neustále se vyvíjí a zdokonaluje. Umožňuje kompresi dat, doplnění o metadata, kombinování různých formátů atd.

Elektronická informace může vzniknout třemi způsoby. Pořízením rovnou v elektronické podobě (například napsáním v textovém editoru, nahráním na pásku), konverzí z jiného zdroje (například OCR skenování tištěných textů) nebo vygenerováním, například různých sestav z dat uložených v databázích. Elektronické informační zdroje mohou být uloženy na různých nosičích, zpočátku se používaly děrné štítky, magnetické pásky, diskety, magnetofonové pásky, VHS, později disky CD\DVD-ROM, pevné disky a flash disky atd. Mohou být přenášeny energií, neboli vysílány online. [14]

Výhody elektronické formy informačních zdrojů jsou zřejmé. Úspora knihovního místa při digitálním uložení dokumentů je ještě daleko větší než při mikrografickém ukládání informací. S tím souvisí i nemalé finanční úspory. Prohledávání elektronických informačních zdrojů a jejich strojové zpracování je značně usnadněno a může probíhat automaticky (například katalogizace, indexování, archivace atd.). Elektronické dokumenty mohou být multimediální a obsahovat textové, obrazové, zvukové informace. Elektronické dokumenty mohou být interaktivní a nutit čtenáře se aktivně zapojovat a měnit tak směr informací, nebo zužovat jejich výběr. Výhodou je i velmi snadné pořízení kopie elektronické informace.

Nevýhodami ukládání informací v elektronické podobě je hlavně nestabilita nosičů a hardwaru vůbec. Archivace a zálohování elektronických dokumentů je zásadní. Ke zničení může dojít snadno, kvůli poruchám technického zařízení, softwarového vybavení, ale hlavně díky selhání lidského faktoru. Problémem jsou i odlišné, často navzájem nekompatibilní, formáty elektronických dokumentů, které ztěžují jejich vytěžování. Se snadnou aplikovatelností informací v elektronické podobě souvisí problémy při ochraně autorských práv a ochraně osobních údajů a dat. Vzhledem ke snadné a masové šiřitelnosti informací po internetu se jedná o závažný problém a informace v elektronické podobě musí být chráněny

speciálními prostředky. Informací je nadměrné množství a ověřování jejich prokazatelnosti a důvěryhodnosti je obtížné.

Zdroje podle způsobu prezentace informací.

Zdroje informací lze bez ohledu na formu, ve které jsou uloženy, rozdělit podle toho, jakým způsobem prezentují informace v nich obsažené na [1]:

- **textové,**
- **numerické,**
- **obrazové,**
- **zvukové,**
- **audiovizuální,**
- **multimediální.**

Textové zdroje

Nejobsáhlejšími zdroji informací jsou zdroje textové. Druhů textových informací je obrovské množství, mezi nejdůležitější patří [14]:

- **monografie**, což jsou publikace komplexně zpracovávající jedno specializované téma. Jedná se většinou o odbornou práci, jejímž původcem je sám autor,
- **sborníky**, což jsou publikace, které v sobě sdružují dva a více samostatných textů příbuzného tématu publikované pod jedním názvem. Sborníky mohou být jednorázové, nebo periodické,
- **slovníky,** které k abecedně řazenému seznamu slov, výrazů, symbolů aj. přiřazují příslušnou vysvětlující informaci, podle zaměření daného díla. Slovníky mohou být výkladové, překladové, etymologické, speciální, a další,
- **encyklopedie**, což jsou díla uspořádaná abecedně nebo systematicky a shrnující základní pojmy lidského vědění,
- **příručky**, které jsou určeny k rychlému vyhledávání základních informací o daném tématu,
- **učebnice**, což jsou učební texty sledující didaktické cíle. Tedy na dané téma vypracované publikace mající za cíl podat příslušný teoretický, či praktický výklad dané problematiky,

- **skripta**, což jsou dočasné učební texty pro vysoké školy, nahrazující učebnice,
- **noviny**, neboli periodicky vycházející publikace (alespoň jednou týdně), obsahující aktuální politické, kulturní, ekonomické, sportovní a jiné zprávy,
- **časopisy**, což jsou taktéž periodicky vycházející publikace (alespoň jednou za půl roku), úžeji, nebo odborně zaměřeny na jedno, či více příbuzných témat,
- **ročenky**, což jsou zpravidla jednou ročně vycházející souborné publikace obsahující přehledné informace různých druhů, jako například výsledky činnosti za kalendářní rok. Například výroční zprávy, události roku, atd.,
- **technické normy** neboli technická literatura jednotně stanovující charakteristiky výrobků, činností, bezpečnostních metod atd. na světové, státní, oblastní, či podnikové úrovni,
- **patentové dokumenty**, což jsou dokumenty týkající se ochrany intelektuálního vlastnictví. Jedná se o popisy vynálezů, ochranných známek, užitných a průmyslových vzorů atd.,
- **firemní propagační literatura**, což je literatura obvykle vydávaná firmami pro propagaci výrobků a služeb. Například prospekty, příručky, katalogy, výroční zprávy,
- **legislativní dokumenty**, jako jsou zákony, vyhlášky, předpisy a nařízení,
- **výzkumné zprávy** vznikající v souvislosti s řešením nějakého úkolu a informující o jeho výsledku, či průběhu,
- **disertace**, což jsou práce vypracované za účelem získání určité kvalifikace vědecké, akademické, pedagogické (diplomové práce, kandidátské práce, doktorské práce, habilitační práce, rigorózní práce),
- **materiály z konferencí**, jako jsou sepsané příspěvky a prezentace uvedené na určité konferenci či kongresu,
- **a další.**

Všechny zmíněné textové zdroje a mnoho dalších slouží, nebo mohou sloužit jako důležitý zdroj informací. Dalšími textovými zdroji, využitelnými především v procesu detektivního rozpracování, budou zejména vnitřní i vnější firemní korespondence, psané zprávy, reporty a výkazy a další dokumenty, zejména pak dokumenty v elektronické podobě. V procesu detektivního rozpracování najdou využití především jako důkazní prostředky.

Numerické zdroje

Zdroje informací, v kterých převažují údaje zahrnující čísla nebo číselnou reprezentaci, jsou zdroje numerické, mezi které patří [14]:

- **statistiky**, což jsou kvantitativní údaje o studované oblasti zájmu. Nejvýznamnější institucí provádějící statistické průzkumy u nás je Český statistický úřad HTTP://WWW.CZSO.CZ,
- **ceníky**, které přiřazují zájmovým položkám údaje o jejich finanční hodnotě,
- **kursovní lístky** informující o poměrech měn jednotlivých zemí HTTP://WWW.KURZY.CZ/KURZY-MEN/,
- **meteorologické a hydrometeorologické údaje**, což jsou měřené údaje o chování atmosféry. Sleduje se teplota, vlhkost, tlak, průtok řek atd. HTTP://HYDRO.CHMI.CZ/HPPS,
- **jízdní a letové řády** informující o odjezdech a příjezdech, respektive odletech a příletech dopravních prostředků HTTP://JIZDNIRADY.IDNES.CZ,
- **tabulky matematické, fyzikální, chemické, astronomické** obsahující vzorce, definice, převodní vztahy, tabulky hodnot veličin aj. potřebné v daném oboru HTTP://WWW.ETABULKY.CZ/,
- **kalendáře**, což jsou abstraktní způsob dělení času pro usnadnění orientace. HTTP://IKALENDAR.INFO/,
- **a další.**

Numerické zdroje informací jsou využívány především k odhalování hospodářské kriminality. Jako důkazy zde slouží především účetní výkazy, různé statistiky a přehledy.

Obrazové zdroje

Zdroje informací, které svůj obsah vyjadřují obrazově nebo prostřednictvím symbolů a značek, se nazývají obrazové. Jejich výhodou je, že jsou srozumitelné i při neznalosti jazyka nebo písma. Jedná se např. o [1]:

- **grafiky, ilustrace**, které jsou druhem výtvarného umění, můžou být ručně řemeslně rozmnoženy. Za druhé se může jednat o produkt vytvořený počítačovou aplikací HTTP://ARTLIST.CZ/?ID=315,
- **fotografie**, což jsou dokumenty vytvořené reakcí citlivého média na dopadající světlo HTTP://WWW.PIXMAC.CZ/,

- **mapy**, neboli zmenšené zjednodušené zobrazení Zemského povrchu, hvězdné oblohy aj. HTTP://MAPS.GOOGLE.CZ/,
- **výkresy, plány, schémata** znázorňující strukturu nějakého území, budov, strojů, problémů atd. HTTP://WWW.URM.CZ/CS/GRAFICKA_CAST,
- **diagramy**, což jsou grafická vyjádření různých vztahů, myšlenek apod. HTTP://VDB.CZSO.CZ/VDBVO/MAKLIST.JSP?&VO=GRAF&,
- **a další.**

Jako důkazní prostředek slouží již dlouhá léta fotografie, ať již se jedná o zdokumentování místa spáchaného činu, nebo k dokumentování činů sledované osoby, či jen pro kontrolní účely. Obrazové informace jsou dnes sdružovány do obrazových databází a informačních systémů HTTP://WWW.GOOGLE.CZ/IMGHP?HL=CS&TAB=WI. Příkladem mohou být geografické informační systémy, tedy systémy informací vztažených k povrchu země. Tyto systémy mohou být pak použity k plánování předpovědí počasí, budoucí výstavby, vytyčení oblastí ohrožených záplavami, plánování prevence kriminality, předvídání výskytu pohřešovaných osob atd.

Vyhledávání grafických informací není jednoduchá záležitost. Již celkem starý způsob vyhledávání podle popisu obrázku, kterého využívají běžné vyhledávače, je časově náročný, vzhledem k pracné indexaci každého obrázku. Nový přístup vyhledávání podle obsahu, tedy podle podobnosti tvarů a barev, může být nepřesný. Jako nejlepší se zatím jeví kompromis, neboli různé hybridní systémy, využívající jak indexace, tak tvarové podobnosti. Nový přístup využívající metod dataminingu je image mining. Využívá poznatků z oblasti vytěžování dat a zpracování obrazu a aplikuje je na celé databáze obrazových informací [6].

Systémy pro automatické zpracování obrazu a vyhledávání grafických informací jsou v detektivní práci velmi prospěšné. Využívají se například při pátrání po odcizených uměleckých dílech, pomáhají při identifikaci osob, dá se pomocí nich odhalit zneužití ochranných známek a průmyslových vzorů. Příkladem může být policií dlouhodobě využívaný systém LOOK, který je schopen automaticky separovat z fotografií projíždějících automobilů poznávací značku a porovnávat ji s evidencí centrálního registru řidičů, či s policejním registrem pátrání po motorových vozidlech.

Zvukové zdroje

Další skupinou informačních zdrojů jsou zvukové informační zdroje obsahující zvukové informace zaznamenané a reprodukované pomocí technického zařízení. Patří mezi ně [1]:

- **zvuky** HTTP://WWW.FINDSOUNDS.COM/,
- **hudba** je složená povětšinou z tónů (periodické kmitání) HTTP://MP3S.NADRUHOU.NET/,
- **mluvené slovo** HTTP://WWW.DATABAZE-HLASU.CZ/,
- **rozhlasové vysílání a zprávy** HTTP://WWW.PLAY.CZ/,
- **a další.**

Podobně jako u ostatních druhů jsou i zvukové informace sdružovány do různých audiodatabází. Vyhledávání zvukových informací funguje na podobném principu jako u informací obrazových, také existují dva přístupy, vyhledávání podle popisu, neboli indexace, a vyhledávání podle podobnosti. Při vyhledávání podle podobnosti se robotovi přehraje nahrávka jako vstupní vzorek, on ji zpracuje a vyhledává nahrávky podobné. V detektivní oblasti je využití hlavně při identifikaci hlasů osob například při vydírání, pátrání po osobách atd. Nutno poznamenat, že hlasová nahrávka slouží pouze jako podpůrný důkaz v soudním řízení.

Audiovizuální zdroje

Sloučením informací ve zvukové podobě a dynamických obrazových informací vznikají audiovizuální dokumenty a tedy i audiovizuální zdroje informací. Sdružují v sobě výhody obou forem, tedy srozumitelnost, lepší zapamatovatelnost a rychlost vnímání nových informací [1].

- **filmy hrané, nebo dokumentární** HTTP://WWW.IMDB.COM/,
- **televizní vysílání a zprávy** HTTP://WWW.TVLINK.CZ/,
- **video** HTTP://WWW.YOUTUBE.COM/,
- **A další.**

Z audiovizuálních zdrojů se v detektivní práci využívají především kamerové záznamy různých bezpečnostních systémů, nahrávajících sledovanou oblast. Kamerové záznamy se dále využívají například k zaznamenání výpovědí svědků, lepšímu zadokumentování místa činu atd. U soudu slouží také pouze jako podpůrné důkazy. Z tohoto důvodu vznikl nový

kriminalistický obor využívající kamerové záznamy – biomechanika. Biomechanika umožňuje z pohybů člověka, především z chůze, přesně identifikovat konkrétní osobu. Metoda má ovšem značná omezení. Vyžaduje nahrání člověka z více úhlů a především pak pořízení kontrolní nahrávky podezřelého pro srovnání.

Multimediální zdroje

Sloučením více druhů informací do jednoho informačního zdroje vznikají multimediální zdroje informací. Tedy zdroje, které v sobě mohou kombinovat informace textové, numerické, obrazové, zvukové i audiovizuální. Vytěžována bývá obvykle každá složka zvlášť svým speciálním nástrojem a výsledky jsou pak navzájem kombinovány. Nejvíce vytěžovanou složkou je pochopitelně složka textová.

Vznik multimediálních zdrojů informací umožnil až nástup osobních počítačů natolik výkonných, že byly schopné zpracovat i audiovizuální složku. V roce 1991 vzniklo konsorcium pod vedením firmy Microsoft, které rozhodlo o první konfiguraci multimediálního počítače MPC. Ta byla v dalších letech několikrát aktualizována na MPC level 2, MPC level 3 atd. [10].

3 Významné informační zdroje v ČR

Při vyhledávání informací je důležité si uvědomit, zda někdo požadované informace neeviduje, nebo jestli je již někdo nevyhledal a nestrukturalizoval. Pak by bylo zbytečné a nákladné je vlastními prostředky shromažďovat. Existuje množství databází vedených státem nebo komerčními organizacemi, které zdarma či za úplatu poskytují různé druhy informací využitelných při detektivní a zpravodajské práci.

Jedním z nejdůležitějších informačních zdrojů v ČR jsou centrální databáze a registry státní správy. Mezi nejvyužívanější dle všeobecného povědomí patří:

- **Obchodní rejstřík** - Obchodní rejstřík a sbírka listin Ministerstva spravedlnosti je evidence, kde jsou vedeny podnikatelské subjekty v něm zaregistrované. HTTP://WWW.JUSTICE.CZ/OR/,

- **Registr ekonomických subjektů** – Registr ekonomických subjektů je rejstřík vedený Českým statistickým úřadem, evidující všechny právní formy ekonomických subjektů v ČR HTTP://REGISTRY.CZSO.CZ/IRSW/,
- **Registr živnostenského podnikání** - Registr živnostenského podnikání uchovává veřejné informace o podnikatelských subjektech a jeho provozovatelem je Ministerstvo průmyslu a obchodu. HTTP://WWW.RZP.CZ/,
- **Katastr nemovitostí** – Katastr nemovitostí je evidence vedoucí údaje o vlastnictví parcel, staveb a jednotek a o stavech řízení vedených katastrálním pracovištěm. HTTP://NAHLIZENIDOKN.CUZK.CZ/,
- **Evidence úpadců** - Evidence úpadců je registr ministerstva spravedlnosti, který vede informace o konkursech a vyrovnáních. HTTP://WWW.JUSTICE.CZ/CGI-BIN/SQW1250.CGI/UPKUK/S_I8.SQW,
- **Insolvenční rejstřík** – Insolvenční rejstřík je souběžně vedený rejstřík s Evidencí úpadců. Vznikl 1. Ledna 2008 a shrnuje informace o insolvenčních řízeních. HTTPS://ISIR.JUSTICE.CZ/ISIR/COMMON/INDEX.DO,
- **ARES** – Administrativní registr ekonomických subjektů je provozován Ministerstvem financí a umožňuje vyhledávání všech ekonomických subjektů v ČR. Umožněno je to tím, že slučuje informace z mnoha zdrojových centrálních registrů. HTTP://WWWINFO.MFCR.CZ/ARES/,
- **a mnoho dalších** evidencí, vedených jednotlivými ministerstvy.

Vedle ministerstev vedou své agendy například i Česká národní banka, nebo Středisko cenných papírů. Centrální registry státní správy jsou relativně objektivními zdroji informací. Jsou dostupné ze zákona 106/1999 Sb., buď zdarma, nebo za úplatu. Nevýhodou se může jevit relativně pomalá doba aktualizace údajů. Rejstříky jsou dostupné přes Portál veřejné správy ČR na adrese HTTP://PORTAL.GOV.CZ.

Kromě registrů státní správy publikují informace komerční společnosti, které mají poskytování informací v předmětu činnosti. Mezi nejvýznamnější patří:

- **ČTK** – Česká tisková kancelář je veřejnoprávní instituce zřízená zákonem č. 517/1992 Sb. Poskytuje nezávislé zpravodajství HTTP://WWW.CTK.CZ,
- **ČIA** - Česká informační agentura se zaměřuje na vybrané skupiny oborů HTTP://WWW.CIANEWS.CZ,

- **ČEKIA** – Česká kapitálová informační agentura se zaměřuje na ekonomické informace o firmách HTTP://WWW.CEKIA.CZ,
- **Creditinfo Albertina** – Databáze firem Albertina HTTP://WWW.ALBERTINA.CZ,
- **HBI** – HBI je online databáze firem HTTP://WWW.HBI.CZ,
- **Economia** - Economia je největším vydavatelstvím ekonomických a odborných periodik v České republice HTTP://ECONOMIA.IHNED.CZ/,
- **Anopress IT** – Působí na trhu jako dodavatel profesionálního monitoringu médií a mediálních analýz HTTP://WWW.ANOPRESS.CZ,
- **Newton Media** – Newton Media se zabývá monitoringem tisku a analýzou médií HTTP://WWW.NEWTONMEDIA.CZ,
- **Reuters** – Reuters je mezinárodní zpravodajská agentura. HTTP://WWW.REUTERS.COM/,
- **Bloomberg** – Zpravodajská agentura poskytující ekonomické a finanční zpravodajství HTTP://WWW.BLOOMBERG.COM,
- **A další.**

Vytěžování zpravodajských portálů a rejstříků veřejné správy je v detektivní a zejména zpravodajské práci velmi využíváno. To, že někdo shromažďuje informace, třídí je a zpřístupňuje, značně šetří detektivovi čas. Policie a státní zpravodajské služby mají všechny státní evidence automaticky k dispozici. Soukromé detektivní agentury a komerční zpravodajské služby si však musí pomoci jinou cestou, a sice vytěžováním pouze dostupných rejstříků a otevřených zdrojů, především internetu.

4 Internet jako otevřený zdroj informací

Největší zdroj různých druhů informací je dnes samozřejmě celosvětová síť internet. Nicméně internet sám o sobě žádné informace neposkytuje, jedná se o obrovskou soustavu navzájem propojených počítačových sítí. Prostřednictvím internetu třetí strany pouze nabízí informace na serverech. Na internetu jsou dnes všechny volně dostupné rejstříky státní správy, na internetu publikují všechny tiskové agentury, na internetu se prostřednictvím blogů a diskuzí vyjadřují jednotlivci, internet je médium, které dnes nelze opomíjet.

4.1 Služby sítě internet

Informace na internetu jsou publikovány různými internetovými službami, které komunikují pomocí svých protokolů. Jak služby a protokoly internetu vznikaly, objevovaly se ke každé službě i specifické vyhledávací nástroje, které využívaly přesně výhody dané služby. Postupem času však došlo k integraci těchto nástrojů do webového prostoru a webových vyhledávačů.

Nejdůležitějšími službami internetu dle širokého povědomí jsou:

- **WWW** je soustava propojených hypertextových dokumentů. Jedná se o nejrozsáhlejší a nejpoužívanější internetovou službu. Touto službou je publikována většina dokumentů a na její vytěžování se zaměřuje detektivní a zpravodajská práce především. Komunikační protokoly jsou HTTP a zabezpečená verze HTTPS,
- **Gopher** je dnes již téměř nepoužívaný předchůdce služby WWW. Pro nedostupnost grafických rozhraní v době jeho vzniku byl textově orientovaný. Protokol se nazýval stejně, tedy gopher. Protokol Gopher již není v žádné nejnovější verzi prohlížeče podporován a touto službou již nejsou publikovány žádné nové informace,
- **E-mail**, neboli elektronická pošta, je způsob odesílání, doručování a přijímání zpráv přes elektronický komunikační systém. Spolu se službou WWW nejpoužívanější internetová služba. E-mail je ještě starší než samotný internet, neboť vznikl již na mainframových počítačích jako způsob komunikace mezi jeho uživateli. Protokolem pro odesílání elektronické pošty je protokol SMTP, protokoly pro stahování/přijímání pošty jsou POP a IMAP. Jako otevřený zdroj informací lze považovat snad jen mailové konference, jinak je e-mailová schránka většinou soukromá a veřejně nedostupná. Lze však využít znalost syntaxe emailové adresy konkrétního člověka k vyhledávání informací o něm na internetu, kde je e-mailová adresa často používána místo loginu, nebo jako doplňková informace k přezdívce, např. na diskusních fórech. Existuje spousta vyhledávačů emailových adres. Např. HTTP://MY.EMAIL.ADDRESS.IS/ HTTP://WWW.PIPL.COM/ a další,
- **FTP** služba, jmenující se stejně jako protokol, je služba zabezpečující přenos souborů po síti. Služba se hlavně používá ke správě účtů webových stránek a přístup k datům. Právě data na anonymních FTP serverech mohou dobře sloužit jako zdroj informací. Integrace přenosu souborů do protokolu HTTP pomocí rozšíření MIME ovšem dnes

odsouvá protokol FTP trochu do pozadí, když nabízí více možností a je jednodušší. Opět existuje řada vyhledávačů ftp serverů jako je např. HTTP://WWW.SEARCHFTPS.COM/, HTTP://WWW.FTPSEARCHENGINES.COM/ a další,

- **Připojení ke vzdálenému počítači** je služba, kterou lze realizovat pomocí různých protokolů. Telnet a SSH jsou dva nejpoužívanější protokoly terminálových přístupů na servery. Nezabezpečený telnet byl později nahrazen bezpečnějším SSH. Slouží k přímému připojení k příkazové řádce počítače a jeho ovládání. Připojení ke grafickému rozhraní slouží například univerzální program VNC a firmou Microsoft do jejich operačního systému implementovaný přístup ke vzdálené ploše. Jako zdroj informací sloužily především při připojování k různým knihovním systémům a databázím. Dnes jsou služby poskytované pomocí těchto protokolů nahrazované přístupem přes službu WWW, takže jejich vytěžování pozbývá smysl. Používají se hlavně v technické sféře, při vzdáleném řešení problémů s počítači, mohou však sloužit i k sledování počítače.
- **Instant messaging** je internetová služba umožňující online komunikaci mezi uživateli v reálném čase. Dá se pomocí ní sledovat, zda jsou ostatní uživatelé právě připojeni. Existuje celá řada IM klientů, které komunikují přes své protokoly, a většinou neumožňovaly vzájemné propojení. Až klienty podporující více protokolů umožnily spojení těchto IM sítí. S nástupem sociálních sítí nastal odliv uživatelů od IM klientů a chatování se přeneslo na webové stránky. Dnes již IM klienty umožňují vzájemné propojení a především propojení se sociálními sítěmi a opět se používají pro svou jednoduchost a to nejenom na počítačích, ale i mobilních zařízeních. Vytěžovat služby IM nebylo až na odposlouchávání komunikace možné. S nástupem sociálních sítí, které zpřístupňují historii chatování a při nízkém povědomí uživatelů o vlastní bezpečnosti, se naopak vytěžování chatů stává důležitou součástí detektivní a zpravodajské práce. Nejvyužívanější IM klienty jsou ICQ, SKYPE, AIM, XMPP, Windows Live Messenger a další,
- **VoIP** je služba umožňující přenos hlasu přes počítačovou síť. Protože předpokladem jejího správného fungování je velmi dobrá kvalita internetového připojení, vznikla později než ostatní internetové služby. Jako vytěžovatelný zdroj informací, bez odposlechnutí přenosu a následného zpracování, nefunguje. Nejvýznamnějším VoIP klientem je Skype.

Publikovat informace na internetu může dnes každý. Neexistuje žádná internetová cenzura. Nikomu není bráněno publikovat cokoli, tedy i informace nepravdivé, lživé, záměrně zkreslené apod. Jedinou obranou je s tím počítat a informace pečlivě ověřovat. Ověřit informace lze nejlépe ověřením věrohodnosti jejich autora. Autor nese zodpovědnost za obsah, který na internet vyslal. Často je ale těžké, ne-li nemožné, autora identifikovat.

Poskytování informací na internetu se nemusí nikomu oznamovat, nemusí je nikdo schvalovat a nikde není napsáno, jakým způsobem mají být informace poskytované a jak mají být strukturované. To všechno vede ke špatné utříděnosti a organizovanosti informací. Vzhledem k nenákladnosti publikování na internetu se na něj dostává obrovské množství informací, které ale nejsou dostatečně kvalitní. Vyhledávání relevantních a kvalitních informací se pak stává velmi složitým.

4.2 Vyhledávání informací na internetu

Přístupy k vyhledávání na internetu jsou v podstatě dva. První se snaží zpracovat a indexovat co možná největší množství informací za cenu opomíjení kvality. Tyto vyhledávače vlastně procházejí všechny webové stránky internetu a indexují je. Zpracování je plně automatizované a obsah není hodnocen. Jedním z prvních takovýchto vyhledávačů byla HTTP://ALTAVISTA.DIGITAL.COM. Druhý přístup se zaměřuje na obsah a kvalitu za cenu toho, že nedokáže zpracovat takové množství informací. Obsah stránek musí někdo nebo něco ohodnotit a zařadit do správné části. Vyhledávač katalogového typu je např. HTTP://WWW.YAHOO.COM. Vyhledávače katalogového typu jsou vhodné pro hledající, kteří zpočátku vlastně neví, co hledají a kategorie katalogu je navedou správným směrem [16].

Potenciál vyhledávačů lze maximalizovat naučením se jak fungují a jak je používat rychle a efektivně. Účelem je položit svůj dotaz správným způsobem tak, abychom neskončili zahlceni informacemi, nebo nezůstali bez relevantního výsledku.

Před vlastním vyhledáváním je důležité si určit téma co nejúplněji a co nejstručněji. Zapsat si přesně jaké informace uživatel hledá, proč, a jaké informace ho naopak nezajímají. Tímto postupem lze nejlépe objevit klíčová slova, která vyhledávání usnadní. S výjimkou vyhledávačů jako je AskJeeves.com, který sám aktivně pokládá dotazy ohledně vyhledávání, je nutné mít připravená klíčová slova. Vyhledávače doporučují jako maximum šest až osm klíčových slov, většinou podstatných jmen. Slovesům, spojkám, zájmenům, předložkám je

lépe se vyhnout a přídavná jména používat jen tehdy, pokud upřesňují hledaný cíl. Například **detektivní rozpracování** (google nalezl 14 400 výsledků) spíše než **rozpracování** (googlem bylo vyhledáno 887 000 výsledků) [26].

Pokud má uživatel vybrána klíčová slova, může vytvořit frázi. Fráze jsou kombinace dvou nebo více slov, které musí dokument obsahovat. Pokud jsou napsány v uvozovkách, musí být v dokumentu v přesném pořadí. Záleží na použitém vyhledávači a nejlepší je si před vlastním vyhledáváním pročíst help daného vyhledávače. Většina vyhledávačů nerozlišuje při vyhledávání velká a malá písmena. Nejspolehlivější je psát dotaz malými písmeny.

Vhodným nástrojem při vyhledávání je užití booleovských logických spojek. Umožní zúžit vyhledávání na rozumný počet výsledků a zároveň výběr zpřesnit a zvýšit tak šance na získání relevantních výsledků. Nejčastěji používané logické spojky jsou AND, OR, AND NOT. Většina vyhledávačů vyžaduje psát tyto spojky velkými písmeny[26].

AND znamená, že jsou hledány pouze dokumenty, které obsahují všechna slova, tedy zpřesňuje vyhledávání. Například vyhledávání **korupce AND státní AND správa** vrátí dokumenty, které obsahují všechna tři slova nebo fráze. Tedy dokumenty o korupci ve státní správě. Google vyhledal 568 000 výsledků oproti vyhledávání pojmu korupce, kdy vyhledal 3 080 000. Většina vyhledávačů spojku AND dává mezi slova automaticky, pokud nejsou v uvozovkách. Pokud jsou slova v uvozovkách, bere je vyhledávač jako přesnou frázi.

OR znamená, že jsou hledány pouze dokumenty, které obsahují alespoň jedno ze slov, je jedno jaké. Například vyhledávání **korupce OR „finanční kriminalita“** vrátí dokumenty, které obsahují alespoň jednu ze dvou frází, tedy daleko větší rozsah dokumentů. Google nalezl 3 120 000 dokumentů. Spojka OR se hodí při použití se dvěma synonymy.

AND NOT znamená, že jsou hledány dokumenty, které obsahují dané slovo, ale neobsahují slovo jiné. Například **korupce AND NOT finanční AND NOT kriminalita** vrátí dokumenty o korupci, kde se nevyskytují slova finanční a kriminalita. Google vyhledal 2 820 000 dokumentů. Většina vyhledávačů spojku AND NOT podporuje. Někdy je nazývána BUT NOT, nebo jenom NOT, nebo je (například Google) nahrazována minusovým znamínkem. Před použitím spojky AND NOT je dobré zkusit si vyhledávání bez ní, aby bylo vidět, jaké výsledky lze dostat a jestli je třeba ještě výběr dále omezovat.

Další používanou spojkou je ~. Znamená, že se hledá dané slovo nebo fráze, nebo jejich synonyma. Například na dotaz **~stealing** vyhledá Google 724 000 000 dokumentů obsahující i synonyma z databáze, jako třeba **theft**. Vyhledávání pouze výrazu **stealing** vrátilo 89 300 000 výsledků. Google má databázi synonym pochopitelně dokonalejší v anglickém prostředí.

+ před hledaným slovem vnutí vyhledávání slova nebo fráze tak, jak jsou napsané a vyloučí automatické použití synonym. Některá slova, například členy nebo spojky jsou vyhledávači automaticky ignorovány. Tímto způsobem mohou být vynucena.

* je operátor nahrazující libovolný počet znaků v hledaném slově nebo celé jedno slovo. Při vyhledávání Googlem nahrazuje pouze celé slovo. Tato funkce je užitečná, pokud si uživatel třeba nemůže vzpomenout na celé jméno hledané osoby, nebo pokud si nepamatuje přesné jméno obchodní společnosti atd. Vyhledané výsledky mu mohou pomoci si vzpomenout.

NEXT vyhledá dokumenty, v nichž jsou daná slova nebo fráze vedle sebe v textu.

NEAR vyhledá dokumenty v nichž jsou daná slova nebo fráze blízko sebe v textu.

Další operátory omezují vyhledávání ne obsahově, ale umístěním, jazykem, časem atd. Mohou se u různých vyhledávačů lišit názvem i použitím.

SITE omezuje vyhledávání na konkrétní webovou stránku

INTITLE vyhledává slovo nebo frázi v názvu dokumentu nebo stránky

INURL vyhledává v URL daného dokumentu.

LINK vrací stránky, které obsahují zadaný odkaz

LANGUAGE omezuje vyhledávání na dokumenty napsané ve zvoleném jazyce

FILETYPE vrací pouze dokumenty určitého formátu. Například **zpravodajství filetype:pdf** vyhledá pouze dokumenty o zpravodajství s příponou pdf

BEFORE vyhledává jen v dokumentech vytvořených nebo upravených před určitým datem

AFTER vyhledává jen v dokumentech vytvořených nebo upravených po určitém datu.

Většina vyhledávačů má na toto rozšířené vyhledávání vytvořen formulář, kde lze z rozbalovacích nabídek vybrat i různé další možnosti vyhledávání, například regionální. Nutno dodat, že se vyhledávače snaží situace předvídat a některé výše zmíněné možnosti uplatňují automaticky, například vynechání zbytečných slov jenom, když je to žádoucí, automatické použití synonym, odhadování překlepů a nabízení jejich korekce, automatické použití některých výše popsaných spojek atd. Bez vyhledávacích služeb by internet sice zůstal největším informačním zdrojem, ale v podstatě nepoužitelným, protože by se uživatelům nepodařilo tyto informace nalézt a získat.

4.3 Příklad využití vyhledávačů v detektivní práci

Ve světě na poli internetových vyhledávačů jasně vede Google, následován portálem Yahoo, MSN a dalšími. V Českém prostředí dlouhodobě pozici Googlu odolává Seznam, který je nejpoužívanějším vyhledávacím portálem, na druhém místě je Google, následují Centrum a Atlas. Je nutné říct, že Seznam používá vlastní technologii pouze k prohledávání českého internetu. K hledání ve světě využívá vyhledávání Googlu. Centrum používá některé technologie Googlu, některé vlastní [2].

Jedním z častých úkolů detektivních kanceláří v procesu detektivního rozpracování je zjišťování informací o zájmové osobě. Zkusíme tedy porovnat výsledky tří vyhledávačů v českém prostředí, při získávání informací o osobě Petr Kališ. Je nutno poznamenat, že se Petr Kališ snaží na internet informace o své osobě poskytovat co nejméně. Nejprve bylo vyzkoušeno vyhledávání jména a příjmení.

Seznam vrátil 7290 výsledků na vyhledávání jména **Petr Kališ**. Na prvních 20ti stranách se nacházela spousta informací o jmenovcích, ale žádná o hledané osobě. Na dalších stranách se již začala objevovat i jiná jména. Pokud se vyhledávání omezilo uvozovkami **„Petr Kališ“** , tedy na přesnou frázi, Seznam vrátil pouze 51 výsledků, z toho dva se týkaly zájmové osoby a sice pořadí na výsledkové listině Mistrovství ČR v dlouhém kouření dýmky a záznam Gymnázia Jiřího z Poděbrad, kde hledaná osoba studovala. Google vrátil 305 000 výsledků vyhledání jména. Již na šesté stránce se na portálu firmy-lidé.cz objevilo zájmové jméno v souvislosti se Společenstvím pro dům čp. 804 v Poděbradech. Informace zřejmě pocházely z obchodního rejstříku. Druhým výsledkem byl opět záznam Gymnázia o absolventech. Centrum počet výsledků neuvedlo, zmínku o osobě našlo Centrum taktéž na výsledkové listině Mistrovství.

Nutno říci, že bez dodatečných informací k jménu je hledání konkrétní osoby velmi těžké, až nemožné. Dají se ale alespoň natipovat některé konkrétní osoby a na ty se poté zaměřit. Pokud své vyhledávání zpřesníme o město pobytu, výsledky jsou následující. Na dotaz **Petr Kališ Poděbrady** nalezl Seznam 144 výsledků a hned první dva relevantní. Google nalezl 2750 výsledků a jako první tři všechny již zmíněné výskyty.

Centrum nalezlo taktéž tyto tři výsledky. Google a Centrum automaticky vyhledávaly i skloňované verze jména a vrátily výsledky i slov **Petra**, **Kališe** a **Kališová**. Seznam vyhledával pouze skloňovanou variantu křestního jména **Petra**.

Z takto vyhledaných informací již lze poměrně dobře identifikovat osobu. Z výsledků prohledávání Obchodního rejstříku lze zjistit přesná adresa, z výsledkové listiny Pipeclubu jedna ze zálib a z absolventského přehledu Gymnázia bývalí spolužáci. Právě na spolužáky je dobré se zaměřit, protože jsou to zřejmě přátelé zájmové osoby, což je výhodné zejména při vyhledávání na sociální síti Facebook. Jelikož Petr Kališ z Poděbrad profil na Facebooku pod svým jménem nemá, musíme se vydat jinou cestou.

Pokud se zaměříme na zjištěnou zálibu, tak na vyhledané výsledkové listině lze zjistit tým, za který Petr Kališ startoval a sice Dymka.net D. Dá se snadno zjistit, že Dymka.net je internetový dýmkařský klub se svou diskusní skupinou. Budeme tedy vyhledávat **„dymka.net D“** v uvozovkách, aby vyhledávače zachovaly celek jako přesnou frázi. Seznam nenalezl nic použitelného, dokonce ani po zpřesnění vyhledávání na **„dymka.net D“ site:dymka.net**. Z nějakého důvodu nebral slova v uvozovkách jako přesnou frázi, i když tuto funkci podporuje. Google při hledání **„dymka.net D“** hned na první stránce zobrazil příspěvek na diskuzním fóru, kde uživatel Fénix píše: *„v Poděbradech za Dymku.net D startovali twineryy a wolfrick*“, což jsou evidentně internetové přezdívky. Centrum tento výsledek vyhledalo až po upřesnění vyhledávání na konkrétní stránku fóra **„dymka.net D“ site:dymka.net.**

Teď už je tedy snadné zjistit, že Petr Kališ z Poděbrad má internetovou přezdívku twineryy. Vyhledáváme klíčová slova **twineryy –site:dymka.net**, abychom eliminovali prohledávání příspěvků na diskusním fóru, kterých bude jistě mnoho. Google opět vyhledal nejvíce relevantních výsledků, Centrum nevyhledalo přezdívku na aukčním portálu Aukro a Seznam vyhledal pouze některé výsledky, které přinesl Google. Na portálu Aukro se dá například zjistit, co uživatel twineryy v poslední době kupoval. Z detailního vytěžení diskusní skupiny se dá zjistit mnoho dalších důležitých údajů, které o sobě dotyčný napsal.

Z výsledků vyhledávání vyplývá, že Google i v českém prostředí vyhledal největší množství relevantních výsledků. Centrum (využívající technologií Googlu) přineslo podobné výsledky, jen menší množství nebo méně přesné, a Seznam ve výsledcích hodně zaostal.

Z tohoto příkladu je zřejmé, jak silnou zbraní jsou internetové vyhledávače a kolik informací lze na základě pár údajů zjistit o zájmové osobě, která je alespoň trochu aktivní na internetu. Nejvíce informací lze zjistit z přezdívky, která je ve většině případů používána opakovaně a to i u emailu, nebo jako přihlašovací login do různých systémů. I když zájmová osoba na internet žádné informace o sobě neposkytuje, neznamená to, že je tam neposkytne někdo jiný (třeba škola, kde dotyčná osoba studovala, nebo zaměstnavatel, kamarád apod.). Z tohoto důvodu je vhodné sledovat informace i o zjištěných přátelích, spolupracovnících, rodině a dalších osobách, které mohou se zájmovou osobou mít nějaký vztah.

Takovéto vyhledávání je ovšem velmi pracné a náročné na čas. Získané výsledky hledání je nutné pečlivě prostudovat a hledat text nebo údaje, které nás mohou zajímat a navést dále. Zautomatizování tohoto postupu je ovšem velmi složité a náročné a využívá například metod nového oboru zvaného Web mining. Těchto metod využívají například některé zpravodajské technologie umožňující pokročilé vyhledávání a analýzu informací.

4.4 Web mining jako nástroj získávání informací z webu

Rozvoj informační společnosti umožnil ukládání a třídění obrovského množství dat. Aby se z těchto dat daly získat potřebné informace a znalosti, musí se ovšem následně zpracovat. Je tedy nezbytné, aby byly vyvíjeny technologie umožňující automatické vyhledání a zpracování dat a informací. Proto došlo k velkému rozvoji vědeckého oboru dolování informací a znalostí z dat - dataminingu.

Datamining se zabývá především dolováním znalostí z databází, tedy z dat spíše relačního typu. V případě dolování znalostí z internetu se však musí metody a techniky dataminingu velmi přizpůsobit dynamickým a nestrukturovaným (v lepším případě semistrukturovaným) datům na webu. Vznikla tak vlastně plnohodnotná vědní disciplína zvaná Web mining. U všech metod můžeme vidět shodu v hlavních částech, a sice že nejprve dochází ke sběru dat, poté dochází k jejich předzpracování, následně k dolování a na závěr k vyhodnocení [23].

Obecnou definicí pojmu dolování na webu – web miningu je vlastně použití algoritmů pro dolování znalostí na webových stránkách. Protože se jedná o velmi širokou oblast, bylo třeba vnést do této problematiky pořádek a zavést ucelené názvosloví. O to se v roce 1997 pokusili

v díle Web mining: Information and pattern discovery on the world wide web pánové Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Web mining je podle nich rozdělen na: Web Content Mining (dolování v obsahu webu) a Web Usage Mining (dolování v datech o využívání webu mezi uživateli). Dnes je ještě přidávána třetí oblast a sice Web Structure Mining (dolování ve struktuře webu) [23].

Zdroji dat pro Web Mining jsou data o chování uživatele, automaticky ukládaná v logovacích souborech na serverech, proxy serverech, nebo u vlastního klienta, vlastní zobrazované obsahy stránek, jejich meta popis, webové odkazy a URL a jeho struktura [23].

Web mining je mocný nástroj umožňující například sledovat počty uživatelů, kteří opustili jednu webovou stránku a přešly na jinou a znázornit jejich „cesty“ graficky. Lze sledovat chování různých skupin uživatelů, například kolik procent uživatelů z oblasti státní správy se zajímalo o kterou stránku, nebo sledovat informace o preferenci produktu na konkrétních webových stránkách atd. Web mining využívají například pojišťovny k předcházení pojišťovacím podvodům, banky v odhadování bonity klientů, či cílení poplatků na konkrétní klienty.

Pro zpravodajskou a detektivní práci mají tyto pokročilé nástroje veliké využití. Automatické vyhodnocování informací z webu a jejich strukturování umožňuje zpracování většího množství adekvátních informací. Kvalitnější vyhledávání informací, jejich filtrace a strukturování umožňují rychlejší a kvalitnější přeměnu informací na znalosti.

Web Content mining

Web Content Mining, neboli v překladu „dolování z obsahu webu“, znamená získávání informací a znalostí z obsahu webových stránek. Znamená to zpracovávat informace textové, obrazové, zvukové, či video, uvádět je do vzájemných souvislostí a extrahovat z nich potřebné údaje. Nejvíce zpracovávanou složkou je pochopitelně část textová, proto je někdy Web Content Mining nazýván Web Text Miningem.

Data se pro potřeby extrakce informací mohou rozdělit na tři typy. Prvním typem dat jsou data strukturovaná, což jsou data, ke kterým je k dispozici jejich popis. Může se jednat například o XML dokumenty s přiloženým DTD nebo XML schématem. Druhým typem jsou semistrukturovaná data. Může se jednat například o Html dokumenty. Zpracování těchto dat je pak založeno na extračních vzorech založených na tzv. omezovačích, což mohou být například posloupnosti Html tagů. Třetím typem dat jsou data nestrukturovaná, neboli čistě

textová data. Zde neexistuje žádná struktura a extrakce informací je založena na technikách Text miningu, především zpracování přirozeného jazyka [12].

Klasické internetové vyhledávače, jak bylo popsáno výše, poskytují služby vyhledání a indexace webových dokumentů, avšak nedokáží s vyhledanými dokumenty dále pracovat, strukturovat je, filtrovat a získávat z nich potřebné znalosti. K tomu slouží inteligentní nástroje pro sbírání informací, jako jsou wrappery, weboví agenti, nebo rozšířené databázové přístupy a na nich pak aplikované techniky dataminingu [12].

Wrappery jsou programové rutiny, které automaticky provádějí extrakci dat z webu a jejich převod do strukturované podoby, umožňující jejich další zpracování. Wrappery nebo jejich sady bývají naprogramovány pro konkrétní webové stránky nebo strukturou podobné. Weboví agenti v sobě oproti tomu zahrnují systémy umělé inteligence, které jim umožňují vyhledat relevantní dokumenty podle okruhu zájmu, otevírat je, třídit a kategorizovat a následně jejich obsah hodnotit na základě samoučení, například z předvoleb uživatelů. Rozšířené databázové přístupy jsou zaměřeny na organizování polostrukturovaných dat z webu a jejich ukládání do strukturovaných souborů, které jsou dále zpracovávány v relačních databázích [12].

Web Structure Mining

Web structure mining, neboli „dolování ze struktury webu“, extrahuje informace z hypertextových odkazů mezi stránkami, které vlastně vytvářejí síť příbuzných stránek, neboli Webový graf. Webový graf se skládá z Webových stránek, uzlů a jednotlivých hyperlinků. Hyperlinky slouží buď pro přímou navigaci, nebo k propojení stránek se stejným tématem. Při zkoumání uzlů a propojení se využívá teorie grafů [13].

Dva hlavní algoritmy pro vyhledávání jsou HITS (Hypertext Induced Search Topic) a Page Rank. Algoritmus HITS vybere ze sítě stránek podgraf na základě dotazu uživatele a na něj aplikuje spojovací analýzu, aby nalezl autoritní stránky, neboli stránky, na které vede nejvíce odkazů a rozcestníky, které odkazují na velké množství stránek. Algoritmus HITS byl implementován například v programu ARC, jehož výsledky vytvoření adresáře příbuzných stránek byly porovnatelné s výsledky, kterých by dosáhl člověk. PageRank je vlastně číslo, které si vyhledávač Google přiřazuje každému URL a označuje jím věrohodnost, nebo důležitost té dané stránky. Struktura hypertextových odkazů zde slouží jako hodnocení dané stránky, což je systém podobný systému hodnocení vědeckých prací podle počtu citací. Ohodnocení závisí na množství a hodnocení stránek, které na danou stránku odkazují.

Hodnota PageRanku je mezi 0 a1. Zjednodušeně se dá říci, že stránka předává část svého PageRanku stránkám, na které sama odkazuje, aniž o svůj PageRank přichází. PageRank má každá jednotlivá stránka, nikoli celý Web dohromady a hodnota PageRanku nezávisí na hledaném slově [13].

Vzorec pro výpočet vypadá zhruba takto [13]:

$$R(a) = \sum_{u \in B_a} \frac{R(u)}{N_u}$$

Ba - množina všech stran, které odkazují na a

Nu - počet odkazů, které vedou z u.

Web structure mining využívají především internetové vyhledávače, při vyhledávání stránek s příbuznými tématy.

Web Usage Mining

Web Usage Mining, v překladu „dolování z uživatelského chování na webu“, získává informace ze záznamu chování uživatelů na webu. Vytváří modely chování uživatelů, snaží se odhalit závislosti a zákonitosti (např. kolik procent uživatelů, kteří navštívili jednu webovou stránku, přešlo na jinou, v kterou hodinu se tak dělo nejčastěji apod.) Využívá asociační pravidla a statistické metody a analyzuje automaticky generované log soubory, které ukládají webové servery [23].

Analýza je rozdělena do tří kategorií. První je předzpracování dat, kde dochází k očištění od irelevantních informací, které logy obsahují, druhou pak seskupení logů do sekcí podle odkazů na stránky a vyhledávání struktury, třetí kategorií je analýza nalezené struktury a interpretace výsledků. Jednou z technik Web Usage Miningu je metoda FSG – frequent generalized sessions. Metoda sbírá z webových logů obvyklé vzory chování uživatele a vytváří takzvaný uživatelský profil. Ten je pak použit k predikci chování podobných skupin uživatelů [23].

Sledování chování uživatelů internetu a predikce jejich chování budoucího je velmi cennou informační komoditou. Web usage mining může pomoci firmám při strategii křížového marketingu, k cíleným reklamním kampaním, při analýzách úspěchu/neúspěchu internetového obchodu (proč uživatelé opouštějí stránky a jaká byla cesta opuštění stránek). Pro

zpravodajské účely lze pomocí těchto technik získat velmi cenné informace, například z chování určitých zájmových skupin na internetu odhadnout jejich budoucí jednání. Tímto způsobem by se mohlo například předcházet teroristickým útokům, nebo odhadnout vývoj politické situace v určité zemi a předcházet nepokojům.

4.5 Google trends

Google trends je služba vyhledávacího portálu Google, která sleduje chování uživatelů internetu, konkrétně tedy jejich vyhledávací preference. Umí porovnávat četnosti vyhledávání jednotlivých topiků z hlediska časového i z hlediska geografického.

Google trends analyzuje počty hledání požadovaného topiku a porovnává ho s celkovým počtem vyhledávání za dané časové období. Z toho pak vypočítá a zobrazí graf nazvaný Search Volume Index graph znázorněný na obrázku 2. Tento graf ukazuje relativní počty hledání prvního topiku vzhledem k průměrnému hledání topiku za určité časové období. Průměrné hledání prvního topiku má Search Volume Index roven jedné. Relativní počty hledání dalších topiků jsou pak vztaženy k počtům hledání toho prvního. Měřítko grafu je pak automaticky upraveno, aby bylo vše dobře viditelné [11].

Pod tímto grafem se nachází další graf nazvaný News reference volume graph, který ukazuje, jak často se hledaný topik vyskytoval na Google News, což je zpravodajský portál Googlu. Graf je pouze ilustrativní a na ose Y nejsou vyneseny žádné konkrétní počty. Při velkém výskytu topiku na Google News, což je na grafu znázorněno špičkou, Google Trends tuto špičku označí a vypíše o jakou událost se jednalo a je možné ji porovnat v Search Volume Index grafu, jestli bylo zaznamenáno i zvýšené vyhledávání topiku [11].

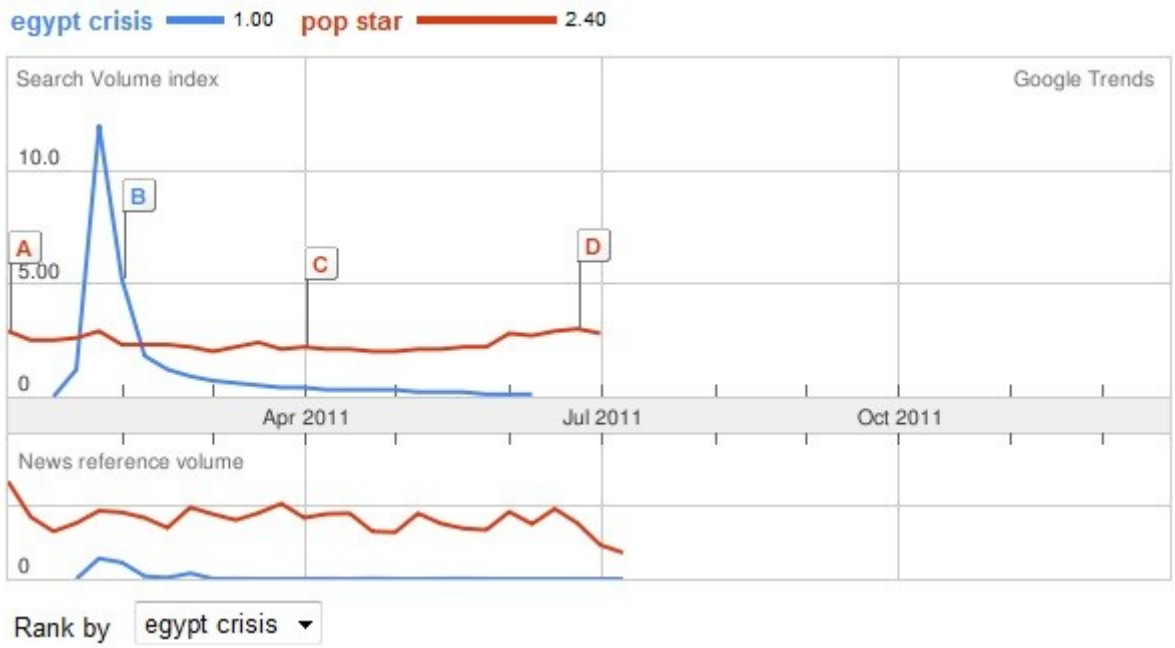
Na obrázku 2 dole se nachází geografické porovnání četnosti hledaných topiků, které lze seřadit podle konkrétního topiku. Ukazuje, v kterých světových lokalitách se hledaný topik vyhledával nejčastěji. Hledání lze zaměřit na konkrétní oblast světa, nebo i stát, pokud se v něm hledaný topik vyskytoval dostatečně často k vypočítání Search Volume Indexu.

Google trends umí také zobrazit průběh návštěvnosti zadaných webových stránek za časové období a ukázat, odkud návštěvníci těchto stránek pocházeli. Toto opět funguje pouze u větších stránek s velikou návštěvností, aby měl Google dostatek dat ke zpracování.

Obrázek 2 Prostředí Google Trends [11]

Na obrázku 3 je vidět porovnání dvou topiků a sice **pop star** a **Egypt crisis** v roce 2011. Vyhledávání topiku Egypt crisis výrazně předčilo koncem ledna a začátkem února roku 2011 vyhledávání topiku pop star, který je dlouhodobě stabilně vyhledávaný. Před 16. lednem 2011 a po 19. červnu 2011 jsou výsledky vyhledávání topiku Egypt crisis opět minimální a nenaplnily ani počet dostatečný k vypočítání Search Volume Indexu. Na grafu je vidět, jak po začátku Egyptské revoluce stoupal zájem světa o Egyptskou krizi, až převýšil zájem o každodenně vyhledávaný topik, jako je pop star. Při rezignaci Egyptského prezidenta Mubaraka začátkem února, zájem světa o Egyptskou krizi již upadal a po jeho rezignaci postupně zanikl úplně.

Na News reference volume grafu je vidět, jak reagovala média (tedy Google News) na situaci v Egyptě. Počet článků s tímto tématem nepřevýšil články o popových hvězdách, které jsou dlouhodobě v zorném poli médií. Je možné také vypozorovat stoupající trend ve vyhledávání topiku Egyptská krize, ještě před tím, než zareagovala média a začala o Egyptské krizi informovat. Čísla u popisu trendů obou topiků znamenají, že topik pop star je 2.4 krát vyhledávanější v roce 2011, než topik Egypt crisis.

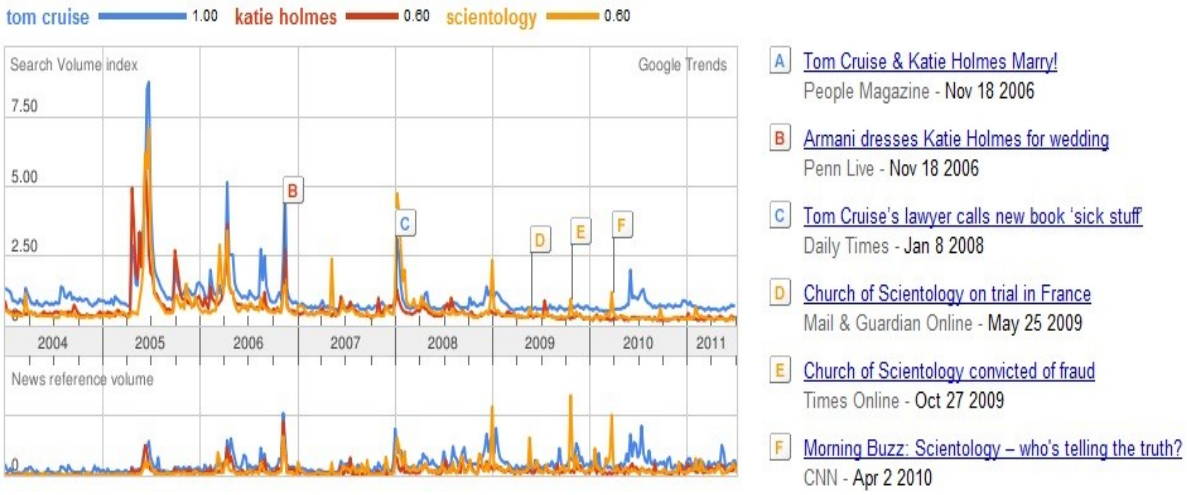


Obrázek 3 Srovnání topiků egypt crisis, pop star [11]

Na obrázku 4 je znázorněn příklad porovnávající vyhledávání jmen dvou hollywoodských hvězd **Toma Cruise** a **Katie Holmes** a topiku **scientology**, což je kontroverzní náboženství, které Tom Cruise vyznává.

Je vidět, že zvýšený výskyt těchto jmen v médiích se v podstatě kryje se zvýšeným počtem vyhledávání těchto topiků. Tedy lidé se o těchto dvou jménech dozvídají především z médií. Před rokem 2005, kdy požádal Tom Cruise Katie Holmes o ruku, spolu křivky těchto jmen nesouvisely a média se o Katie Holmes v podstatě nezmiňovala. Po tomto datu reagují křivky obou jmen ve stejném trendu. Je tedy vidět, že lidé od té doby vnímají tato dvě jména dohromady. Je také zřejmé, že Tom Cruise se těší dlouhodobě většímu zájmu (jeho jméno je 1.7 krát vyhledávanější), než jeho manželka a také se o něm více zmiňují média.

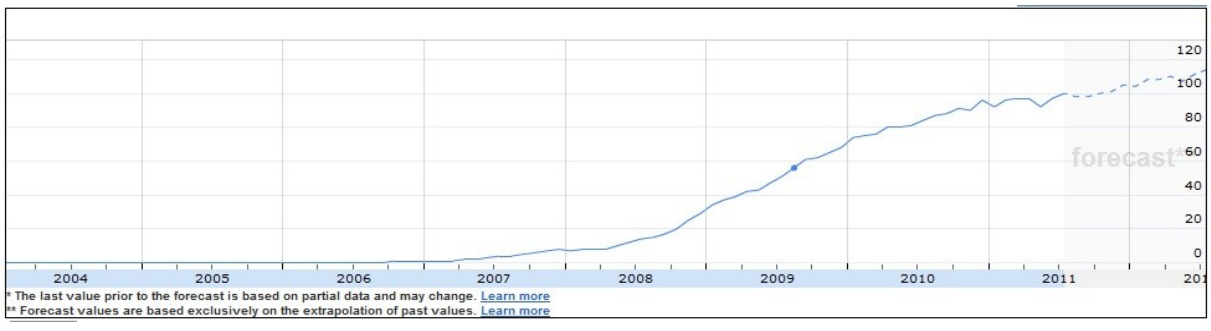
Křivka vyhledávání termu scientologie víceméně kopíruje křivku Toma Cruise, takže je vidět, že lidé mají jeho jméno spojeno se scientologickou církví. Ostatní výkyvy křivky termu scientology jsou pochopitelně dány jinými událostmi, kdy se o scientologické církvi mluvilo, nebo psalo v médiích a nebylo zmíněno jméno Toma Cruise.



Obrázek 4 Srovnání topiků Tom Cruise, Katie Holmes, scientology [11]

Tato jména jsem zvolil proto, že jsou dobře známá v anglicky mluvícím světě a těší se velkému zájmu novinářů a veřejnosti, právě i díky jejich kontroverzní víře. Proto o nich má Google velké množství záznamů. Na průběhu těchto křivek lze dobře ukázat souvislosti. Vypozorované potencionální souvislosti mezi topiky by se daly využít v jiných případech při detektivní a zpravodajské práci. Trendy v tom, co lidé vyhledávají lze využít ke krátkodobým předpovědím a modelům. Vysledovaná nálada lidí by mohla pomoci vládám k předvídání nepokojů, sledování zájmů určité skupiny lidí by mohlo předejít teroristickým a kybernetickým útokům apod.

Součástí Google Trends je i služba Google Insides for Search, která umožňuje další rozšířené funkce. Kromě funkcí zmíněných výše ještě dokáže z předchozího vývoje odhadnout, jak bude křivka topiku vypadat v budoucnu. Na obrázku 5 je vidět odhad vývoje vyhledávání topiku **facebook** do konce roku 2011.



Obrázek 5 Odhad vývoje vyhledávání topiku facebook [11]

Google Insides for Search využívá funkce Google trends a rozvíjí je dále. Další funkcí je grafické znázornění geografického porovnání četností vyhledávání topiků na přehledné mapě.

Obrázek 6 znázorňuje mapu světa se zvýrazněnými místy, kde byla četnost vyhledávání topiku **facebook** vysoká začátkem roku 2011. Čím tmavší barva, tím vyšší Search Index. Mapa se dá přiblížit v podstatě až na úroveň velkých měst, pokud má Google dostatek dat z dané oblasti. Mapa umožňuje sledovat vývoj četnosti vyhledávání topiku v různých regionech, za dané časové období. Lze tak vypozorovat, kde se o daném topiku začalo mluvit nejdříve a jak se informace šířily.

Obrázek 6 Mapa četností vyhledávání topiku facebook [11]

Google Insides for Search dále umožňuje zjistit, které topiky byly nejvyhledávanější v dané části světa, státu, regionu, či města v určitém časovém období, opět pokud má dostatek dat z dané oblasti. Mezi nejvyhledávanější termy patří obecně facebook, google a další lokální vyhledávače, dále pak počasí, názvy velkých měst atd. Google Trends funguje zatím pouze v angličtině a čínštině, ale další jazyky jsou plánované. Prohledávat nicméně umí všechny jazyky včetně češtiny, ale nejvěrohodnějších výsledků dosahuje samozřejmě v angličtině a ve velkých zemích s odpovídajícími počty hledajících uživatelů, používajících k vyhledávání Google.

Google je sice největším a nejvyužívanějším vyhledávačem na světě, nicméně každý region světa má svá specifika a lidé mohou preferovat k vyhledávání určitých témat jiné vyhledávače, jiná klíčová slova apod. Proto je třeba na výsledky Google Trends pohlížet s tímto vědomím. Nejpřesnějších výsledků dosáhne Google Trends pochopitelně v anglicky mluvících zemích s velkým počtem obyvatel, kde má vedoucí roli mezi internetovými vyhledávači.

5 Sociální sítě jako zdroj informací

Sociální síť je uměle vytvořená komunita lidí, která sdružuje společný zájem, přátelství, či práce, nebo jiný společný rys. Sociální sítě jsou bezesporu fenoménem dnešní doby. Každým dnem narůstá počet jejich uživatelů. Prvotním záměrem sociálních sítí je otevřené sdílení informací mezi lidmi. Právě otevřenost sociálních sítí z nich dělá ideální nástroj pro získávání informací o zájmových subjektech. Vytěžování sociálních sítí je novou érou pro zpravodajskou a detektivní práci, protože zde lze nalézt relativně levně a bezpečně obrovské množství informací o zájmových osobách, firmách, postojích veřejnosti k různým tématům, výrobkům a značkám a mnoho dalších informací.

První sociální komunita vznikla již v roce 1985 a je jí HTTP://WWW.WELL.COM . Několik let poté se přidaly další jako HTTP://WWW.TRIPOD.LYCOS.COM/ a HTTP://WWW.THEGLOBE.COM/ . Postupně vznikaly další sociální sítě jako HTTP://WWW.LINKEDIN.COM/ (odborně a profesně zaměřená), HTTP://WWW.MYSPACE.COM/ , HTTP://WWW.BEBO.COM/, HTTP://TWITTER.COM/ a konečně HTTP://WWW.FACEBOOK.COM/. Popularita sociálních sítí celosvětově rapidně vzrostla roku 2005 právě s nástupem Facebooku [9].

Facebook založil student Harvardské univerzity Mark Zuckerberg v únoru 2004 a prvotně vznikl pro účely studentů Harvardské univerzity. Nedlouho nato se rozšířil na ostatní univerzity ve Spojených státech a Kanadě a 26. srpna 2006 byl otevřen pro celý svět pro každého, kdo má platnou emailovou adresu a je mu více než 13 let. V lednu 2011 měl Facebook více než 600 milionů aktivních uživatelů [9].

Facebook má skvělé vyhlídky, ovšem i své stinné stránky. Velkým problémem se stává otázka ochrany soukromí. Ve standardním nastavení Facebooku zjistí o uživateli kdokoli cokoli. Ani přísným nastavením soukromí se však nelze tomuto ubránit. K informacím lze proniknout oklikou přes přátele, kteří nemají tak přísné nastavení soukromí. Uživatelé mohou být kýmkoli lustrováni. Žárlivý partner může potichu sledovat aktivity svého protějšku. Bytař si natipuje, kdy je nejlepší čas vykrást byt, jehož rodina jede na dovolenou. Dítě se může stát obětí kyberšikany, žena zase obětí stalkera. S osobními daty, která jsou na Facebook poskytnuta, může společnost kreativně nakládat, mohou být strojově zpracovávána například dataminingovými nástroji a cizí firmy mohou na uživatele cílit marketing. Účet na Facebooku navíc nešel donedávna bezezbytku smazat.

Twitter, který už registruje přes 200 milionů uživatelů, takovéhle problémy nemá, protože žádnou ochranu soukromí nenabízí a nenabízel. Od svého vzniku je plně veřejný a kdo na něm publikuje si je toho vědom. Poslední dvě velké revoluce, íránská a egyptská, informačně stály z velké části právě na Twitteru. Twitter je skvěle využitelný pro seriózní, vědecké a komerční účely. Z Twitteru se získávají odhady v různých oborech, od odhadu úspěšnosti filmu, přes krátkodobou předpověď počasí, až po analýzy chování celého trhu. Facebook díky své uzavřenosti nic takového neumožňuje. Právě kvůli konkurenceschopnosti na tomto poli, kde Facebooku unikají zisky, se snaží za každou cenu otevřít a tvrzení Zuckerberga *„Soukromí je přežitek dvacátého století“* to jenom dokládá. Je asi jenom otázkou času, kdy se Facebook stane plně veřejným a sociální síť vlastně splyne s internetem. Osobní data budou předávána dál vyhledávačům a cíleným službám [8].

S nástupem sociálních sítí roste počet trestných činů sociální sítě využívající. Sociální sítě vyvolávají klamný dojem přátelskosti a důvěryhodnosti a snižují obezřetnost uživatelů. Ti se tak stávají náchylnější k útokům. Může jít o klasické útoky hackerů a spammerů. Ale může jít i o závažnější věci, jako například pedofily a sexuální násilníky, hledající na síti své oběti. Již zmínění bytaři a další podvodníci, využívající lidskou naivitu. Tak jako se sociální sítě stávají vhodným nástrojem těmto živlům, mohly by se stát významným nástrojem pro detektivní a zpravodajské práce. V USA a zemích, kde jsou sociální sítě rozvinutější, jsou ve velké míře využívány vládními úřady k vyšetřování trestných činů. Tedy získávání důkazů, vyvracení alibi, určování polohy osob (pokud tuto informaci poskytují), objasňování motivů, odposlouchávání komunikace. Facebook jako takový se nebrání tyto informace poskytovat. V České republice zatím sociální sítě policií vytěžovány, bez přímého podnětu, nejsou. I když by to mnohdy bylo přínosnější, levnější a účelnější, než přímé odposlechy. Je to dáno i tím, že u nás sociální sítě teprve svůj boom prožívají a lidé jsou k nim, po zkušenostech ve světě, mnohdy skeptičtí.

Programy pro monitoring a prohledávání sociálních sítí.

Až do konce dvacátého století fungoval jednoduchý mediální model. Vydavatelé, coby autorita, a čtenáři, coby konzumenti informací. Od nástupu internetu a sociálních sítí však tento model přestává platit. Bez schvalovací autority na internetu každý může být vydavatelem i čtenářem. Informace se šíří a mění obrovskou rychlostí a není možné čekat, až se objeví v tištěné formě, nebo až bude natočena reportáž v televizi. Informace na internetu

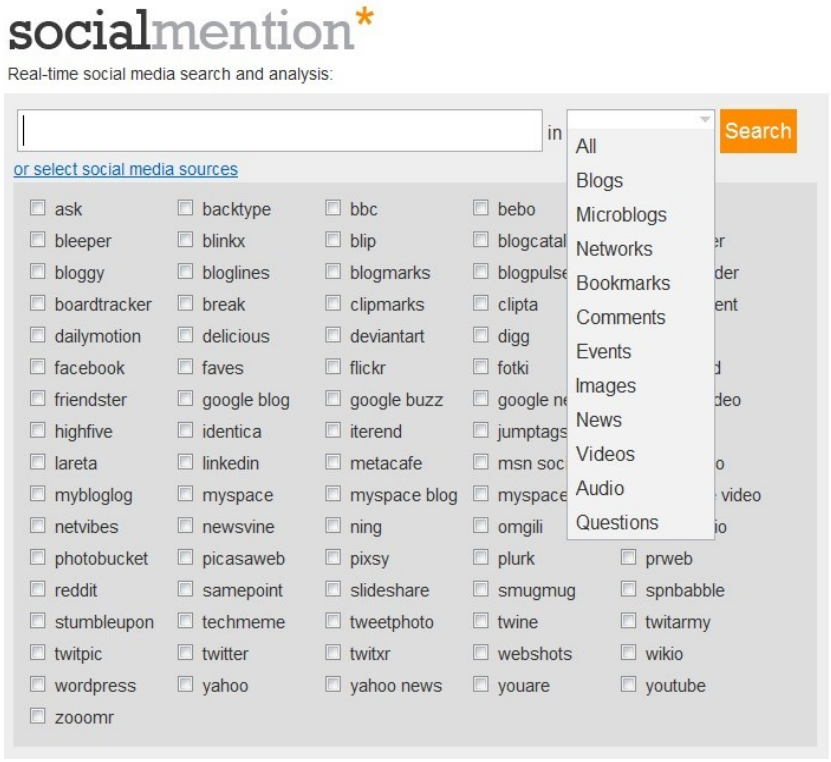
tečou po sociálních sítích, blozích, diskuzních fórech, prostě po celém webovém veřejném prostoru.

Při vyhledávání informací není možné se zaměřit pouze na jednu oblast, stejně tak není možné pokrýt celý webový prostor. Důležité je tedy najít v celé struktuře místa a uzly, kde se může mluvit o tématu, který hledám, osobě, která mě zajímá, a zde spustit tzv. monitoring. To znamená použít správné nástroje a zadat správná klíčová slova a fráze. Pak už jen čekat, až se někdo o nich zmíní. Monitorovací nástroje poté sbírají informace, shromažďují je a mohou je i samy analyzovat, nebo je předají na vstup dalším metodám, například web miningovým, nebo text miningovým.

Zde jsou uvedeny některé programy pro monitoring sociálních sítí, dle povědomí široké veřejnosti:

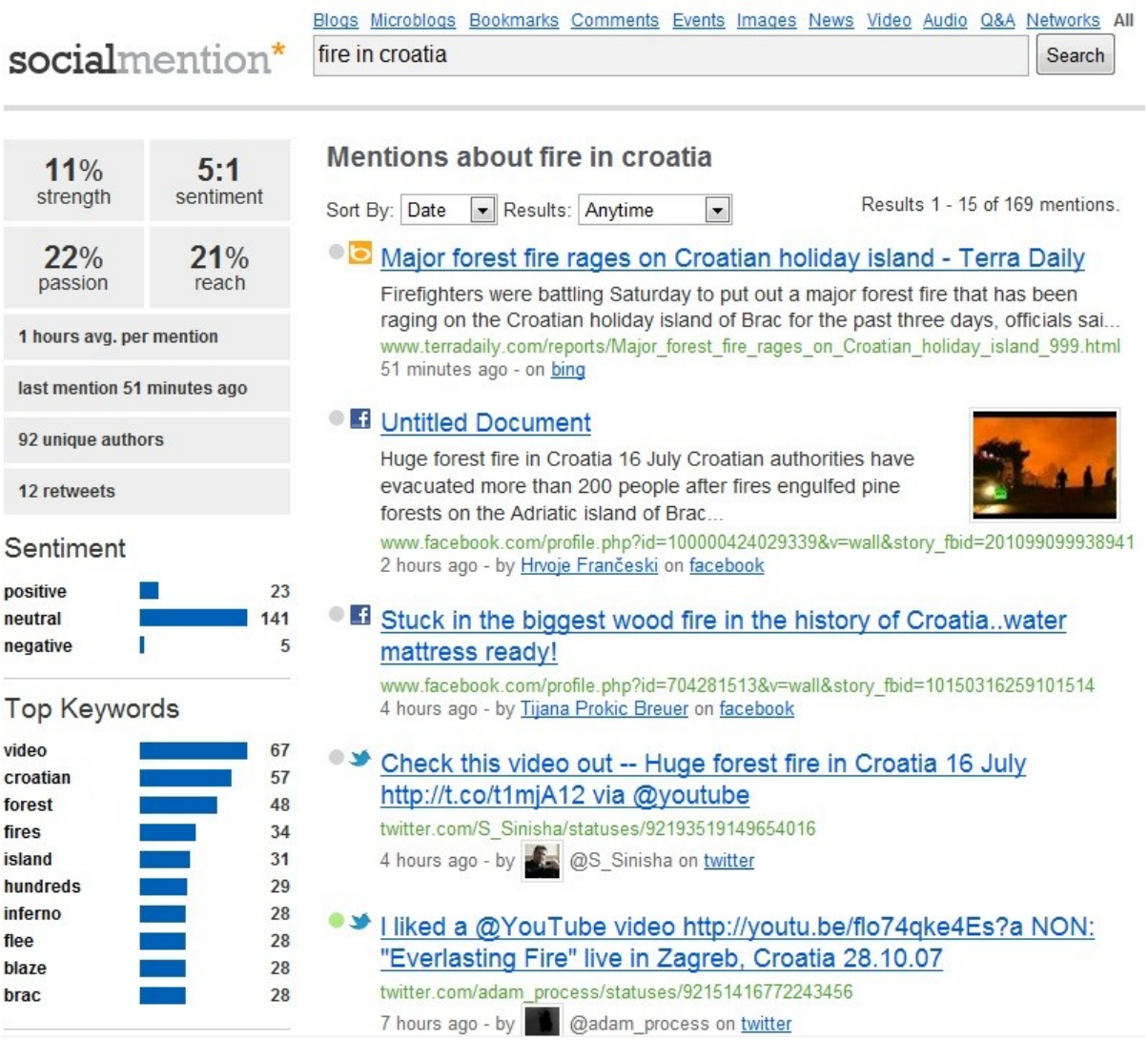
- **Tweetdeck, Seesmic, HootSuite** – jsou desktopové klienty pro sociální síť Twitter, které ale spolupracují i s jinými sociálními sítěmi. Na zvolených soc. sítích naslouchají a vyhledávají dané téma, umožňují i rovnou odpovídat, sledují nálady pro dané téma, porovnávají výsledky mezi sebou a mnoho dalšího. Sledovaných témat může být více, výsledky jsou zobrazeny v přehledných sloupcích,
- **Gogle alerts –** je již stará služba Googlu používající klasické vyhledávání, která ale umožňuje nechat vyhledávání takzvaně „běžet“ a informuje o nových nálezech emailem. S nalezenými údaji již ale neumí dále pracovat, analyzovat je, ani kategorizovat,
- **Google social search** – se souhrnně nazývají funkce Googlu Blogs, Realtime a Discussions. Jedná se vlastně o sociální vyhledávání. První funkce vyhledává v obsahu blogů, druhá v aktuálních zprávách, především na Twitteru, třetí prohledává diskuzní skupiny. Protože Google vyhledává především podle významnosti, kterou sám vyhodnocuje, tak české prostředí s pár miliony uživateli pro něj není příliš relevantní. Výsledky sociálního vyhledávání v českém prostředí tedy nebudou plnohodnotné. Služba Google Realtime byla 4. července 2011 dočasně zastavena a bude zakomponována do velikého sociálního projektu Google+,
- **Google Sparks** – za zmínku slouží nová chystaná služba Googlu, umožňující stanovit si své zájmové oblasti a nechat Google, aby prohledával internet a automaticky přinášel uživateli aktuální požadované zprávy,

- **Socialmention** – je platforma umožňující prohledávání velkého množství sociálních sítí, zpravodajských serverů, blogů, diskusních fór a dalších zdrojů, jak lze vidět na obrázku 7,



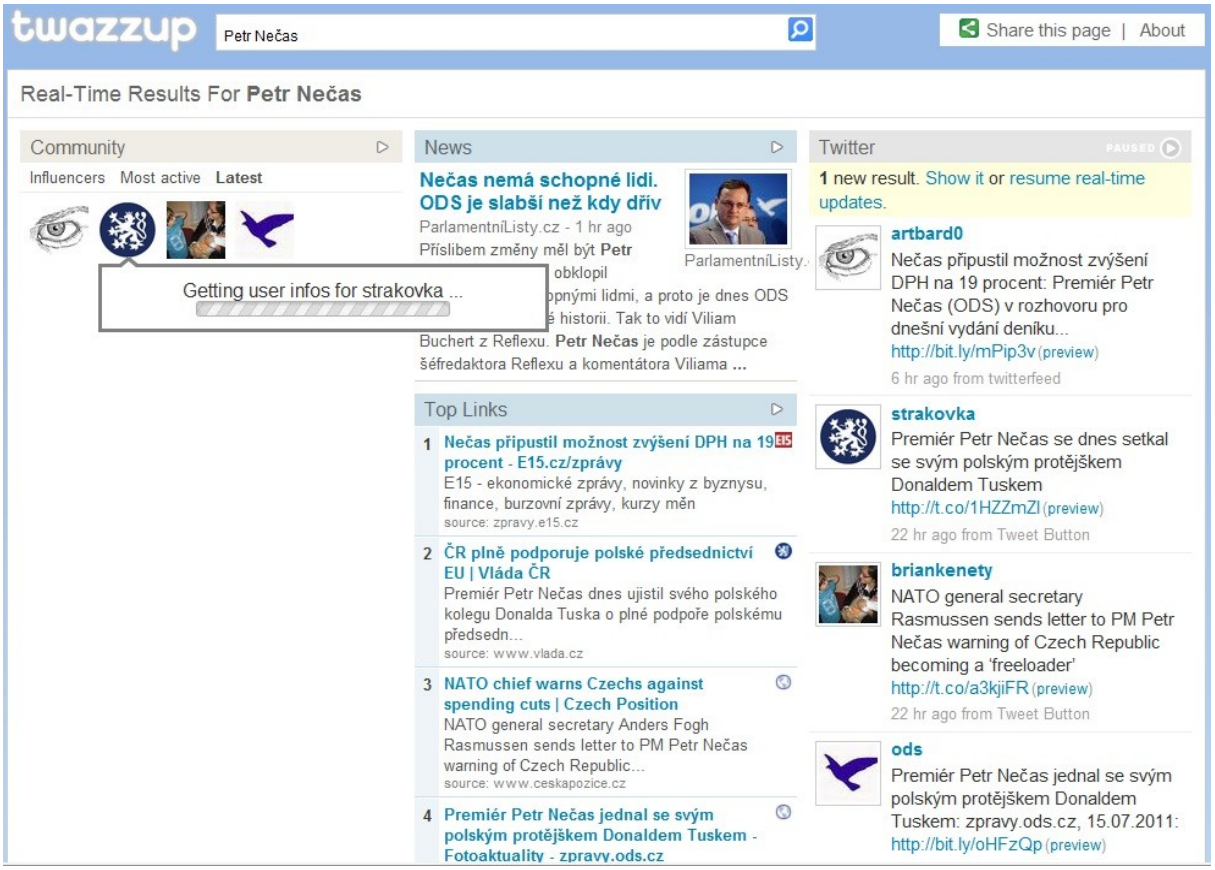
Obrázek 7 Prostředí vyhledávače socialmention [18]

Vyhledává podle klíčových slov a umožňuje provádět jednoduché analýzy, jako odlišit pozitivnost, negativnost reakcí na hledané téma, zobrazit jak moc se o daném tématu mluví, na jakých zdrojích je nejvíce zmínek o tématu a další drobné analýzy. V Českém prostředí prakticky nefunguje, celosvětově se však jedná o velmi silný nástroj. Obrázek 8 ukazuje prostředí vyhledávače socialmention při vyhledávání klíčového slova **fire in croatia**, což je relativně aktuální téma o požáru na ostrově Brač, o čemž svědčí i celkem velká síla topiku 11%, počítaná jako poměr zmínek o topiku ke všem zmiňovaným topikům za posledních 24 hodin. Zobrazeny jsou příspěvky z Facebooku a Twitteru, zprávy ze zpravodajských portálů, videa a obrázky této přírodní katastrofy.



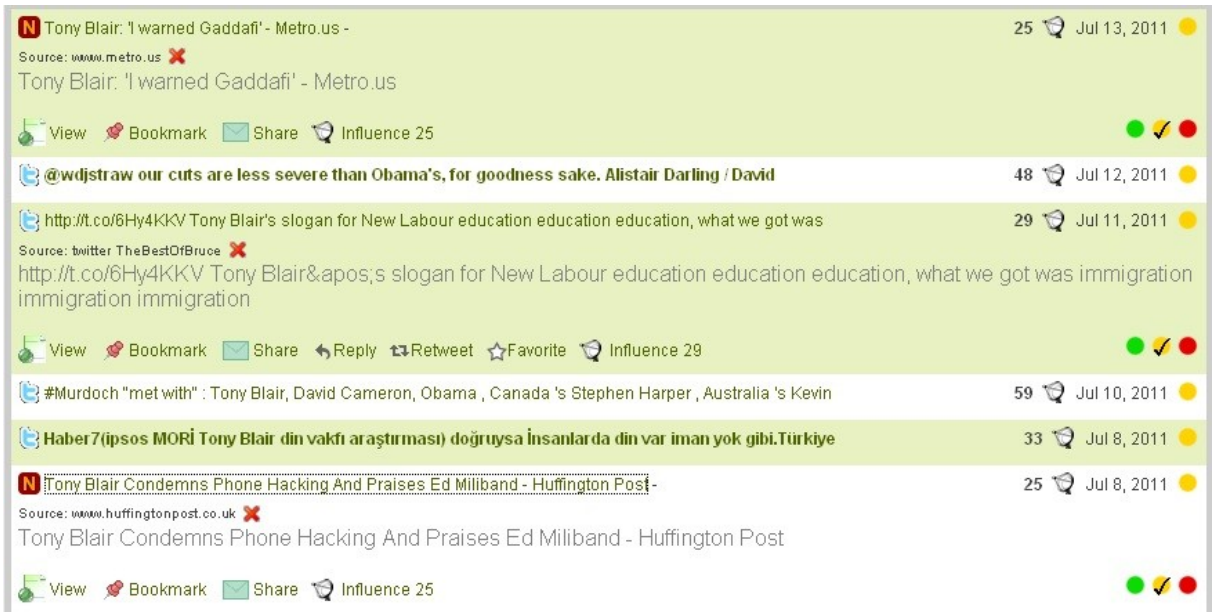
Obrázek 8 Výsledek vyhledávání topiku fire in croatia [18]

- **Twazzup** – je webový monitor, který zjišťuje zprávy a nové informace v reálném čase. Stačí zadat požadované klíčové slovo a engine prohledává zpravodajské portály a novinky na Twitteru. Vyhledávání běží neustále a na stránce se objevují nové informace, tak jak se objevují v médiích a na Twitteru. Služba je zdarma a při zkušebním vyhledání jména českého premiéra Petra Nečase, což je vidět na obrázku 9, nalezla celkem odpovídající výsledky a s českým prostředím neměla služba problém. Vyhledávání běží, dokud se stránka nezavře.



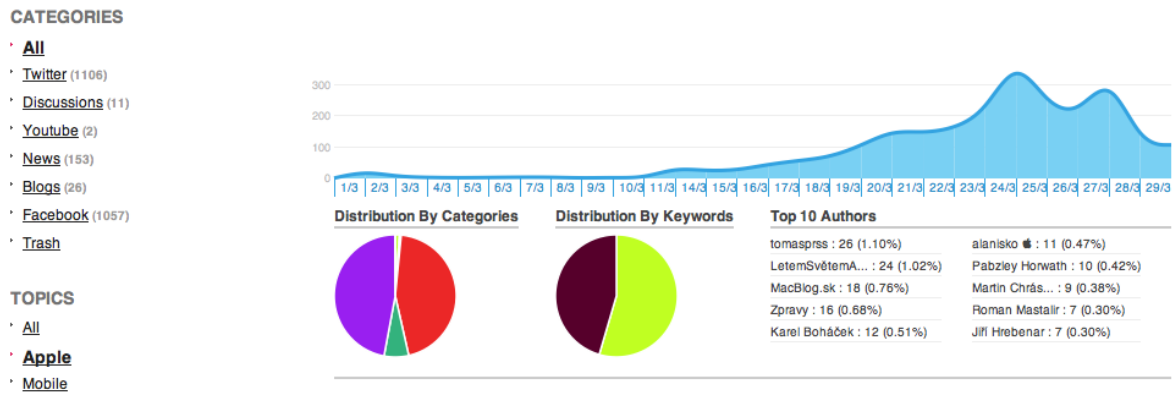
Obrázek 9 Prostředí vyhledávače twazzup [22]

- **Trackur** – je další z řady sociálních vyhledávačů, umí prohledávat sociální sítě Twitter, Facebook, Delicious a další, umí vyhledávat obrázky na obrázkových databázích, videa na youtube a dalších portálech, umí prohledávat blogy a zpravodajské servery. Umí nalezená data i analyzovat a graficky znázorňovat, ale pouze v placené verzi. Při dotazu na jméno českého premiéra ovšem nic nenašel a to ani po upřesnění vyhledávání na Českou republiku. Na obrázku 10 je vidět prostředí trackuru zobrazující výsledky vyhledávání jména bývalého britského premiéra Tonyho Blaira.



Obrázek 10 Prostředí vyhledávače trackur [21]

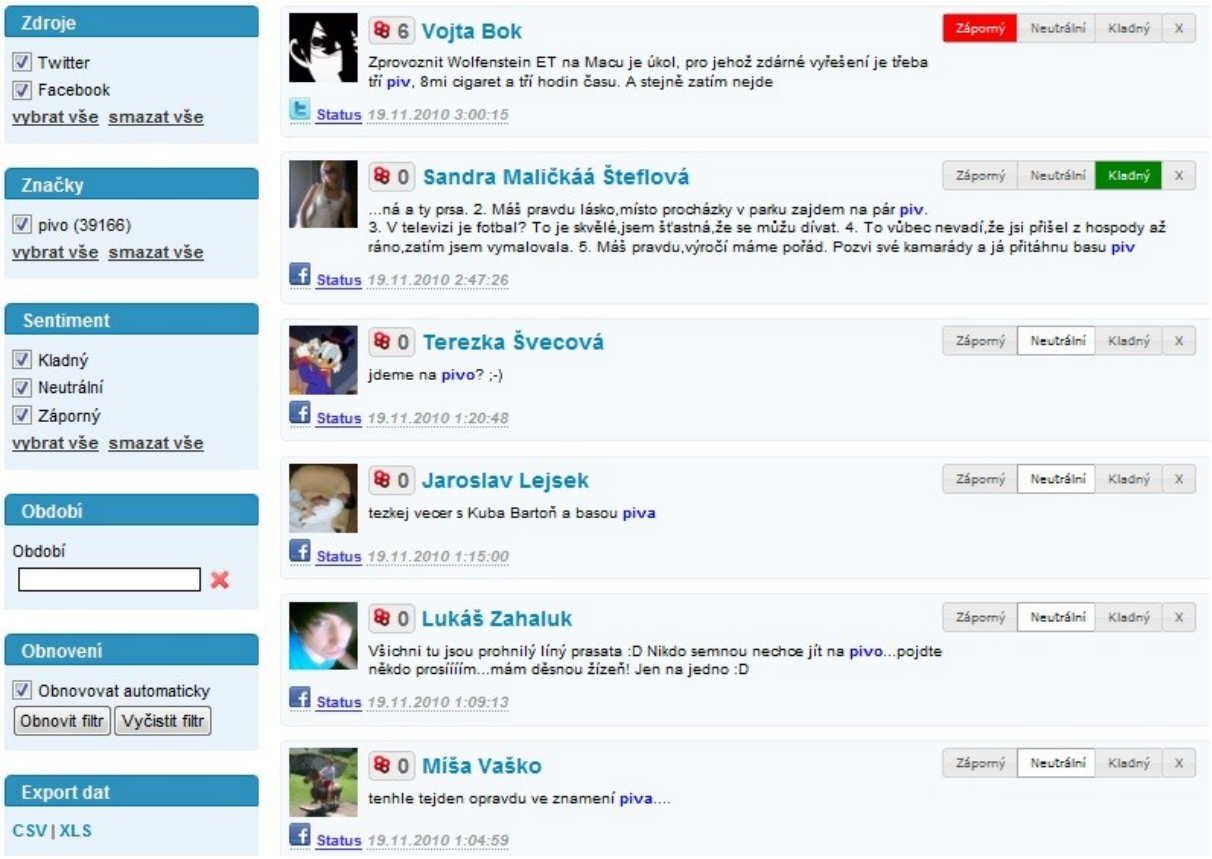
- **Ataxo Social Insider** – Ataxo je firma specializující se na marketing a reklamu. Vyvíjí jeden ze dvou placených nástrojů pro monitoring a analýzu orientovaných přímo na české prostředí. Umí prohledávat a monitorovat v Česku a Slovensku nejvyužívanější sociální sítě, důležité blogy a diskuzní skupiny. Provádí analýzy výsledků a jejich grafickou vizualizaci v reálném čase. Umí informace exportovat pro další využití a vytvářet každodenní reporty svých výsledků. Na obrázku 11 je znázorněna jedna z analýz -graf vývoje množství příspěvků na téma **apple** za určité časové období.



Obrázek 11 Graf vývoje množství příspěvků v prostředí Social Insider [2]

- **Newton Social Media Monitoring** – Newton Media je firma přímo se zabývající monitoringem médií a tisku. Kromě klasických zdrojů, vyvíjí i druhý placený nástroj, orientovaný na české, slovenské a polské prostředí – Social Media Monitoring. Nabízí

podobné funkce a možnosti jako Ataxo Social Insider. Na obrázku 12 je vidět v prostředí programu výstup vyhledávání klíčového slova **pivo.**



Obrázek 12 Prostředí Newton Social Media Monitoringu [15]

- **Vyhledávače integrované přímo v prostředí konkrétní sociální sítě**
- **a další.**

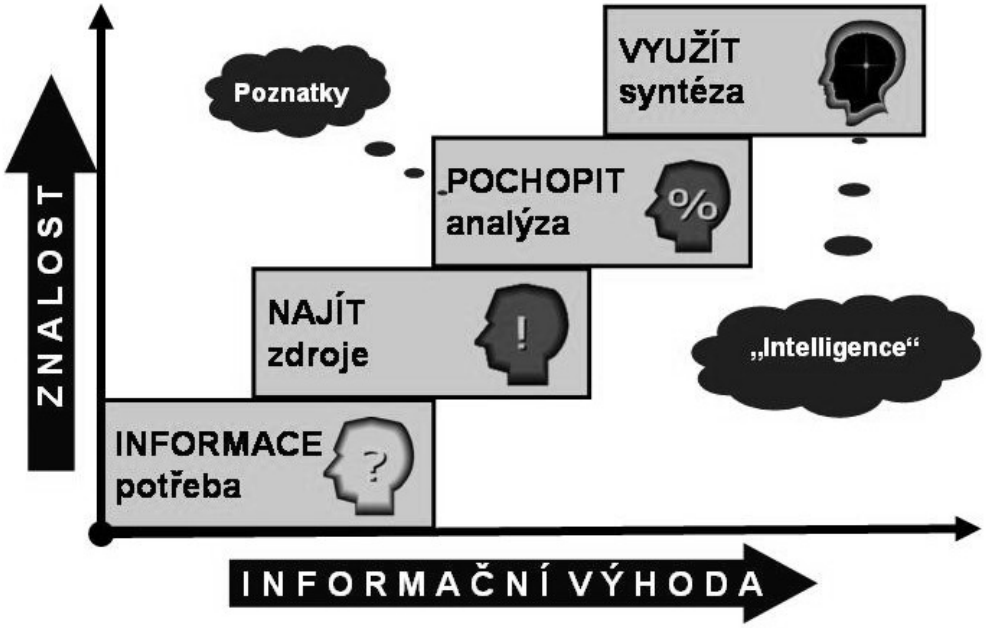
Srovnání vyhledávačů a monitorovacích portálů v prostředí sociálních sítí zatím nelze příliš realizovat, protože každý z nich přináší velmi odlišné výsledky. Nejlepších výsledků se dosáhne při jejich kombinaci. V českém prostředí fungoval pouze monitor Twazzup a samozřejmě i dva placené vyhledávače zaměřené na sociální sítě a orientované přímo na český trh - Social Insider a Social Media Monitoring. V celosvětovém měřítku jsou velmi silným nástrojem vyhledávač Socialmention, nabízející i drobné analýzy výsledků a samozřejmě služby vyhledávacího portálu Google.

6 Zpravodajství

Činnost, při které dochází k cílenému a specifickými postupy prováděnému sběru, dopravě a vyhodnocování takových informací, které jsou jinými cestami a v potřebném čase

nezjistitelné, nazýváme zpravodajskou činností. Právě zpravodajství jako takové nejvíce využívá vytěžování otevřených zdrojů. Zpravodajská činnost funguje na státní bázi, provozovaná státními zpravodajskými službami, ale zhruba od poloviny 20. století se stále více přesouvá do nestátní sféry a vzniká nestátní, komerční, či konkurenční zpravodajství a zpravodajský servis. Nestátní zpravodajství je realizováno prostřednictvím soukromé detektivní činnosti jako jedné z forem detektivní činnosti, zvané detektivní zpravodajství [4].

Zpravodajství je vlastně 4 fázový informační proces, jehož fáze jsou znázorněny na obrázku 13. V první fázi je třeba identifikovat informační potřebu, tedy co je předmětem hledání, v druhé fázi dochází k vyhledávání, sběru a třídění informací, ve třetí fázi dochází k interpretaci informací, jejich vyhodnocování, uvádění do souvislostí, tvorbě hypotéz a závěrů a v závěrečné fázi jsou výsledky předchozích tří fází převáděny do formy vhodné pro zadavatele a předání výsledků zadavateli k jeho využití. Vyhledávání informací pomocí zpravodajských technologií zajišťuje zpravodajská operativa. Pochopení informací, jejich třídění, analýze a interpretaci se věnuje zpravodajská taktika. Sledování znalostí ve vzájemných souvislostech je předmětem zpravodajské strategie [4].



Obrázek 13 Zpravodajský informační proces [4]

Jak již bylo řečeno, pro detektivní a zpravodajskou činnost je charakteristická práce s informacemi a informačními zdroji. Informačních zdrojů je veliké množství a každý vyžaduje odlišný přístup a postup při vytěžování. Každý informační zdroj má také určitou úroveň spolehlivosti.

V tabulce 1 je znázorněno využití prostředků na získání informací z různých typů informačních zdrojů z pohledu detektivní práce a zpravodajství. Je zřejmé, že 95 % informací lze získat z otevřených zdrojů a z toho 80 % za použití běžně dostupných prostředků [24].

Tabulka 1 Využití prostředků na získání informací [24]

		Běžně dostupné prostředky	**Prostředky dostupné organizaci**	**Nelegální prostředky**
Uzavřené informační zdroje 5%		Zakázané	Zakázané	Špionáž
Otevřené informační zdroje 95%	**Dostupné 15%**	Zpravodajství	Zpravodajství	Nebezpečné
	Publikované 80%	Zpravodajství	Drahé	Nebezpečné

Jen samotný sběr informací je nedostatečný, je nutné nalezené informace ověřit a stanovit jejich spolehlivost. Ve zpravodajské oblasti se využívá kodifikace 4X4, na základě které se každé informaci přiřazuje kód složený z písmene a číslice, podle následujících tabulek 2,3 [24].

Tabulka 2 Kódy používané pro ohodnocení zdroje [24]

A	Nejsou žádné pochyby o věrohodnosti, pravdivosti a kvalifikovanosti zdroje NEBO zdroj byl ve všech předchozích případech spolehlivý.
B	Zdroj byl ve většině předchozích případů spolehlivý.
C	Zdroj byl ve většině předchozích případů nespolehlivý.
D	Dosud neověřený zdroj NEBO jsou pochyby o věrohodnosti, pravdivosti a kvalifikovanosti zdroje.

Tabulka 3 Kódy používané pro ohodnocení informace [24]

1	Informace je bez výhrad známá jako pravdivá.
2	Informace je známá osobně zdroji, ale ne osobně tomu, kdo ji pořídil.
3	Informace není osobně známá zdroji, ale je potvrzena jinou již získanou informací.
4	Informace není osobně známá zdroji a v dané chvíli nemůže být nijak potvrzena.

Stanoví se tím vlastně jakýsi rating informace a informačního zdroje. Hodnotné jsou pouze informace s vysokým ratingem, což mohou být paradoxně informace s nejnižšími náklady na pořízení. Naopak informace drahé na pořízení mohou být z hlediska informační hodnoty bezcenné. Důležitější než vyhledání a ohodnocení informací je ale jejich analýza a využití.

Právě kvalita využití informací odděluje zpravodajství od prostého monitoringu informací. Monitoring pouze distribuuje sebrané informace zadavateli. Zpravodajským produktem je naproti tomu analytikem interpretovaný význam informací. Pro snadnější analýzy informací jsou používány různé vizualizační prostředky, pomocí kterých se dají sledovat vazby mezi jednotlivými informacemi [24].

Vedle vyhledání, třídění a analýzy informací je velmi důležitá také distribuce závěrů. Tedy přetvoření znalostí do formy vhodné, pochopitelné a využitelné pro zadavatele a doručení znalostí zadavateli v potřebném čase. Obecně platí zásada doručit výsledky raději nedokonale a včas, než dokonale a pozdě na to, aby mohly být využity ke konkrétnímu rozhodnutí [24].

Zpravodajské technologie

Dnešní informační systémy podniku jsou z hlediska zpravodajství omezené svou strukturou databáze a předem danými typy dat, které databáze umí zpracovat. Umožňují pouze předem definované analýzy a pohledy na data. Při potřebě zpracování nového typu dat je potřeba měnit strukturu databáze. Pro zpravodajství jsou ale velmi důležité nejen informace ze strukturovaných zdrojů firmy, ale především z nestrukturovaných dokumentů, což mohou být různé zápisy z porad, emailové komunikace, smlouvy atd. a hlavně z otevřených zdrojů, kde se jedná o tiskové zprávy, výstupy z různých rejstříků a informace z internetu obecně. Takováto dynamicky se měnící data podnikový informační systém zpracovat neumí. Je

potřeba nad podnikový informační systém aplikovat nadstavbu – systém pro podporu konkurenčního zpravodajství, který umí kromě interních strukturovaných dat podniku pracovat i s otevřenými zdroji v podniku, a hlavně mimo něj. Tento systém není prakticky nikdy dokončen a mění se tak, jak se mění data, která zpracovává [24].

Většina zpravodajských informačních systémů stojí v současnosti na dvou technologiích [24]:

- **Fulltextová technologie** – například technologie firmy Verity inc., která byla vytvořena na základě požadavku americké CIA.
- **Technologie vizuální analýzy** – například technologie firmy i2 inc., vycházející z metodiky ANCAPA používané americkou FBI.

Technologie jsou pro zpravodajskou práci velmi důležité, ale jsou vždy pouze nástrojem. Nejdůležitější pro zpravodajství bude vždy člověk – analytik efektivně využívající metodiky práce se zpravodajskými technologiemi.

Fulltextová technologie Verity

Technologie Verity vznikla na základě zadání amerických tajných služeb, a tedy již od svého vzniku byla vyvíjena speciálně pro zpravodajské účely vládní agenturou Advanced Decission Systems. Technologie Verity umožňuje stejně jako ostatní fulltextové technologie vyhledávat dokumenty podle zadaných klíčových slov. Rozdíly mezi jednotlivými technologiemi se pak projevují v dalších pokročilých funkcích, jako jsou [24]:

- formulace složitějších dotazů, kdy dochází ke kombinaci více klíčových slov,
- řazení dokumentů podle relevance obsahu vůči zadanému dotazu,
- shlukování dokumentů podle různých vlastností, například tematicky.

Technologie Verity se díky vysoké kvalitě a stále probíhajícímu vývoji prosadila i v soukromé sféře a je dnes v podstatě nejlepší technologií v oboru. Z hlediska využití pro zpravodajství jsou nejdůležitějšími následující vlastnosti technologie Verity [24]:

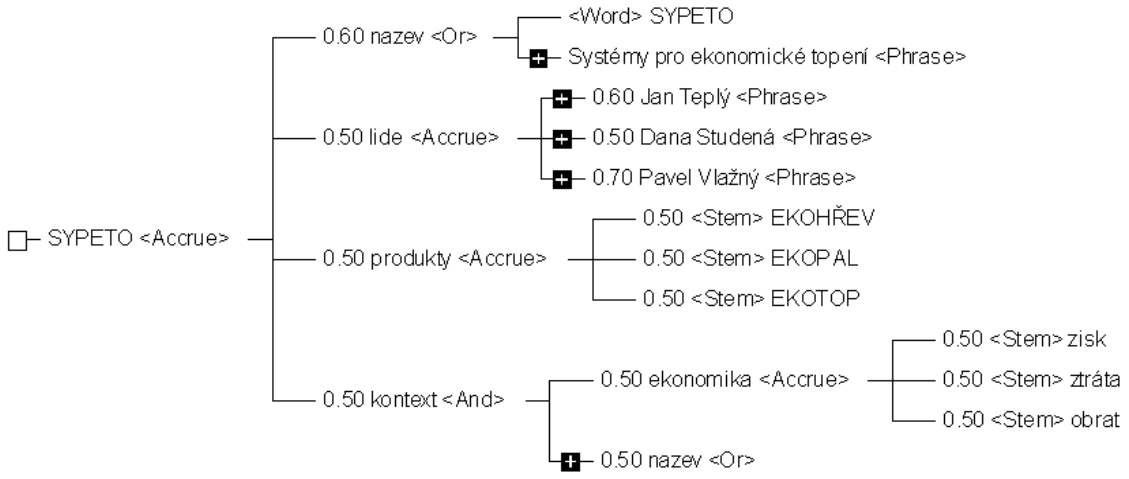
- **concept retrieval**, neboli pojmové vyhledávání pomocí strukturovaných dotazů,
- **relevance rating**, čili hodnocení důležitosti vyhledaných dokumentů vzhledem k dotazu,

- **clustering, summarization**, což znamená statistické vyhodnocování obsahu dokumentů**.**

Pojmové vyhledávání

Při vyhledávání informací o nějakém tématu je potřeba mít dobře rozmyšlena klíčová slova, tak jak bylo již zmíněno v části o vyhledávačích. Pokud je klíčových slov mnoho a postupně se rozvíjejí s postupem hledání, tvoří vlastně jakousi myšlenkovou mapu (soubor klíčových slov) postupu hledání. Technologie Verity funguje na podobném principu. Přiděluje skupinám klíčových slov určité subjektivní ohodnocení a při prohledávání dokumentu identifikuje tato klíčová slova a jejich kombinace a podle toho určuje, jak je dokument relevantní vůči dotazu [24].

Strukturovaný dotaz může vypadat například takto (obrázek 14) [24]:



Obrázek 14 Strukturovaný dotaz na firmu SYPETO; zdroj [24]

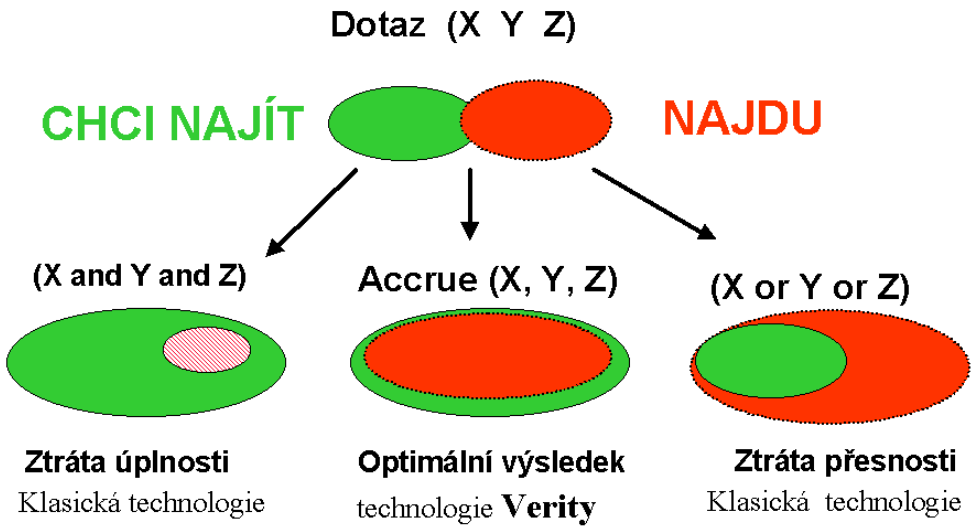
Jedná se o dotaz na firmu SYPETO a její ekonomické výsledky. Takovémuto strukturovanému dotazu se říká TOPIC a z obrázku 14 je zřejmé, že je tvořen [24]:

- **klíčovými slovy**, neboli listy,
- **váhami,** označujícími míru významnosti výskytu slov při hodnocení dokumentu,
- **listovými operátory**, určujícími jak budou slova z dotazu porovnávána se slovy v dokumentu,
- **logickými operátory**, určujícími způsob hodnocení výskytu skupin slov v dokumentu,
- **strukturou.**

Jak již bylo zmíněno, při stavbě TOPICu jsou důležitá klíčová slova. Podle složitosti TOPICu jich mohou být až stovky, někdy i tisíce. Výsledek vyhledání je právě tak kvalitní, jak kvalitní je TOPIC a klíčová slova v něm obsažená. TOPICy lze sdružovat do znalostních bází pro další využití, kde musí být uveden popis TOPICu a co lze jeho pomocí vyhledat. Tyto báze pak umožňují sdílení znalostí mezi uživateli a zpřístupňují tyto pokročilé metody i méně zkušeným uživatelům [24].

Určování důležitosti dokumentů

Určování důležitosti dokumentu vzhledem k dotazu je další vlastností fulltextové technologie Verity. Jak vyhledaný dokument vyhovuje dotazu, nebo složitému TOPICu určují především operátory, které jsou použity ke spojení klíčových slov. Vlastnosti jednotlivých booleovských operátorů byly zmíněny v části práce o vyhledávačích. Vyhledávání pomocí klasických booleovských operátorů vede k dilematu mezi přesností a úplností, jak je znázorněno na obrázku 15.



Obrázek 15 Dilema přesnosti a úplnosti [24]

Hledání pomocí operátoru AND sice najde důležité dokumenty, ale mnoho potenciálně důležitých dokumentů bude ignorováno, dochází tedy ke ztrátě úplnosti. Při hledání pomocí operátoru OR dostaneme mnoho potenciálně důležitých dokumentů, ale nevíme, které jsou nejdůležitější, dochází tedy ke ztrátě přesnosti. Technologie Verity toto dilema řeší operátorem ACCRUE, zapisovaným jako čárka mezi klíčovými slovy. Operátor ACCRUE funguje na principu "čím více různých klíčových slov je nalezeno, tím je dokument důležitější". Vyhledané dokumenty jsou pak řazeny podle důležitosti. Na prvním místě jsou dokumenty obsahující všechny, nebo nejvíce klíčových slov, na dalších pak dokumenty

obsahující méně a méně klíčových slov. Důležitost dokumentu je pak vyhodnocována podle váhy klíčových slov a jejich četnosti [24].

Nutno poznamenat, že podobné řešení dilematu mezi přesností a úplností dnes automaticky nabízejí i internetové vyhledávače. Podobnou technologii používá i Google, který řadí vyhledané dokumenty stejným způsobem.

Shlukování dokumentů podle obsahu

Pokud jsou při hledání známy pouze základní informace a uživatel vlastně ještě neví, co hledá, další klíčová slova se objevují až v nalezených dokumentech a tím se vlastně hledání zpřesňuje. Podobným způsobem bylo postupováno i v příkladu vyhledávání informací o zájmové osobě v části této práce. Tento postup je ale zdlouhavý a je nutné vždy nalezené dokumenty prostudovat a další klíčová slova nalézt.

Technologie Verity tento proces výrazně ulehčuje shlukováním dokumentů podle společného obsahu. Již při indexaci obsahu dokumentu je prováděna statistická analýza. Pro každý dokument je vybrána řada klíčových slov, které s jistou pravděpodobností vystihují jeho obsah. Tato skupina slov se nazývá významový vektor. Dokumenty jsou pak shlukovány podle počtu shodných slov v jejich významových vektorech [24].

Problém shlukování dokumentů je také řešen metodami text miningu. Umožňuje snadnější orientaci ve vyhledaných dokumentech, rychlejší hledání nových klíčových slov a také objevování skrytých souvislostí mezi vyhledanými dokumenty.

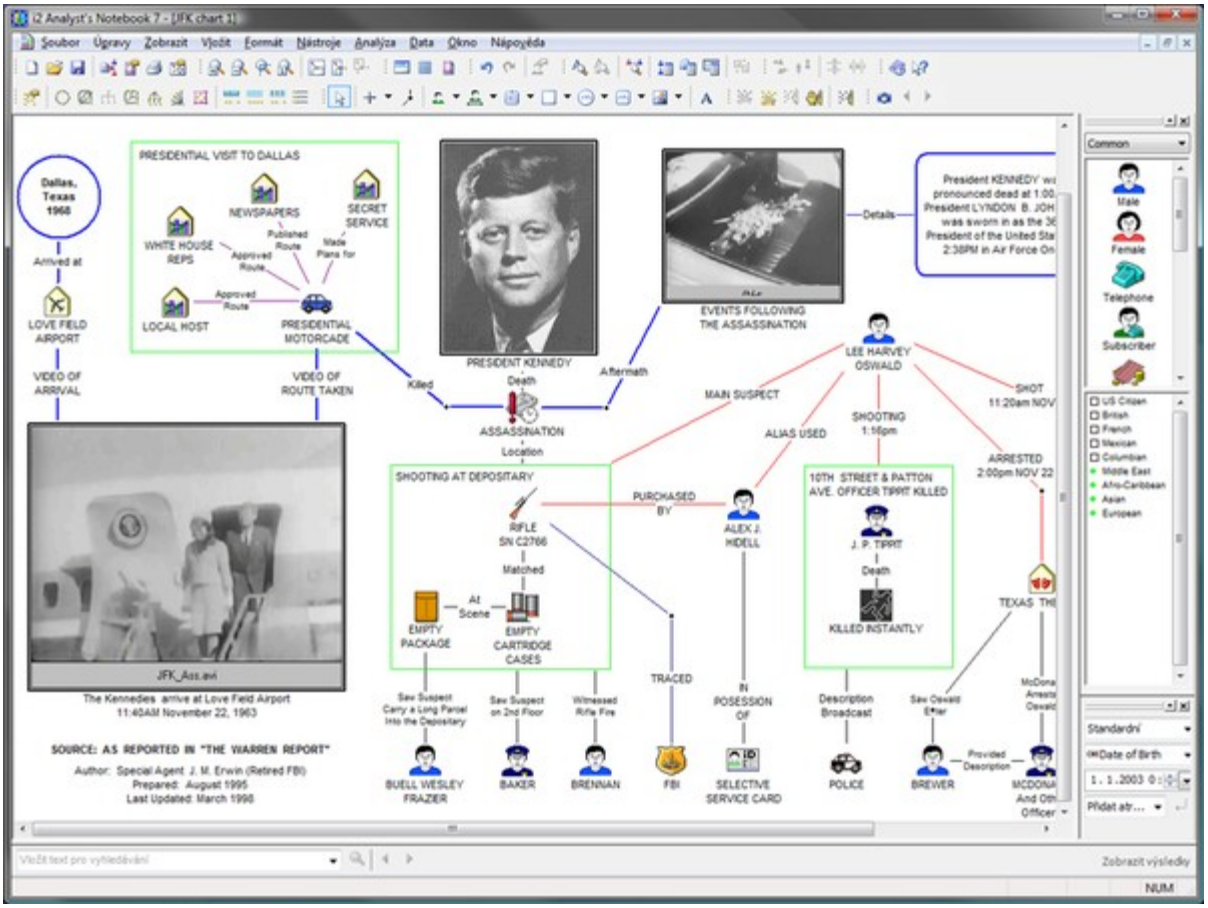
Technologie vizuální analýzy firmy i2 inc.

Lidský mozek umí pracovat různě s různými typy prezentace informací. Zatímco informace v textové podobě dokáže zpracovat pouze tak, jak jdou za sebou, grafické informace umí zpracovávat paralelně, tedy více informací najednou. Pochopení textu a vztahů a znalostí v něm obsažených zabere tedy mnohem více času a soustředění, než pochopení například vizuálního schématu. K usnadnění práce analytika a rychlejšímu odhalení těchto vztahů jsou vyvíjeny metody vizuální analýzy, mezi které patří asi nejznámější a nejpoužívanější technologie vizuální analýzy firmy i2.

Základní produkt firmy i2 pro vizuální analýzu informací se nazývá Analyst Notebook a vychází z dnes již standardizované metody ANACAPA, používané americkou FBI. Pro vizuální prezentaci a analýzu informací jsou používána různá schémata (například vztahová,

časová, prostorová, tabulková a další), vytvářená podle daných pravidel. Aplikace pravidel je důležitá, aby bylo možné schémata navzájem propojovat a analyzovat tak informace v širších souvislostech [24].

Na obrázku 16 je v prostředí Analyst Notebook zobrazeno vztahové schéma atentátu na prezidenta Kennedyho. Každý objekt, neboli entita a každá spojnice, neboli link, zobrazené na tomto schématu, mohou obsahovat svoje atributy a další textové informace, rozvíjející grafickou část. Sdružují tak v sobě přehlednost grafického rozhraní s podrobností rozhraní textového.



Obrázek 16 Vztahové schéma v prostředí Analyst Notebook [20]

Komplexním nástrojem, který integruje produkty firmy i2 do jednotného analytického prostředí, je **Analyst Workstation**. Analyst Workstation obsahuje již zmíněnou aplikaci Analyst notebook, rozšířenou o databázovou aplikaci iBase s dataminingovými nástroji, dále nástroj k propojení s externími databázemi iBridge a nástroj k propojení s Geografickými informačními systémy GIS Interface [20].

Produkty firmy i2 se staly doslova standardem v oblasti nástrojů pro vizuální analýzu. Používá je více než 80 procent všech státních zpravodajských služeb po celém světě. 25 členských států NATO využívá jejich produkty k účelům obrany státu. Tyto nástroje každodenně slouží k odhalování teroristických útoků, podvodů a dalších trestných činů.

Závěr

V této práci jsem se pokusil uvést do problematiky vytěžování databází a dokumentů s ohledem na využití informací v detektivní a zpravodajské práci. V první části je představena jedna z forem detektivní činnosti – detektivní rozpracování, při které jsou zejména využívány metody detektivního vyhodnocování dokumentů a vedení a vytěžování databází, které jsou dále rozpracovány. Práce se zaměřuje na vytěžování elektronické formy dokumentů, způsob jejich indexace a krátce se zmiňuje o nově se rozvíjejících metodách extrakce informací z nestrukturovaných dokumentů – metodách text miningu.

V další části se práce snaží vysvětlit základní typologii informačních zdrojů a nastínit výhody a nevýhody některých typů, a jejich případné využití v detektivní práci. Dále jsou v práci zmíněny významné informační zdroje v ČR, zejména rejstříky státní správy a databáze informačních agentur a organizací, zabývajících se monitoringem médií.

Ve čtvrté části se práce zaměřuje na vyhledávání informací na internetu, jakožto největším otevřeném zdroji informací. Zmíněny jsou základní služby a protokoly internetu a v krátkosti možnosti jejich vytěžování. Dále se práce věnuje vyhledávání pomocí vyhledávacích služeb, ukazuje možnosti používání speciálních operátorů a vše rozvíjí na konkrétním příkladu vyhledávání informací o zájmové osobě. Pokročilé metody dolování informací a znalostí z webových stránek jsou zmíněny v části práce o web miningu, jakožto oboru nabízejícímu nové možnosti pro detektivní a zpravodajskou práci. Ukázky konkrétního chování uživatelů internetu jsou rozpracovány v prostředí služby Google Trends.

V páté části práce popisuje fenomén dnešní doby – sociální sítě, zejména nejrozšířenější z nich, sociální síť Facebook. Sociální sítě již dnes slouží jako obrovský zdroj informací využitelný v detektivní práci. Práce ukazuje funkce a možnosti některých vyhledávací a monitorovacích nástrojů použitelných v prostředí sociálních sítí.

V poslední části se práce věnuje zpravodajství, jakožto detektivní činnosti nejvíce využívající vytěžování otevřených zdrojů. Definuje zpravodajství jako čtyřfázový proces, krátce se

zabývá metodou hodnocení informací a informačních zdrojů. Na závěr jsou v práci rozebrány dvě nejdůležitější zpravodajské technologie pro vyhledávání a analýzu informací – fulltextová technologie Verity a technologie vizuální analýzy firmy i2.

Seznam použité literatury

[1] Akademická knihovna JU. *AKADEMICKÁ KNIHOVNA JIHOČESKÉ UNIVERZITY* [online]. 2011 [cit. 2011-05-31]. Druhy dokumentů a kde je hledat. Dostupné z WWW: <www.lib.jcu.cz/docs/ak-ikurz-druhy-dokumentu.pps>.

[2] *Ataxo Czech : internetová reklama a SEO optimalizace pro vyhledávače* [online]. c2010 [cit. 2011-07-15]. Dostupné z WWW: <http://www.ataxo.cz/>.

[3] BRABEC, F., et al. *Soukromé detektivní služby* . [s.l.] : Eurounion, 1995. 263 s. ISBN 80-85858-16-9.

[4] BRABEC, F. *Technologie detektivní činnosti*. Zlín : Univerzita Tomáše Bati ve Zlíně, 2009. ISBN 978-807318-780-4.

[5] BLAŽEK J. Srovnání automatické a intelektuální indexace. *Inflow: information journal* [online]. 2008, roč. 1, č. 4 [cit. 2011-08-01]. Dostupný z WWW: <http://www.inflow.cz/srovnani-automaticke-intelektualni-indexace>. ISSN 1802-9736.

[6] BLAŽEK J. Systémy vyhledávání obrazových informací. Část II.: Problematika vyhledávání. *Inflow: information journal* [online]. 2010, roč. 3, č. 3 [cit. 2011-08-01]. Dostupný z WWW: <http://www.inflow.cz/systemy-vyhledavani-obrazovych-informaci-cast-i-problematika-vyhledavani>. ISSN 1802-9736 .

[7] BÍLKOVÁ, E. Vznik a vývoj automatické indexace. *Knihovnictví a informační věda informuje* [online]. 06.10.2004, [cit. 2011-05-15]. Dostupný z WWW: <http://www.phil.muni.cz/kivi/clanky.php?cl=44>. ISSN 1214-7265.

[8] BOČEK, J. Ochrana soukromí na internetu. *Computer*. 2011, 3, s. 64-65. ISSN 1210-8790.

[9] BUGNER, M. Sociální síť, dobrý sluha, zlý pán. *Internet pro všechny* [online]. 10. 11. 2009, [cit. 2011-06-10]. Dostupný z WWW: <http://www.internetprovsechny.cz/socialni-sit-dobry-sluha-zly-pan/>. ISSN 1801-1160.

[10] *GATEWAY : Multimedia PC Level 3 Spec* [online]. c2011 [cit. 2011-05-24]. Support Documents. Dostupné z WWW: <http://support.gateway.com/s/SOUND/U00647/U0064725.shtml>.

[11] *Google* [online]. c2011 [cit. 2011-07-11]. Dostupné z WWW: <http://www.google.com/>.

[12] HALAMÍČEK, M. *Transformace různých webových zdrojů do RDF*. Praha, 2008. 86 s. Diplomová práce. ČVUT. Dostupné z WWW: <https://dip.felk.cvut.cz/browse/pdfcache/halamm1_2008dipl.pdf>.

[13] CHAKRABARTI, S. *Mining the web : Discovering Knowledge from Hypertext Data*. USA : Morgan Kaufmann, 2003. 345 s. ISBN 1-55860-754-4.

[14] KUČEROVÁ, H. *Zpracování informací a znalostí* [online]. 2009 [cit. 2011-05-31]. Dostupné z WWW: <http://web.sks.cz/users/ku/ZIZ/ziz.htm>.

[15] *Newton Media : Media analysis and monitoring of media* [online]. c2010 [cit. 2011-07-17]. Dostupné z WWW: <http://www.newtonmedia.eu/>.

[16] PETERKA, J. *Earchiv.cz* [online]. 1996 [cit. 2011-05-31]. Internet jako informační zdroj. Dostupné z WWW: <http://www.earchiv.cz/arevue/a803r200.php3>.

[17] SKLENÁK, V.*Vyhledávání informací v prostředí webu – mírný pokrok v mezích zákona*[online]. 2005 [cit.2011-6-2]. Dostupný z WWW: < http://www.akvs.cz/akp-2005/10-sklenak.pdf>.

[18] *Socialmention : Real-time social media search and analysis* [online]. c2010 [cit. 2011-07-14]. Dostupné z WWW: <http://www.socialmention.com/>.

[19] *Text mining a jeho možnosti* (aplikace). *FI MUNI* [online]. [cit. 2011-6-5]. Dostupný z WWW: <http://www.fi.muni.cz/usr/jkucera/pv109/2003p/xsedlac5.htm>.

[20] *Tovek, spol. s r.o.* [online]. c2011 [cit. 2011-07-22]. Dostupné z WWW: <http://www.tovek.cz>.

[21] *Tracur : Social Media Monitoring Tools Made Easy* [online]. c2011 [cit. 2011-07-12]. Dostupné z WWW: <http://www.trackur.com/>.

[22] *Twazzup : Realtime News* [online]. c2010 [cit. 2011-07-11]. Dostupné z WWW: <http://www.twazzup.com/>.

[23] VEČEŘA, M.*Dobývání znalostí z webu-DP*[online].2007 [cit. 2011-06-10]. Dostupný z WWW: < http://is.muni.cz/th/72839/fi_m/dp.pdf>.

[24] VEJLUPEK, T. *Firemní zpravodajský informační systém*[online]. 2001 [cit. 2011-07-12]. Dostupný z WWW: <http://www.inforum.cz/archiv/inforum2001/prispevky/vejlupek.htm>.

[25] VŠCHT. *Výkladový slovník české terminologie z oblasti informační vědy a knihovnictví* [online]. Praha : Vydavatelství VŠCHT, 2006 [cit. 2011-05-31]. Dostupné z WWW: <http://vydavatelstvi.vscht.cz/knihy/uid_es-005/motor/main.obsah.html>. ISBN 80-7080-599-4.

[26] *Yahoo* [online]. c2011 [cit. 2011-07-11]. Dostupné z WWW: <http://www.yahoo.com/>.

[27] Zákon č. 412/2005 Sb., o ochraně utajovaných informací a o bezpečnostní způsobilosti, ve znění pozdějších předpisů.

Seznam obrázků

Seznam tabulek