Moderné vzdelávanie pre vedomostnú spoločnosť/
Projekt je spolufinancovaný zo zdrojov EÚ

METÓDY RIEŠENIA ÚLOH SOCIÁLNEHO WEBU (SÉMANTICKÝ A SOCIÁLNY WEB)

Fakulta elektrotechniky a informatiky

Kristína Machová

Európska únia
Európsky sociálny fond

Agentúra Ministerstva školstva, vedy, výskumu a športu SR pre štrukturálne fondy EÚ

Operačný program VZDELÁVANIE

Táto publikácia vznikla za finančnej podpory z **Európskeho sociálneho fondu** v rámci Operačného programu **VZDELÁVANIE.**

Prioritná os 1 Reforma vzdelávania a odbornej prípravy

Opatrenie 1.2 Vysoké školy a výskum a vývoj ako motory rozvoja vedomostnej spoločnosti.

Názov projektu: **Balík doplnkov pre ďalšiu reformu vzdelávania na TUKE**

ITMS 2611020093

NÁZOV: Sémantický a sociálny web
AUTOR: doc. Ing. Kristína Machová, PhD.
RECENZENTI: Ing. Martin Sarnovský, PhD., doc. Ing. Marián Mach, CSc.
VYDAVATEĽ: Technická univerzita v Košiciach
ROK: 2015
ROZSAH: 141 strán
NÁKLAD: 70 ks
VYDANIE: prvé
ISBN: **978-80-553-1974-2**

Rukopis neprešiel jazykovou úpravou.
Za odbornú a obsahovú stránku zodpovedajú autori.

Predslov

Predkladaná učebnica je zameraná na problematiku sémantického webu a jeho vzťahu k sociálnemu webu, pričom väčší priestor je venovaný práve sociálnemu webu. Tieto témy sú dnes veľmi aktuálne najmä v súvislosti s dolovaním konverzačného obsahu.

V rámci sémantického webu je hlavná pozornosť venovaná problémom a technológiám sémantického webu. Osobitná podkapitola je venovaná otázke integrácie webu pomocou modelu úložišťa informácií. Sú uvedené rozličné systémy a aplikácie založené na sémantických technológiách.

Druhá rozsiahlejšia časť učebnice je venovaná sociálnemu webu a obsahuje širokú škálu oblastí súvisiacich s rozličnými platformami sociálneho webu, ako sú sociálne siete a ich vizualizácia, analýza sociálnych sietí, notácie statickej aj dynamickej sociálnej siete a sieťové štatistiky (3. a 4. kapitola). Potom nasledujú časti venované práve analýze a dolovaniu konverzačného obsahu (5.kapitola). V rámci dolovania sociálneho webu je hlavná pozornosť venovaná klasifikácii názorov založenej na slovníkovom prístupe, na dynamickom koeficiente a na použití n-gramov. Dolovaniu konverzačného obsahu musí predchádzať extrakcia dát z webových zdrojov (6. kapitola) a napokon v konverzačnom obsahu je možné objavovať nielen sumarizované názory ale aj informácie o prirodzených autoritách webových diskusií (7.kapitola).

Práca sa nesnaží byť vyčerpávajúcim textom z predkladanej oblasti a nerobí si ani nárok na úplnosť. Je vhodná pre tých, ktorí chcú získať základný prehľad oblasti, ale aj pre tých, ktorí majú hlbší záujem o prezentovanú problematiku a detaily experimentov, na ktorých participovala autorka publikácie. Predložená publikácia je vhodná ako základná literatúra na výučbu problematiky sémantického a sociálneho webu na technických univerzitách a vypĺňuje medzeru, ktorá v oblasti informatickej literatúry v našich podmienkach neustále trvá, a ktorá sa stále citeľne prejavuje ako nedostatok lekčných fondov vo vzdelávacom procese v pôsobnosti vedeckej, technickej a vysokoškolskej komunity Slovenska.

Predkladaná učebnica vznikla na Katedre kybernetiky a umelej inteligencie Fakulty elektrotechniky a informatiky Technickej univerzity v Košiciach. Pri jej zostavovaní vychádzala autorka zo skúseností, získaných počas riešenia viacerých výskumných projektov, vedenia záverečných prác a zabezpečovania výuky predmetu Sémantický a sociálny web, ktorý je ponúkaný v dvoch študijných programoch a to Inteligentné systémy a Hospodárska informatika.

Autorkina vďaka patrí recenzentom Ing. Martinovi Sarnovskému, PhD a doc. Ing. Mariánovi Machovi, CSc. za recenziu predkladanej publikácie ako aj jej kolegom z Katedry kybernetiky a umelej inteligencie FEI TU v Košiciach za cenné rady a podnety.

Košice, február 2015 autorka

Obsah

1 SÉMANTICKÝ WEB

1.1 Úvod

Web ako taký je fenoménom dnešnej doby. Je preň charakteristický dynamický rozvoj a rýchla expanzia. Táto expanzia sa prejavuje kontinuálnym nárastom počtu serverových staníc a rozličných zdrojov. Rýchly rast so sebou prináša nové problémy. Keďže web obsahuje obrovský počet stránok, výsledkom klasického prehľadávania pomocou kľúčových slov môže byť trochu menšia ale stále priveľká množina webových stránok. Priveľká na prečítanie a nájdenie konkrétnej informácie. Vynára sa tu potom rekurzívny problém ako vyhľadať špecializované články v množine už vyhľadaných dokumentov. Toto je problém privysokej návratnosti prehľadávania webu. Ale problémom môže byť aj opačný prípad, keď je návratnosť príliš nízka alebo žiadna, čo môže byť spôsobené tým, že používateľ používa iný slovník pri zadávaní kľúčových slov, ako je slovník ostatných používateľov, ktorý je reprezentovaný indexmi stránok. A tak sa hľadaná informácia nenájde, aj keď je na stránkach prítomná. Problémy spôsobuje aj „hypertextovosť“ webu, ktorá je spôsobená často chaotickou reprezentáciou dát na webe, keď jednotlivé stránky sú poprepájane aj mnohonásobne, odkazujú na seba navzájom, vytvárajú sľučky a tak komplikujú hľadanie.

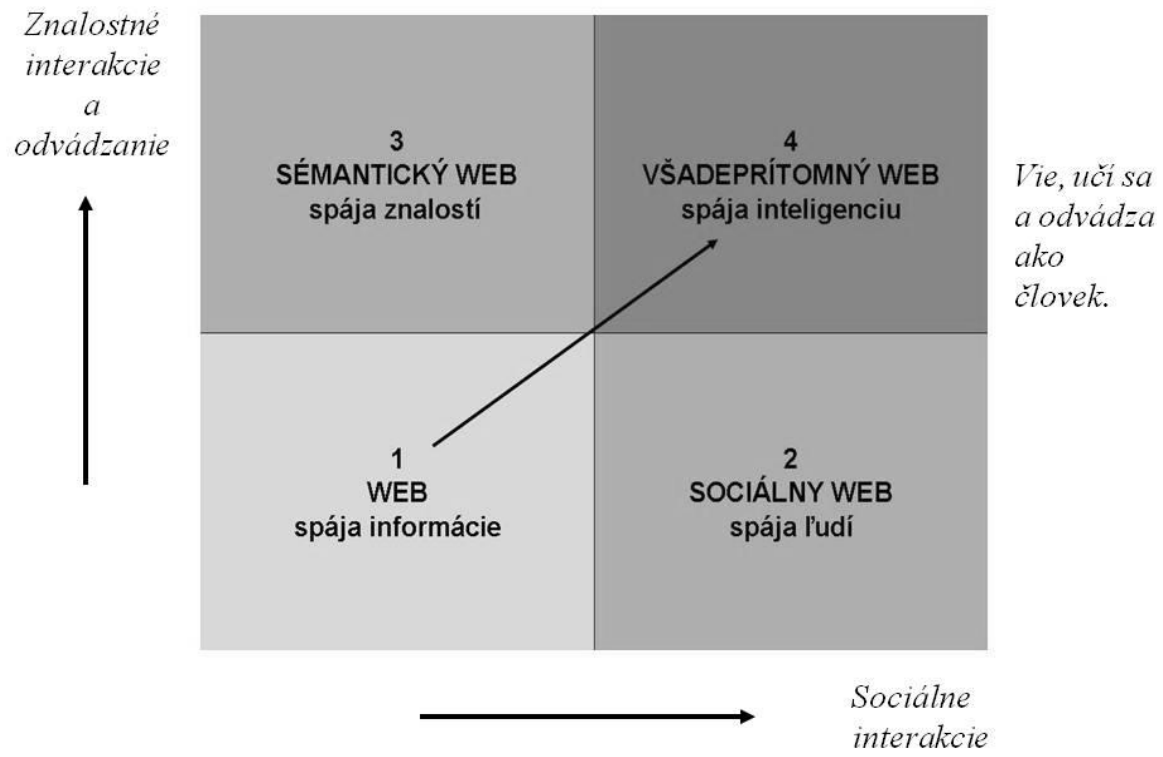
Riešením by mohol byť inteligentný prehľadávací stroj, ktorý by bol schopný navigovať používateľa pri surfovaní webu a ktorý by bol sémantický v tom zmysle, že by do určitej miery rozumel zmyslu požiadavky - otázky používateľa. To vyžaduje obohatenie webového prehľadávača o strojové vnímanie, ktoré umožní chápať požiadavku používateľa nie ako reťazec kľúčových slov pospájaných logickými operáciami, ale ako určitú formu reprezentácie hľadanej informácie, ktorej významu rozumie.

1.2 Sémantický verzus sociálny web

Keďže sa tento učebný text zaoberá nielen sémantickým ale aj sociálnym webom, bude vhodné hneď na začiatku definovať základný rozdiel medzi mini. Zatiaľ čo sémantický web posilňuje interakcie medzi znalosťami, sociálny web posilňuje sociálne interakcie medzi používateľmi, čo je ilustrované na Obr.1.1.

Sémantický web sa snaží rozumieť zmyslu - sémantike požiadavky používateľa na prehľadávací stroj a teda má ambíciu rozpoznať tú istú znalosť v každom jej výskyte na webe bez ohľadu na rozličnú formu reprezentácie, na rôznorodosť použitého slovníka alebo programovacieho jazyka či technológie. Ale je tu ešte jeden aspekt. Sémantický web dokáže vyhľadať aj znalosti, ktoré nie sú vyjadrené explicitne, iba implicitne a ktoré je potrebné odvodiť z tých explicitných znalostí a faktov, ktoré sú priamo reprezentované a teda vyhľadateľné na konkrétnych webových stránkach. Teda, sémantický web posilňuje nielen znalostné interakcie ale aj odvádzanie.

Sociálny web posilňuje niečo celkom iné a to sociálne interakcie medzi používateľmi webu v rámci rozličných platforiem sociálneho webu ako: sociálne siete, diskusné fóra, blogy microblogy, chaty, chatrooms a pod. Tieto platformy mimoriadne uľahčili a urýchlili komunikáciu medzi ľuďmi v porovnaní s interakciami v reálnom svete. Veľká vzdialenosť nie je prekážkou. Vďaka sociálnemu webu sa počet producentov webového obsahu v porovnaní s počtom jeho konzumentov prudko zvýšil.



Obr. 1.1 Ilustrácia divergentného vývoja sémantického a sociálneho webu.

Obr.1.1 ilustruje divergentný vývoj sémantického a sociálneho webu, pretože každý z nich posilňuje celkom iný aspekt rozvoja webu. Je tu určitá predstava, že niekedy v budúcnosti sa bude sémantický a sociálny web rozvíjať takým spôsobom, ktorý ich dovedie k jednotnému takzvanému všadeprítomnému webu, ktorý bude vedieť, učiť sa a odvádzať nové znalosti tak ako človek.

1.3 Problémy sémantického webu

Web predstavuje vhodný a široký informačný priestor pre hľadanie informácií pomocou vyhľadávacích strojov, ktoré sú veľmi úspešné, hlavne Google. Väčšina dnes známych prehľadávačov webu: ako napríklad AltaVista, Yahoo, Google, vyhľadávajú stránky na základe kľúčových slov. Avšak s hľadaním informácií na webe založenom na kľúčových slovách sú spojené mnohé problémy, ktoré už boli zmienené v úvode. Najčastejšie sa stretávame s vysokou návratnosťou spojenou s nízkou presnosťou. Ak sú výsledkom vyhľadávania tisícky stredne relevantných a irelevantných stránok, tak ani nemôžeme hovoriť o vyhľadávaní. Ide skôr o redukciu webového priestoru na webový priestor danej domény. Inokedy je návratnosť naopak nízka alebo žiadna, keď nie je vrátená žiadna stránka, alebo to málo čo sa vráti používateľa neuspokojí. Tieto problémy sa ešte znásobujú citlivosťou výsledkov vyhľadávania na používaný slovník. Môže sa stať, že budeme neúspešní iba preto, lebo sme ako kľúčové slovo použili synonymum slova, ktoré sa na stránke nachádza. Vyhľadávanie na základe kľúčových slov fakticky nie je vyhľadávaním informácií ale web stránok. Web stránka môže byť dynamická alebo je to iba

zoznam odkazov na iné stránky. A tak sa môže stať, že aj keď máme relevantné zdroje ako výsledok hľadania, informácia sa vlastne nachádza na viacerých stránkach a hľadanú informáciu z tejto kolekcie stránok musíme abstrahovať.

Otázkou je ako by mohli byť tieto problémy vyriešené. To čo by bolo možné urobiť je na súčasný web aplikovať sofistikované metódy prehľadávania webového priestoru, ktoré budú založené na umelej inteligencii a výpočtovej lingvistike.

Iný prístup je reprezentovať obsah webu vo forme, ktorá je jednoduchšie spracovateľná strojom, lebo stroj dokáže rozlíšiť povahu a doménu hľadanej informácie. Tento prístup smeruje k sémantickému webu. Taktiež je potrebné zabezpečiť podporu slabo štruktúrovaných textových zdrojov veľkého počtu. S tým súvisí časová náročnosť. Na udržiavanie konzistencie, správnosti a aktuálnosti informácií potrebujeme mechanizmy reprezentácie sémantiky a obmedzení, ktoré je možné využiť pri detekcii nekonzistentných údajov.

V podstate je potrebné do chaoticky usporiadaného webového priestoru vniesť usporiadanie aspoň do určitej miery. Núkajú sa rôzne konkrétne existujúce prístupy k riešeniu nahromadených problémov:

- **Metóda hrubej sily.** Často používané riešenie, ktoré spočíva vo zvýšení výkonu serverov a priepustnosti komunikačných liniek. Nerieši situáciu dlhodobo a v konečnom dôsledku vedie k vyšším nárokom používateľov a k opätovnému nedostatku výkonu.

- **Automatická manipulácia s dátami.** Ide o začlenenie „meta znalostí“ do webového dokumentu vo forme takzvaných „meta tagov“, ktoré umožňujú štruktúrovať text vložením informácie o druhu obsahu, ako napríklad kľúčové slova, informácie o autorovi, názov alebo popis, čím sa konkrétna stránka zaoberá. Obsah tagov nie je zobrazovaný na stránke, takže pre bežného používateľa ostáva skrytý. Využívať ich môžu predovšetkým agenti - sofistikované programy, ktoré používajú vyhľadávacie služby na indexovanie webových stránok. Agenti na základe kľúčových slov, alebo popisu stránky, dokážu presnejšie určiť oblasť, ktorou sa konkrétna stránka zaoberá. Nápomocné sú pri tom kaskádové štýly, ktoré sa snažia o odlíšenie štruktúry dokumentu od rozloženia dokumentu a následnú špecifikáciu určitej vlastnosti časti dokumentu ako formátovanie, informačný obsah a pod.

- **Dolovanie z webu.** Tento prístup zvláda chaotickú štruktúru Internetu lepšie ako iné prístupy. V princípe je možné použiť dolovanie (web mining) nielen na vyhľadávanie stránok a informácii ale aj na vyhľadávanie autoritatívnych stránok. Stránky obsahujúce informácie relevantné k dotazu sa označujú ako autoritatívne stránky. Prepojovacie stránky, ktoré obsahujú prevažne iba odkazy na iné stránky, nie sú autoritatívnymi stránkami. Stránka, na ktorú sa odkazuje veľa iných stránok, je pravdepodobne autoritatívnou stránkou. Metódy dolovania z webu je možné rozdeliť do nasledovných oblastí: dolovanie z obsahu webu, dolovanie z používania webu a dolovanie zo štruktúry webu.

- **Multiagentové systémy.** V prostrediach s chaotickou a neorganizovanou štruktúrou sa agentové technológie javia ako výhodné riešenie, ale riešeniam založeným na multiagentových systémoch často chýba dostatočná podpora pre znalosti a inteligenciu. Pri správnom použití však agenti môžu chápať formalizované znalosti podobne ako ľudia.

- **Sémantický web.** Internet ako strojovo spracovateľná sieť „smart“ dát sa javí ako najvhodnejšie z uvádzaných riešení. Tento prístup k riešeniu problémov Internetu by mohol úspešne riešiť jeho informačné preťaženie. Taktiež by mohol dať odpoveď na riešenie problémov agregácie obrovského množstva dát v dátových skladoch a s tým súvisiacich problémov závislosti aplikácií na danej štruktúre dát. Riešením by mohol byť vývoj prístupu k objavovaniu informácií a relácií v organizovaných úložiskách.

Nemalo by zmysel začať vytvárať sémantický web ako niečo celkom nové paralelné k existujúcemu webu. Nie je možné zrušiť čo bolo zatiaľ vytvorené. Sémantický web budovaný nad existujúcim webom a dopĺňajúci ho je propagovaný aj W3C (World Wide Web Consortium). Je všeobecna známe, že zakladateľom webu je Tim Berners Lee, ktorý v 80-tych rokoch vymyslel základný princíp, ktorý umožnil návrh a implementáciu webu. Podrobnejšie o sémantickom webe sa pojednáva okrem iného aj v zdroji [Antoniu-vanHarmelen,
2004].

1.4 Možnosti reálneho uplatnenia sémantického webu

Sémantický web je možné použiť pri riešení takmer všetkých problémov, ktoré súvisia s používaním webu. Príkladom môže byť znalostný manažment a rôzne druhy elektronického obchodu ako aj vývoj personálneho sémantického agenta. Hlavne táto posledná možnosť predstavuje silnú motiváciu pre rozvoj sémantického webu.

Úlohou sémantického webu v oblasti znalostného manažmentu je umožniť rozvoj pokrokovejších systémov znalostného manažmentu, ktoré umožnia organizáciu znalostí v konceptuálnom priestore podľa ich významu a podporu udržiavania znalostí prostredníctvom kontroly konzistentnosti ale aj extrakciou nových znalostí. V tomto prípade sa môže vyhľadávanie pomocou kľúčových slov nahradiť takzvaným „query answering“, v rámci ktorého požadované znalosti budú navrátené vo forme presnej odpovede, extrahovanej z viacerých navrátených stránok vo forme priateľskej používateľovi. Teda bude podporovaná odpoveď z viacerých dokumentov a naviac bude možné definovať, kto môže byť príjemcom určitej informácie, čo je v oblasti manažmentu veľmi dôležité.

Rozlišujeme dva druhy elektronického obchodu a to B2C a B2B. B2C je obchod spájajúci biznis a konzumenta (Businiess-to-consumer) Ide o dominantnú komerčnú aktivitu používateľa Internetu. V tomto prípade používateľ navštevuje on-line obchody a prehľadáva ich. Potom usporadúva produkty na základe manuálne zhromaždených informácií o cenách, podmienkach, termínoch s účelom vybrať najlepšiu ponuku. To je časovo náročná aktivita, preto obvykle používateľ nenavštívi všetky internetové obchody. Riešením tohto problému by mohli byť softvéroví agenti „shopbots“, ktorí by navštívili všetky dostupné internetové obchody a extrahovali z nich informácie o produktoch. Aby to tak fungovalo, museli mať k dispozícii takzvané „wrappers“ – obálky, teda programy extrahujúce informácie z on-line obchodov. Tento prístup má svoje nedostatky: cena sa nájde iba ak je nasledovaná znakom „$“, získané informácie sú limitované, môže chýbať napr. čas dodania, úroveň bezpečnosti, reštrikcie krajiny dodávateľa a naviac „wrappers“ sú konzumenti času. Sémantický web by umožnil vývoj softvérových agentov, ktorí by nahradili nutnosť programovania obálok a interpretovali by informácie o produktoch, o vlastnostiach a cenách produktov, ktoré by boli extrahované korektne a nasledované informáciami o dodacích podmienkach a poistení. Tieto informácie by boli porovnávané s požiadavkami používateľa Internetu, ktorý by dostal k dispozícii aj informácie o reputácii on-line obchodov od nezávislých agentúr. Softvérovi nákupní agenti by realizovali automatické vyjednávania namiesto kupujúceho priamo s obchodnými agentmi.

B2B je obchod vedúci priamo od biznisu k biznisu (business-to-business). Tradične sa na takýto typ obchodu používa prístup EDI (Electronic Data Interchange), ktorý je komplikovaný a zrozumiteľný iba expertom, náročný tak na programovanie ako aj na udržiavanie a náchylný k chybovosti. Komunikáciu predražuje nutnosť programovať každý B2B obchod osobitne. Realizácia sémantického webu by mohla umožniť zúčastneným stranám vstupovať do partnerstva bez prílišných ťažkostí. Rozdiely v terminológii by mohli byť rozpustené použitím abstraktného doménového modelu OWL (Ontology Web Language), ktorý patrí do skupiny jazykov na reprezentáciu znalostí. Dáta by boli zdieľané použitím prenosových služieb. Aukcie, negociácie a koncepty, resp. návrhy kontraktov by boli vybavované semi - automaticky softvérovým agentom.

1.4.1 Sémantický agent – scenár budúcnosti

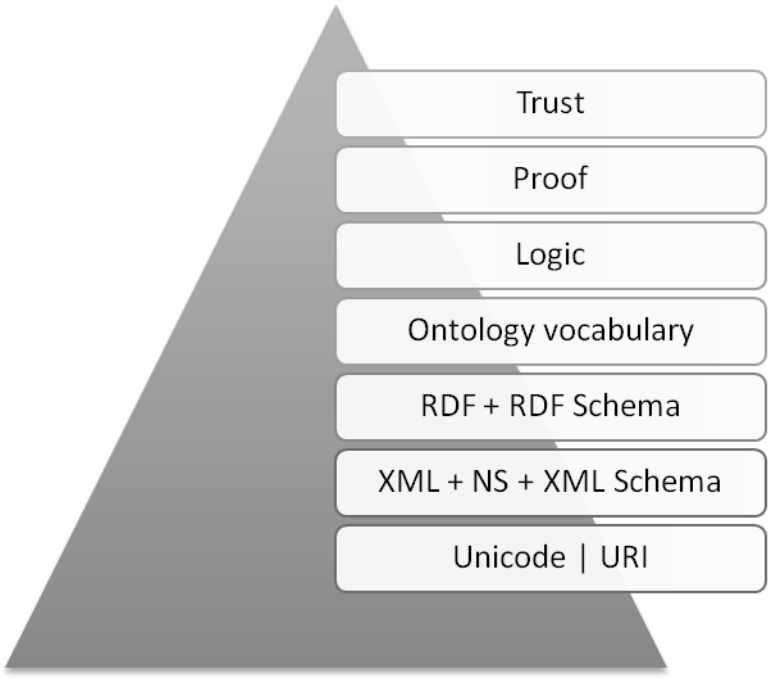
František dostal ponuku prednášať na renomovanej francúzskej univerzite. Potrebuje v krátkom čase zlepšiť úroveň znalostí francúzskeho jazyka. Preto požiada svojho sémantického webového agenta o sémantické vyhľadanie možností absolvovať kurz francúzskeho jazyka. Sémantický agent získa najprv detaily o požadovanej úrovni francúzštiny z webových stránok univerzity, ktorú František navštívi. Potom agent vyhľadá zoznam jazykových škôl v materskom meste Františka a vyberie tie, ktoré sú vo vzdialenosti do 15 minút autom od Františkovej kancelárie alebo domova. Taktiež, na základe informácií o reputácii týchto zvolených jazykových škôl, pripraví ich rebríček. Informácie k reputácii získa od renomovaných ratingových agentúr, ktoré zverejňujú svoje zistenia na webe. Z tohto rebríčka vylúči všetky jazykové školy, ktoré nedisponujú lektormi s francúzštinou ako rodným jazykom („native lector“). Potom sa sémantický agent pokúsi zladiť hodiny kurzov francúzštiny s Františkovým kalendárom. Výsledkom budú tri alternatívy. No Františkovi nevyhovujú. Jeden kurz sa začína až o mesiac a on nechce strácať čas. K druhému by musel chodiť autom počas špičky a na tretí by musel chodiť peši a cesta by trvala až 45 minút. Preto sa František rozhodne modifikovať časové obmedzenia a požiada agenta o iné alternatívy. Sémantický agent skutočne nájde novú alternatívu a to s „native“ lektorom, ktorý je k dispozícii už zajtra, ale v čase, keď má František v práci niektoré menej dôležité povinnosti. František je ochotný si tieto povinnosti preplánovať. Avšak kurz by presahoval časový interval, ktorý má František do odchodu do Francúzska. Sémantický agent Františka sa dohodne s agentom lektora na individuálnych hodinách, pričom si bude účtovať trochu vyššiu cenu. František je ochotný zaplatiť väčšiu sumu, keďže kurz bude intenzívny, takže sa stihne pripraviť na zahraničnú cestu včas. Avšak, František si chce byť úplne istý, že toto riešenie je najlepšie z tých čo boli k dispozícii, preto požiada agenta o vysvetlenia niektorých jeho tvrdení. Ako získal informácie o reputácii lektora? Či sa dalo vyhnúť preplánovaniu niektorých povinností? Ako bolo vedené vyjednávanie ceny? Agent svojim vysvetlením rozptýlil Františkove obavy a ten požiadal svojho sémantického agenta aby finalizoval úlohu.

1.4.2 Vrstvový prístup k budovaniu sémantického webu

Už bolo povedané, že nemôžeme zahodiť existujúci web a povedať, že my teraz vybudujeme celkom od začiatku úplne nový, lepší, sémantický web. Sémantický web sa musí budovať v akýchsi stupňoch, pričom každý stupeň predstavuje vybudovanie novej vrstvy na vrchole predchádzajúcej. Preto hovoríme o vrstvovom prístupe. Pri budovaní každej ďalšej vrstvy sémantického webu nad inou, musia byť dodržané dva princípy a to klesajúca kompatibilita a čiastočné porozumenie smerom nahor.

- **Klesajúca kompatibilita** by mala zabezpečiť schopnosť „plne“ interpretovať a používať informácie zapísané v nižšej vrstve. Napríklad, dokument v OWL úplne rozumie informáciám zapísaným v RDF a RDF Schema.
- **Čiastočné porozumenie smerom nahor**. Agenti by mali mať aspoň čiastočný prospech z informácií na vyššej vrstve. Napríklad, dokument RDF (a RDF Schema) môže čiastočne interpretovať znalosti zapísané v OWL vrstve s výnimkou tých elementov, ktoré idú nad rámec RDF.

Obr.1.2 znázorňuje vrstvovú schému (podľa Tima Berners-Lee), ktorá obsahuje návrh a víziu hlavných vrstiev sémantického webu.



Obr. 1.2 Vrstvový prístup k budovaniu sémantického webu.

Hneď nad vrstvou URI sa nachádza XML (eXtensible Markup Language) značkovací jazyk, ktorý bol vyvinutý a štandardizovaný konzorciom W3C (World Wide Web Consortium). Jazyk XML umožňuje písať web dokumenty, ktoré obsahujú štruktúrované informácie, pričom slovník „tagov“ si definuje používateľ sám.

Nad touto vrstvou je možné vybudovať RDF vrstvu, keďže RDF je možné považovať za rozšírenie XML. RDF je základný dátový model, reprezentujúci vzťahy medzi entitami Aj v RDF je možné zapisovať jednoduché výroky o webových zdrojoch. RDF dátový model sa nespolieha plne na XML, ale má syntax založenú na XML. RDF Schema umožňuje organizáciu webových objektov do hierarchie. Kľúčové primitíva sú triedy a vlastnosti, podtriedy a vlastnostné vzťahy, ako aj doména a rozsah vlastností. RDF Schema je založená na RDF. RDF Schema sa berie ako jednoduchý jazyk na písanie ontológií.

OWL je jazyk, ktorý sa špecializuje na budovanie ontologických modelov, ktoré môžu slúžiť okrem iného ako slovník termov využívaných v značkovacích jazykoch, konkrétne v „tagoch“.

Logická vrstva – „Logic“ sa používa na zlepšenie hľadanie, odvádzanie a následne zápis aplikačno - špecifických deklaratívnych znalostí.

Vrstva dôkazu a kontroly - „Proof“ využíva deduktívne procesy na formuláciu dôkazov vo web jazykoch, formuláciu vysvetlení ako aj kontrolu platnosti dôkazov.

Posledná vrstva „Trust“ - dôvery sa dá realizovať napríklad prostredníctvom používania elektronického podpisu ale aj informácií - odporučení od ratingových a certifikačných agentúr a spotrebiteľských združení. Táto dôvera je usporiadaná distribuovaným a chaotickým spôsobom, ktorý je vlastný webu a je najťažšie dosiahnuteľná. Sémantický web napokon môže byť používateľmi webu prijatý keď títo používatelia budú môcť dôverovať vlastným sémantickým agentom a kvalite získaných informácií.

1.5 Technológie sémantického webu

Problémom pri presadzovaní idey sémantického webu je formálna rôznorodosť webu, ktorá komplikuje úlohu nájdenia nejakej jednotnej stratégie prehľadávania, ktorá by bola schopná sprostredkovať stroju (programu) sémantiku, teda význam formálne odlišných dokumentov v rôznych jazykoch, aj keď človek s tým nemá problém. Realizácia vízie sémantického webu je dosiahnuteľná aj bez revolučného objavu, pretože ide skôr o inžiniersku a technologickú výzvu ako vedeckú. Faktom totiž je, že poznáme dosť čiastočných riešení, využiteľných na integráciu. Integrované by mohli byť známe sémantické technológie za účelom vývoja prostriedkov a aplikácií spôsobom, ktorý by zabezpečil prijatie používateľom. Ale samozrejme ďalší technologický vývoj je vítaný a mohol by viesť k pokrokovejšiemu sémantickému webu. Základné už existujúce technológie sémantického webu sú značkovacie jazyky, ontológie, logiky a agentové systémy.

1.5.1 Značkovacie jazyky

Súčasný web je dobre čitateľný človekom ale stroju (softvéru) je viac menej nečitateľný. Prevládajúcim jazykom web stránok je totiž HTML. Nasleduje fragment web stránky v jazyku HTML:

```
<h1>Francúzska aliancia</h1>

Vitajte na domovských stránkach Francúzskej aliancie. Potrebuje
rýchlo vylepšiť úroveň vedomostí francúzskeho jazyka?. K dispozícii
máme skvelých lektorov, ktorých materinským jazykom je
francúzština. Obráťte sa na nás a naši zamestnanci Petriša
Tirpáková (naša sekretárka), Francois-Xavier Demarty and Wandrille
de Cambrai sa Vám budú ochotne venovať.

<h2>Konzultačné hodiny</h2>
Pondelok   14:00 – 19:00<br>
Utorok     13:00 – 18:00<br>
Streda     14:00 – 19:00<br>
Štvrtok    13:00 – 18:00<br>
Piatok     11:00 – 14:00<br>
V týždňoch,ktoré sú vymenované na nasledujúcej stránke nie sú
lekcie <a href=”…”>Voľné týždne</a>.
```

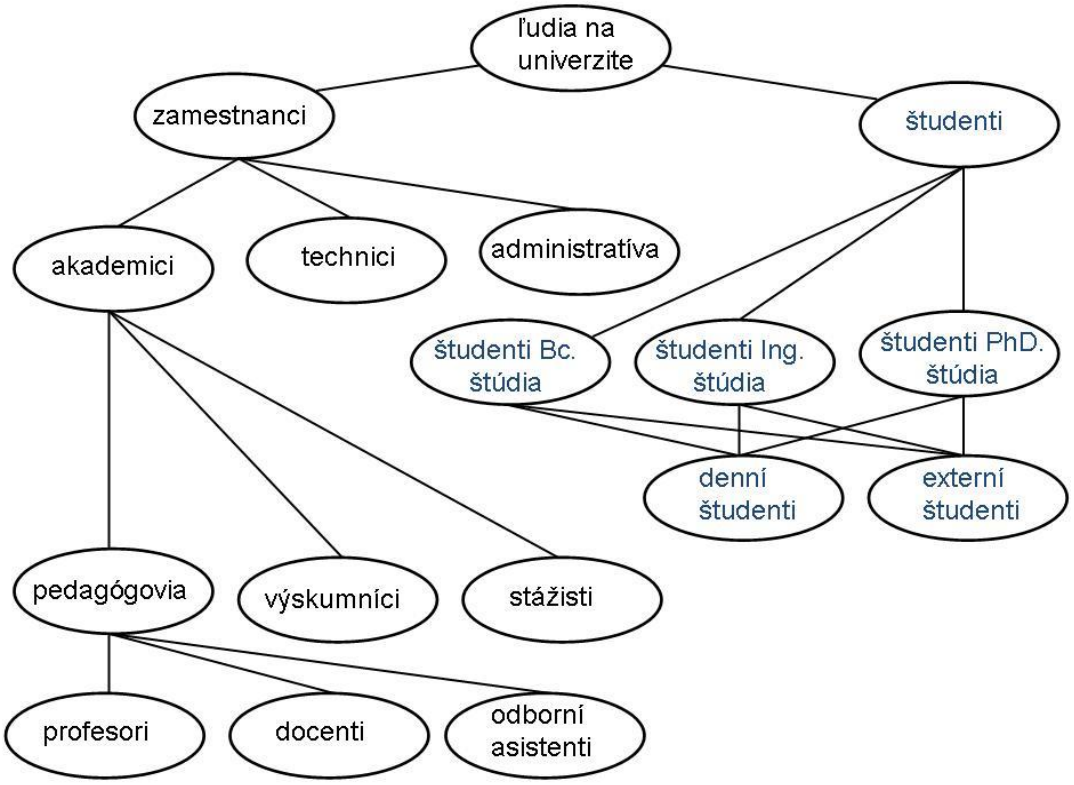
Človek takýmto informáciám rozumie, ale stroj by mal s ich pochopením problémy. Ak by agent použil vyhľadávanie podľa kľúčových slov, dokázal by identifikovať niektoré charakteristické slová: jazykové centrum a konzultačné hodiny. Inteligentný agent by bol možno schopný identifikovať zamestnancov centra, no nevedel by odlíšiť lektorov od sekretárky. Taktiež by mal problém s identifikáciou dodatočnej informácie ku konzultačným hodinám na linke nazvanej „Voľné týždne“. Podstata sémantického webu nespočíva ani tak vo vývoji super inteligentného agenta. Najprv je tu snaha zdokonaliť reprezentáciu informácií na webových stránkach tak, aby boli ľahšie čitateľné strojom. To vedie ku potrebe náhrady HTML vhodnejšími jazykmi, ktoré by umožnili reprezentovať informácie na webových stránkach v štruktúrovanej forme. Môžeme si pomôcť predstavou akýchsi informačných puzdier. Príklad takýchto puzdier (pre názov spoločnosti, pre ponúkané služby pre zamestnancov sekretariátu a zamestnancov lektorov) v rámci vyššie uvedeného fragmentu stránky by bol nasledovný:

```
<spoločnosť>
      <názov>Francúzska aliancia</názov>
      <ponúkané služby>lekcie francúštiny</ponúkané služby>
      <zamestnanci>
            <sekretárka> Petriša Tirpáková</sekretárka>
            <lektor>Francois-Xavier Demarty</lektor>
            <lektor>Wandrille de Cambrai</lektor>
      </zamestnanci>
</spoločnosť>
```

Takýmto spôsobom reprezentovaným informáciám dokáže stroj oveľa ľahšie porozumieť a spracovať ich. Aj keď je pravdou, že pre človeka je takáto štruktúrovaná reprezentácia menej čitateľná ako obyčajný text, teda vlastne človeku, ktorý nie je veľmi zručný v programovaní, sme to skomplikovali. Štruktúra dát je definovaná usporiadaním a obsahom jednotlivých „tagov“, ktoré predstavujú metadáta. Metadáta sú dáta o dátach. Reprezentujú sémantiku (význam) dát. V takom prípade je možné priamo z web stránok automaticky získať informácie o detailoch kurzov, o kalendári, prístupnosti, cenách a popise služieb. Je otázne, či by sa bežný používateľ bol ochotný vzdať HTML, ktorý je jednoduchý. Prví používatelia sa rozhodli prijať HTML v čase, keď to bol nový a jediný štandard na tvorbu webových stránok. Podobne môžeme očakávať dobré prijatie ďalších štandardov ako XML a RDF, keď používatelia pochopia, že im to prináša výhody. To by mohlo viesť k prijatiu ďalších nových technológií a to bude rozhodujúci krok smerom k sémantickému webu.

1.5.2 Ontológie

Termín „ontológia“ pochádza z gréckej filozofie. Napríklad, pozorovanie „že svet pozostáva zo špecifických objektov, ktoré môžu byť zoskupované do abstraktných tried, založených na zdieľaní vlastností“ je typická ontologická úloha. Avšak, v posledných rokoch tomuto termínu dala počítačová veda technický význam. T.R.Gruberova [Gruber, 1993] definícia bola spresnená R. Studerom: „*Ontológia je explicitná a formálna špecifikácia konceptualizácie*.“ [Studer-at al., 1998]. Vo všeobecnosti ontológia formálne popisuje doménu, predstavuje konečný zoznam termínov a vzťahov medzi nimi. Termíny reprezentujú objekty domény respektíve triedy objektov domény, teda pojmy. V súvislosti s ontológiami sa používa skôr koncept ako pojem. Ak by sme chceli špecifikovať takéto objekty v univerzitnom prostredí, boli by to také pojmy ako: zamestnanci, študenti, kurzy, prednáškové miestnosti, predmety a pod. Vzťahy typicky zahŕňajú hierarchiu tried (viď Obr.1.3). Hierarchia špecifikuje, že trieda *C* je podtriedou inej triedy *C´,* ak každý objekt z *C* je zahrnutý aj do *C´*.



Obr. 1.3 Hierarchia univerzitnej domény.

Okrem vzťahov medzi triedami a podtriedami, ontológia môže zahŕňať aj nasledovné informácie: v*lastnosti* (*X* vyučuje *Y*), o*bmedzenia hodnôt*, obmedzenia kardinality (katedra musí mať aspoň päť zamestnancov), v*ýroky o disjunktnosti*, *špecifikácie a logické vzťahy medzi objektmi.*

V kontexte sémantického webu, ontológie napomáhajú porozumeniu domény tým že vyjasňujú rozdiely v terminológii. Napríklad slovo kmeň predstavuje v doméne biológie časť rastliny – stromu. Ale slovo poistný kmeň v ekonomickej doméne predstavuje skupinu ľudí poistených v tej istej poisťovni, v doméne potravinárstva pojem kokosový kmeň reprezentuje sladkosť a v doméne antropológie slovo kmeň reprezentuje spoločenstvo ľudí žijúcich pomerne izolovane a vyznačujúcich sa špecifickou kultúrou (kmeň v pralese). Takéto rozdiely je možné prekonať mapovaním partikulárnej terminológie na zdieľanú ontológiu alebo definíciou priameho mapovania medzi ontológiami. Prekonávaním rozdielov v terminológii je podporovaná sémantická interoperabilita.

Pre informatiku je ontológia zaujímavá kvôli svojej užitočnosti v organizácii web stránok a následne pri navigácii používateľa v procese hľadania web stránok alebo konkrétnych informácií. Ontológie môžu byť užitočné aj pri zlepšovaní presnosti prehľadávania webu, ktoré sa tak stáva sémantickým prehľadávaním. Vyhľadávacie stroje môžu hľadať stránky, ktoré sa týkajú presného konceptu v ontológii, namiesto zhromažďovania všetkých stránok, v ktorých sa vyskytujú určité všeobecne nejednoznačné kľúčové slová. Prehľadávanie webu môže využívať zovšeobecnenie, resp. špecifikáciu dopytu. Ak dopyt zlyhá v hľadaní relevantných dokumentov, vyhľadávací stroj môže navrhnúť používateľovi všeobecnejší dopyt. Ak je navrátených príliš veľa odpovedí, vyhľadávací stroj môže ponúknuť používateľovi niektorú špecifikáciu dopytu.

V umelej inteligencii existuje dlhá tradícia vývoja a používania ontologických jazykov. Je to základ, na ktorom môže budovať výskum sémantického webu. V súčasnosti najdôležitejšie ontologické jazyky pre web sú *RDF Schema* ako slovníkový popisný jazyk pre definíciu vlastností a tried RDF zdrojov, ktorý definuje aj hierarchiu tried a vlastností. Hovoríme, že *RDF Schema* je primitívny ontologický jazyk. Typickým jazykom na tvorbu ontológií je *OWL*. *OWL* je bohatý slovníkový popisný jazyk pre definíciu vlastností a tried ako aj relácií medzi triedami danej domény, pre definíciu mohutnosti, rovnosti, disponujúci bohatým typovaním vlastností a taktiež charakteristikami vlastností a vymenovaných tried.

1.5.3 Logiky

Logika študuje princípy myslenia a ako taká siaha až po Aristotela. Vo všeobecnosti logika ponúka:

- **formálne jazyky** pre vyjadrenie vedomostí,
- **zrozumiteľnú formálnu sémantiku** - vo väčšine logík je význam viet definovaný bez potreby operovať s vedomosťami, hovoríme o deklaratívnych vedomostiach, definujeme čo platí a nestaráme sa ako sa to dá odvodiť,
- **automatických zdôvodňovateľov (reasoners)** - môžu dedukovať závery z daných vedomostí, teda z implicitných vedomostí vytvárajú explicitné.

To čo je z logiky podstatné pre sémantický web, je inferencia. Predpokladajme, že vieme, že všetci profesori sú pedagógovia, že všetci pedagógovia sú zamestnanci, a že Adam je profesor. V predikátovej logike je možné tieto informácie vyjadriť nasledovne:

```
profesor (X) → pedagóg (X)
pedagóg (X) → akademik (X)
profesor (Adam)
```

Potom môžeme dedukovať nasledovné fakty a jednu implikáciu (naučenú znalosť):

```
pedagóg (Adam)
akademik (Adam)
profesor (X) → akademik (X).
```

Tento príklad zahŕňa také typy znalosti, ktoré by sa mohli vyskytnúť v ontológii. Z toho vyplýva spôsob, ako by sa logika dala použiť na odhalenie nových znalostí, ktoré nie sú v ontológii vyjadrené explicitne ale sú v nej takpovediac schované (implicitné znalostí). Okrem toho, že logika môže viesť k odhaleniu neočakávaných vzťahov, môže napomôcť aj k odhaleniu inkonzistencií. Sémantický agent ju môže používať na tvorbu rozhodnutí. Napríklad, agent predajca sa môže rozhodnúť poskytnúť zľavu zákazníkovi na základe nasledujúceho pravidla, pričom informácie o lojalite zákazníkov môžu byť dostupné v zdieľanej databáze.

```
lojálny zákazník (X) → zľava (5%)
```

Logíky sú pre sémantický web dôležité aj preto, že dokážu poskytnúť vysvetlenie záverov vo forme spätne generovanej série vykonaných inferenčných krokov. Také vysvetlenie môže byť spätne vystopované („retrased“) a následne využité hlavne vo vrstve dôkazu a kontroly v rámci vrstvového prístupu k budovaniu sémantického webu. Vysvetlenie je pre sémantický web dôležité, pretože zvyšuje dôveru používateľa v sémantického webového agenta. Také vysvetlenie je dôležité aj kvôli komunikácii rôznych sémantických agentov medzi sebou. Umožňuje spoluprácu jednotlivých agentov. Zatiaľ čo jeden agent môže načrtnúť logické závery, iní agenti

môžu potvrdiť dôkazy, teda skontrolovať či tvrdenie iného agenta je správne.

1.5.4 Agentové systémy

Agenti sú súčasťou softvéru, ktorý pracuje autonómne a proaktívne. Presahujú koncept objektovo – orientovaného programovania a vývoja softvéru založeného na komponentoch. Personálny sémantický agent prijme niektoré úlohy a preferencie, resp. obmedzenia od osoby – používateľa, neúplné informácie z webových zdrojov, komunikuje s inými agentmi, porovnáva informácie o používateľových požiadavkách a obmedzeniach, vyberá určité možnosti a dáva odpovede používateľovi. Sémantický agent nemôže nahradiť ľudského používateľa webu. Vo väčšine prípadov je úloha agenta iba v zbieraní a organizovaní informácií a napokon v predkladaní možností na výber pre používateľa. Sémantický webový agent môže používať všetky načrtnuté technológie: metadáta na identifikáciu a extrakciu informácií, ontológie pri interpretácii navrátených informácií a komunikáciu s ostatnými agentami, logiku na spracovanie navrátených informácií a na výber vhodných záverov – alternatív riešenia problému.

1.6 Integrácia webu

Porozumenie potrebám používateľa v rámci sémantického webu vyžaduje zvládnuť okrem iného aj integráciu informácií [Finin at al., 2006]. Integrácia informácií je súčasťou sémantického prehľadávania, keďže umožňuje rozpoznať dve informácie prezentované v rozličnej forme ako v postate tú istú, majúcu ten istý význam. Táto integrácia informácií sa dá najlepšie vykonávať nad štruktúrovanou formou RDF dokumentov. Informácie z týchto dokumentov sú zhromažďované v takzvanom úložišti informácií, kde sa každá informácia nachádza iba raz. Taktiež reprezentácia informácií v tomto úložišti uľahčuje detekciu relácií medzi informáciami.

1.6.1 Model úložišťa informácií

Model úložišťa informácií je systém, ktorý umožňuje uskladňovanie informácií a následné prístupy k nim použitím dotazov. Tieto dotazy môžu byť založené na špecializácii (ktorý objekt zodpovedá podmienkam z dotazu) ale aj zovšeobecneniu (ktoré znalosti platia pre všetky objekty odpovedajúce podmienkam dotazu). Model úložišťa používa binárnu maticu pre reprezentáciu atribútov a ich hodnôt, čo je ilustrované na Obr.1.4.

Uvažujme dáta zo zdroja *z* pochádzajúceho z množiny zdrojov *Z*. Predpokladajme schému zdroja pre každý zdroj $S_z = (A_z , F_z)$, ktorá pokrýva aspoň zoznam atribútov A_z a funktionálne relácie F_z (A_z x A_z) medzi atribútmi. Predpokladajme, že tieto dáta sú reprezentované vo forme párov: atribút a hodnota z rozsahu hodnôt A_z x D_z, kde D_z reprezentuje všetky hodnoty – etalóny (*e* z *E*) pokryté zdrojom *z*.

Dáta z tohto zdroja môžu byť reprezentované vo forme funkčnej relácie *f* z F_z pomocou implikácie $e_i \rightarrow e_j$ medzi elementmi. Tieto implikácie pre každý zdroj môžu byť reprezentované v binárnej matici úložišťa $\Phi_I = [\ \Phi^I_{ij}\ ]$ definovanej v rovnici (1):

$$\Phi^I_{ij} = 1 \ if \ z_I covers \ e_i \rightarrow e_j \ (0 \ else) \qquad (1)$$

Analogicky, matica Δ_l = [δ^l_{ij}] domény atribútov zdroja z_l je definovaná rovnicou (2):

$$\delta^l_{ij} = 1 \; if \; e_i = A_{jl} \; (0 \; else) \qquad (2)$$

	elements								capital city		state					currency			EU	Middle		Vyseg Europ	
Prague	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1	1	0	1	0
Brno	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	1	0	1	0
Kosice	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	1	1	0	1	0
Bratislava	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	1	1	0	1	0
Poprad	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	0	1	1	0	1	0
Wiene	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	1	1	0	0	1
Budapest	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	1	1	0	1	0
Paris	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1	0	1	0	1	0	1
yes	1	0	0	1	0	1	1	1	1	0	-2	-2	-2	-2	-2	-2	-2	-2	1	-2	-2	-2	-2
no	0	1	1	0	1	0	0	0	0	1	-2	-2	-2	-2	-2	-2	-2	-2	1	1	0	1	0
Czech rep.	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	1	0	1	0
Slovakia	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	1	0
Austria	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	0	1
Hungary	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	1	1	0	1	0
France	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1
CZK	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0
Euro	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1	0	1	-2	-2	-2	-2
Florint	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0
yes	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	-2	-2	-2	-2
yes	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	-2	-2
no	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
yes	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
no	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Obr. 1.4 Príklad binárnej matice úložišťa informácií.

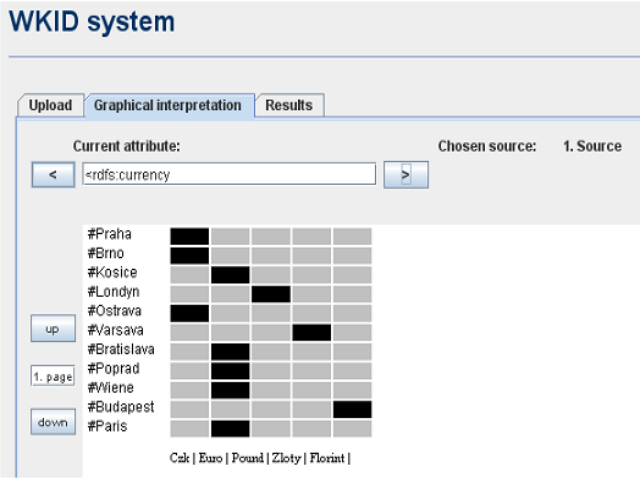
Každá nie nulová hodnota matice $\Phi^l_{ij} > 0$ reprezentuje príklad nejakej funkčnej relácie $f = (A_i \rightarrow A_j) \in F_l$.

Model úložiska informácií (viď Obr.1.4) je modelom získaných dát a reprezentovaných informácií. Tieto dáta môžu byť získavané z RDF a XML zdrojových dokumentov. Dáta z každého zdroja môžu byť reprezentované binárnou maticou v rámci úložiska. Z tejto matice vieme vyčítať hodnoty všetkých atribútov z aktuálneho zdroja. Napríklad (viď Obr.1.4), atribút ”elements” má hodnoty Prague, Brno, Košice, Bratislava, Poprad, Vienna, Budapest a Paris. Atribút “capital city” má hodnoty áno a nie. Atribút “currency” má hodnoty: CZK, Euro a Forint a pod. Teda názvy stĺpcov reprezentujú atribúty a v riadkoch sú názvy hodnôt. Ak niektorá bunka matice je „1“ potom atribút zo stĺpca, v ktorom sa bunka nachádza má hodnotu príslušného riadku. Inak bunka matice obsahuje “0”.

V Rámci práce [Machová-Fodorová, 2011] bol navrhnutý systém WKID na prácu s úložiskom informácií. Sub-matice pozdĺž hlavnej diagonály sú generované systémom WKID aby reprezentovali meta informácie o atribútoch a o ich hodnotách. Všetky sub-matice alokované mimo hlavnej diagonály reprezentujú následne dedukované informácie.

Po ukončení procesu importovania dát, určité časti matice úložiska ostanú prázdne. Tieto časti sa postupne počas spracovania informácií zapĺňajú novými dedukovanými informáciami a reláciami medzi nimi a tak sa tieto časti postupne

menia na aktívne sub-matice, ktoré sa môžu použiť na objavovanie if–then pravidiel. To sa deje na prvej úrovni spracovania. Tieto aktívne časti úložiska sú základom pre následné efektívne spracovanie a tvorbu grafickej reprezentácie informácií. Táto grafická reprezentácia je implementovaná vo WKID systéme a ilustrovaná na Obr.1.5.



Obr. 1.5 Grafická reprezentácia dát v systéme WKID.

1.6.2 Objavovanie znalostí v úložisku webových informácií

V práci [Paralič at al., 2010] je uvedený prístup k dolovaniu znalostí z textových dokumentov, ktoré pochádzajú z webu. Prístup použitý vo WKID je iný, pretože tento systém transformuje RDF dokumenty z webu do binárnej matice úložiska a následne v tejto matici objavuje znalosti a robí to na rôznych úrovniach

Na prvej úrovni spracovania sa generujú if-then pravidlá v rámci metódy nazvanej „*tvorba modelu úložiska*“. Táto metóda je schopná spracovať iba jeden zdroj a používa sa na vyplnenie prázdnych sub-matíc po importe dát z aktuálneho zdroja. Tieto sub-matice reprezentujú relácie medzi atribútmi odpovedajúcimi pozícii v matici. Táto pozícia (bunka matice) obsahuje informáciu o dvoch atribútoch, ktoré sú práve porovnávané. Ak všetky elementy (napríklad mestá v príklade na Obr.1.4), ktoré majú tú istú hodnotu prvého atribútu, sú v tej istej skupine, ktorá reprezentuje hodnoty druhého atribútu, potom je importovaná hodnota “1” do aktuálnej pozície (bunky) vo zvolenej prázdnej sub-matici. Hodnota “1” je zapísaná do riadku s aktuálnou pozíciou skupiny a stĺpa reprezentujúceho hodnotu prvého atribútu. Hodnota “0” je vpísaná do všetkých ostatných buniek v tom istom riadku sub-matice. Napokon sa hodnoty “-2” vpíšu do buniek, kde nie sú žiadne relácie. Táto hodnota reprezentuje fakt, že algoritmus nie je schopný nájsť v procese porovnávania atribútov žiadnu reláciu. Výsledkom tejto metódy je generovanie if-then pravidiel, ktoré sú primárne určené na použitie vo forme vstupov pre sofistikovanejšie metódy.

Na druhej úrovni spracovania bola navrhnutá metóda nazvaná “*presné porovnávanie*”, ktorá taktiež pracuje iba nad jedným informačným zdrojom. Vstupom tejto metódy sú výsledky spracovania metódou prvej úrovne, preto nie je potrebné znova načítavať zdroje, čo šetrí čas. Táto metóda odhaľuje závislosti medzi dvoma atribútmi monitorovaním ich chovania v rozličných štartovacích podmienkach. Táto

metóda sa zameriava na nájdenie toho istého chovania porovnávaním sub-matíc, ktoré boli vytvorené v rámci spracovania metódou prvej úrovne čo ilustruje Obr.1.6.

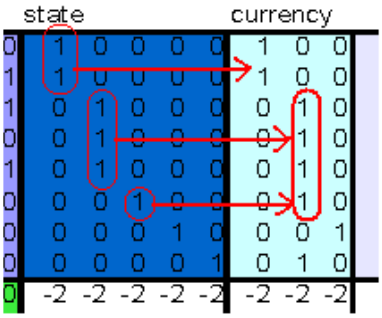
1	0	1	0	1
1	-2	-2	-2	-2
1	1	0	1	0
1	1	0	1	0

1	0	1	0	1
1	1	0	1	0
1	-2	-2	-2	-2
1	1	0	1	0
1	-2	-2	-2	-2

Obr. 1.6 Ilustrácia metódy „presné porovnávanie. V tomto prípade bola dedukovaná relácia medzi „Middle Europe“ a „Vinegar“.

Zo skutočnosti, že posledné dva stĺpce v oboch sub-maticiach, reprezentujúcich atribúty „Middle Europe“ a „Vinegar“, sú totožné, vyplýva vzťah medzi týmito dvoma atribútmi. Táto relácia je ľudskému čitateľovi zrejmá, ale nie je explicitne vyjadrená v zdroji ani v úložisku. Nevýhodou tejto metódy je to, že vyžaduje presnú zhodu, čo nemusí byť splnené, aj keď relácia existuje. Preto bola navrhnutá ďalšia metóda nazvaná „dynamické porovnávanie“

Na tretej úrovni spracovania sa používa spomenutá metóda „dynamického porovnávania“, používajúca špecifickú skupinu dát z matice úložiska. V tomto prípade môžu byť porovnávané sub-matice jednotlivých atribútov rozličných rozmerov. Nehľadá sa presná zhoda ale iba podobnosť vo výskyte a počte výskytov hodnôt „1“ v oboch sub-maticiach. Hľadajú sa podobné vzory, čo ilustruje Obr.1.7.



Obr. 1.7 Ilustrácia metódy „dynamického porovnávania.

Je faktom, že v metóde dynamického porovnania musia byť sledované počty dovolených odchýlok porovnávaných sub-matíc. Hodnota tohto limitu musí byť adekvátna dimenzii použitých zdrojov.

1.7 Systémy sémantického webu

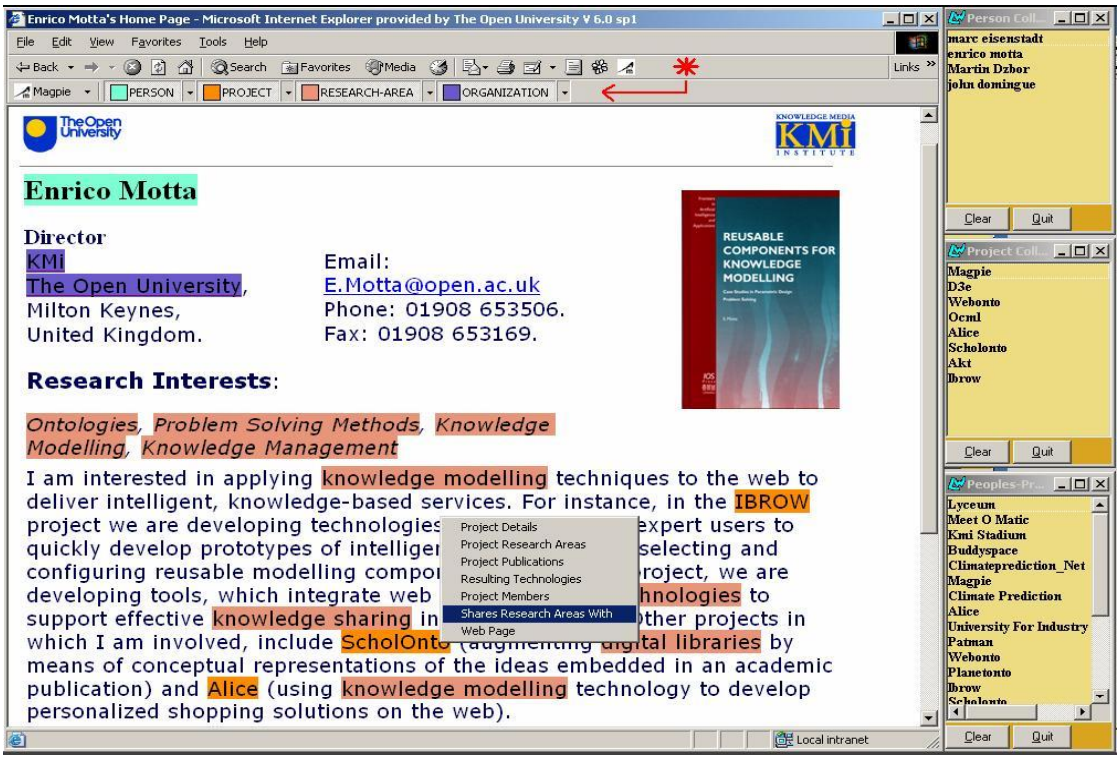
Dnes existuje rada systémov, ktorých činnosť je sémanticky obohatená a sú schopné vnímať sémantiku webových informácií. Nasledujú stručné charakteristiky niektorých z týchto systémov.

1.7.1 Magpie

Systém Magpie sa vyznačuje schopnosťou interpretácie obsahu webových stránok. Bol vyvinutý na Knowleadge Media Institut (KMi) na Open University vo Veľkej Británii [Domingue-Dzbor-Motta, 2004]. Magpie bol navrhnutý vo forme rozšírenia

k Internet Explorer-u, ktoré automaticky vytvára sémantickú vrstvu pre webové stránky použitím ontológie. Sémantická vrstva predstavuje vysvetlivky (komentáre, poznámky) k webovej stránke, s menu užitočných sémantických služieb. To je dôležité nielen pre nájdenie správneho webového zdroja ale aj pre pochopenie zmyslu jeho obsahu. Pochopeniu obsahu napomáhajú komentáre, ktoré spájajú meta informácie s webovými zdrojmi. Komentáre poskytujú užitočný spôsob podpory skupinovo založenej a zdieľanej interpretácie. Systém Magpie predstavuje jeden z prvých krokov k sémantickému prehliadania webu. Všetky schopnosti Magpie sa opierajú o ontologické rozhodovanie, ktoré umožňuje používateľom používať históriu nie len ako záznam aktivít používateľov, ale taktiež vidieť aj ich sémantiku.

Fungovanie tohto systému ilustruje Obr.1.8. Ak sa otvorí web stránka patriaca do domény pre ktorú má systém k dispozícii ontológiu, tak na tejto stránke sa farebne vyznačia všetky termíny, ktoré sa nachádzajú v danej ontológii. Na tieto podfarbené slová je možné kliknúť pravým tlačidlom, čo sprístupni menu, kde sa okrem iného nachádza možnosť vysvetlenia tohto slova.



Obr. 1.8 Ilustrácia činnosti systému Magpie.

Magpie potrebuje doplňujúci informačný zdroj vo forme ontológie s relevantnými informáciami. Magpie automaticky spojí vyhľadaný webový zdroj s jeho sémantickým obsahom, ktorý je obsiahnutý v ontologickom modeli. Dostupnosť sémantickej vrstvy teda závisí na dostupnosti ontológie pre daný webový zdroj. Magpie predstavuje nástroj na využívanie služieb sémantického webu a dokáže sa flexibilne prispôsobiť rôznym potrebám používateľov [Dzbor-Motta-Domingue, 2004].

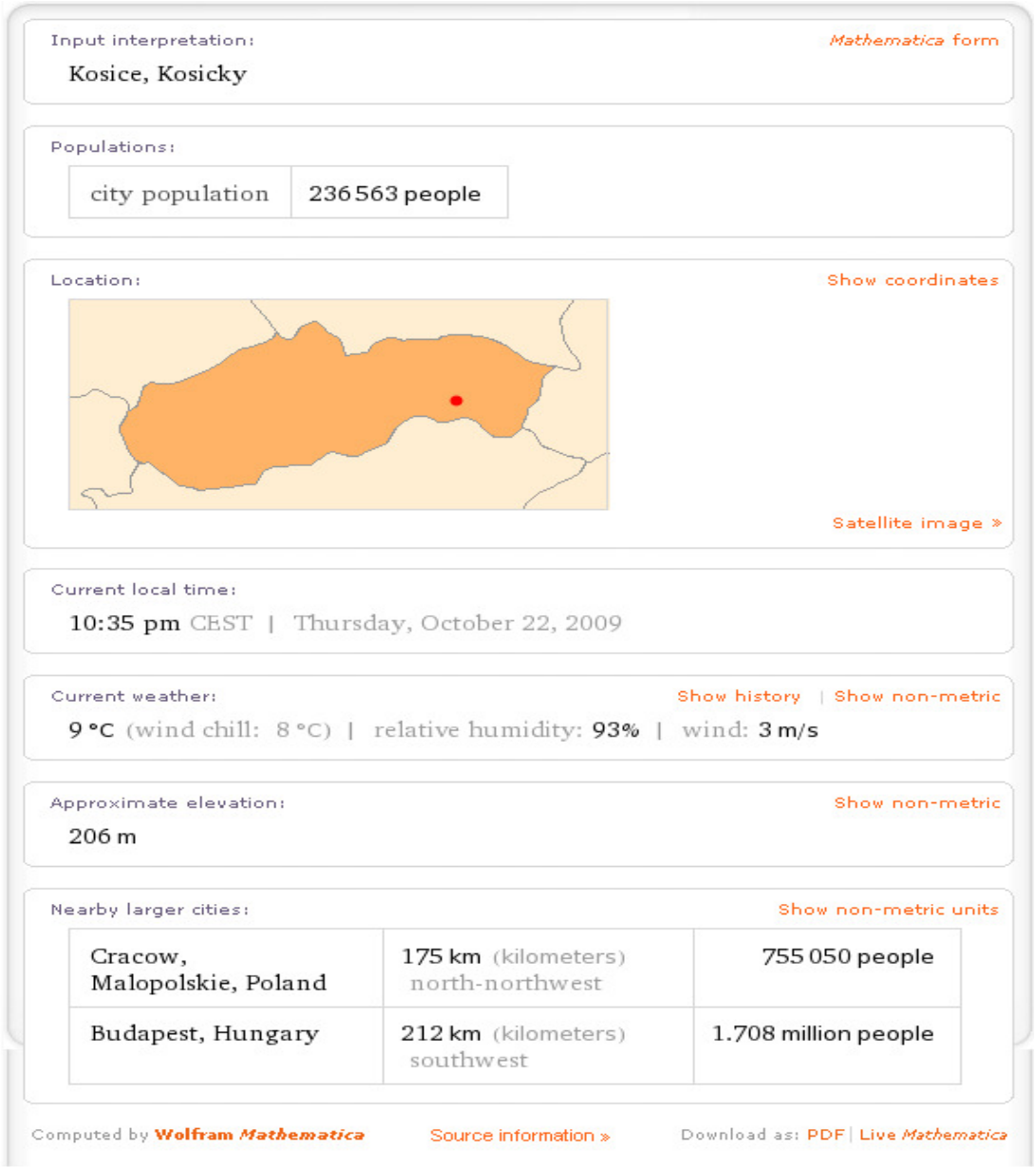
1.7.2 WolframAlpha

WolframAlpha je sémantický vyhľadávač vedomostí a informácií, keďže dokáže do určitej miery pochopiť otázku používateľa formulovanú v anglickom jazyku. Jeho tvorca ho však charakterizuje ako výpočtovo založený vedomostný stroj (Computational Knowledge Engine), keďže výsledok vyhľadávania systémom WolframAlpha nie je webová stránka ale priamo štruktúrovaná odpoveď. Je to akási zmes sémantického vyhľadávača a encyklopédie, ktorá môže fungovať iba vďaka používaniu vlastných obsiahlych databáz. Tvorcom tohto systému je britský matematik a fyzik Stephan Wolfran. Systém je silný hlavne v matematike, geografii, ekonomike a vede. Dokáže do určitej miery spracovať prirodzený jazyk, preto môže používateľ formulovať dotazy na informácie ako pri bežnej konverzácii v angličtine.

Používanie vyhľadávača sa rozšírilo s uvedením inteligentnej hlasovej asistentky SIRI (Speech Interpretation and Recognition Interface) v modeli mobilného telefónu spoločnosti Apple iPhone 4S. Od uvedenia služby prichádza veľká časť dopytov práve prostredníctvom tohto rozhrania. Okrem zadávania otázok podporuje WolframAlpha aj priame vloženie súboru dát, obrazovej informácie, alebo upload celého súboru, nad ktorými vykoná operácie a výsledok zobrazí.

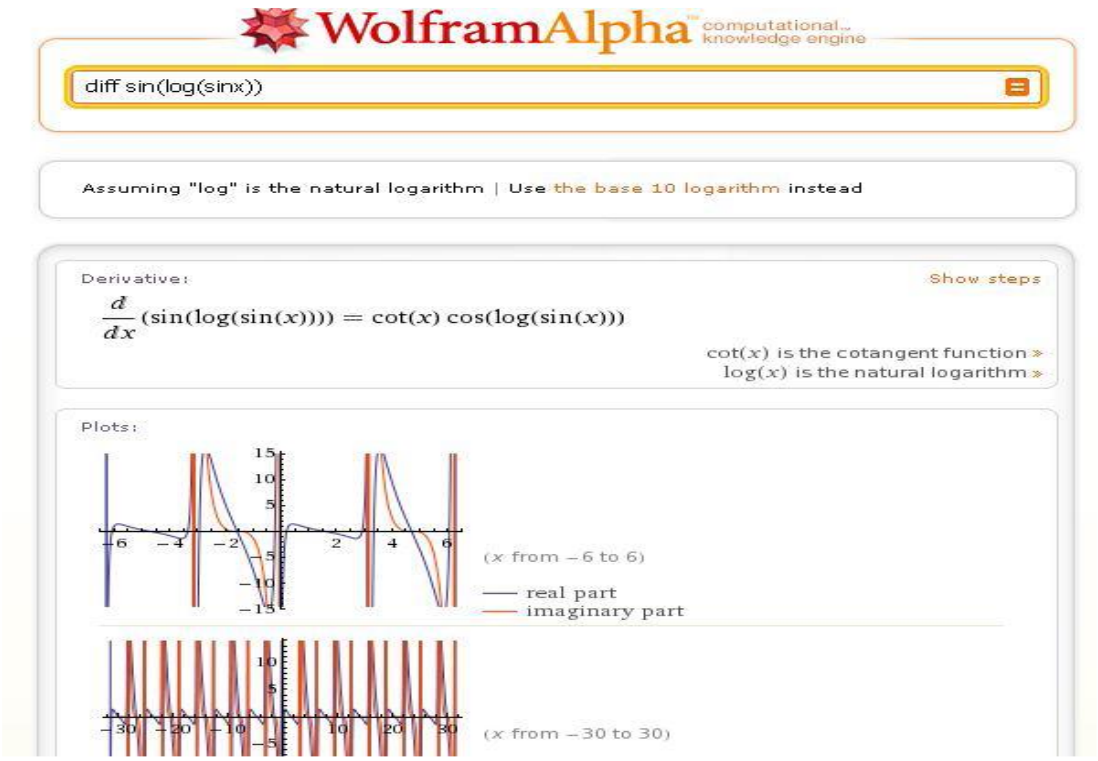
Jadro tohto systému tvorí systém *Mathematica,* ktorý bol navrhnutý pôvodne hlavne na vedecké výpočty. Samotný WolframAlfa preberá značnú časť funkcií práve zo systému Mathematica. Tieto funkcie sú založené na pavučine algoritmov, metód a modelov z rôznych oblastí vedy aj umelej inteligencie, ktoré boli vyvinuté v rámci takmer 20 ročného rozvoja robustných technológií vo Wolfram Research. WolframAlpha obsahuje viac ako 10 biliónov dát, 50 000 typov algoritmov, 5 miliónov riadkov kódu.

Ak sa systému zadá dotaz vo forme názvu mesta, je používateľovi vrátená mapa s lokalizáciou danej obce, počtom obyvateľov, aktuálnym počasím, časom, kedy zapadlo slnko a pod., čo je ilustrované na Obr.1.9.



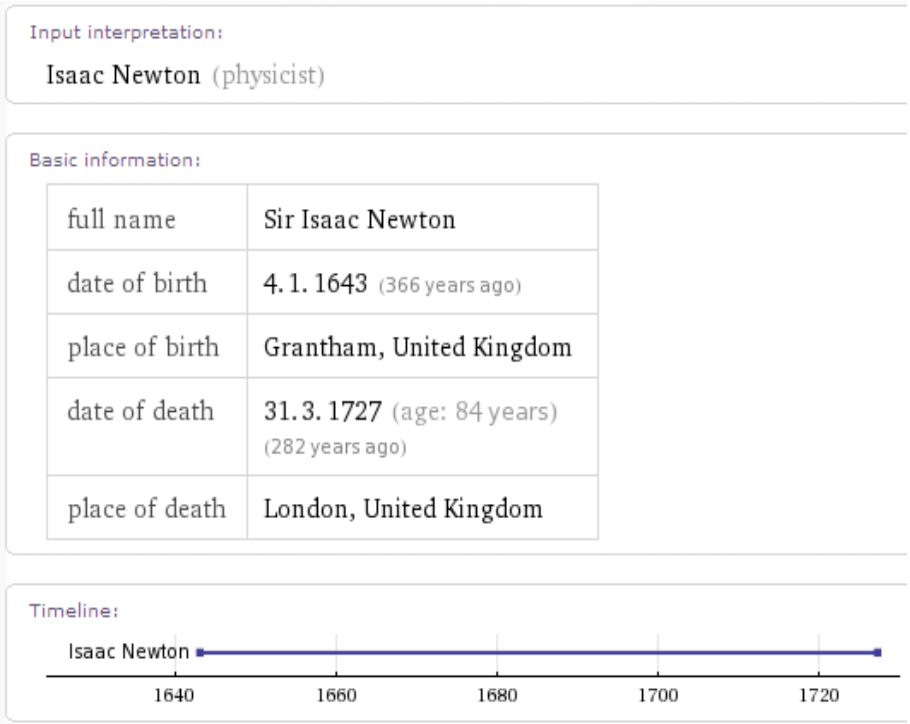
Obr. 1.9 Štruktúrovaná odpoveď systému WolframAlpha na dotaz „Košice, Košický“.

Systém poskytuje aj rôzne informácie z oblasti chémie: o chemických zlúčeninách, ľudskom genóme, zlate a pod. Taktiež prevody mien, fyzikálne prevody, prepočty molových hmotností, riešenie kvadratických rovníc, diferenciálnych rovníc, integrálov a iných komplikovaných matematických rovníc, čo dokumentuje obrázok Obr.1.10.

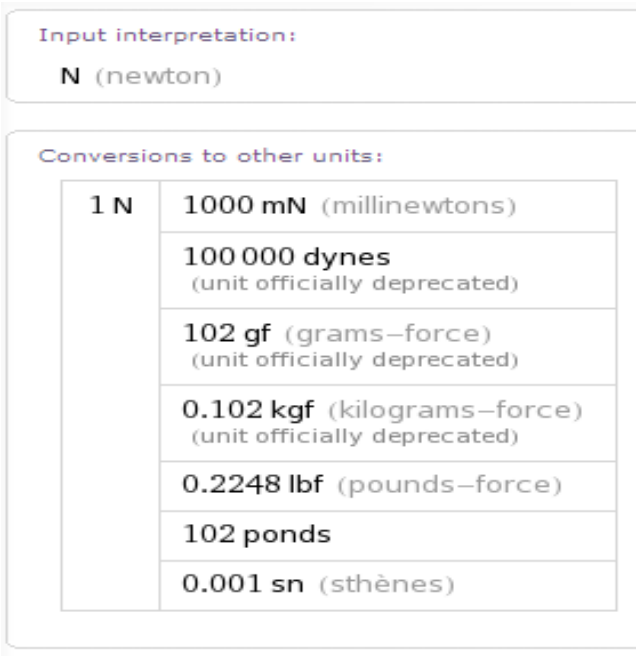


Obr. 1.10 Komplikovaná matematická rovnica.

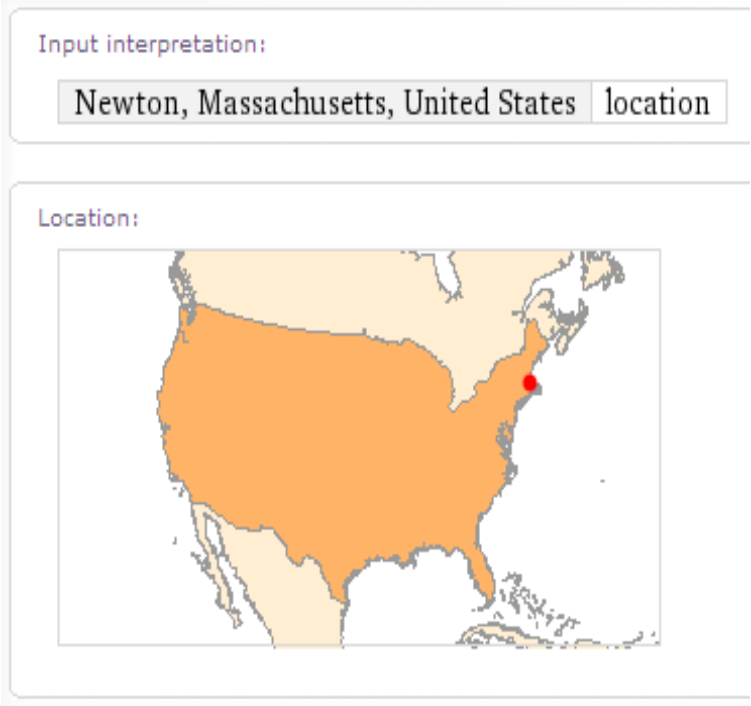
Systém WolframAlpha dokáže rozlíšiť otázky typu: kto?, čo? a kde?, čo ilustrujú nasledovné obrázky (Obr.1.11, Obr.1.12 a Obr.1.13).



Obr. 1.11 Odpoveď WolframAlpha na dotaz: Kto je Newton?



Obr. 1.12 Odpoveď WolframAlpha na dotaz: Čo je Newton?



Obr. 1.13 Odpoveď WolframAlpha na dotaz: Kde je Newton?

Systém WolframAlpha dokáže dávať odpovede aj na kombinované otázky, ako napríklad „Za aký čas by sme aktuálnym tempom zabehli maratón?“. Odpoveď na takúto otázku vyžaduje zistiť aktuálne tempo z nejakých meracích prístrojov, ktoré má bežec so sebou, ďalej je potrebné z internetu získať presnú dĺžku trate maratónskeho behu a napokon vypočítať odpoveď.

Ak by sme mali porovnať WolframAlpha a Google, ťažko by sme mohli dať

jednoznačnú odpoveď, ktorý systém je lepší. WolframAlpha prechádza pomocou robotov web a získané údaje ukladá do databáz. Tieto databázy neobsahujú iba indexy stránok, ale aj všetky užitočné informácie prebraté z obsahu, ktoré sa potom v štruktúrovanej forme predkladajú používateľovi. S týmito informáciami sa ďalej pracuje a dedukujú (respektíve počítajú) sa z nich nové dáta. WolframAlpha čerpá odpovede z databáz, nie priamo z Internetu. Odpovede nie sú webové stránky ale priamo informácie, ktoré používateľa zaujímajú.

Na druhej strane, Google čerpá informácie priamo z Internetu. Vracia linky na webové stránky, na ktorých používateľ pravdepodobne nájde, čo hľadá. Nepočíta matematické príklady ako WolframAlpha ani neupresňuje odpoveď pomocnými výpočtami. Google indexuje webové stránky na rozdiel od WolframAlpha. Čo majú spoločné je to, že sa snažia preraziť na mobilných platformách.

Klady a zápory vedomostného stroja WolframAlpha. Výhodou tohto systému je, že poskytuje overené a pravdivé informácie, odpovede sú do určitej miery vypočítané, umožňuje klasické matematické výpočty s grafmi, poskytuje revolučné vyhľadávanie doplňujúce Google, má encyklopedický charakter a výsledná odpoveď býva doplnená odkazmi na Wikipediu a Google. Na druhej strane nevýhodou je dlho trvajúce získavanie nevyhnutných záznamov, používanie databáz, potreba pravidelnej aktualizácie databáz, akceptácia otázok iba v anglickom jazyku a dlhodobosť vývoja.

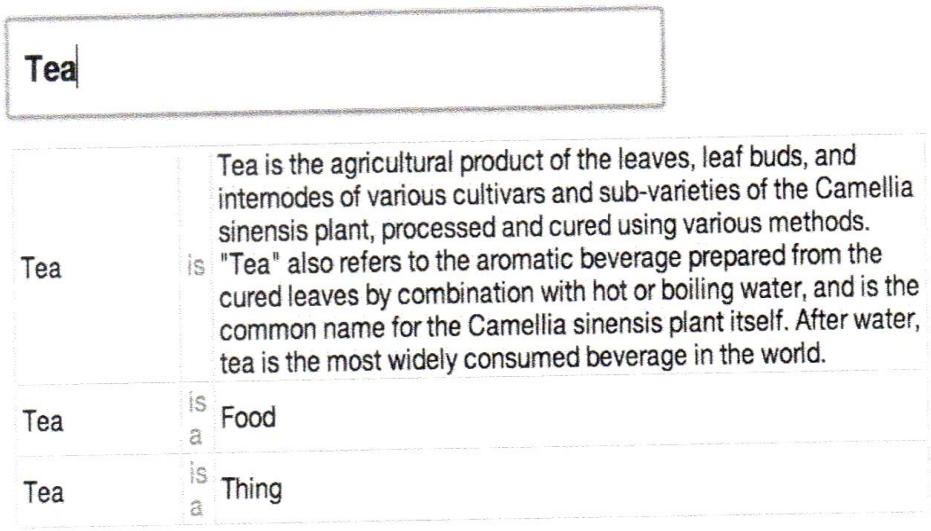
Neobvyklé a nové na tomto systéme je, že sa snaží niekedy odpoveď vypočítať, čo je naozaj v niektorých prípadoch lepšie. Skúsme si predstaviť, že namiesto toho, aby sme vypočítali súčin dvoch čísel, mali by sme tabuľku, v ktorej by boli uložené všetky kombinácie súčinov dvoch čísel a v nej by sme realizovali vyhľadávanie. Keďže čísel je nekonečne veľa, bolo by nemožné takúto tabuľku zostrojiť. Softvér Mathematica je významným a rešpektovaným nástrojom na vysoko úrovňové spracovávanie a modelovanie rôznych matematických, inžinierskych, finančných a iných údajov.

1.7.3 Sémantický vyhľadávač SWSaDS

Sémantický web musí sprostredkovať informácie spolu s ich sémantikou nie len ľuďom ale aj strojom. Potrebuje vyhľadať, priniesť a integrovať množstvo dostupných sémantických dát roztrúsených po rôznych doménach a dátových skladoch. Preto je potrebné riešiť problém efektívneho prehľadávania týchto dát ako celku. Naviac sa od neho očakáva, že bude schopný fungovať aj za hranicami dátového skladu, nad ktorým bol implementovaný.

Systém SWS*a*DS (Semantic Web Search with the *aid* of Data Storage) predstavuje webovú aplikáciu ktorá vyhľadáva informácie na webe za pomoci existujúcich sémantických dátových skladov tretích strán a poskytuje výsledok tohto hľadania vo forme štruktúrovanej sémantickej informácie, teda nie webovej stránky, podobne ako WolframAlpha. Výsledok nie je závislý na dátovom sklade a teda presahuje hľadanie v dátovom sklade. Ak zlyhá prístup k jednému skladu, aplikácia môže fungovať ďalej pomocou zvyšných skladov.

Ako je ilustrované na Obr.1.14, aplikácia ponúkne používateľovi možnosť zaznamenať hľadaný pojem, ktorý je najprv vyhľadaný v niektorej z existujúcich databáz, ako DBPedia alebo Freebase. Tak sa získa korešpondujúci RDF súbor, ktorý je následne parsovaný na RDF výroky. Tieto výroky sú zoradené tak, aby na prvých miestach boli výroky, ktoré zrozumiteľne popisujú zadaný pojem.



Obr. 1.14 Používateľské rozhranie vyhľadávača SWSaDS.

Výhodou tohto sémantického vyhľadávača je, že ako štartovací bod využíva veľkú databázu sémantických dát, no kedykoľvek narazí na výrok, ktorý odkazuje na zdroj mimo tejto databázy, aplikácia funguje ďalej nezávisle od tejto databázy ako webový prehliadač.

1.7.4 Hry s Účelom

Objavovanie sémantiky a vzťahov medzi pojmami je možné aj pomocou zaujímavých počítačových hier, ktorým sa hovorí „hry s účelom“ („games with a purpose“). S týmto názvom sa spája aj termín “human computation”, keďže sa využíva potenciál ľudského mozgu na riešenie problémov, ktoré nedokážu riešiť počítače. Dokážu spojiť dve na prvý pohľad protichodné oblasti a to zábavu a riešenie úloh. Príkladom takej hry je získavanie anotácií obrázkov.

Web sa síce postupne vyvíja, ale signifikantná väčšina internetových stránok súčasnosti zatiaľ ešte nie je pripravená pre sémantické vyhľadávanie, nakoľko ich tvorí iba text bez potrebných metadát. Vyhľadávanie založené na indexácii stránok je z tohto dôvodu v súčasnosti najrozšírenejším typom vyhľadávania. Z toho vyplýva, že určenie správnych kľúčových slov je veľmi dôležité pre získanie relevantných výsledkov vyhľadávania. Z tohto dôvodu je dôležité zaoberať sa sémantickou anotáciou webového obsahu. Ak autor stránky optimalizuje stránku pre vyhľadávanie, získava tým výhodu oproti stránkam, ktoré takto optimalizované nie sú.

Problém nastáva pri obsahu, ktorý nie je možné indexovať a priradiť mu anotáciu. Často krát sa jedná o multimediálny obsah, ktorý je potom pre vyhľadávacie stroje neznámy. Takýto obsah je potrebné anotovať manuálne, čo je zdĺhavá a finančne náročná činnosť. Z tohto dôvodu sa rozšírila metóda získavania sémantiky na webe prostredníctvom hier. Jej hlavnými výhodami je získavanie anotácií prevažne multimediálneho obsahu formou hry, ktorá je zábavná pre hráča, ale zároveň vytvára užitočné anotácie pre jej sprostredkovateľa. Tento spôsob získavania

anotácií má vysoký potenciál, nakoľko denne sa odohrá veľké množstvo hodín počítačových hier. Tieto hry pritom nie sú iba konzolové, ale postupne s masívnym rozšírením Internetu aj takzvané on-line hry. Na základe zložitosti implementácie môžeme rozdeliť hry s účelom na hry, ktoré sa pripájajú na server, a na hry, ktoré bežia lokálne na strane klienta.

Pri prvom type hier prebieha celá logika hry na serveri a jednotliví hráči sa pomocou internetového prehliadača do tohto virtuálneho herného sveta pripájajú. Takýto typ hier sa označuje skratkou MMOG („Massively Multiplayer Online Game“). Dovoľuje veľkému počtu hráčov podieľať sa na hre a vytvárať tak virtuálny svet.

Logika druhého typu hier prebieha priamo na strane klienta bez komunikácie zo serverom a často sa jedná o hry, ktorých implementácia je jednoduchšia. Predmetom týchto hier je často logické myslenie, ako aj skúška šikovnosti a postrehu hráčov. Tieto hry používajú na svoj beh HTML skriptovacie jazyky, ako napríklad PHP, alebo JavaScript. S rozmachom sociálnych sietí sa spopularizovali aj hry, ktoré tieto siete ponúkajú. Ponúkajú totiž možnosť spoločného hrania s ostatnými členmi sociálnej siete.

Z jednoduchej dostupnosti on-line hier vyplýva aj ich kritizovaná nevýhoda, ktorou je hranie na úkor produktívnej činnosti, napríklad v zamestnaní. Spravodajský portál CNN upozornil, že pri nasadení jednoduchej on-line hry Pac-Man namiesto loga na stránke vyhľadávača Google, bola týmto spôsobená strata vo výške 120 483 800 dolárov, nakoľko sa používatelia zdržali istý čas hraním tejto hry. Pridaním účelu do hier tohto typu, by sa dal tento ľudský výpočtový potenciál využiť v prospešnú vec.

1.7.5 Návrh hier s účelom

Dnes existujú sémantické vyhľadávače, ktoré využívajú schopnosť používateľov internetu, prirodzene chápať sémantiku. Preto nechávajú človeka definovať sémantiku objektov akosi mimochodom ako vedľajší produkt nejakej zaujímavej počítačovej hry. Tomu sa hovorí „*hra s účelom*“. Takúto hru je potrebné navrhnúť tak, aby bola zábavná. Chybovosť hry môže vyvolať nezáujem u hráčov o celú hru. Taktiež, ak je zavedený systém hodnotenia hráčov nespravodlivý, hráč nebude prejavovať záujem o hru. Aby sme dokázali náležite zvážiť všetky prvky hry, je potrebné najprv definovať hru s účelom a jej cieľ. Cieľom každej hry s účelom je v nejakej podobe vytvárať artefakty – produkty ľudskej činnosti, použiteľné potom pri riešení reálnych problémov (napríklad metadáta k obrázkom). Úspešnosť hry závisí od jej kvality (správnosti) a kvantity. Kvalita artefaktov je podmienená samotným herným systémom (dynamika a mechanika) hry, teda akým spôsobom donúti, či motivuje hráčov vytvárať správne artefakty alebo záznamy, z ktorých sa artefakty odvodia dodatočnou automatickou metódou. Kvantita je zasa podmienená schopnosťou hry pritiahnuť a udržať hráčov [Šimko, 2011]. Kľúčové aspekty pre vytvorenie hier s účelom sú nasledovné:

- definícia problému
- priraďovanie úloh hráčom
- validácia artefaktov
- zábavnosť hry
- zamedzenie podvádzaniu.

Návrh hry s účelom pre anotáciu textových dokumentov. Táto hra je navrhnutá pre získavanie kľúčových slov z krátkych textových úryvkov. Prezentovaná hra je

zameraná na použitie v školskom prostredí so zameraním na žiakov základných škôl.

Účelom hry je zapojiť žiakov do vyučovacieho procesu zábavným spôsobom a teda riešiť súčasný pálčivý problém žiakov s porozumením písaného textu. Vedľajším produktom hry je nájdenie charakteristických kľúčových slov textu.

Definícia problému spočíva v kolaborácii dvoch hráčov, ktorí vyberajú na základe krátkeho textového úryvku päticu slov, ktoré najvýstižnejšie reprezentujú význam zadaného textu. Slová na ktorých sa žiaci vo dvojici zhodnú, sú uložené a považované za kľúčové slová pre vyhľadávanie v týchto textoch.

Priraďovanie úloh hráčom je v tomto prípade jednoduché, lebo všetci hráči majú rovnakú úlohu. Hra prebieha vo dvojiciach.

Validácia artefaktov prebieha na základe zhody dvoch hrajúcich hráčov na jednom pojme.

Zábavnosť hry je zabezpečená nepriamou konfrontáciou s protihráčom, ktorý motivuje hráčov podávať čo najlepšie výsledky.

Zamedzenie podvádzaniu je jednoducho zabezpečené tým, že vytvorenie nesprávnych anotácií na základe vopred dohodnutých slov je jednoducho odhaliteľné na základe porovnávania s doménou slov vytvorenou dlhodobým zhromažďovaním výsledkov.

Hra bola implementovaná v programovacom jazyku Ruby. Na komunikáciu medzi dvoma hráčmi sa používa „socket“ komunikácia. Súbežné zadávanie textu a počúvanie na porte pri komunikácii zabezpečuje samostatné vlákno. Získané kľúčové slová ako aj slová, ktoré zadali jednotliví používatelia sú uložené v samostatných textových súboroch. Príklad výsledkov získaných v tejto hre pre text z biológie o stromoch je uvedený v Tab.1.1. Zvýraznené slová sú zvolené ako kľúčové.

Tab. 1.1 Príklad výsledkov získaných v hre pre získavanie kľúčových slov z krátkych textových úryvkov.

Hráč 1	Hráč 2
strom	kmeň
zeleň	strom
listy	les
drevo	drevo
fotosyntéza	fotosyntéza

Návrh hry s účelom pre zdieľanie vedomostí Knowledge portál Hra je určená primárne na zdieľanie vedomostí formou hry na otázky a odpovede. vo vlastnej oblasti záujmu. Odpovede sú hodnotené zadávateľom otázky, čím dochádza k triedeniu a následnej filtrácii odpovedí.

Účelom hry je získať odpovede a popri tom aj kľúčové slová pre implementáciu vyhľadávania. Vyhľadávanie je založené na odpovediach, v ktorých sa nevyskytujú vyhľadávané slová priamo, ale iba slová, ktoré sú s vyhľadávanými slovami spárované. Párovanie kľúčových slov nie je jednoduché ale je to proces, pri ktorom

dochádza ku párovaniu slov so stále vhodnejšími a výstižnejšími odpoveďami. Web predstavuje v súčasnosti jeden z najväčších digitálnych informačných priestorov. Špeciálne postavenie majú digitálne knižnice, ktoré čiastočne môžu byť súčasťou webu. V tomto priestore cestuje záujemca o informáciu. Prezentovaný prístup nanovo otvára potrebu podpory, navigácie, personalizáciu a zohľadnenie kontextu. Cestovateľ v informačnom priestore nachádza a zanecháva značky, anotácie, hodnotenia a odporúčania s cieľom efektívneho prístupu k informáciám [Návrat et al., 2011].

Definícia problému predstavuje urýchlenie prístupu ku relevantným odpovediam na zadanú otázku na základe vytvorených a spárovaných kľúčových slov zadaných spolu s otázkou a odpoveďou. Spolu s otázkou sa zadáva aj kategória a kľúčové slová asociované z otázkou. Takto zozbierané otázky sú rozdelené do jednotlivých kategórií (v našom prípade do kategórií podľa programovacích jazykov). Princíp hry spočíva v hodnotení odpovedí pýtajúcimi sa používateľmi, ktorí hodnotia nakoľko odpovede jednotlivých odpovedajúcich hráčov riešia problém, ktorý formulovali v otázke. Tento prvok motivuje hráčov k vzájomnej súťaživosti v poskytovaní, čo najlepších odpovedí. Pri kontrolovaní svojho skóre majú odpovedajúci možnosť konfrontovať odpovede s ostatnými a tým sa zlepšovať v oblasti svojho záujmu. Odpovedajúci k svojím otázkam priradzujú kľúčové slová súvisiace s odpoveďou (nie otázkou).

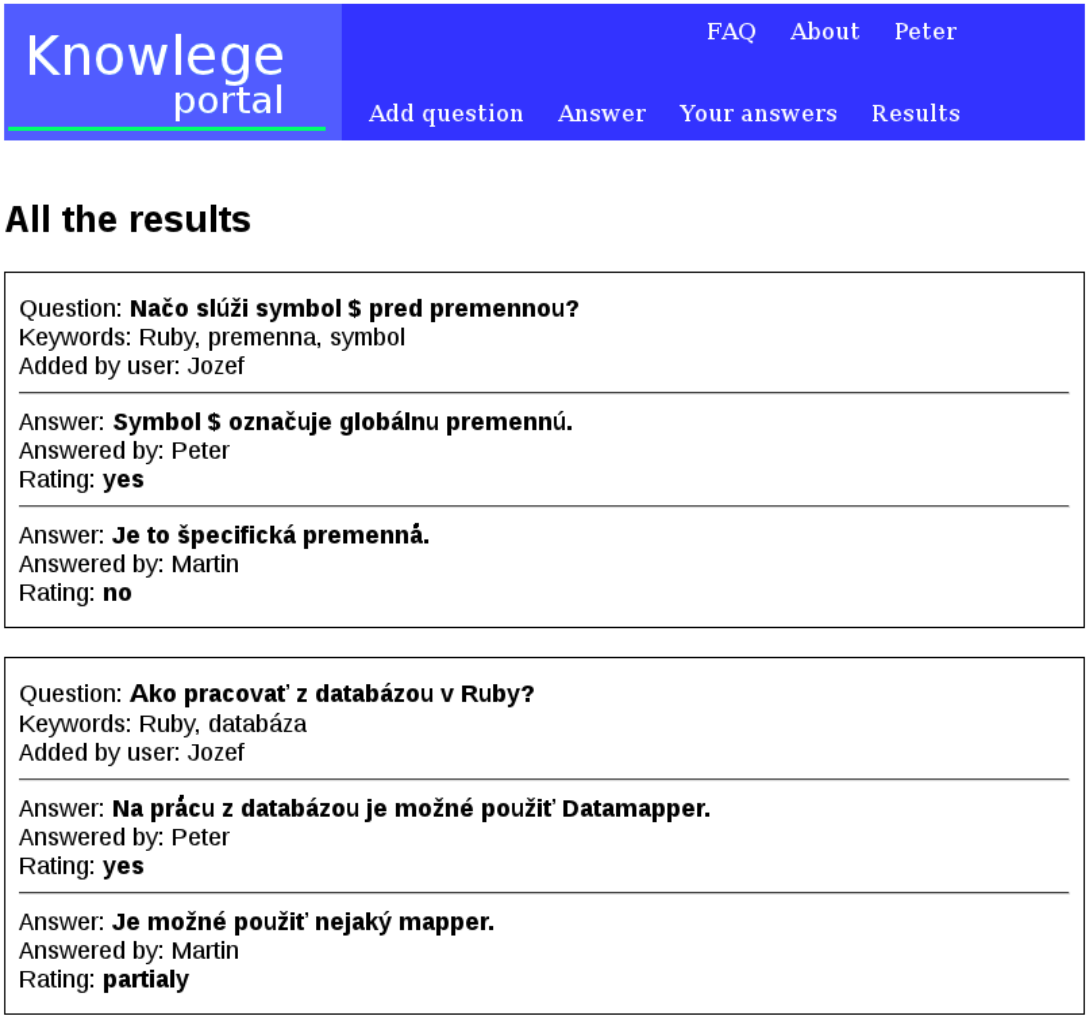
Úlohy sú priradzované automaticky formou požiadavky na vyplnenie kľúčových slov pri otázke a odpovedi.

Validáciu artefaktov vykonáva zadávateľ otázky pri kontrole odpovedí a odpovedajúci pri zadávaní kľúčových slov.

Zábavnosť hry je založená na súťaživosti odpovedajúcich a ich pocite úspechu pri riešení problému ako aj pri dosiahnutí pozitívnych ohlasov na odpoveď. Motiváciou pre hráčov je ochota pomôcť a možnosť získania nových poznatkov.

Zamedzenie podvádzaniu. V tomto prípade používatelia nemajú motiváciu podvádzať.

Hra je implementovaná v jazyku Ruby ako webová aplikácia. Využitý je „Framework Sinatra“ v kombinácii so značkovacím jazykom ERB („Embedded Ruby“). Otázky, odpovede ako aj registrácie užívateľov sú uložené v databáze „sqlite3“. Príklad grafického rozhrania hry je ilustrovaný na Obr.1.15.



Obr. 1.15 Používateľské rozhranie hry pre anotáciu textov.

Ďalším stupňom vývoja hier s účelom by mohlo byť ich širšie nasadzovanie v reálnom živote. Príjemnou formou totiž môžu ponúkať riešenie širšieho spektra úloh.

POUŽITÁ LITERATÚRA

[Antoniu-vanHarmelen, 2004] Antoniu, G., van Harmelen, F.: *A Semantic Web Primer*. Massachusetts Institute of Technology, USA, 2004, 238 pp., ISBN 0-262-01210-3.

[Domingue-Dzbor-Motta, 2004] Domingue, J.B., Dzbor, M., Motta, E.: Collaborative Semantic Web Browsing with Magpie, In Proc. of the 1st European Semantic Web Symposium (ESWS), May 2004, Greece.

[Dzbor-Motta-Domingue, 2004] Dzbor, M., Motta, E., Domingue, J.B. Opening Up Magpie via Semantic Services, In Proc. of the 3rd Intl. Semantic Web Conference, November 2004, Japan.

[Finin at al., 2006] T. Finin, at al., "Information Integration and the Semantic Web", Workshop on Information Integration, 2006.

[Gruber, 1993] Gruber, T.R.: A translation approach to portable ontology specifications. In Knowledge Acquisition, 5(2), 199-220, 1993.

[Machová-Fodorová, 2011] Machová, K., Fodorová, D.: Knowledge Discovery from Repository of Web Information. In: American Journal of Intelligent Systems Vol. 1, no. 1(2011), p. 37-42, ISSN 2165-8978.

[Návrat et al., 2011] Návrat, P. et al.: Od surfovania k personalizovanému cestovaniu v digitálnom priestore. In: 6th Workshop on Intelligent and Knowledge oriented Technologies - WIKT 2011, 24. - 25.11.2011, Herľany, EQUILIBRIA, s.r.o., Košice, 2011, ISBN 978-80-89284-99-3.

[Paralič at al., 2010] J. Paralič, K. Furdík, G. Tutoky, P. Bednár, M. Sarnovský, P. Budka, F. Babič, “Knowledge mining from texts“, Equilibria, s.r.o., Košice, 2010, 80 ps, ISBN 978-80-89284-62-7.

[Studer-at al., 1998] Studer, R., Benjamins, R., Fensel, D.: Knowledge engineering: Principles and methods. Data & Knowledge Engineering, 25(1–2):161–198, 1998.

[Šimko, 2011] Šimko, J.: Perspektívy hier s účelom pre vytváranie metadát. In: 6th Workshop on Intelligent and Knowledge oriented Technologies - WIKT 2011, 24. - 25.11.2011, Herľany, EQUILIBRIA, s.r.o., Košice, 2011, ISBN 978-80-89284-99-3.

2 SOCIÁLNY WEB

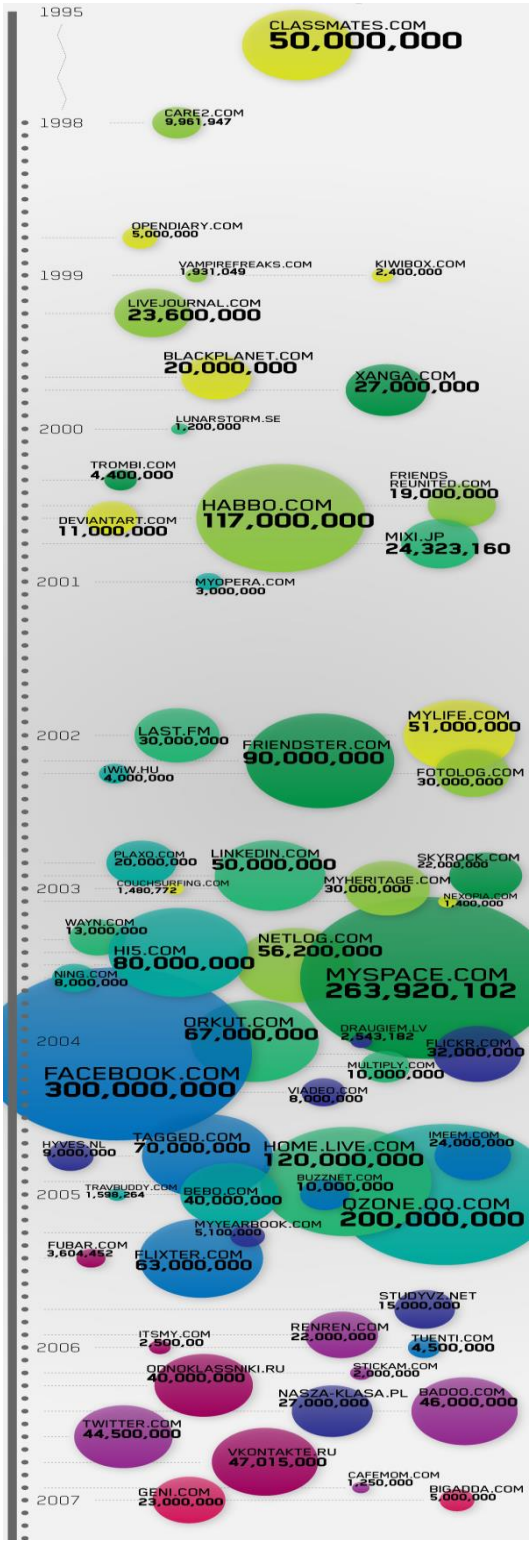
2.1 Úvod

V dnešnej dobe informácií a internetu sa ľudia pri riešení rôznych tak každodenných ako aj profesijných problémov utiekajú k Internetu, aby tam podľa možnosti efektívne vyhľadali potrebné fakty, informácie a vedomosti. Ich zdrojom môžu byť rôzne webové zdroje. S vývojom sociálneho webu začali ľudia okrem toho vyhľadávať aj názory a postoje k nejakej entite (napr. politik, typ počítača či mobilného telefónu a pod.), ktoré môžu zohrať dôležitú úlohu v procese rozhodovania. Zdrojom týchto názorov môžu byť sociálne siete, webové fóra, a pod. Množstvo príspevkov vo webových diskusiách obsahujúcich relevantné názory môže presahovať desiatky až stovky, čo je časovo náročné na prečítanie. Preto sa výskum v oblasti sociálneho webu zameriava okrem iného aj na automatickú softvérovú extrakciu týchto názorov a postojov používateľov webu.

Sociálny web v porovnaní s klasickým webom zrovnoprávňuje počet producentov takzvaného konverzačného obsahu s počtom konzumentov tohto obsahu. V súvislosti s tým sa vynára obava, aby neproduktívne časti sociálneho webu, (povedzme, že balast), nepreválili jeho produktívne časti.

Sociálny web predstavuje prechod od sociológie bežného života k jej reprezentácii prostredníctvom informačných technológií, napríklad vo forme nejakej sociálnej platformy. Prudký nárast sociálnych platforiem v čase demonštruje Obr.2.1.

Typickým reprezentantom sociálneho webu je napríklad sociálna sieť. Pod pojmom sociálna sieť rozumieme štruktúru, ktorá pozostáva z množiny sociálnych subjektov reprezentovaných uzlami siete (používatelia webu alebo organizácie) a prepojení medzi nimi pomocou rozličných relácií, ako sú: priateľstvo, príbuzenstvo, obchod, vízie, nápady atď. [Pénzeš, 2010].



Obr. 2.1. Ilustrácia prudkého nárastu platforiem sociálneho webu.

Typickými predstaviteľmi sociálneho webu sú:

- sociálne siete
- blogy, mikroblogy
- chat, chatrooms, IRC („Internet Relay Chat“)
- sociálne webové aplikácie
- diskusné fóra
- komentáre k príspevkom v internetových novinách
- wiki
- atď.

Sociálny web by mal uľahčovať vznik nových vzťahov medzi svojimi používateľmi, spájať ľudí a tvoriť trebárs aj nové sociálne interakcie s odozvou v reálnom živote. Sociálny web je reprezentovaný aj webovými aplikáciami umožňujúcimi vytváranie „virtuálnych“ sociálnych spoločenstiev. V rámci sociálneho webu vzniká obrovské množstvo väčšinou krátkych príspevkov či už v „point-to-point“ alebo „multicast“ konverzáciách.

Komunikácia, či už verbálna alebo neverbálna je neoddeliteľnou súčasťou života každého človeka. V posledných rokoch sa vďaka sociálnemu webu aj neverbálna komunikácia stala veľmi obľúbenou a záujem o akúkoľvek jej formu či platformu stále rastie. Jednotlivci sa v prostredí virtuálneho sveta spájajú do rozličných skupín a zapájajú sa do rôznych diskusií k aktuálnym témam. Napokon nemôžeme opomenúť, že tento virtuálny svet, či chceme alebo nie, ovplyvňuje ten reálny.

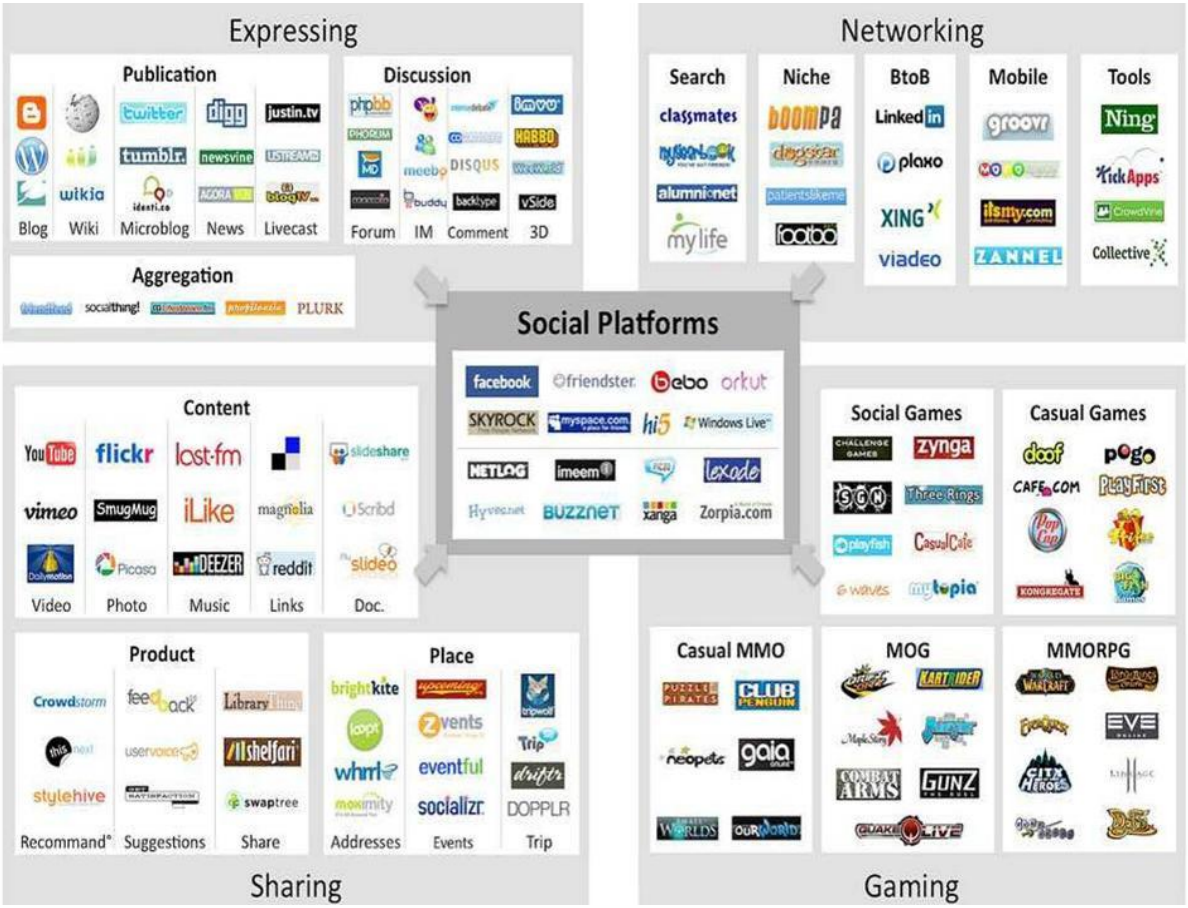
2.2 Platformy sociálneho webu

V rámci sociálneho webu je možné identifikovať rôzne typy sociálnych platforiem. Tieto platformy sa zameriavajú na:

- **vyjadrovanie**:
  - publikovanie („blog, wiki, news, livecast“)
  - diskutovanie (fórum, „instant message“)
  - agregovanie
- **zdieľanie:**
  - obsah (video, fotografie, hudba, dokumenty)
  - produkt (odporučenia, názory, postoje)
  - miesta (adresy, príležitosti, výlety)
- **sieťovanie:**
  - vyhľadávanie
  - „niche“
  - „business to business“
  - mobility
  - nástroje
- **hranie**
  - sociálne hry
  - príležitostné hry

- príležitostné on-line hry pre viacerých hráčov
- on-line hry pre viacerých hráčov
- on-line hry pre viacerých hráčov (hráči majú rozličné úlohy)

Reprezentantov spomenutých platforiem je možné vidieť na Obr.2.2.



Obr. 2.2. Ilustrácia rozličných platforiem sociálneho webu.

Podľa [Boyd-Ellison, 2014] predstavuje sociálny web internetové služby umožňujúce používateľovi nasledovné:

1. Vytvoriť verejný alebo poloverejný profil, ako formu reprezentácie jedinca. Súčasťou tohto profilu môžu byť informácie dotvárajúce individualitu používateľa (obľúbený citát, fráza a pod.).
2. Vytvoriť a editovať zoznam používateľov, zahrnutých vo svojej sociálnej sieti.
3. Prezerať a prechádzať zoznamy priateľstiev iných používateľov.
4. Zdieľať názory, príspevkov, fotografií, videí a komentárov.

Vlastnosti sociálnych webových platforiem, ktoré uľahčujú ich úspešné napredovanie sú nasledovné [Donah, 2014]:

- **životnosť** – Vyjadrenia ostatných používateľov sú ukladané a zobrazované, čo umožňuje tvorbu asynchrónnej komunikácie a tým predlženie komunikačného aktu.

- **dostupnosť** – Uchovávanie vyjadrení od používateľov umožňuje ľahké vyhľadávanie. Nájdenie osoby na webovej sociálnej stránke je otázkou stlačenia pár kláves.
- **replikovanie** – Verejné vyjadrenia sú kopírované na iné miesta bez toho aby sa dal rozoznať originál od kópie.
- **neviditeľnosť** – Prezeranie profilov ostatných používateľov neinformuje majiteľov profilov.

Primárnou úlohou sociálnych sieti je spájanie ľudí a vytváranie nových virtuálnych spoločenstiev ako aj uľahčovanie komunikácie medzi používateľmi a napomáhanie pri udržiavaní vzťahov. Sociálne siete ako také sú bohatým zdrojom informácií a teda poskytujú mnoho príležitostí na dolovanie poznatkov rôzneho druhu.

2.3 Sociálne siete

Sociálnu sieť môžeme najjednoduchším spôsobom definovať ako množinu uzlov (reprezentujú aktérov sociálnej siete), ktoré sú navzájom poprepájané hranami (reprezentujú vzťahy). Aktéri predstavujú členov siete, ktorí sa zapájajú do diskusií v rámci danej sociálnej komunity. Vzťahy, alebo inak povedané vzťahové väzby, zohľadňujú sociálne väzby medzi aktérmi a sú reprezentované reláciami špecifického typu (napríklad priateľstvo). Interakcie každého človeka s okolitým svetom môžu byť použité ako zdroj dát pre sociálnu sieť. Rozpoznávame dve skupiny používateľov sociálnej siete:

- **priateľ** je používateľ, ktorého autor stránky zverejnil na svojej profilovej stránke ako priateľa (pridal si ho k priateľom na sociálnej sieti),
- **influencer** je používateľ, ktorý svojím pôsobením získal pozornosť ostatných používateľov (odkazujú sa na jeho príspevky, na jeho publikačnú činnosť).

Z hľadiska domény rozpoznávame širokú škálu rôznych sociálnych sietí, napríklad technologické, ekonomické, biologické a pod. Oveľa dôležitejšie hľadisko pri delení sociálnych sietí je to, či máme v sieti aktérov jedného druhu alebo viac druhov (množín) aktérov. Z tohto hľadiska môžeme sociálne siete deliť na:

- jedno - módové
- dvoj - módové
- viac - módové.

Dominantný typ sociálnej siete predstavujú práve *jedno-módové siete*, kde vystupuje práve jedna množina aktérov, napríklad iba ľudia, iba organizácie, iba štáty a pod. Relácie môžu byť rôzneho druhu. Príkladom jedného druhu relácií môže byť individuálne ohodnotenie vzťahu, napríklad priateľstvo, sympatia alebo rešpekt. Iné druhy relácií môžu byť: transakcie, presun nehmotných zdrojov, príbuznosť alebo rôzne interakcie.

Dvoj - módové siete pracujú s dvoma množinami aktérov alebo s jednou množinou aktérov a jednou množinou udalostí. V prípade dvoch množín aktérov sa jedná prevažne o siete popisujúce napríklad množinu pedagógov a množinu študentov. Dvoj - módové siete, ktoré zohľadňujú jednu množinu aktérov a druhú množinu udalostí sa označujú ako *pridružujúce alebo členské siete*. Aktéri zo známej množiny aktérov sú vzájomne prepojení na základe vzťahov ku množine udalostí, alebo množine organizácií. Takáto sieť potom vyjadruje napríklad spoločnú účasť na udalostiach alebo spoločnú príslušnosť ku organizácii [Paralič, 2011].

2.3.1 História vzniku sociálnych sietí

Z Sociálne siete, ktoré sa zameriavajú na interakcie medzi používateľmi je možné považovať za webové stránky. Historicky prvou sociálnou sieťou tohto druhu bola on-line zoznamovacia služba „match“. Novšie formy takýchto sietí ponúkajú aj službu organizácie stretnutí v reálnom živote ako „fruehstueckstreff“. Niektoré majú dokonca tendenciu asistovať pri cielenej lokalizácii mobilných telefónov (napríklad „dodgeball“) ak to používateľ dovolí. Boli vytvorené aj sociálne siete umožňujúce udržiavanie kontaktov po ukončení školy, napríklad „friendsreunited“ a „spoluziaci.sk“. Neskôr vznikli aj profesijné sociálne siete, ktoré obsahujú osobné životopisy a pomáhajú ľuďom nájsť si prácu alebo pracovne komunikovať a vymieňať si vedomosti a nápady ako „linkedin“ a „iamResearcher“. Takéto siete môže samozrejme využiť aj nejaký zamestnávateľ na vyhľadávanie vhodných zamestnancov.

Také sociálne siete, ktoré umožňujú pozývať svojich priateľov na rôzne akcie, hrať hry, zdieľať s nimi fotografie, videá, textové správy a pod., majú najväčší úspech. Môže ísť o sociálne siete rôznych kategórií:

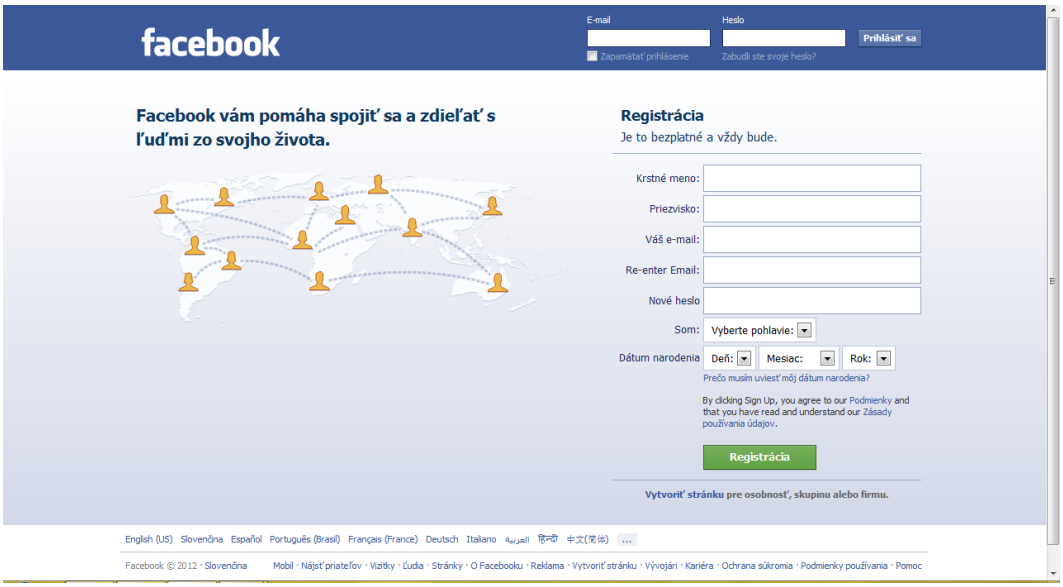
- so všeobecným charakterom „facebook“
- určené pre tínedžerov „bebo“
- určené pre študentov „unister“
- určené na osobnú prezentáciu „myspace“.

Ďalšie sociálne siete („43things“, „care2“, „dontstayin“) poskytujú široké spektrum iných služieb s ich osobitným určením. Novinkou sú stránky na vytvorenie vlastnej sociálnej siete „ning“ [Kostovčíková, 2013]. Nasleduje stručná charakteristika najúspešnejších sociálnych sietí.

2.3.2 Facebook

Facebook (www.facebook.com) je obrovská sociálna sieť združujúca priateľov z celého sveta. Je to projekt Marka Zuckerbera zo štúdia na Harvarde. Sieť sa rozširuje na základe pozvánky od člena siete. Facebook klasifikuje používateľov do jednej z nespočetných skupín podľa miesta bydliska, absolvovanej univerzity a pod. Sieť taktiež analyzuje vzťahy medzi používateľom a priateľmi jeho priateľov a na základe podobnosti ponúka potenciálnych priateľov. Používa vizualizačné prostredie „touchgraph“ na vizualizáciu priateľstiev podľa skupín, do ktorých priatelia patria.

Facebook (Obr.2.3) je sociálna sieť služieb, ktorej webová prezentácia bola spustená vo februári 2004. Táto sociálna sieť je prevádzkovaná a držaná v súkromnom vlastníctve spoločnosťou Facebook Inc.



Obr. 2.3. Registračná stránka sociálnej siete Facebook.

Už vo februári 2012 mal Facebook viac ako 845 miliónov aktívnych používateľov, ktorí navštívili webovú stránku tejto sociálnej siete minimálne raz za posledných 30 dní. Používateľ sa musí pred vstupom zaregistrovať a následne si môže vytvoriť svoj osobný profil, pridať ďalších používateľov ako priateľov a vymieňať si správy s ostanými používateľmi. Medzi poskytovanými službami je aj automatické informovanie v prípade aktualizácie profilov ostatných používateľov - priateľov. Používatelia sa taktiež môžu pripojiť k bežným záujmovým skupinám používateľov, ktoré sú organizované podľa zamestnania, školy, univerzít alebo iných vlastností. Taktiež môžu kategorizovať svojich priateľov do zoznamov, ako sú "ľudia v práci" alebo "blízki priatelia". Na Facebook sa môže registrovať každý používateľ, ktorý má viac ako 13 rokov [Blog, 2014].

Facebook umožňuje jednotlivým používateľom vybrať si vlastné nastavenie ochrany osobných údajov a zvoliť, kto môže vidieť jeho profil, ako aj to, ktoré časti jeho profilu majú byť viditeľné ktorým používateľom. Jediné, čo Facebook vyžaduje od všetkých používateľov je zverejnenie používateľského mena a fotografie profilu všetkým používateľom. Jednotliví používatelia môžu kontrolovať, kto vidí informácie, ktoré zdieľajú, rovnako ako aj kto ich môže pri vyhľadávaní nájsť. Facebook totiž disponuje databázou všetkých používateľov a tí, ktorí nechcú byť do nej zahrnutí si cez nastavenia ochrany osobných údajov môžu zabezpečiť súkromie pred zvyškom sveta.

2.3.3 Friendster

Friendster (www.friendster.com):

- Spoločnosť Friendster získala v San Franciscu v roku 2003 patent on-line sociálnej siete.
- Vynálezca patentu Jonathan Abrams dostal ocenenie za systém, metodiku a aparát pre spájanie používateľov na základe priateľstva.
- Spočiatku bola veľmi expandovaná (9-10 mil. používateľov), no neskôr bola zatienená spoločnosťou MySpace.
- V budúcnosti môže licencovať metodiku on-line spracovanie sociálnej siete.

2.3.4 MySpace

MySpace (www.myspace.com):

- Je sociálna sieť s interaktívnou štruktúrou.
- Je najväčšia na svete čo sa týka objemu prenesených dát a obsahuje medzinárodné osobné profily, blogy, fotografie, hudbu a videá.
- Najpoužívanejšie z mnohých modulov tejto siete sú:
  - *MySpaceIM,*
  - *MySpaceTV,*
  - *MySpaceMobile,*
  - *MySpaceNews* a pod.

2.3.5 Xanga

Xanga (www.xanga.com):

- Bola jedným z najväčších blogovacích portálov s počtom používateľov viac ako 27 mil.
- Vznikla v Newyork-u v roku 1999 spustením služby pre čítanie recenzií na knihy a filmy. Neskôr umožnila pridávanie blogov, čo ju priviedlo k expanzii.

2.3.6 Hi5

Hi5 (www.hi5.com):

- Sociálna sieť, ktorá bola derivovaná z MySpace pre tínedžerov.
- Zaznamenala 40 miliónov používateľov.
- Je to 8. najrozšírenejšia sociálna sieť v USA.
- Je viac profilovaná ako MySpace. Najviac obľúbené sú podsiete: *Hip Hop a R&B*.

2.4 Vizualizácia sociálnych sietí

V súčasnosti najbežnejšou a najmä najpreferovanejšou komunikačnou formou je práve komunikácia prostredníctvom internetu. Aj táto forma komunikácie môže mať niekoľko podôb, tá ktorá bude predovšetkým zaujímať nás prebieha prostredníctvom internetových diskusií, kde majú používatelia možnosť zapojiť sa do rôznych hodnotiacich debát a obohatiť ich tak o svoje komentáre k ľubovoľným témam. Nakoľko však tieto internetové diskusie nie sú kapacitne obmedzené, často je ich prezeranie a taktiež aj vyhľadávanie istých konkrétnych informácií sťažené. Existuje hneď niekoľko možností ako tieto diskusie a ich štruktúru sprehľadniť a jednou z nich je grafická vizualizácia. Cieľom tejto vizualizácie je získať nadhľad nad získanými rozsiahlymi dátami pre účely ich následnej analýzy.

V rámci sociálnych sietí rozlišujeme najčastejšie relácie priateľov a relácie influencerov. Tieto relácie tvoria bázu dát pre vizualizačné metódy. Vizualizácia sociálnej siete poskytuje zjednodušený pohľad nad komplexnou množinou vstupných dát, ktorý zvýrazňuje vlastnosti, ktoré chceme skúmať. V tomto prípade analýza konvenčnými metódami je neefektívna. Vizualizujeme priateľstvá a taktiež vplyv používateľov na komunitu a spätnú reakciu komunity na články daného „influencera“.

2.4.1 Reprezentácia vstupných dát

Existujú dve formy reprezentácie vstupných dát pre vizualizáciu sociálnej siete a to pravouhlé a štvorcové zobrazenie:

Pravouhlé zobrazenie je možné získať vyhodnotením dát z formulárov, ankiet, prieskumov aj štatistických výskumov. Riadky reprezentujú jednotlivé prípady štúdie – entity. Stĺpce reprezentujú vlastnosti týchto entít. Bunky obsahujú hodnotu vlastnosti danej entity (napríklad istej osoby). Príklad pravouhlého zobrazenia štatistických dát je v Tab.2.1.

Tab. 2.1. Pravouhlé zobrazenie dát

	Pohlavie	Vek	Vzdelanie
Dominika	Žena	42	Vysokoškolské
Matúš	Muž	37	Stredoškolské
Jakub	Muž	21	Postgraduálne
Klarisa	Žena	25	Vysokoškolské
Adam	Muž	55	Postgraduálne

Štvorcové zobrazenie resp. sieťová reprezentácia je matica, v ktorej sú stĺpce a riadky reprezentované tými istými entitami. Bunky matice reprezentujú vzťah medzi entitami. Príklad štvorcového zobrazenia štatistických dát je v Tab.2.2.

Tab. 2.2. Štvorcové zobrazenie dát

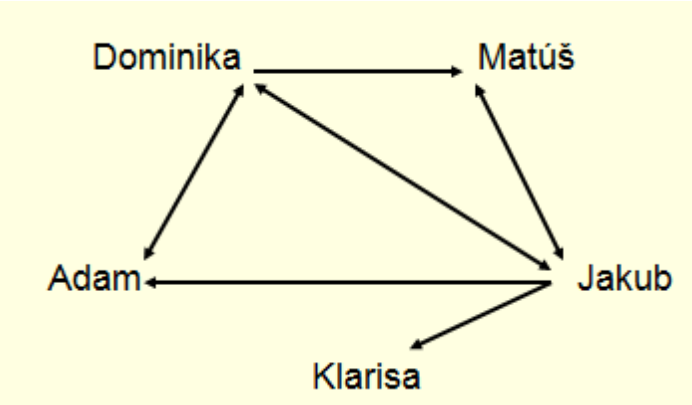
Osoba	Dominika	Matúš	Jakub	Klarisa	Adam
Dominika	-	1	1	0	1
Matúš	0	-	1	0	0
Jakub	1	1	-	1	1
Klarisa	0	0	1	-	0
Adam	1	0	0	0	-

Ako vyplýva z porovnania Tab.2.1 a Tab.2.2, pravouhlé a štvorcové zobrazenie sa od seba podstatne líšia. Zatiaľ čo pravouhlé zobrazenie zachytáva konvenčnú štruktúru dát a odhaľuje kvantitatívne aj kvalitatívne vlastnosti jednotlivých prvkov siete, štvorcové zobrazenie nás informuje iba o vzťahoch medzi jednotlivými prvkami siete navzájom.

Štvorcové zobrazenie môže byť použité na analýzu, ktorá spočíva v porovnávaní tak riadkov ako aj stĺpcov. Porovnávaním riadkov zisťujeme podobnosti medzi osobami, ktoré majú podobných priateľov. Porovnávaním stĺpcov zisťujeme kto je komu podobný, lebo si ich vybrala za priateľa tá istá osoba.

Štvorcové zobrazenie je možné analyzovať: aj iným spôsobom. Ak sa v ňom nachádza približne rovnaký počet „1“ a „0“, potom má sieť priemernú hustotu

obľúbenosti. Ďalej porovnaním stĺpcov a riadkov je možné určiť, či sa vo voľbách priateľov nenachádza reciprocita, ktorá indikuje vzájomné (obojsmerné) priateľské vzťahy. Takým vzťahom je napríklad vzťah medzi Dominikou a Jakubom na Obr.2.4, ktorý predstavuje grafovú vizualizáciu dát z Tab. 2.2.



Obr. 2.4. Vizualizácia vzájomných vzťahov na základe dát z Tab. 2.2.

Sociálna sieť je reprezentovaná incidenčnou maticou *B* vrcholov a hrán (*n x m*, kde *n* je celkový počet vrcholov a *m* je počet hrán). Každý prvok b_{ij} matice *B* je:

$b_{ij} = 1$ ak vrchol *i* je incidentný s hranou *j* v grafe G

$b_{ij} = 0$ inak.

Táto reprezentácia sa nazýva matica susednosti a je to najjednoduchšia a najpoužívanejšia binárna forma. Príkladom takejto matice je aj štvorcové zobrazenie dát.

Vizualizovaná matica susednosti je graf, ktorý reprezentuje sociálnu sieť. Uzly sú entity siete a hrany reprezentujú reláciu medzi uzlami, napr. priateľ, zamestnávateľ, príbuzný, spolužiak a pod.

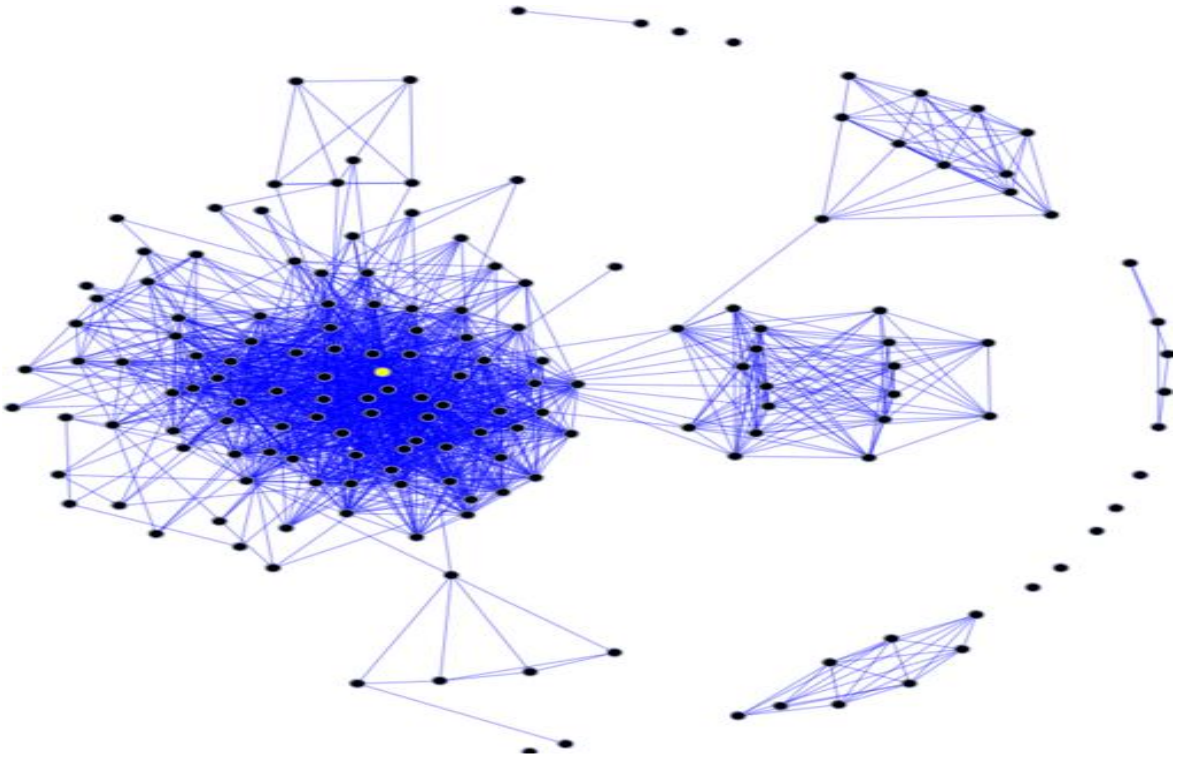
2.5 Vizualizačné techniky

Vizualizačné metódy sa používajú v mnohých vedných oblastiach. Podľa historika Alfreda Crosbyho je práve vizualizácia jeden z dvoch faktorov, ktorým vďačíme za rozmach všetkých vedných oblastí. Tým druhým faktorom je meranie. Bez vizualizácie, teda bez možnosti vykresliť získané údaje, by bol výskum sťažený a v niektorých prípadoch až nemožný. Existuje množstvo rozličných vizualizačných metód. Nasledujú tie, ktoré sa hodia na výskum v oblasti analýzy sociálnych sietí:

- diagram
- oblúkový diagram
- diagram toku údajov
- kruhový centralizovaný diagram
- kruhová konvergencia
- emplicitná implózia
- kruhová hierarchická reprezentácia
- strom
- mrak tagov („tag clouds“).

2.5.1 Diagram

Diagram je klasická grafová reprezentácia v jej najjednoduchšej forme. Pri trochu rozsiahlejších dátach odvodených od často navštevovaných sociálnych sietí je výsledkom tohto druhu vizualizácie dosť neprehľadný graf, čo ilustruje Obr.2.5.



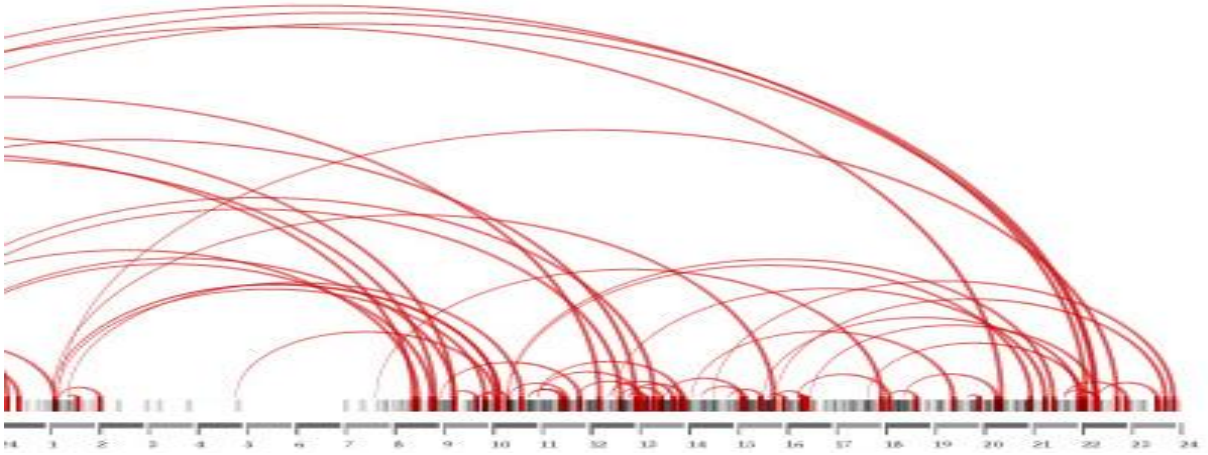
Obr. 2.5. Vizualizácia sociálnej siete diagramom.

V takomto diagrame je možné potom analyzovať centrálnu skupinu používateľov sociálnej siete, satelitné skupiny, a rozličné typy uzlov. Na základe topológie siete a charakteru spojenia medzi uzlami rozlišujeme nasledovné základné typy uzlov:

- **Hviezda** (Star, Snowflake) je charakteristická centrálnym uzlom, na ktorý sa napájajú ostatné uzly. Dostaneme ju ak centrálny uzol umiestnime do stredu kružnice a ostatných používateľov na kružnicu.
- **Most** (Bridge) je v podstate nedominantný uzol čo do počtu spojení. Jeho dôležitosť spočíva v tom, že svojimi hranami sprostredkuje spojenie medzi rôznymi skupinami priateľov v rámci sociálnej siete.

2.5.2 Oblúkový diagram

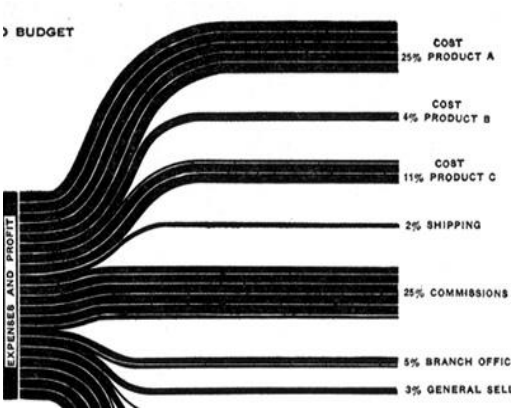
Oblúkový diagram („Arc Diagram“) predstavuje štruktúru v slučke, čo ilustruje Obr.2.6. Táto vizualizačná technika je vhodná v prípade výskytu mnohých sub – sekvencií. Entity sú umiestnené na priamke, čím sa vytvára celkom iná organizácia vizualizácie relácií. Relácie medzi entitami musia byť potom reprezentované pol kružnicami. Túto techniku navrhol nemecký študent University of Applied Sciences Postdam ako riešenie problému vizualizácie odpovedí na maily v čase.



Obr. 2.6. Vizualizácia pomocou oblúkového diagramu.

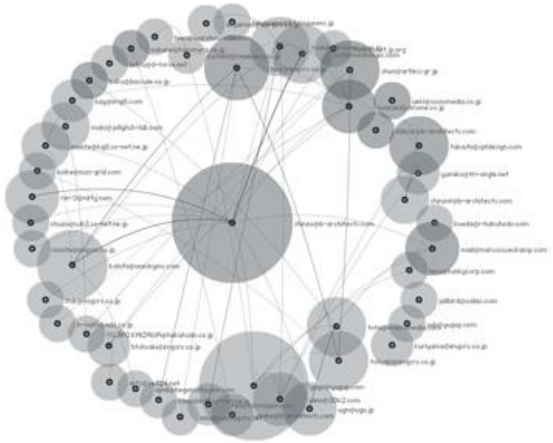
2.5.3 Diagram toku údajov

Diagram toku údajov („Data Flow Diagram“) vizualizuje smer posunu informácie od zdroja informácie k výstupu a má teda sekvenčný charakter. Príklad takého diagramu je znázornený na Obr.2.7.



Obr. 2.7. Ilustrácia techniky diagramu toku údajov.

2.5.4 Kruhový centralizovaný diagram



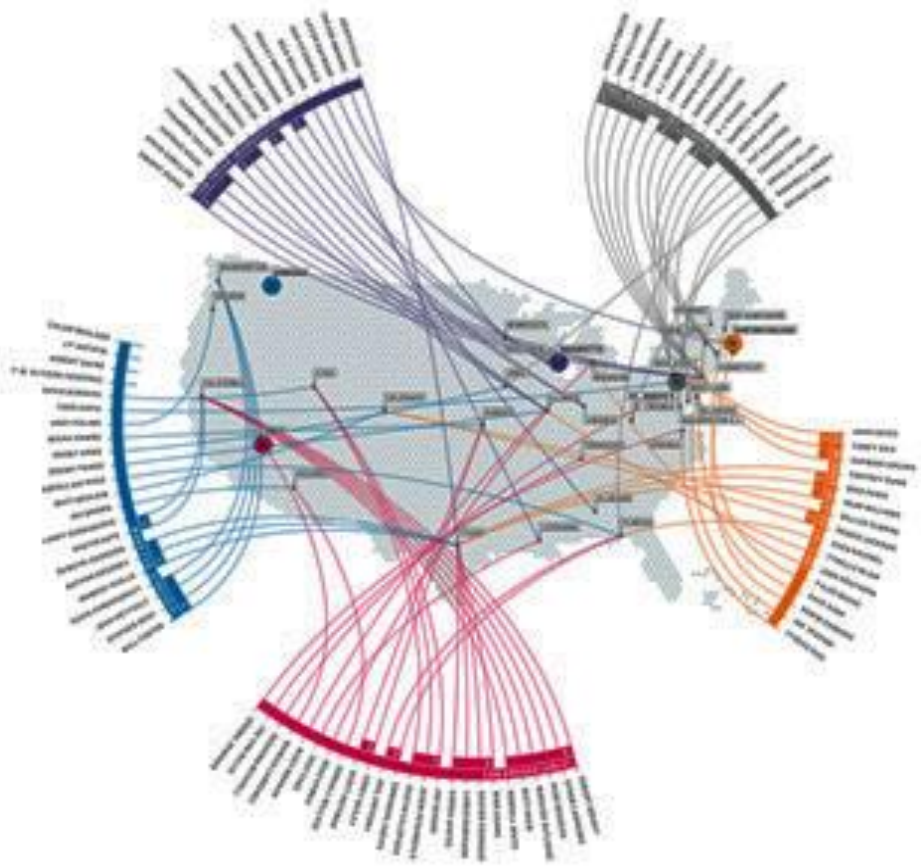
Kruhový centralizovaný diagram („Radial Centralized Diagram“) predstavuje reprezentáciu pomocou kruhového stromu. Všetky uzly v sieti sú reprezentované pomocou kruhov, pričom koreň stromu je centrálnym uzlom (viď Obr.2.8).

Obr. 2.8. Vizualizačná technika kruhový centralizovaný diagram.

2.5.5 Kruhová konvergencia

Kruhová konvergencia („Radial Convergence“) je vizualizačná technika založená na stromovej štruktúre v rámci ktorej sú uzly umiestnené na kružnici. V príklade

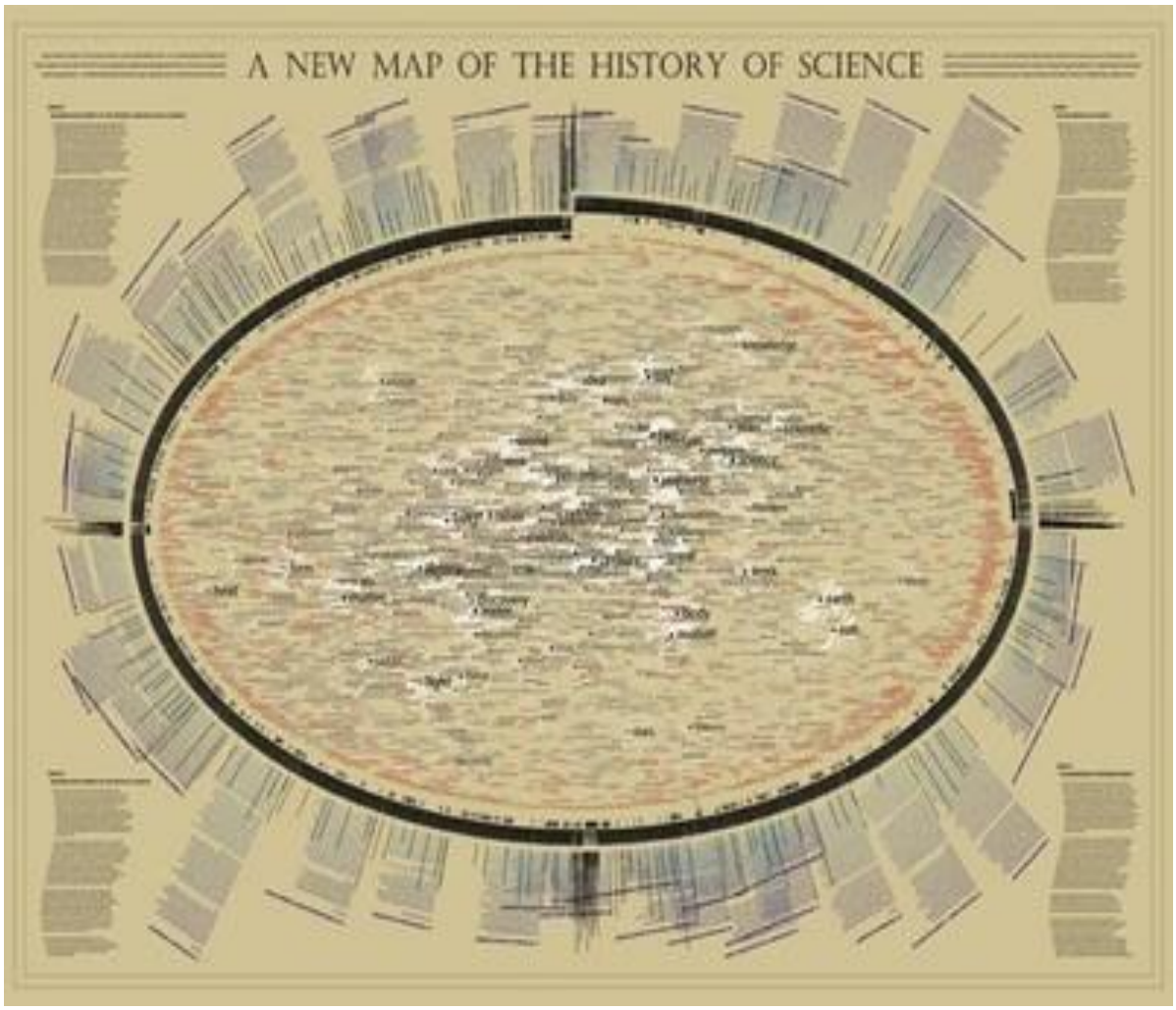
ilustrovanom na Obr.2.9. sú to tréneri baseballu, ktorí v rámci formovania piatich basebalových tímov vyhľadávali talentovaných hráčov po celých Spojených štátoch amerických. Relácie odkazujú na mestá, kde objavili svojich zverencov.



Obr. 2.9. Ilustrácia použitia vizualizačnej techniky kruhová konvergencia.

2.5.6 Eliptická implózia

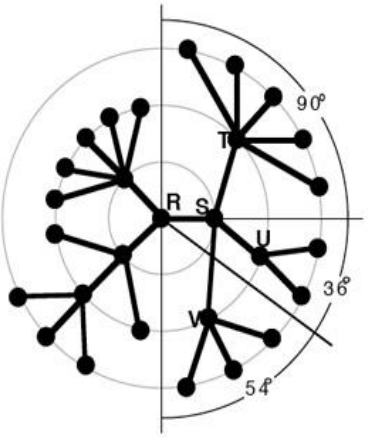
Eliptická implózia („Eliptical Implosion“) je vizualizačná technika, ktorá je dosť odlišná od predchádzajúcich. Uzly sú totiž umiestnené v obsahu kruhu a obvod kružnice (resp. elipsy) reprezentuje časovú os, pričom vzdialenosť uzlu od stredu určuje dôležitosť v čase, ku ktorému ich spojnica stredu a aktuálneho uzlu smeruje. Túto techniku navrhol W. Bradford Paley za účelom vizualizácie histórie vedy pre knižnú publikáciu „The History of Science“. Táto vizualizačná technika je ilustrovaná na Obr.2.10.



Obr. 2.10. Použitie vizualizačnej techniky eliptická implózia v historických vedách.

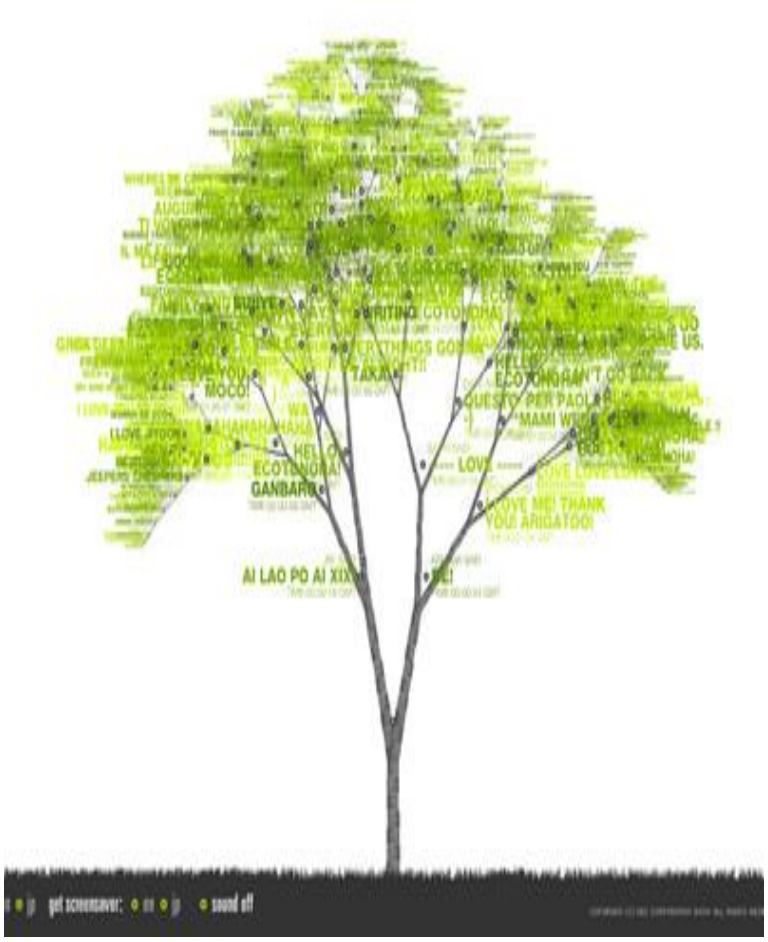
2.5.7 Kruhová hierarchická sieť

Kruhová hierarchická sieť („Radial Hierarchical Network“) je vhodná pre rozsiahle siete (do 10 mil. uzlov). Táto vizualizačná metóda bola vyvinutá spoločnosťou NicheWorks. Umožňuje interaktívne prehľadávanie siete, čo uľahčuje analýzu uzlov ale aj hrán so zameraním sa na výskum a objavovanie relácií. V porovnaní s podobnými metódami je rýchlejšia a produkuje dostatočne výstižný výstup (viď Obr.2.11).



Obr. 2.11. Kruhová hierarchická reprezentácia.

2.5.8 Strom



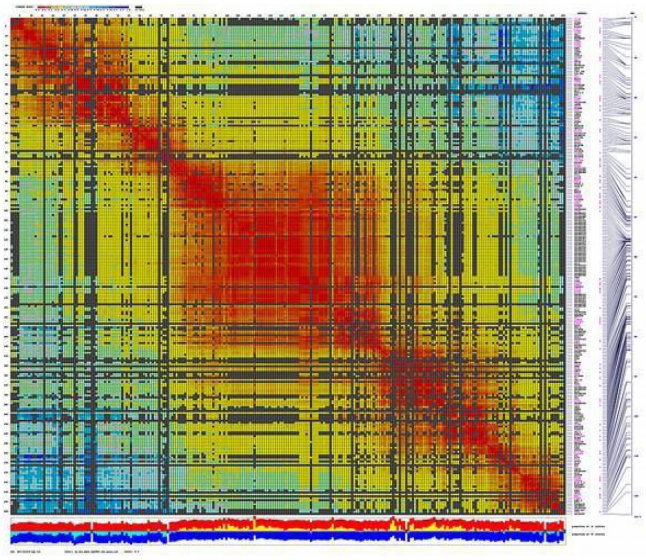
Strom („Tree“) je dlho známa a osvedčená vizualizačná technika (viď Obr.2.12), ktorá sa často využíva v oblastiach ako: matematika, štatistika a automatizácia. V informatike a umelej inteligencii sa osvedčil špeciálny druh stromu ako acyklického grafu a to rozhodovací strom.

Rozhodovacie stromy vyhovujú iba striktne hierarchicky usporiadaným dátam. Koreňový uzol stromu („root“) je najsilnejší bod siete. Reprezentuje všetky pozorovania, resp. takzvanú nultú hypotézu. Ostatné uzly sa hierarchicky napájajú na tento uzol a tak sa postupne generujú medziľahlé uzly, delením priestorov (reprezentujúcich skupinu pozorovaní) na pod priestory. Každá vetva je zakončená listovým uzlom, reprezentujúcim rozhodnutie.

Obr. 2.12. Vizualizačná technika založená na generovaní stromu.

2.5.9 Matica susedností

Ďalšiu významnú vizualizačnú techniku predstavuje reprezentácia maticou susedností (Obr.2.13). „*Podstata tejto vizualizačnej techniky spočíva vo vykreslení vzťahov medzi dátami v matici, ktorej políčka predstavujú vzťahy medzi riadkom pre daný uzol a stĺpcom pre iný uzol, ktorý je s ním spojený* [Gonda, 2014].“ Týmto spôsobom je možné vytvoriť charakteristickú vzorku na základe zhlukovania. Čiže na základe podobnosti riadkov respektíve stĺpcov dochádza k ich spájaniu do zhlukov.



Obr. 2.13. Vizualizácia pomocou matice susednosti.

2.5.10 Oblak tagov

Vizualizačná technika nazvaná Oblak tagov „Tag Clouds“ je v dnešnej dobe veľmi obľúbeným a vyhľadávaným spôsobom vizualizácie informácií rôzneho druhu. Úspešnosť tohto prístupu je založená na vizualizácii dát spojenej s aplikáciou webových dizajnových prvkov, zabezpečujúcich estetiku vyjadrenia, a sociálnych ukazovateľov. Oblak tagov, respektíve mračno tagov, dokáže veľmi intuitívnym spôsobom vyjadriť, o čo sa človek alebo skupina ľudí zaujíma, a táto informácia je sprostredkovaná rýchlo. Dá sa pojať letmým pohľadom, čo ilustuje Obr.2.14.

„Tag Clouds“ zobrazujú používateľom generované tagy v rámci kartézskej súradnicovej sústavy, kde ich poloha je určená x - ovou a y - ovou súradnicou. Zobrazované tagy (slová a frázy) výstižne vyjadrujú obsah webových stránok a umožňujú špecifickú a presnú orientáciu v obsahu stránok. Hlavnou prednosťou je schopnosť vyzdvihnúť najdôležitejšie a najobľúbenejšie témy. Táto technika taktiež podporuje navigáciu.

Obr. 2.14. Vizualizácia pomocou „Tag Clouds“.

Základná Idea tejto techniky spočíva v reprezentácií značiek (slov, fráz) na ploche podľa ich významu a váhy odvodených z frekvencie výskytu týchto značiek. Toto zobrazenie sa realizuje výberom vhodných veľkostí písma a farieb. Tagy sú veľmi často považované aj za jeden z typických prvkov Sociálneho webu.

Zvyčajné umiestnenie „Tag Clouds“ je v bočnom paneli na ľavej alebo pravej strane stránky. Pre nedostatok priestoru sa veľmi triezvo navrhuje veľkosť písma a teda nie sú až tak veľké rozdiely medzi najmenším a najväčším písmom, čo je ilustrované na Obr.2.15. Z týchto dôvodov je niekedy lepšie váhu tagov určiť inak ako veľkosťou písma. Veľmi vďačné je použitie rôznych farieb. Svoju úlohu tu hrá aj kontrast medzi farbou tagu a jeho podkladom. Väčší kontrast predznamenáva aktívnejší tag. Na druhej strane tagy majúce farbu podobnú farbe podkladu reprezentujú pasívne tagy. Je potrebné zvážiť počet použitých farieb. Odporúča sa použiť cca. dve alebo tri farby, aby neboli zmarené šance návštevníka na okamžité zachytenie nosných pojmov.

Obr. 2.15. Rozdiely vo veľkosti tagov.

Existujú rozličné konkrétne implementácie myšlienky „Tag Clouds“:

- Radenie tagov podľa abecedy. Najdôležitejšie, prípadne časté pojmy sú zvýraznené pomocou vhodnej veľkosti písma.
- Všetky slová majú priradenú rovnakú veľkosť písma a váhu a dôležité pojmy sú zvýraznené farbou písma alebo farbou pozadia.
- Radenie tagov podľa dôležitosti a frekvencie, pričom veľkosť písma a farby môžu byť použité pre zdôraznenie významu pojmov.
- Tagy nie sú vôbec zoradené, používa sa však veľkosť písma a farba na vyjadrenie váhy tagu.
- Zoradenie tagov podľa podobnosti, pričom podobné výrazy sa objavujú ako susedia vedľa seba. Je možné použiť rôzne vizuálne formátovanie.

Avšak vo väčšine prípadov sú tagy radené podľa abecedy [Smashing, 2014].

Dnes je k dispozícií rad nástrojov, ktoré pomôžu používateľovi vytvoriť „Tag Clouds“ automaticky, napríklad „Google Tag“ „Cloud Marker“, „TagCrowd“, „TagmyCloud“, „MakeCloud“ a pod.). Hlavná myšlienka týchto služieb spočíva v analýze kľúčových slov a vyčíslení frekvencie ich výskytu, ktorá v prípade potreby je lineárne normovaná podľa rovnice (3).

$$Si = \frac{fmax*(ti-tmin)}{tmax-tmin}, \text{ pre } t_i > t_{min} \text{ , inak } S_i = 1; \qquad (3)$$

Kde:

Si je výsledná veľkosť zobrazeného tagu (slova),

fmax je maximálna prípustná veľkosť slova,

ti je frekvencia výskytu daného tagu v analyzovanom texte,

tmin je minimálny uvažovaný počet výskytov tagu,

tmax je maximálny uvažovaný počet výskytov tagu.

2.6 Využitie vizualizačných techník v analýze sietí

V Pomocou vizualizácie je možné dôjsť k záverom, resp. dopátrať sa k informáciám, ktoré vznikli emergenciou v rámci veľmi zložitého systému. Pri komplikovanej štruktúre vstupných dát je ich analýza zložitá občas až nemožná. Analýza sociálnych sietí je tiež toho druhu. Analýze sociálnych sietí bude venovaná nasledujúca kapitola. V rámci aktuálnej podkapitoly bude venovaná pozornosť analýze Internetu, ako siete počítačov, z hľadiska bezpečnosti.

Analýze počítačových sietí bol venovaný projekt OPTE. Tento projekt sa pokúsil riešiť niekoľko problémov ako napríklad:

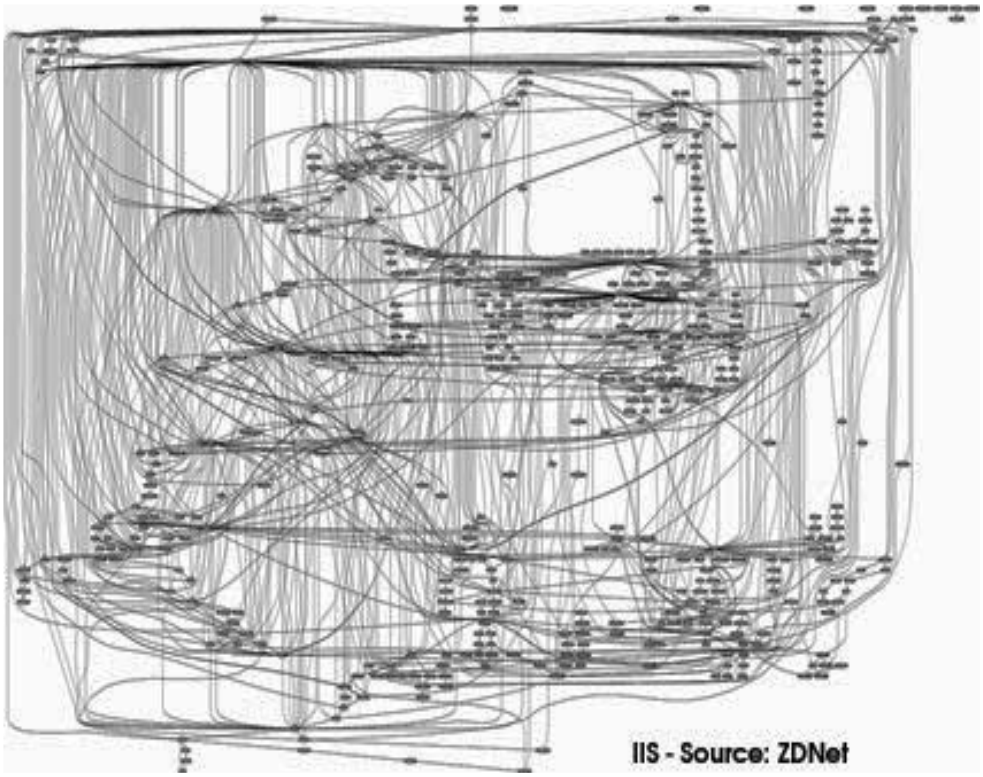
- pokus o vizualizáciu Internetu
- analýza zbytočných rozsahov IP adries
- detekcia prírodných katastrof, počasia, vojny (informácia o vnútornej štruktúre môže poslúžiť predikcii)
- analýza preťažených častí sietí môže inicializovať ich posilnenie, prepracovanie, resp. zrušenie.

Ďalšie problémy, pri riešení ktorých môžu byť vizualizačné techniky nápomocné je kontrola bezpečnosti a optimalizácia kódu.

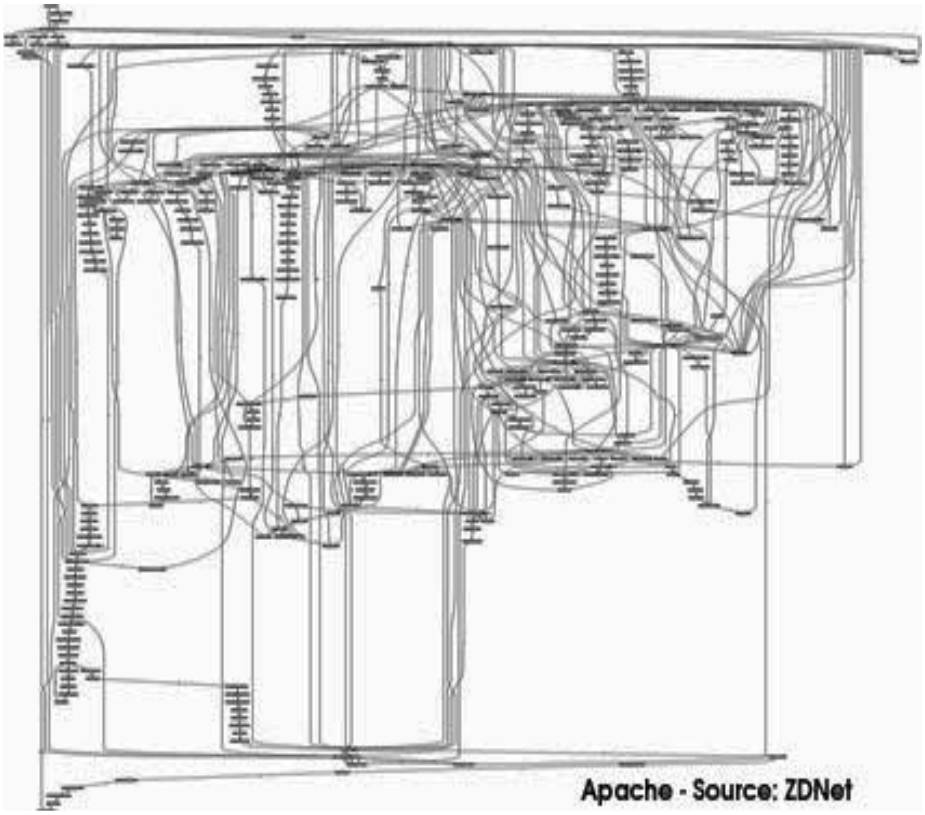
2.6.1 Vizualizácia a kontrola bezpečnosti

E Stiennon povedal o kontrole bezpečnosti nasledovné: „*Systémové volanie je šanca na adresovanie pamäte. Hacker analyzuje každý prístup do pamäte kvôli odhaleniu náchylnosti na útok pomocou pretečenia zásobníku. Vývojár musí všetky tieto prístupy zabezpečiť. Čím viac prístupov, tým väčšia šanca pre útočníka*.“

Preto je pre kontrolu bezpečnosti kľúčovým faktorom možnosť vizualizácie všetkých systémových volaní. Na Obr.2.16 a Obr.2.17 sú uvedené vizualizácie všetkých systémových volaní, ktoré boli uskutočnené v priebehu zobrazenia jedinej a tej istej stránky na dvoch rozličných webových serveroch Apache a Microsoft IIS. Porovnanie bezpečnosti a optimalizácie behu dvoch webových serverov Apache a Microsoft IIS jednoznačne určuje ako bezpečnejší Apache s podstatne menším počtom systémových volaní.



Obr. 2.16. Vizualizácia systémových volaní pri zobrazovaní jednej konkrétnej stránky na webovom serveri IIS.

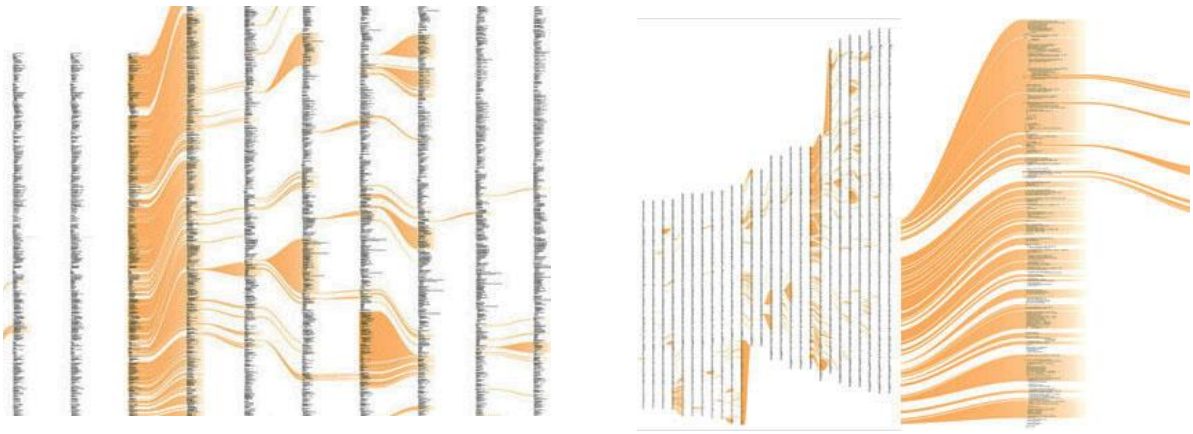


Obr. 2.17. Vizualizácia systémových volaní pri zobrazovaní tej istej konkrétnej stránky na webovom serveri Apache.

2.6.2 Vizualizácia a optimalizácia kódu

Vizualizačné techniky môžu byť nápomocné aj pri optimalizácii kódu. Vizualizáciou môžete získať predstavu o všetkých zmenách vo vyvíjanom kóde a taktiež predstavu o tom, v ktorých etapách bol tento kód najvýraznejšie revidovaný. Ide o dynamickú analýzu kódu vyvíjaného softvéru, ktorý sa dynamicky mení v čase.

Projekt REVIOSIONIST je zameraný na generovanie vizualizácie zmeny v štruktúrach a obsahu stránok. Na Obr.2.18. sú ilustrácie dvoch rozličných výstupov tohto projektu. Kód je zobrazený čiernou farbou a jeho zmeny pomarančovou farbou.



Obr. 2.18. Vizualizácia dynamickej zmeny kódu prostredníctvom služby Revisionist.

Revízia zmien vo vyvíjanom kóde môže byť podkladom pre niektoré manažérske rozhodnutia, týkajúce sa napríklad počtu zapojených vývojárov. Viac vývojárov na tvorbe projektu znamená jeho rýchlejšie dokončenie ale aj častejšie zmeny kódu a možno viac revízií a viac práce pri spájaní čiastkových riešení do celku. Niekedy je orientácia v tak zložitom a dynamickom prostredí problematická, preto metódy vizualizácie sú tu veľmi nápomocné. (Všetky obrázky tejto kapitoly boli prevzaté. Väčšina z Internetu.)

POUŽITÁ LITERATÚRA

[Blog, 2014] Blog Facebook information and news. *All you nedd to know about Facebook: Facebook information and history*. [online]. Powered by WordPress and K2 Entries Feed and Comments Feed, July 2008 [cit 2014-20-03]. Dostupné na internete: <http://fcbkinfo.com/>.

[Boyd-Ellison, 2014] Boyd, D., Ellison, N. B., *Social Network Sites: Definition,History, and Scholarship*. [cit 2014-20-03] Dostupné na internete: http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html

[Donah, 2014] Donah, B., *Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life*. [cit 2014-20-03] Dostupné na internete: http://www.danah.org/papers/WhyYouthHeart.pdf.

[Gonda, 2014] Gonda, P., *Vizualizácia sociálnej siete komunitného serveru Kyberia.sk*. [online]. Univerzita Komenského v Bratislave. Fakulta matematiky, fyziky a informatiky. Katedra aplikovanej informatiky. Bratislava. 2007 [cit. 2014-03-23]. Dostupné na internete: <http://www.gondapeter.sk/files/peter_gonda_bakalarka..pdf>.

[Kostovčíková, 2013] Kostovčíková, V., *Korelácia medzi študijnými výsledkami a komunikáciou s autoritami danej sociálnej siete*. FEI Technická univerzita v Košiciach, 2013, Košice, 1-76.

[Paralič, 2011] Paralič, J.. *Objavovanie znalostí v databázach: Objavovanie a využívanie znalostí v sociálnych sieťach*. [online]. Košice : TUKE-FEI, 2003. Aktualizované 25-11-2011. Dostupné na internete: <http://people.tuke.sk/jan.paralic/prezentacie/OZ/Analyza-socialnych-sieti.pdf>.

[Pénzeš, 2010] Pénzeš, T., *Karma a sentiment vo webových diskusiách*. FEI Technická univerzita v Košiciach, 2010, Košice, 1-62.

[Smashing, 2014] Smashing Magazine, Tag Clouds Galéria: Príklady a osvedčené postupy [online]. [cit. 2014-03-24] Dostupné na internete: <http://www.smashingmagazine.com/2007/11/07/tag-clouds-gallery-examples-and-good-practices/>

3 ANALÝZA SOCIÁLNYCH SIETÍ

3.1 Úvod

Sociálne siete sú fenoménom dnešných dní a ako také sú atraktívne nielen pre bežného používateľa ale aj pre vedeckú verejnosť, ktorú zaujíma práve analýza sociálnych sietí z rôznych hľadísk a to z hľadiska spôsobu fungovania, z hľadiska štruktúry siete, z hľadiska jej rastu respektíve úpadku a z hľadiska obsahu, ktorý sa hromadí v rámci jej konverzácií.

Hlavnou prednosťou sociálnych sietí je, že nám pomáhajú udržiavať kontakt s blízkymi a známymi, bez ohľadu na vzdialenosť či geografickú lokalitu. Ďalšou prednosťou je variabilita ich využitia, pretože nemusia byť využité iba na vymieňanie si noviniek, kuriozít, videí, hudobných skladieb a pod. s blízkymi. Je možné ich využiť aj ako platformu pre riešenie problémov rozličného druhu, od receptov až po riešenie technických a pracovných problémov. Taktiež nám poskytujú výhodu pomerne jednoduchého vstupu do skupiny odborníkov.

Sociálnu sieť ako takú môžeme vnímať dvojako. Prvoplánový pohľad na sociálne siete nám ich odhaľuje ako webové aplikácie, ktoré poskytujú možnosť udržiavať kontakt s blízkymi a priateľmi, čo napokon bolo hlavným dôvodom ich vzniku. Nie prvoplánový pohľad sa zameriava na ich analýzu a teda možnosť skúmať sociálne väzby medzi používateľmi, ktorí svojou prítomnosťou na sociálnej sieti sledujú iné zámery ako ju rozvíjať. Ide teda o siete, ktoré sa rozvíjajú bez zásluh niektorého, resp. niektorých konkrétnych používateľov. Môžeme povedať, že vznikajú emergenciou nezávisle od vôle používateľov a reflektujú potrebné a účelné diskusie medzi používateľmi.

Ak sa pozeráme na sociálnu sieť ako na sociálnu štruktúru v istom čase, hovoríme o statickej sociálnej sieti. Zapojením dynamickej zložky do procesu mapovania sociálnej siete máme možnosť sledovať jej evolúciu, rast a charakteristiky správania sa diskutujúcich. Takáto analýza poskytne čitateľovi možnosť oboznámiť sa s rôznymi usporiadaniami entít v sociálnej sieti ako aj sledovať dynamické parametre sociálnej siete, ktoré môžu byť nápomocné pri lepšom pochopení sociálnych štruktúr vznikajúcich v kyberpriestore.

Sociálna sieť môže byť reprezentovaná akoukoľvek dátovou štruktúrou, ktorá zachytáva interakciu medzi entitami skúmanej oblasti. Ako každá iná grafická reprezentácia aj grafová reprezentácia sociálnej siete sa skladá z hrán a uzlov. Uzly predstavujú jednotlivé skúmané entity a hrany reprezentujú prepojenia uzlov v sieti.

Podľa [Rakuščinec, 2009] ak existuje v sociálnej sieti akákoľvek relačná spojitosť, je ju možné vyjadriť prostredníctvom grafu jednoduchou vizualizáciou dát zozbieraných z danej siete. Takýto graf sa skonštruuje hranovým prepojením bodov siete (aktérov). Takýto, spravidla orientovaný graf nám umožní globálny a niekedy aj zjednodušený pohľad nad komplexnou a neusporiadanou bázou dát. Takáto vizualizácia je relatívne nenáročný proces, ktorý produkuje vysoko informatívny pohľad do dátovej štruktúry. Jeho analytické dolovanie by vyžadovalo o mnoho viac času a bolo by o mnoho výpočtovo náročnejšie.

3.2 Sociálna sieť

Dnes je možné stretnúť sa z mnohými definíciami pojmu sociálna sieť. Definícia tohto pojmu zaujíma nielen informatikov, ale aj psychológov a špecialistov mnohých

ďalších prírodných vied. Klasická definícia definuje sociálnu sieť ako množinu sociálne príslušných uzlov spojených jedným alebo viacerými typmi vzťahov (hrán). Uzly alebo členovia siete sú základné jednotky siete spojené vzťahmi, na ktoré sa väčšinou sústreďuje analýza sociálnych sietí. Spomenuté jednotky zvyčajne predstavujú osoby alebo organizácie a každá jednotka môže byť prepojená na jednu alebo viacero iných jednotiek [Marin-Wellman, 209]. Zjednodušene povedané, sociálna sieť je množina uzlov (členov siete - aktérov), ktoré sú vzájomne prepojené jedným, alebo viacerými typmi vzťahov [Wasserman-Faust, 2009].

Informatik by skôr prijal definíciu že sociálna sieť je heterogénna a mnoho vzťahová dátová množina reprezentovaná graficky, pričom graf je zvyčajne veľmi veľký s uzlami odpovedajúcimi objektom a hranami odpovedajúcim spojeniam predstavujúcim vzťahy alebo interakcie medzi objektmi. Oboje, uzly aj hrany majú atribúty a objekty môžu byť zaradené do tried. Vzťahy môžu byť orientované a nemusia byť výlučne binárne [Han, 2003].

Sociálne siete môžu byť rôzneho druhu: technologické, obchodné, ekonomické, biologické, elektrické rozvodné siete, komunikačné siete, šírenie počítačových vírusov, kolaboračné siete (napr. spoluautorstvo), citačné siete, webové siete, siete spoločného nákupu, internetové peer-to-peer siete a mnohé iné.

Opäť, informatik by sociálne siete rozdelil na: orientované, neorientované, bipartitné a multigrafové. Z pohľadu dostupných informácii o sieti na časové a značkované. Z analytického hľadiska je možné deliť sociálne siete na jednomódové, dvojmódové, či viacmódové. Podľa špecifickosti ich vzniku môžeme sociálne siete deliť napríklad na reálnesiete, náhodné siete, siete malého sveta a bezškálové siete [Repka, 2011].

3.2.1 Základné pojmy

Pre potreby analýzy sociálnej siete je účelné zaviesť niektoré fundamentálne pojmy z danej oblasti a stručne ich charakterizovať. Podľa [Wasserman-Faust, 2009] to majú byť predovšetkým nasledovné základné koncepty:

Entita. Sociálna sieť je zložená z entít (v najjednoduchšom prípade reálnych používateľov siete) a ich vzájomných prepojení. Entity sú v grafe reprezentované uzlom alebo nódom (výraz prevzatý z angličtiny a zdomácnený). Vzájomné prepojenia entít (v najjednoduchšom prípade ide o rozličné typy vzťahov medzi používateľmi) sú v grafe reprezentované hranami (spojnicami) a práve tie je vhodné skúmať. Entita predstavuje diskrétnu individualitu voči širšej reprezentácii. Môže reprezentovať jeden základný stavebný prvok sociálnej siete. V závislosti od skúmanej množiny, môžu entity existovať aj ako korporátna alebo kolektívna sociálna jednotka. Entitou môže byt individualita v spojení so skupinou, oddelenie v spojení s korporáciou. Napríklad mestská časť v spojení s mestom alebo štáty v spojení s nadnárodnou organizáciou.

Väzba. Spojenia medzi entitami môžeme chápať ako väzby. Existuje mnoho druhov väzieb, avšak v zjednodušenej forme vieme väzbu definovať ako logické spojenie dvoch entít. Väzby v reálnom svete môžu predstavovať rôzne spojitosti:

- Evaluácia jednej osoby voči inej osobe, napr. vyjadrené priateľstvo, obľúbenosť alebo rešpekt.
- Prenos materiálnych zdrojov, napr. obchodné transakcie, požičiavanie alebo prenájom vecí.
- Asociácia a afiliácia, napr. účasť na nejakom sociálnom podujatí, resp. príslušnosť ku skupine.

- Behaviorálna interakcia, napr. konverzácia alebo výmena správ.
- Fyzické prepojenie, napr. cesta, rieka alebo most spájajúci dva body.
- Formálne relácie, napr. autorita, podradenosť a pod.
- Biologický vzťah, napr. rodičovstvo, potomstvo a pod.

Táto publikácia sa zameriava na väzby vo forme prenosov správ, behaviorálne väzby a evaluačné väzby.

Dyáda. Diáda predstavuje najjednoduchší model sociálnej siete tvorený práve dvoma entitami. Takáto väzba medzi dvoma entitami je prirodzene súčasťou dvojice a preto je nemožné prisúdiť túto väzbu jednej alebo druhej entite. Existuje mnoho analýz sociálnych sietí, ktoré sa zaoberajú hlavne týmito najjednoduchšími modelmi a to z hľadiska dominancie vyplývajúcej z orientácie väzby, resp. jej cyklickej varianty. Teoreticky môže dyádu tvoriť aj taký pár entít, medzi ktorými neexistuje žiadna väzba. Príklady dyád je možné zaradiť do troch izomorfných tried rovnocennosti. Prvá trieda predstavuje obojstrannú - recipročnú dyádu, ktorá obsahuje obidve opačne orientované hrany. Druhá trieda – asymetrická dyáda má iba jednu orientovanú hranu. Treťou triedou je nulová dyáda, ktorá neobsahuje ani jednu hranu medzi párom vrcholov. Tieto triedy sú v analýze sietí dôležité hlavne pri skúmaní tendencií reciprocity a asymetrie sociálnych sietí. Táto analýza využíva informácie o počtoch dyád spadajúcich do jednotlivých tried [Wasserman-Faust, 2009].

Triáda. Analýza sietí väčšinou nie je založená iba na skúmaní dvojíc ale aj na skúmaní komplikovanejších podmnožín sociálnej siete. Mnoho dôležitých analýz sociálnych sietí sa opiera práve o skúmanie trojíc, teda podmnožín sociálnej siete tvorených práve troma entitami a ich (ak existujú) väzbami.

Podskupina. Ďalším rozšírením trojice je podskupina, ktorá je definovaná ako nejaká podmnožina všetkých entít a ich spojení. Vyhľadávanie a identifikácia podskupín na základe istých kritérií tvorí dôležitú súčasť analýzy sociálnych sietí.

Skupina. Analýza sociálnych sietí sa neobmedzuje iba na analýzu dvojíc, trojíc alebo podskupín. V širšom kontexte, analýza siete spočíva v schopnosti modelovať vzťahy medzi systémami entít. Systém pozostáva z väzieb medzi entitami v rámci nejakej ohraničenej skupiny. Skupinou môžeme nazývať všetky entity, ktorých väzby sme sa rozhodli skúmať.

Relácia. Súbor väzieb určitého typu medzi jednotlivými entitami skupiny sa nazýva relácia. Príkladom môže byt množina priateľstiev medzi pármi detí v triede. Isté druhy relácií je možné aj ohodnotiť, napríklad pri relácii obchodných prepojení medzi štátmi napr. Eurozóny je vhodné uvádzať aj výšku finančných prostriedkov, ktoré boli prevedené. Relácia môže existovať iba nad množinou väzieb istého druhu nad množinou entít z istej špecifickej skupiny.

Sociálna sieť. Vyššie spomenuté definície entity, skupiny a relácie poskytujú priestor pre definíciu sociálnej siete ako konečnej množiny entít a vzťahov medzi nimi. Prítomnosť relácií je nutná podmienka sociálnej siete a definuje celú jej funkčnosť.

3.2.2 Typy sociálnych sietí

Sociálne siete delíme buď podľa počtu väzieb, podľa spôsobu ohodnotenia hrán alebo podľa typov aktérov. Podľa povahy a počtu väzieb (hrán) delíme sociálne siete na:

- *orientované* – Záleží na orientácii hrany (smere väzby), ktorá má špecifický význam napríklad pri transakciách.
- *neorientované* – Nezáleží na orientácii hrany. Väzba nemá orientáciu alebo je obojsmerná a má rovnakú váhu v oboch smeroch.
- *multihranové* – Sieť obsahuje viacpočetné hrany medzi párom uzlov. Príkladom takej siete je multigraf.
- *so slučkami* – Sieť obsahuje hrany vychádzajúce a vstupujúce do toho istého vrcholu. Príkladom je pseudograf.

Podľa spôsobu ohodnotenia hrán delíme sociálne siete na:

- *nevážené* – V sieti sú všetky hrany rovnocenné alebo ich váha nie je podstatná.
- *vážené* – Ak sa v sieti nachádzajú údaje reprezentujúce váhy jednotlivých hrán, potom sú tieto hrany vážené. Častým príkladom je cestná sieť, kde väzby sú hodnotené reálnou vzdialenosťou dvoch miest.
- *značkované siete* – Predstavujú špeciálnu formu váženej siete, kde hodnota hrany môže nadobúdať dve hodnoty: kladnú hodnotu (+) a zápornú hodnotu (-).
- *časové (temporálne)* – Také siete obsahujú pre každú hranu záznam o čase, kedy vznikla.

Podľa počtu rôznych typov aktérov delíme sociálne siete na:

- *jednodómové* – Jednomódová sociálna sieť je sieť, ktorá obsahuje iba jeden typ aktérov, napr. ľudia, podskupiny zamestnancov organizácie, kolektívy, národy, štáty a pod., čiže zobrazuje väzby medzi rovnakými typmi aktérov. Je to najčastejšie sa vyskytujúci typ sociálnej siete. V takej sociálnej sieti je zvyčajne iba jeden typ relácii, ale v prípade multi-relačnej jednodómovej siete ich môže byť aj viac. Podmienkou jednomódových sietí je iba jeden typ aktérov, nie aj jeden typ relácie.
- *dvojdómové* – V tomto prípade je možné množinu aktérov rozdeliť na dve podmnožiny aktérov. Príkladom takejto siete je spoločnosť a jej zamestnanci alebo sieť autorov a ich článkov. V týchto sieťach sa typicky analyzujú vzťahy, kde aktéri jedného typu môžu mať vzťah iba s aktérmi druhého typu. Zvyčajne v takom prípade môže vytvoriť väzbu len jeden typ aktérov, ktorý sa označuje ako „odosielateľ" a druhý typ môže väzbu len akceptovať a označuje sa ako „prijímateľ". Príkladom takejto siete je sieť, ktorá obsahuje množinu aktérov a množinu udalostí. Takáto dvojmódová sieť sa nazýva pridružujúca sieť [Marin-Wellman, 2009], vyjadrujúca príslušnosť ľudí k udalostiam alebo organizáciám, správnym radám, firmám respektíve príslušnosť štátov k medzinárodným organizáciám.

Viac informácií o delení sociálnych sietí je možné nájsť v [Repka, 2011].

3.3 Notácie v sociálnych sieťach

3.3.1 Notácia statickej siete

Statická sociálna sieť predstavuje taký stav (väčšinou konečný), ktorý je tvorený množinou dát v danom čase. Množina dát o sociálnej sieti, ktorú chceme podrobiť skúmaniu a analýze, musí obsahovať informácie o väzbách (reláciách), ktoré existujú medzi jednotlivými entitami. Preto je potrebné zaviesť notácie, ktoré nám jednotlivé entity a relácie pomôžu jednoznačne zapísať.

Najjednoduchšie notácie sú tie, ktoré sú určené pre základnú dychtonomickú reláciu. Komplikovanejším reláciám sa venuje publikácia [Wasserman-Faust, 2009].

Existuje mnoho spôsobov, ako popísať dáta pre sociálnu sieť matematicky, ale väčšinou sa používajú tri základné notácie. Tieto notácie je možné adaptovať na široké spektrum sociálnych sietí. V špecifických typoch sociálnych sietí môže dôjsť k preferencii niektorej z týchto notácií.

Tri základné notácie sú:

- Teória grafov
- Sociometrika
- Algebraická notácia

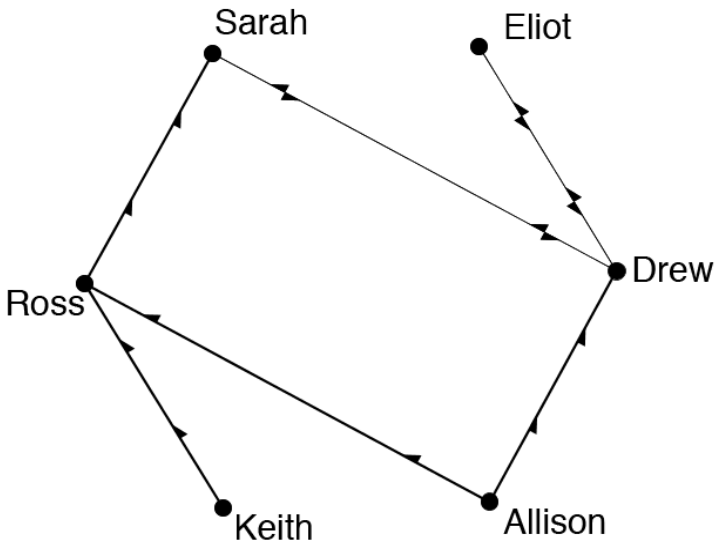
Metóda notácie pomocou **teórie grafov** je považovaná za elementárny spôsob reprezentácie entít a relácií medzi nimi. Je to jednoduchá notácia, vyjadrujúca entity ako vrcholy a samotné relácie ako hrany grafu. Matematici a štatistici ako Bock, Harary, Katz a Luce zaviedli ako prví pojem orientovaného resp. neorientovaného grafu. Využitie orientácie hrán v grafoch poskytuje možnosť zobraziť jednostranné relácie (resp. obojstranné, ak sú spoločné). Majme množinu entít N s ich početnosťou g a notáciou $N = \{n_1, n_2, \ldots, n_g\}$. Pre prípad množiny $g = 6$ detí z prvého stupňa základnej školy (viď Tab.3.1), by mohla byť použitá nasledovná notácia:

Tab. 3.1. Notácia pomocou teórie grafov [Wasserman-Faust, 2009].

Relácia 1	Relácia 2	Relácia 3
Priateľstvo na začiatku	Priateľstvo na konci	Neďaleké bydlisko
<Allison, Drew>	<Allison, Drew>	(Allison, Ross)
<Allison, Ross>	<Allison, Ross>	(Allison, Sarah)
<Drew, Sarah>	<Drew, Sarah>	(Drew, Eliot)
<Drew, Eliotl>	<Drew, Eliot>	(Keith, Ross)
<Eliot, Drew>	<Drew, Ross>	(Keith, Sarah)
<Keith, Ross>	<Eliot, Ross>	(Ross, Sarah)
<Ross, Sarah>	<Keith, Drew>	
<Sarah, Drew>	<Keith, Ross>	
	<Ross, Keith>	
	<Ross, Sarah>	
	<Sarah, Drew>	

N = {*Allison, Drew, Eliot, Keith, Ross, Sarah*}. Súbor týchto šiestich entít je možné reprezentovať nasledovne *n*1 = *Allison, n*2 = *Drew, n*3 = *Eliot, n*4 = *Keith, n*5 = *Ross a n*6 = *Sarah*.

Predpokladajme usporiadanú dvojicu entít n_i *a* n_j pričom platí, že prvá entita je vo vzťahu s druhou entitou. Ak relácia existuje, môže byť nazvaná *L*. Všetky relácie medzi entitami sú reprezentované daným počtom usporiadaných dvojíc *L*. Grafické znázornenie pozostáva z množiny entít *N* a množiny relácií *L*, a je možné ho matematicky vyjadriť pomocou týchto dvoch množín ako *G(N, L)*. Príklad notácie pomocou teórie grafov uvádza Tab.3.1. Odvodená grafická reprezentácia je ilustrovaná na Obr.3.1.



Obr. 3.1. Grafická reprezentácia notácie podľa teórie grafov.

Druhá možnosť je **sociometrická notácia**, ktorá je v súčasnosti v odbornej literatúre najviac používaná. Notácia pomocou sociometrickej matice (čo je ekvivalent incidenčnej matice) je tvorená dvojrozmernou maticou, v ktorej stĺpcoch a riadkoch sú umiestnené entity, ktoré vytvárajú páry. V tejto notácii nám ale zaniká možnosť zaznamenať smer spojenia. Príklad sociometrickej notácie množiny entít z príkladu v Tab.3.1 ilustruje Tab.3.2, Tab.3.3 a Tab.3.4.

Tab. 3.2. Sociometrická notácia – PRIATEĽSTVO na ZAČIATKU [Wasserman-Faust, 2009].

	Allison	Drew	Eliot	Keith	Ross	Sarah
Allison	-	1	0	0	1	0
Drew	0	-	1	0	0	1
Eliot	0	1	-	0	0	0
Keith	0	0	0	-	1	0
Ross	0	0	0	0	-	1
Sarah	0	1	0	0	0	-

Tab. 3.3. Sociometrická notácia – PRIATEĽSTVO na KONCI [Wasserman-Faust, 2009].

	Allison	Drew	Eliot	Keith	Ross	Sarah
Allison	-	1	0	0	1	0
Drew	0	-	1	0	1	1
Eliot	0	0	-	0	1	0
Keith	0	1	0	-	1	0
Ross	0	0	0	1	-	1
Sarah	0	1	0	0	0	-

Tab. 3.4. Sociometrická notácia – NEĎALEKÉ BYDLISKO [Wasserman-Faust, 2009].

	Allison	Drew	Eliot	Keith	Ross	Sarah
Allison	-	0	0	0	1	1
Drew	0	-	1	0	0	0
Eliot	0	1	-	0	0	0
Keith	0	0	0	-	1	1
Ross	1	0	0	1	-	1
Sarah	1	0	0	1	1	-

Tretia notácia - **algebraická notácia**, býva často využívaná pri skúmaní takých sociálnych sietí, ktoré sa vyznačujú početnými reláciami. Táto notácia je vhodná na skúmanie rolí jednotlivých entít v sociálnej sieti využitím rôznych algebraických prostriedkov, ktoré sa dajú dobre využiť na porovnávanie a kategorizáciu jednotlivých relácií. Ak máme k dispozícii sociálnu sieť, ktorá je vyjadrená reláciami “A je priateľ B” a alebo “A je nepriateľ B”, je možné pomocou algebraickej notácie detekovať sadu relácií “priateľ” a “nepriateľ”.

3.3.2 Notácia dynamickej siete

Skúmanie sociálnych sietí v dynamike umožňuje sledovať zmeny v sociálnej sieti vo vymedzenom časovom rozmedzí a následne ich analyzovať. Stačí obohatiť vyššie uvedené notácie predchádzajúcej kapitoly o časovú zložku. Potom dokážeme jednoznačne kategorizovať jednotlivé entity a ich väzby v konkrétnom čase. Časová zložka v notácii môže byt vyjadrená viacerými spôsobmi. Príkladom môže byť notácia pomocou teórie grafov, kde každá dvojica predstavujúca istú väzbu, je zároveň označená časovým údajom. Taká notácia je príznačná pri použití metódy **vzorkovania**, keď sa stav siete zaznamenáva (vzorkuje) v pravidelných časových intervaloch. Iným spôsobom notácie dynamiky sociálnej siete je vyznačenie **trvania existencie** danej väzby. V takom prípade sa zaznamenáva, v akom časovom rozmedzí sledovaná väzba trvá, v akom čase sa väzba pretrhla a ak sa prerušila, kedy vznikla znova.

3.4 Sieťové štatistiky

Sieťové štatistiky sú základom analýzy sociálnych sietí. Na účely definície vlastností sociálnych sietí, dôležitých pre ich analýzu, je vhodné použiť notáciu podľa teórie grafov. Teória grafov, ako matematický systém obsahujúci množinu entít a množinu väzieb medzi týmito entitami je veľmi vhodnou reprezentáciou sociálnej siete. Grafické zobrazenie sociálnej siete je hlavným východiskom pri jej analýze. Vizualizačný potenciál grafov dobre poslúži zjednodušeniu a čitateľnosti reprezentácie sociálnej siete a napomôže odhaliť také spojenia, ktoré by mohli ostať pri inej notácii skryté. Graf, ako model sociálnej siete, pracuje s neorientovanými dichtonomickými reláciami v zmysle existencie, resp. neexistencie väzby. Aj neorientované relácie môžu byť nositeľmi zaujímavých informácií o spolupráci entít, ich ekvivalentnej príslušnosti, ako napríklad "je príbuzný", "býva blízko", "pracuje s" a pod. Graf *G* je reprezentovaný dvoma množinami: množinou entít $N = \{n_1, n_2, \ldots, n_g\}$ a množinou hrán $L = \{l_1, l_2, \ldots, n_L\}$. Existuje teda *g* entít a *L* hrán. V grafickej reprezentácii je každá hrana reprezentovaná neusporiadanou dvojicou $l_k = (n_i, n_j)$. Kedže ide o neorientovanú dvojicu, každá hrana medzi entitami n_i, n_j je ekvivalentná hrane medzi n_j ,n_i. Slučky, keď existuje hrana medzi entitou samotnou (n_i, n_i), nie sú uvažované. Teda platí, že $l_k = (n_i, n_j) = (n_j, n_i)$ a graf je *G(N, L)*. Jednoduchú reprezentáciu pomocou teórie grafov ilustruje Tab.3.5, ktorá korešponduje s Tab.3.2 až Tab.3.4. Tento príklad reprezentuje analýzu väzby "býva blízko" nad skupinkou žiakov, pre ktorú platí $g = 6$ a $L = 6$.

Tab. 3.5. Reprezentácia relácie „Neďaleké bydlisko" [Wasserman-Faust, 2009].

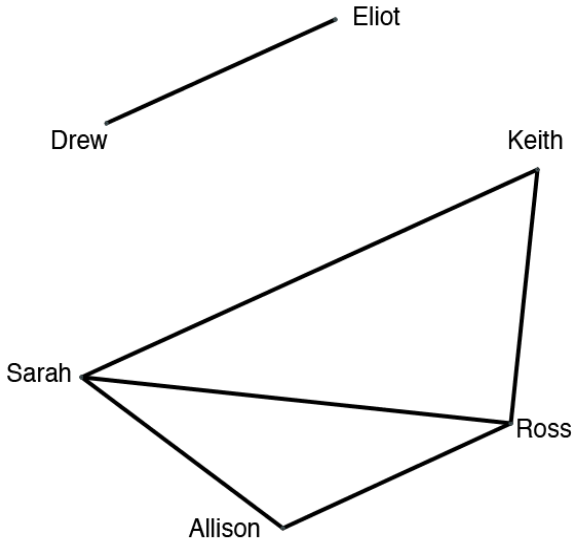
	Entita	Neďaleké bydlisko
n_1	Allison	Ross, Sarah
n_2	Drew	Eliot
n_3	Eliot	Drew
n_4	Keith	Ross, Sarah
n_5	Ross	Allison, Keith, Sarah
n_6	Sarah	Allison, Keith, Ross
$l_1 = (n_1, n_5)$		
$l_2 = (n_1, n_6)$		
$l_3 = (n_2, n_3)$		
$l_4 = (n_4, n_5)$		
$l_5 = (n_4, n_6)$		
$l_6 = (n_5, n_6)$		

Analýza sociálnej siete sa sústreďuje na vyčíslenie a skúmanie vlastností tejto siete, ktoré je možné odhadnúť v rámci danej notácie. Na to sa využívajú rozličné sieťové štatistiky. V ďalšom budú tieto štatistiky charakterizované. Budú uvedené konkrétne veličiny popisujúce tak lokálne ako aj globálne vlastnosti najčastejšie používané v analýze sietí rozličného druhu, ako napríklad stupeň módu, najkratšia cesta,

konektivita, koeficient zhlukovania, centralita a prestíž ako aj globálna charakteristika - separácia uzlov.

3.4.1 Stupeň módu

Stupeň nódu resp. uzlu v grafe, $d(n_i)$ predstavuje počet väzieb, ktoré mu bezprostredne prislúchajú, teda je vyjadrením počtu entít, s ktorými je skúmaná entita v koincidencii. Rozsah hodnôt stupňa nódu sa pohybuje od hodnoty 0 (nód nemá žiadnu väzbu – hovoríme o izolovanom nóde) až po g −1(nód má väzby so všetkými možnými ostatnými nódmi). Vychádzajúc z Tab.3.5 je možné vyčísliť stupne jednotlivých entít nasledovne: $d(n_1) = 2$, $d(n_2) = 1$, $d(n_3) = 1$, $d(n_4) = 2$, $d(n_5) = 3$ a $d(n_6) = 3$, čo je možné ľahko overiť na Obr.3.2.



Obr. 3.2. Grafická reprezentácia – stupne nódov.

Aj najjednoduchšia charakteristika ako je výpočet stupňa uzla môže byť veľmi nápomocná pri analýze sociálnych sietí. Napríklad, keď sa vyskytuje v sociálnej sieti entita s veľmi malým stupňom, môže to indikovať používateľa so slabým záujmom o spoločnosť. Na druhej strane entita s vysokým stupňom, môže predstavovať známu osobnosť s mnohými priateľmi.

V zdroji [Dorogovstev-Mendes, 2003] je definovaný stupeň nódu k ako súčet všetkých vstupných a výstupných väzieb daného nódu. Z fyzikálneho hľadiska stupeň nódu vyjadruje konektivitu nódu, pričom vstupný stupeň k_i reprezentuje počet vstupných konektivít, tj. väzieb ktoré sa pripájajú k sledovanému nódu. Výstupný stupeň k_o reprezentuje počet odchádzajúcich konektivít, resp. väzieb, ktoré inicioval daný nód smerom k inému nódu. Výsledný stupeň k je tvorený ich súčtom (4):

$$k = k_i + k_o \tag{4}$$

Stupeň k teda vyjadruje počet najbližších susedov, ktorí sa nachádzajú v okolí aktuálneho nódu n_i. Celková distribúcia stupňov uzlov danej siete $P(k_i,k_o)$ je daná ako súčet vstupných a výstupných spojení celej siete $P(k)$. Z daného vyplýva jú nasledovné vzťahy (5), (6) a (7):

$$P(k) = \sum P(k_i, k - k_i) = \sum P(k - k_o, k_o), \quad (5)$$

$$P_i(k_i) = \sum P(k_i, k_o), \quad (6)$$

$$P_o(k_o) = \sum P(k_i, k_o). \quad (7)$$

Kvôli prehľadnosti sa miesto notácie $P_i(k_i)$, resp. $P_o(k_o)$ zvyčajne používa notácia $P(k_i)$, resp. $P(k_o)$.

3.4.2 Konektivita siete

Ak sieť nemá žiadne spojenia s okolitým svetom, potom priemerný vstupný a výstupný stupeň všetkých nódov siete sa rovnajú. Pri analýze sociálnych sietí, dôležitú úlohu zohráva priemerný stupeň celej siete, ktorý určuje takzvanú konektivitu siete. Tento štatistický parameter je možné určiť pomocou vzťahu (8):

$$d = [\sum d(n_i)] / g = 2L / g \quad (8)$$

Hodnota d môže byť užitočná napríklad na odhalenie pravidelnej siete, pretože všetky nódy takejto siete majú rovnaký stupeň. Teda uniformná sieť má konštantný stupeň nódu pre všetky jej nódy.

3.4.3 Najkratšia cesta

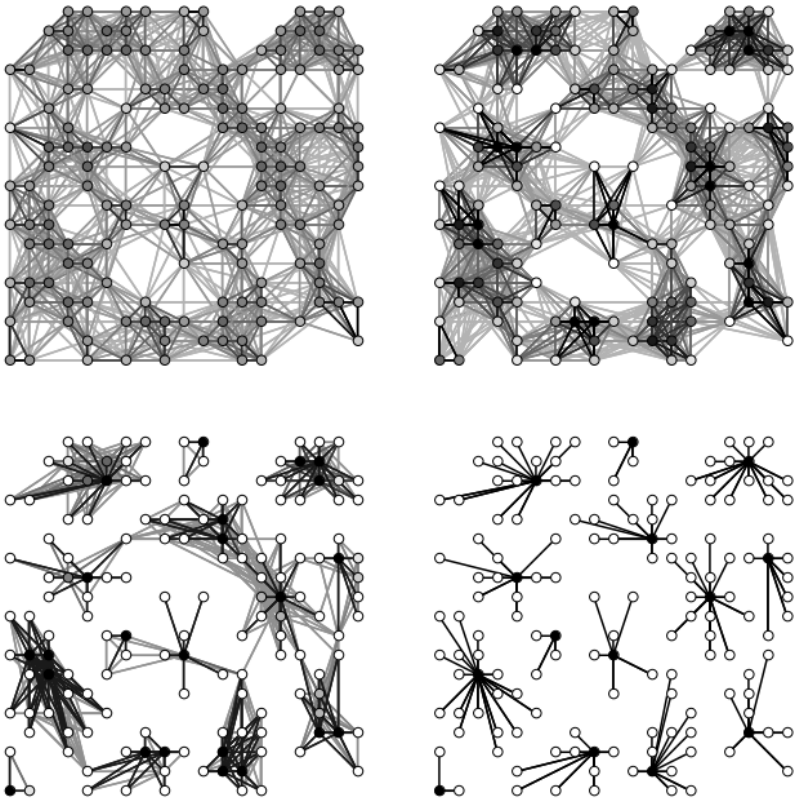
Predpokladajme, že geodetická vzdialenosť dvoch bodov n_i *a* n_j je definovaná ako najkratšia možná vzdialenosť medzi danými nesusednými bodmi $p(n_i,n_j)$. Za povšimnutie stojí, že $p(n_i,n_j)$ nemusí byť to isté ako $p(n_j,n_i)$. Potom priemernú najkratšiu vzdialenosť siete je možné definovať ako priemer všetkých možných najkratších ciest medzi entitami. Táto priemerná najkratšia vzdialenosť siete je často citovaná ako „priemer“ (diameter) siete.

3.4.4 Koeficient zhlukovania

[Dorogovstev-Mendes, 2003] popisuje koeficient zhlukovania ako funkciu závislú na počte väzieb entity a jej najbližšieho okolia. Pre siete s neorientovanými spojnicami môžeme zjednodušiť výpočet všetkých možných spojníc aktuálneho uzlu so susednými uzlami nódu µ *na* $z_1^{(\mu)}(z_1^{(\mu)}-1)/2$. Ak berieme do úvahy iba istý počet z nich $y^{(\mu)}$, potom koeficient zhlukovania je daný vzťahom (9):

$$C^{(\mu)} = 2y^{(\mu)} / z_1^{(\mu)}(z_1^{(\mu)} - 1) \quad (9)$$

Spriemerovaním $C^{(\mu)}$ nad celou sieťou dostaneme koeficient zhlukovania C. Tento koeficient vyjadruje pravdepodobnosť, že dva najbližšie uzly siete sú zároveň bodom s najkratšou cestou aspoň jedného ďalšieho uzlu. “*Orientovaný graf je možné označiť za schopný zhlukovania, ak je možné rozdeliť uzly grafu do konečného počtu podskupín a to tak, že pozitívne označené väzby spájajú dva uzly v jednom zhluku a negatívne označené väzby spájajú dva uzly z rôznych zhlukov. Takto vytvorené zhluky nazývame aj partície.*” [Wasserman-Faust, 2009] Algoritmus tohto zhlukovania uzlov siete a zhlukovú analýzu ilustruje Obr. 3.3.



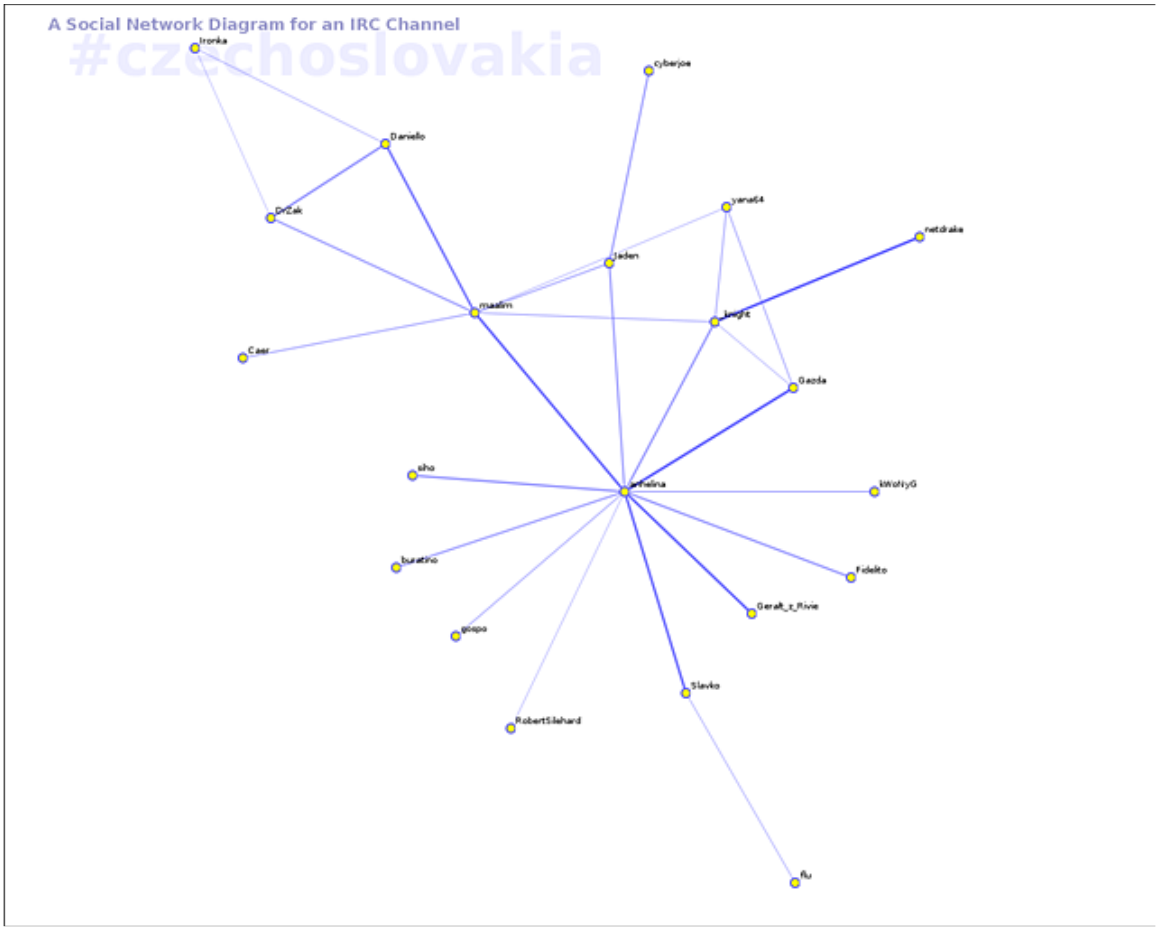
Obr. 3.3. Vyhľadávanie zhlukov podľa Markovovho algoritmu. Obrázok vľavo hore reprezentuje východzí stav siete a obrázok v pravo dole stav siete po aplikácii Markovovho algoritmu. [Dongen, 2000]

3.4.5 Vážený graf

Pri analýze sociálnych sietí sa často používa množina dát, v ktorej sú k dispozícii väzby medzi jednotlivými entitami siete spolu s ich ohodnotením. V bežnej dichtonomickej sieti je váha relácie ohodnotená iba hodnotou 1 pri existencii väzby alebo hodnotou 0 pri jej absencii. V praxi sa často stretávame s vážením väzieb medzi entitami na základe istej porovnateľnej zložky. Príkladom môže byt frekvencia interakcií medzi dvoma používateľmi sociálnej siete teda jej entitami. Podľa [Wasserman-Faust, 2009] môžeme takúto váhu nazývať aj *magnitúda* alebo jednoducho *hodnota hrany*. *Vážený (orientovaný) graf* je graf, v ktorom každá hrana (väzba) má svoju hodnotu. Vážený graf je možné reprezentovať trojicou množín a to nie len množinou vrcholov (entít) $N = \{n_1,n_2, \ldots ,n_g\}$ a hrán $L = \{l_1, l_2, \ldots , l_L\}$ ale aj množinou váh $V = \{v_1,v_2, \ldots ,n_L\}$. Orientovaný respektíve vážený graf je teda možné vyjadriť ako VG (N,L,V). Na ohodnotenie väzieb medzi entitami môže byť použitý stupeň nódu. Iný spôsob váhovania väzieb môže byť založený na dĺžke trvania vzťahu (priateľstva, komunikácie, chatu) medzi entitami sociálnej siete.

3.4.6 Centralita a prestíž

V poslednom čase sa analýza sociálnych sietí zameriava aj na identifikáciu autoritatívnych aktérov sociálnej siete. Títo aktéri spravidla v notácii pomocou teórie grafov predstavujú centrálny uzol časti (respektíve celej) sociálnej siete. Odpovedajúca časť grafu je potom zoskupená do hviezdy, viď Obr.3.4.

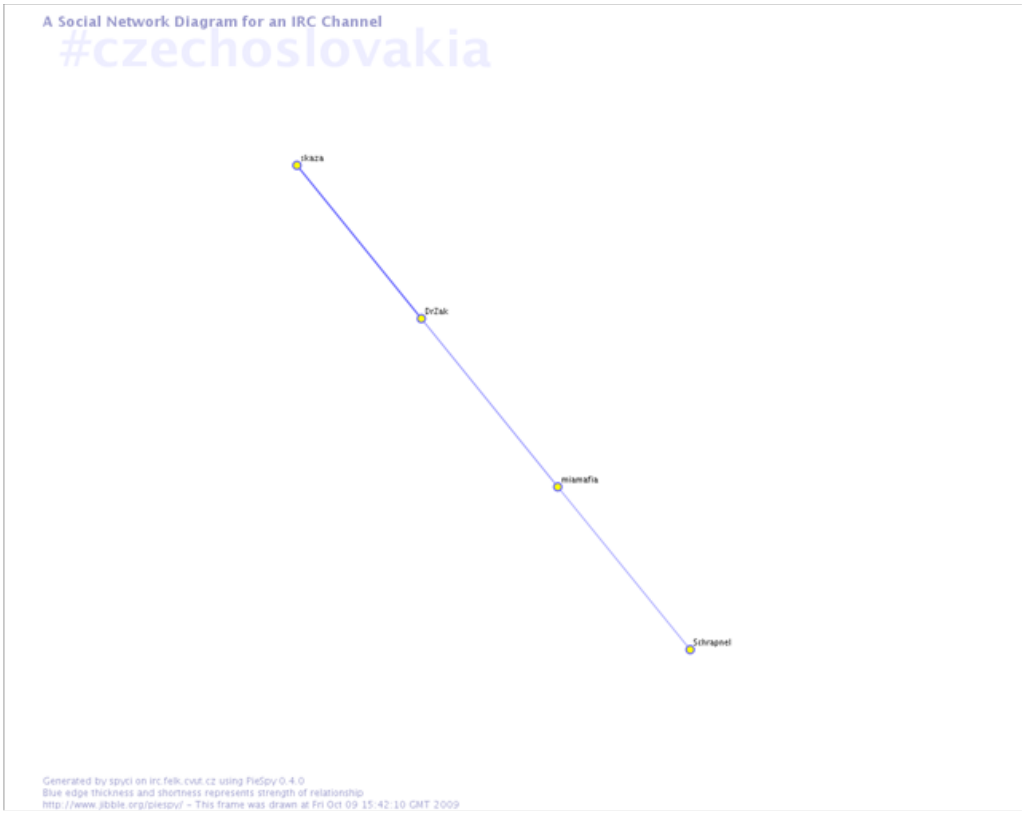


Obr. 3.4. Zoskupenie uzlov sociálnej siete IRC Chanel do útvaru typu hviezda v dolnej časti obrázku.

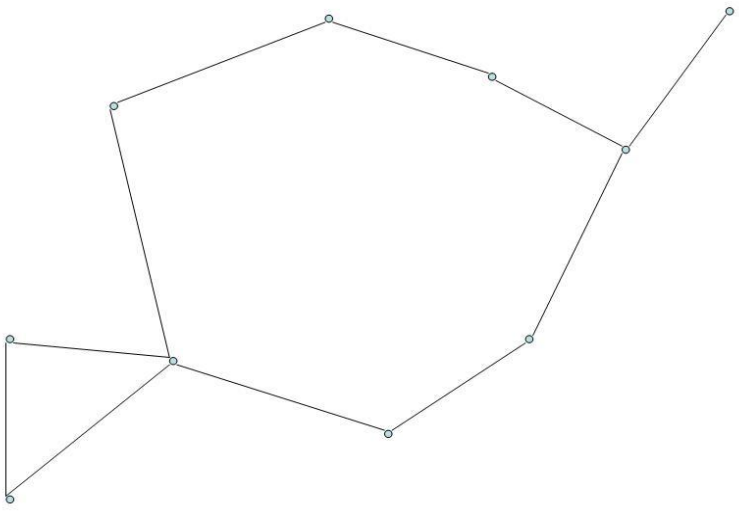
Ďalšie dva typické útvary v grafovej notácii sociálnej siete, ktoré sú zaujímavé z hľadiska jej analýzy, sú čiara (viď Obr.3.5) a kruh (viď Obr.3.6). Tab.3.6 obsahuje incidenčné matice pre zoskupenie typu hviezda, čiara a kruh. Už z týchto matíc je možné vyčítať o aký druh zoskupenia ide.

Autoritatívni aktéri respektíve prominenti vytvárajú v sociálnej sieti akési centrá, okolo ktorých sa zhlukujú ostatní používatelia. Takíto prominentní aktéri sa výrazne podieľajú na vzájomnom vzťahu iných entít a naopak iní aktéri zvyšujú svojim záujmom kredibilitu resp. viditeľnosť prominenta. V analýze sociálnych sietí sa najčastejšie používajú dva druhy viditeľnosti a to centralita a prestíž.

Centralita. Hlavným znakom prominentného aktéra v sociálnej sieti je jeho vysoký počet väzieb s okolitým svetom. Pri pojme centralita a použití orientovaných spojení je málo dôležité, či majorita týchto spojení spočíva v prichádzajúcich väzbách (aktér je prijímateľ) alebo v odchádzajúcich väzbách (aktér je odosielateľ). Skúma sa nakoľko silné je spojenie autora s inými aktérmi. Skúmanie centrality má veľký význam pri skúmaní štruktúry skupín hlavne v komunikačných sieťach.



Obr. 3.5. Zoskupenie uzlov sociálnej siete IRC Chanel do útvaru typu čiara.



Obr. 3.6. Zoskupenie uzlov sociálnej siete IRC Chanel do útvaru typu kruh.

Tab. 3.6. Incidenčné matice pre zoskupenie typu hviezda, čiara a kruh.

Hviezda	0	1	1	1	1	1	1
	1	0	0	0	0	0	0
	1	0	0	0	0	0	0
	1	0	0	0	0	0	0
	1	0	0	0	0	0	0
	1	0	0	0	0	0	0
	1	0	0	0	0	0	0
Čiara	0	1	1	0	0	0	0
	1	0	0	1	0	0	0
	1	0	0	0	1	0	0
	0	1	0	0	0	1	0
	0	0	1	0	0	0	1
	0	0	0	1	0	0	0
	0	0	0	0	1	0	0
Kruh	0	1	0	0	0	0	1
	1	0	1	0	0	0	0
	0	1	0	1	0	0	0
	0	0	1	0	1	0	0
	0	0	0	1	0	1	0
	0	0	0	0	1	0	1
	1	0	0	0	0	1	0

Prestíž. Ak sa v rámci skúmania centrality zameriame na rozlíšenie, či majoritná časť väzieb aktéra s vysokou centralitou je tvorená prichádzajúcimi alebo odchádzajúcimi väzbami, potom môžeme definovať prestížneho aktéra ako aktéra s vysokým počtom väzieb, ktoré smerujú k nemu, teda aktér je prevažne prijímateľ takýchto väzieb. Prestíž má vyššiu vypovedaciu hodnotu ako centralita ale niekedy nie je možné ju úspešne merať. Taktiež platí, že v prípade nárastu väzieb aktéra sa síce zvyšuje jeho centralita, ale prestíž nemusí rásť, ak iniciátorom väzieb bol prevažne aktér samotný. Kvantifikácia prestíže predpokladá použitie orientovaných väzieb a možnosť merať vstupný stupeň uzla.

3.4.7 Separácia uzlov siete

Separácia uzlov siete predstavuje sieťovú štatistiku zameranú na globálne vlastnosti siete. Pomocou veličiny zvanej separácia uzlov siete je možné merať blízkosť uzlov na základe priemernej najkratšej vzdialenosti. Jej definícia je zrejmá už z názvu. Vyberáme náhodne dvojice uzlov v sieti a spočítame, aký najmenší počet hrán musíme precestovať, aby sme sa od uzla i dostali k uzlu j. Priemer týchto vzdialeností (určitého počtu náhodne vybraných uzlov) nazývame **separáciou uzlov danej siete**. Takže pod separáciou uzlov Z rozumieme priemernú najkratšiu

vzdialenosť medzi dvoma náhodne zvolenými uzlami siete väčšinou meranú pomocou vyššie definovanej sieťovej štatistiky - najkratšia cesta v grafe.

3.4.8 Ďalšie charakteristiky

Medzi ďalšie charakteristiky používané na analýzu sociálnych sietí patria podľa [Repka, 2011] nasledovné atribúty:

- *Počet uzlov* (Nodes) v sieti
- *Počet hrán* (Edges) v sieti
- *Počet uzlov v najväčšom slabo prepojenom komponente* (Nodes in largest WCC)
- *Počet hrán v najväčšom slabo prepojenom komponente* (Edges in largest WCC)
- *Počet uzlov v najväčšom silne prepojenom komponente* (Nodes in largest SCC)
- *Počet hrán v najväčšom silne prepojenom komponente* (Edges in largest SCC)
- *Počet trojuholníkov* (Number of triangles) t.j. počet trojíc uzlov navzájom prepojených, pričom sa uvažuje neorientovaná sieť.
- *Podiel uzavretých trojuholníkov* (Fraction of closed triangles) t.j. podiel počtu trojuholníkov a počtu (neorientovaných) ciest dĺžky 2.
- *Priemer siete* (najdlhšia najkratšia cesta) t.j. dĺžka maximálnej neorientovanej najkratšej cesty, pri veľkých sieťach väčšinou určená odhadom (analýzou vzorky náhodných uzlov).
- *Efektívny priemer 90. percentilu* t.j. 90. percentil distribúcie dĺžiek neorientovaných najkratších ciest (vzorkovaný z určitého počtu náhodne vybratých uzlov).

3.5 Kohézne podskupiny

Identifikácia a analýza kohéznych podskupín patrí medzi bežné ciele analýzy sociálnych sietí. Kohézne podskupiny sú podmnožiny aktérov medzi ktorými existujú relatívne silné, priame, intenzívne alebo inak povedané frekventované väzby [Wasserman-Faust, 2009]. Je známych viacero techník na nájdenie týchto podskupín. Tie sa delia podľa typu dát, ktoré majú byť analyzované. Inak sa spracovávajú dáta získané z jednodómových sietí a inak dáta z dvojmódových sietí. Podskupiny jednomódových sietí sa totiž zameriavajú na vlastnosti párových väzieb, zatiaľ čo podskupiny dvojmódových sietí sa zameriavajú na väzby aktérov podľa ich príslušnosti k organizáciám (kolektívom).

Kohézne podskupiny sú identifikované na základe skúmania nasledovných charakteristík:

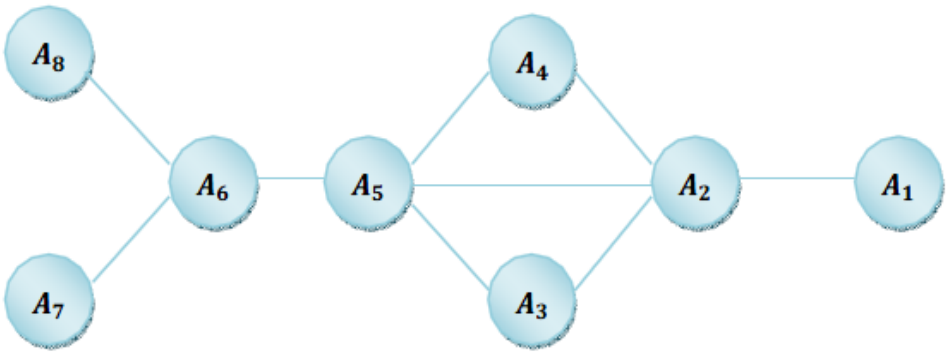
- počet alebo kompletnosť vzájomných väzieb,
- blízkosť alebo dosiahnuteľnosť aktérov podskupiny,
- frekvencia väzieb medzi aktérmi [Tutoky, 2010].

Klika. Do kategórie podskupín identifikovaných na základe vzájomnosti väzieb sa radia hlavne podskupiny identifikované pomocou kliky. Definícia kliky dovoľuje rozpoznať viacero podskupín sociálnej siete, ktoré sa môžu prekrývať. V sieti na Obr.3.7 je možné nájsť dve kliky: $A_{SG1} = \{A_2, A_3, A_5\}$ a $A_{SG2} = \{A_2, A_4, A_5\}$. Definícia

kliky je striktná a tak sa v reálnych sieťach nenachádza veľké množstvo kompletne prepojených častí siete. Častejší prípad je, keď sa v sieti nájde malý počet klík s nízkym počtom vrcholov.

N-klika. Kohézne podskupiny je možné nájsť pomocou n-kliky na základe známej dosiahnuťeľnosti a priemeru siete. N-klika má vlastnosti, ktoré znižujú prísnosť základnej definície kliky tak, že umožňujú, aby aktéri boli v podskupine aj keď nie sú priamo prepojení. Na príslušnosť do skupiny stačí podmienka dosiahnuteľnosti prostredníctvom iných väzieb s inými aktérmi, pričom sa uvažujú iba krátke cesty.

Stupeň vrcholu. Ďalšou možnosťou je identifikovať podskupiny pomocou stupňa vrcholu. Tento prístup skúma priľahlosť jednotlivých členov tak, že zisťuje či každý aktér skúmanej podskupiny má aspoň minimálny preddefinovaný počet priľahlých aktérov, ktorí sú tiež členmi danej podskupiny. Mierou tohto minimálneho počtu priľahlých aktérov je vlastne stupeň vrcholu. Takto definovaná kohézna podskupina sa vyznačuje vysokou robustnosťou, čo znamená, že odobratím náhodného aktéra nie je tak ľahko negovateľná, ako pri predchádzajúcich typoch podskupín. Najznámejší reprezentanti takýchto kohéznych podskupín založených na stupni vrcholu sú k-plexa a k-jadro [Tutoky, 2010].



Obr. 3.7. Príklad jednoduchej siete. [Repka, 2011]

3.6 Analýza reálnych sietí

Reálne siete sú siete, ktoré vznikajú tzv. samo organizovaným spôsobom. Takéto siete vznikajú a rozširujú sa pomocou mechanizmu preferenčného pripájania uzlov. Uzly a hrany sa v rámci tohto mechanizmu pridávajú tam, kde sa to podľa istých vnútorných pravidiel siete najviac očakáva. Väčšinou sú to bezškálové siete, ktoré sa vyznačujú robustnosťou a teda dobre odolávajú náhodným zmenám v sieti. Taká náhodná zmena môže byť reprezentovaná napríklad náhodným vymazaním uzlov v dôsledku poruchy. Robustnosť zaručuje stabilitu štruktúry siete aj v prípade náhlej zmeny podmienok. Tvar distribučnej funkcie stupňov vrcholov v reálnych sieťach často odzrkadľuje aj vlastnosti sietí malého sveta. Typické vlastnosti týchto sietí sú malá separácia uzlov a veľký koeficient klasterizácie. Procesy samo organizácie optimalizujú nasledovné:

- zachovanie lokálnej štruktúry siete,
- zachovanie dobrej komunikácie medzi uzlami siete,
- dobrú odolnosť voči náhodným poruchám.

Príkladmi reálnych sietí sú: internet, siete profesionálnych kontaktov, komunikačné siete a podobne.

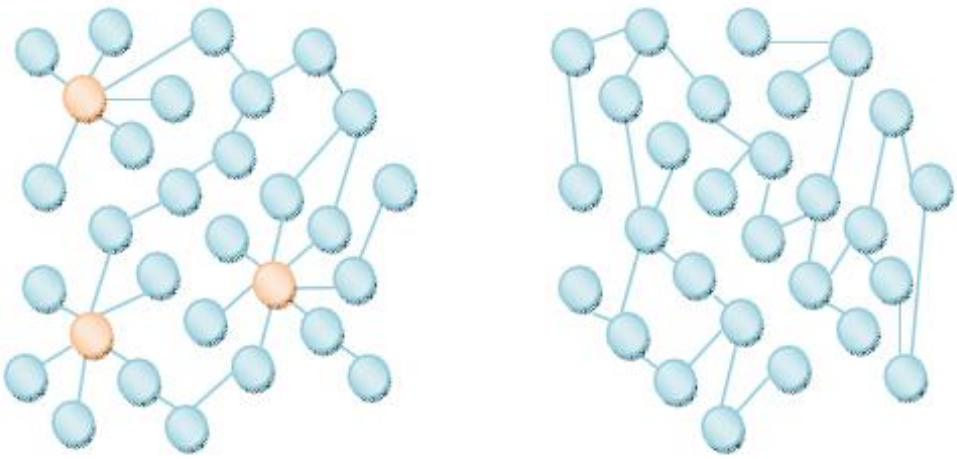
3.6.1 Siete malého sveta

Sieť malého sveta je taká sieť, ktorej väčšina uzlov nie je priamo prepojená ale je dosiahnuteľná z ostatných uzlov prechodom cez malý počet hrán, takzvanou krátkou cestou. Formálne je takáto sieť definovaná ako sieť, kde vzdialenosť $d(i,j)$ medzi dvoma ľubovoľnými aktérmi i a j rastie proporcionálne s logaritmom $log(N)$, kde N je počet aktérov v sieti.

Vo [Watts-Strogatz, 1998] bola identifikovaná kategória sietí malého sveta ako náhodne generované grafy s dvoma podstatnými nezávislými atribútmi, čo je tzv. Watt-Strogatzov model. Čisto náhodné siete generované pomocou Erdős-Rényi modelu vykazujú krátku priemernú dĺžku najkratšej cesty s malým klasterizačným koeficientom. Siete malého sveta majú krátke najkratšie cesty ale na druhej strane klasterizačný koeficient je značne väčší ako u čisto náhodných sietí. Z toho vyplýva, že siete malého sveta je možné charakterizovať kombináciou veľkého klasterizačného koeficientu s malou separáciou uzlov. Watt-Strogatzov model sa snaží emulovať tieto vlastností do generovaných grafov. Siete malého sveta sú čiastočne usporiadané, s niekoľkým hranami, ktoré sú prepojené akoby náhodným spôsobom. Medzi zaujímavé vlastnosti sietí malých svetov patrí tendencia obsahovať kliky a n-kliky (dôsledok vysokého klasterizačného koeficientu) a ďalej nadbytok „hubs“ (uzlov s vysokým stupňom vrcholu).

3.6.2 Bezškálové siete

Bezškálové siete sú siete ktorých distribúcia stupňa je minimálne asymptoticky závislá od tzv. zákona sily. Znamená to, že podiel aktérov $P(k)$ v sieti s k väzbami závisí od $ck^{-\gamma}$, kde c je konštanta a γ je zvyčajne z intervalu (2, 3). Pre modelovanie javu spôsobeného distribúciou stupňa podľa zákona sily sa používa preferenčný spôsob pripájania uzlov do siete. Tento spôsob však generuje len špecifickú podskupinu bezškálových sietí, preto bolo navrhnutých mnoho alternatívnych algoritmov. Snaha formálne popísať pojem bezškálový viedla k tzv. bezškálovej metrike. Obr.3.8 ilustruje bezškálovú sieť (ľavá časť obrázka).



Obr. 3.8. Príklad bezškálovej a náhodnej siete. [Repka, 2011]

3.6.3 Náhodné a mriežkové siete

Náhodné grafy (viď pravú časť Obr.3.8) majú klasterizačný koeficient blízky k nule a separáciu uzlov, ktorá rastie s počtom uzlov pomaly. Teda *d(i, j)* sa blíži *log(N)*. Usporiadané, mriežkové grafy oproti tomu môžu mať veľký klasterizačný koeficient a separácia uzlov rastie s počtom uzlov rýchlo, pokiaľ je dimenzia mriežky malá. Pre separáciu uzlov v mriežkových grafoch platí, že separácia uzlov *L* sa blíži N^{-d}, kde *d* je dimenzia mriežky [Markošová, 2010].

3.6.4 Sieť typu Sociálne kruhy

Generatívny model siete Sociálne kruhy je modelom tzv. siete so spätnou väzbou [Douglas et al., 2006]. Tento model popisuje triedu náhodných grafov generovaných jednoduchými procesmi formovania väzieb a spätných slučiek v sociálnych kruhoch. Táto skupina sietí sa síce odlišuje od sietí malého sveta a bezškálových sietí ale na druhej strane modeluje mnoho charakteristík reálnych sietí.

Sieťou so spätnými väzbami sa označuje sieť, ktorej aktívne uzly (aktéri, agenti) komunikujú cez sieť, aby koordinovali pripájanie nových aktérov. Snažia sa vytvoriť nové väzby. Proces prebieha tak, že ak zatiaľ nepripojený aktér splní isté podmienky (napr. teoretická vzdialenosť), vytvorí sa nová väzba s aktérom a taktiež slučka so spätnou väzbou. Ak sa nepripojenému aktérovi nepodarí nájsť ďalšieho partnera - aktéra vo vnútri siete, tak aktívny aktér v rámci siete (ktorý už bol zapojený vo výbere, ale nesplnil podmienky) zavolá nového partnera - aktéra mimo siete a vytvorí väzbu. Siete modelované týmto procesom reprezentujú evolúciu veľkých sietí s kohéznymi väzbami, teda tzv. sociálne kruhy. Touto evolúciou nakoniec môžeme nagenerovať aj siete malého sveta, bezškálové siete alebo siete iných topológií.

Generatívny model siete môže byť nápomocný pri hľadaní odpovedí na otázky: ako sú v sieti vytvárané huby, ako vznikajú dôležité hrany a taktiež vysvetlenie, ako sa tvoria uzavreté cesty (kruhy, cykly). Model sociálne kruhy vytvára sieť na základe informácii o rozptyle a kohéznosti siete, ktoré používa ako funkciu aktivity a spätnej väzby. Tento model sietí môže simulovať topológie mnohých typov reálnych sietí, pričom tieto topológie budú odrážať základné sieťové štatistiky ako je hustota siete, klasterizácia, kohéznosť a distribúcia stupňa aktérov.

POUŽITÁ LITERATÚRA

[Dongen, 2000] Dongen, van, S.: *Graph clustering*. 2000, ISBN 90-393-2408-5.

[Dorogovstev-Mendes, 2003] Dorogovstev S. N., Mendes J. F. F.: *Evolution of networks*. Oxford, 2003.

[Douglas et al., 2006] Douglas, R. W., Kejzar, N., Tsallis, C., Farmer, J. D., Scott, D.W.: *Generative Model for Feedback Networks*. In Physical Review E, 2006.

[Han, 2003] Han, W.P.: *Hyperlink Network Analysis*: A New Method for the Study of Social Structure on the Web Royal Netherlands Academy of Arts & Sciences (KNAW) Amsterdam, The Netherlands, CONNECTIONS 25(1). 2003. 49-61

[Marin-Wellman, 2009] Marin, A., Wellman, B.: *Social Network Analysis: An Introduction*. London: Forthcoming in Handbook of Social Network Analysis, 2009 [online]. [cit. 2014-04-10] Dostupné na internete: http://www.chass.utoronto.ca/~wellman/publications/newbies/newbies. pdf.

[Markošová, 2010] Markošová M. *Networks Dynamics*. In: Kvasnička et al. (ed.) Artificial Intelligence and Cognitive Sciences II. Bratislava (SK): STU, 2010, p. 321-377.

[Rakuščinec, 2009] Rakuščinec, T.: *Vizualizácia sociálnych sietí.* Košice, Technická univerzita v Košiciach, Fakulta elektrotechniky a Informatiky, 2009, 1-30.

[Repka, 2011] Repka, M.: *Analýza určitých typov sociálnych sietí.* Košice, Technická univerzita v Košiciach, Fakulta elektrotechniky a Informatiky, 2011. 1-75.

[Tutoky, 2010] TUTOKY, G.: *Objavovanie a využívanie znalostí v sociálnych sieťach*. Fakultaelektrotechniky a informatiky. TUKE. Košice. 2010.

[Wasserman-Faust, 2009] Wasserman, S., Faust, K.: *Social Network Analysis*. Cambridge, Cambridge University Press, 2009, 1-825, ISBN 9178-0-521-38707-1.

[Watts-Strogatz, 1998] Watts, D. J., Strogatz, S. H: *Collective dynamics of 'small-world' networks*. Nature 393 (6684), 1998, 440–442, [online]. [cit. 2014-04-28] Dostupné na internete: <http://tam.cornell.edu/tam/cms/manage/upload/SS_nature_smallworld.pdf>

4 Dynamická analýza sociálnych sietí

4.1 Úvod

Všetky druhy sietí, teda aj sociálne siete sa vyvíjajú v čase. Preto okrem statických vlastností je možné ich analyzovať aj z hľadiska dynamiky. Klasická analýza sociálnych sietí sa sústreďuje na štrukturálnu časť siete a vyhodnocovanie statických charakteristík resp. sieťových štatistík. Na druhej strane, dynamická analýza sa sústreďuje na dynamické zmeny v sieti a teda vyžaduje okrem klasických statických dát aj dáta zachytávajúce časovú dimenziu. Preto sa pri dynamickej analýze pridávajú na vstup ďalšie atribúty, ktoré reprezentujú temporálne dáta, ako napríklad časové údaje o vzniku, zmene, zániku jednotlivých aktérov a väzieb. Dynamická analýza vyššej úrovne dokáže spracovávať aj dáta s určitou úrovňou nejasnosti, napr. pravdepodobnostné modely. V tomto prípade hovoríme o procese učenia na základe sledovania neustálych zmien týkajúcich sa aktérov siete ako aj zmien ich vlastností - atribútov. Tento proces učenia vedie ku znalostiam o tom, akým smerom sa sieť vyvíja. Dynamická analýza môže používať širokú škálu metód a techník ako napríklad multi-agentové modelovanie, strojové učenie, vizualizačné techniky a pod. Podľa [Repka, 2011], typickými problémami, ktorými sa dynamická analýza zaoberá, sú:

- hľadanie vhodnej metriky pre dynamickú analýzu,
- predikcia zmien v sociálnych sieťach,
- dynamická vizualizácia siete,
- vývoj algoritmov pre monitorovanie sociálnych sietí,
- štúdium sietí v čase.

Napríklad multi-agentové systémy sa dajú s úspechom použiť na učenie, hlavne ak sa jedná o siete s aktívnymi uzlami. V tomto prípade, na základe dobre známych sociálnych a kognitívnych procesov, je možné využívať agentov ako aktívne zložky v sieti. Ich aktívnosť spočíva v tom, že sa v rámci siete dokážu učiť, nadväzovať vzťahy a zúčastňovať sa na akciách. Takíto agenti potom môžu dynamicky meniť sieť pomocou jednoduchých mechanizmov. Aj pri dynamickej analýze, ak má byť úspešná, je potrebné zvoliť resp. objaviť vhodné metriky na popis siete, ktoré umožnia zachytiť pozíciu, postoj, či rolu aktéra v nejakej preddefinovanej skupine

4.2 Dynamika sociálnych sietí

Dynamika v kontexte nejakej sociálnej siete predstavuje procesy, pri ktorých uzly do sociálnych sietí pribúdajú, zanikajú, sú vymazávané pod. V dôsledku toho sa menia aj vlastnosti siete. V mnohých reálnych prípadoch je možné zanikanie uzlov zanedbať. Takúto reálnu sieť je možné potom modelovať pomocou modelov rastúcich sietí [Barabási-Albert, 1999]. Príklady rastúcich sietí sú práve sociálne siete osôb ale aj internetová sieť, ekologické siete, neurónové siete a podobne. Platí, že ak siete rastú samo organizovaným spôsobom, majú mnohé vlastnosti spoločné. K rastúcim sieťam často patria siete malého sveta alebo bezškálové siete. Vlastnosti sietí malého sveta zväčša v sebe integrujú vlastnosti náhodných grafov a usporiadaných mriežkových grafov [Dorogovstev-Mendes, 2003].

Štruktúra siete samozrejme odráža dynamické procesy, ktoré v nej prebiehajú, napríklad spôsob ich rastu. Spätným odhadom niektorých charakteristík siete je možné získať znalosti o histórii tejto siete. Ak ide o siete, v ktorých uzly zanikajú iba

raritne alebo vôbec a oveľa viac uzlov vzniká, hovoríme o dynamike rastúcich sietí. Dôležitú úlohu hrá aj spôsob, akým sú aktéri pripájaní do siete. Tento spôsob pripájania určuje finálnu štruktúru siete. Rozoznávame dva hlavné spôsoby pripájania aktérov do siete a to náhodné pripájanie aktérov a preferenčné pripájanie aktérov [Albert-Barabási, 2001].

4.2.1 Rast siete náhodným pripájaním uzlov

Náhodné pripájanie aktérov do siete je proces tvorby siete diskrétnym spôsobom postupným napájaním aktérov. Teda v každej časovej jednotke pribudne do siete jeden aktér. Tento aktér sa spojí s niektorým starým aktérom jedinou väzbou. Pravdepodobnosť vytvorenia väzby so starým aktérom, je rovnaká pri všetkých už existujúcich aktéroch. Každý aktér je teda spojený s konkrétnym časom príchodu do siete. Predpokladajme, že aktér *s=1* prišiel do siete v čase *.t=1*, aktér *s=2* prišiel do siete v čase *t=2* a podobne. Potom v čase *t* má daná sieť práve *t* aktérov. Tento proces je možné modelovať nasledujúcim vzťahom (10):

$$p(k,s,t+1) = \frac{1}{t+1}\, p(k-1,s,t) + (1+\frac{1}{t+1})\, p(k,s,t) \qquad (10)$$

Pravdepodobnosť toho, že nejaký aktér (prišiel do siete v čase *t* a pozorujeme ho v čase *t+1*) bude mať práve *k* susedov je rovná pravdepodobnosti toho, že daný uzol, ktorý reprezentuje daného aktéra má v čase *t* práve *k-1* susedov a zachytí hranu vytváranú novým uzlom. Z uvedenej rovnice úpravami získame výraz (11) pre priemerný stupeň uzla, ktorý prišiel do siete v čase *t* a pozorujeme ho v čase *t+1*:

$$k(s,t+1) = k(s,t) + \frac{1}{t+1} \qquad (11)$$

Takéto siete nazývame exponenciálnymi sieťami, pretože distribučná funkcia klesá so zvyšujúcim sa stupňom uzla exponenciálne. Veľkosť stupňa uzla je zhora ohraničená, uzly s veľkým stupňom sa prakticky nevyskytujú [Markošová-Náther, 2010].

4.2.2 Rast siete preferenčným pripájaním uzlov

Aj v tomto prípade začneme s jedným aktérom. Preferenčné pripájanie uzlov do siete predstavuje taký proces rastu siete, keď k prvému aktérovi každú časovú jednotku pribudne nový aktér a pripojí sa jednou (alebo aj viacerými) väzbami k existujúcim aktérom. Pravdepodobnosť toho, že starý aktér zachytí novú hranu je však úmerná stupňu starého aktéra. Nový aktér má teda takú stratégiu, že si preferenčne vyberá skôr aktérov s vysokým stupňom, teda takých, ktorí sú dobre prepojení s ostatnými aktérmi. Tento proces je možné modelovať pomocou (12):

$$p(k,s,t+1) = \frac{k-1}{2t}\, p(k-1,s,t) + (1-\frac{k}{2t})\, p(k,s,t) \qquad (12)$$

Tak vzniká podobná rovnica ako v prípade náhodného pripájania ale s rozdielnymi konštantami. Použitá normovacia konštanta je v tomto prípade *2t*, pretože každá väzba prispieva ku zvýšeniu stupňa u dvoch aktérov. Podobne aj rovnicu vyššie možno upraviť tak, aby sme dostali vzťah pre priemerný stupeň uzla, ktorý prišiel do

siete v čase *s* a pozorujeme ho v čase *t*. Rovnica (13) pre tento vzťah vyzerá nasledovne: .

$$\frac{\partial k(s,t)}{\partial t} = \frac{k(s,t)}{2t} \tag{13}$$

Kde *k(s,t)* predstavuje priemerný stupeň uzla, ktorý prišiel do siete v čase *s* a pozorujeme ho v čase *t*.

V takejto sieti sú aktéri väčšinou navzájom veľmi dobre prepojení ale existujú aj takí aktéri, ktorí majú málo spojení s inými uzlami. Takáto sieť sa nazýva bezškálovou sieťou a popísaný model je Barabási-Albert model (takzvaný BA model) s preferenčným pripájaním uzlov. BA model je modelom pre bezškálové siete. Existujú aj mnohé iné procesy, ktorých výsledkom sú bezškálové siete. Avšak tie zväčšia predstavujú iba rôzne variácie základného BA modelu. Ak môžeme sledovať vývoj siete dostatočne dlhú dobu a potom ho zastavíme, dostaneme sieť s mnohými väzbami a aktérmi. Jej distribučnú funkciu *P(k)* môžeme zmerať. Analýzou jej tvaru, môžeme dospieť k záverom o histórii jej vývoja, teda, či sa uzly do siete pridávali preferenčným, alebo náhodným spôsobom [Markošová-Náther, 2010].

4.3 Analýza dynamiky diskusného kanála

Kľúčom k štúdiu dynamiky nejakej sociálnej siete je vhodná vizualizačná technika. Analýza dynamiky je v tomto dokumente zameraná na vizualizáciu sociálnej siete tvorenej účastníkmi privátneho diskusného kanála *www.kyberia.sk*. Boli vykonané niektoré prípravné úkony nad touto sociálnou sieťou a prvotná vizualizácia dynamiky tejto siete. Bežný export dát zo živej databázy často obsahuje mnohé šumy a nezaujímavé informácie, ktoré majú negatívny charakter na možnosti vizualizácie sociálnej siete. Často sa exportujú aj informácie obsiahnuté v uzloch siete, ktoré majú príliš málo hrán a zbytočne zaberajú miesto vo vizualizovanom priestore, čím prispievajú k nepriehľadnosti prostredia. K šumom je možné započítať aj uzly, ktoré v časovom intervale, v rámci ktorého prebieha vizualizácia, vykazujú nekonzistentné chovanie alebo sú charakteristické malým celkovým počtom hrán v čase. Ak chceme dospieť k prehľadnejšej vizualizácii sociálnej siete, je potrebné data predspracovať pomocou zhlukovania, resp. clusterizácie, v ktorom je vhodným výberom eliminovaná istá skupina uzlov, ktoré nehrajú dôležitú úlohu v sieti.

Zapojenie dynamiky do vizualizačného procesu sociálnych sietí poskytuje nový pohľad na túto problematiku. Sociálna sieť spravidla reprezentuje spojenia medzi rôznymi entitami v dvoch alebo troch dimenziách. Pridaním ďalšej dimenzie, a to časovej osi, môžeme sledovať ako sa daná sieť vyvíjala, a akými procesmi došlo k sformovaniu finálneho stavu siete. Je potrebné podotknúť, že sociálna sieť je dynamická štruktúra a v čase vizualizácie jej reálna podoba už môže byť odlišná.

Máme k dispozícii databázu priateľstiev, ktorá je reprezentovaná troma údajmi a to iniciátorom priateľstva, pridaným priateľom a časovým trvaním tejto akcie. Dáta sú uložené v databáze „mysql“ v nasledovnom formáte:

```
mysql> select ID_A as Kto, id_B as Skym, kedy as Cas from friends
limit 3;

   +------+------+------------+
   | Kto  | Skym | Cas        |
   +------+------+------------+
   |  406 |  369 | 2004-02-25 |
   | 2434 | 2884 | 2004-02-25 |
   | 2434 | 2884 | 2004-02-25 |
   +------+------+------------+
   3 rows in set (0.00 sec)
```

4.3.1 Štatistické údaje o kompletnej sociálnej sieti

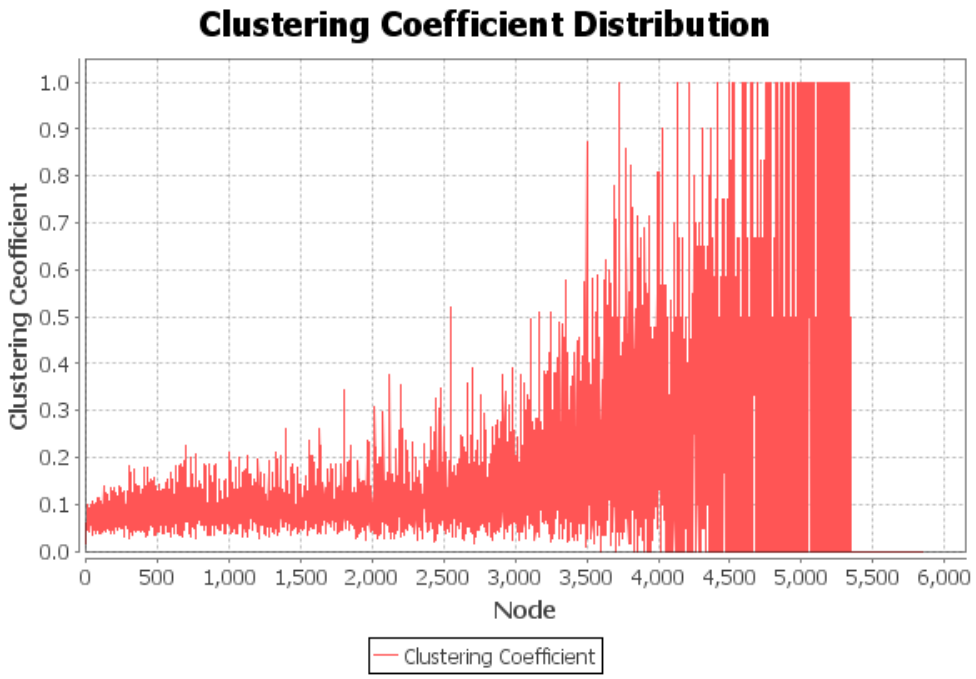
Úplná báza dát týkajúcich sa sociálnej siete – diskusného kanála „kyberia“, pred akoukoľvek modifikáciou nižšie spomínanými metódami, má nasledovné vlastnosti:

Počet uzlov:	5852
Počet hrán:	150392
Časové rozmedzie:	25.2.2004 – 16.3.2009
Priemerný stupeň siete:	51,3984

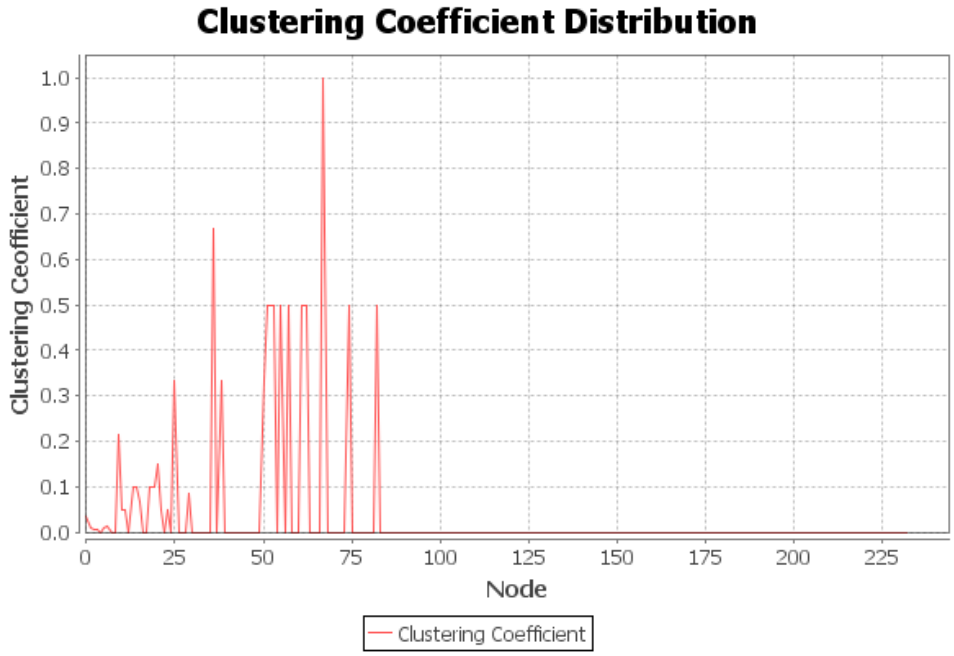
Je zrejmé, že vizualizovať sociálnu sieť s 5852 uzlami je relatívne zložité a výpovedná hodnota vizualizácie takto preplnenej siete neposlúži skúmaným účelom. Preto bol vyselektovaný časový interval tridsať dní. Báza dát o tejto sociálnej sieti v danom intervale obsahuje nasledovné vlastnosti:

Počet uzlov:	233
Počet hrán:	291
Časové rozhranie:	25.2.2004 – 27.3.2004
Priemerný stupeň siete:	5,78623

Pre porovnanie môže čitateľ sledovať vývoj koeficientu zhlukovania na celej sociálnej sieti (viď Obr.4.1) a na jej jednomesačnej selekcii (viď Obr.4.2).

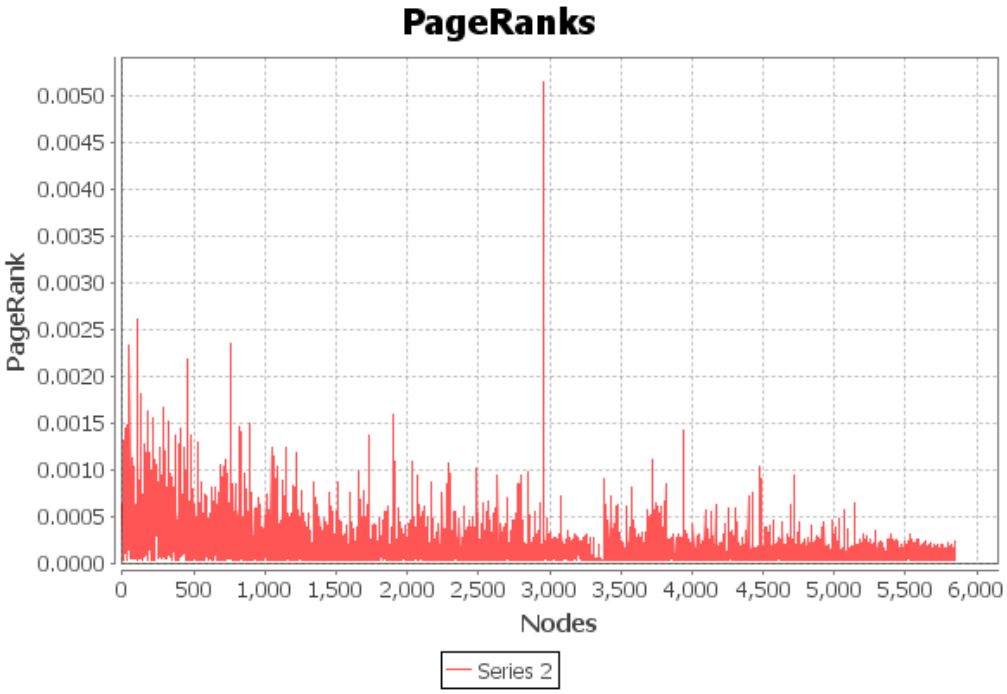


Obr. 4.1. Koeficient zhlukovania pre celú sociálnu sieť „kyberia“.

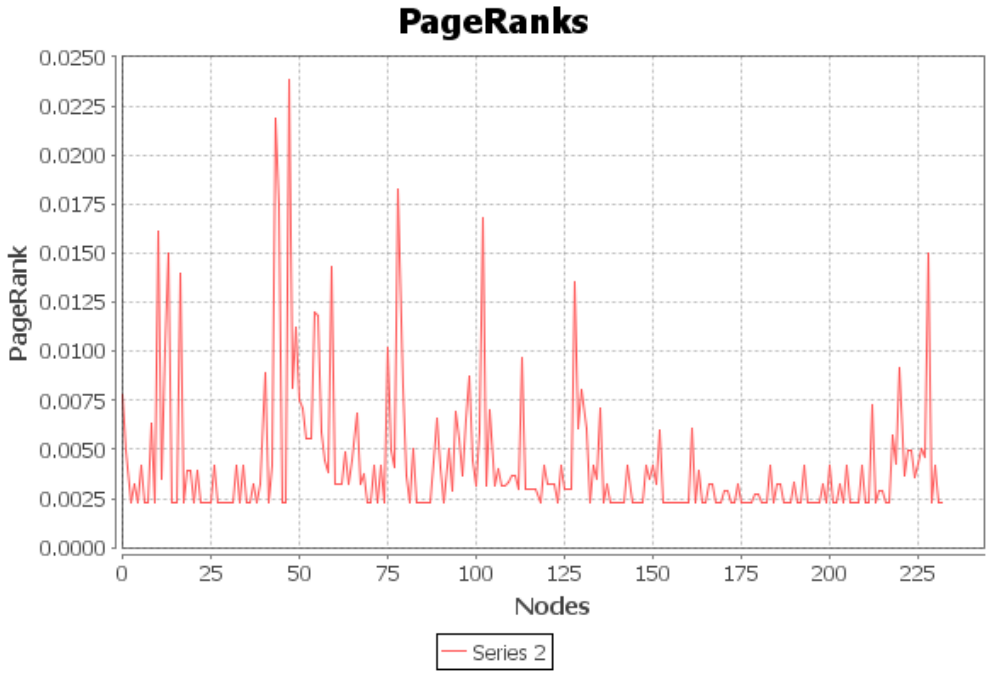


Obr. 4.2. Koeficient zhlukovania pre jednomesačnú selekciu sociálnej siete „kyberia“.

Pre zaujímavosť môžeme porovnať dynamiku Koeficientu zhlukovania s dynamikou vývoja koeficientu PageRank taktiež v dvoch vizualizáciách a to na celej sociálnej sieti (viď Obr.4.3) a na jej jednomesačnej selekcii (viď Obr.4.4).



Obr. 4.3. Dynamika vývoja PageRanku v rámci celej sociálnej siete „kyberia“.



Obr. 4.4. PageRank pre jednomesačnú selekciu sociálnej siete „kyberia“.

Orezanie sociálnej siete na menšie časové úseky nám teda dáva štatisticky prehľadnejšie dáta. Tieto selekcie sa môžu ďalej spracovávať do normovaných zhlukov dát a postupne v kratších časových úsekoch. Takto je teda možné dosiahnuť vizualizáciu rozsiahlej sociálnej siete aj v relatívne veľkom časovom rozmedzí (rádovo roky) vo forme série vizualizácií z kratších úsekov.

4.3.2 Jednomesačná selekcia sociálnej siete

Pre potreby vizualizácie budú skúmané vlastnosti siete, ktorá bola vytvorená z celkovej bázy dát selektovaním jednomesačného intervalu. Prvotná sociálna sieť vytvára relatívne nejednotný obraz, keďže je tvorená z veľkej časti uzlami s nízkym počtom hrán. Prvotnú vizualizáciu tejto siete ilustruje Obr.4.5. Aplikovaním Fruchterman Reingoldovho algoritmu [Fruchterman-Reingold, 1991]. Použitím tohto algoritmu bola sledovaná sieť preorganizovaná do podoby, akú znázorňuje Obr.4.6. N tomto obrázku sú už viditeľné niektoré centrálne uzly so silnými hranami a slabé uzly s nízkou centralitou a malým počtom hrán sú umiestnené na okraji sociálnej siete. Ich početnosť je relatívne vysoká ale nenesú skoro žiadnu informáciu o sociálnej sieti okrem tej týkajúcej sa svojej početnosti. V tejto fáze bol aplikovaný zhlukovací mechanizmus NCM [Markov, 2014], pomocou ktorého boli generované nasledovné zhluky:

Zhluk 1 – 320 uzlov

Zhluk 2 – 12 uzlov

Zhluk 3 – 123 uzlov

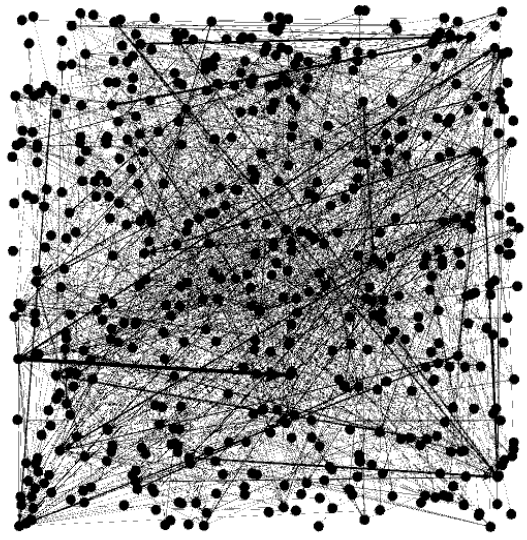
Zhluk 4 – 8 uzlov

Zhluk 5 – 8 uzlov

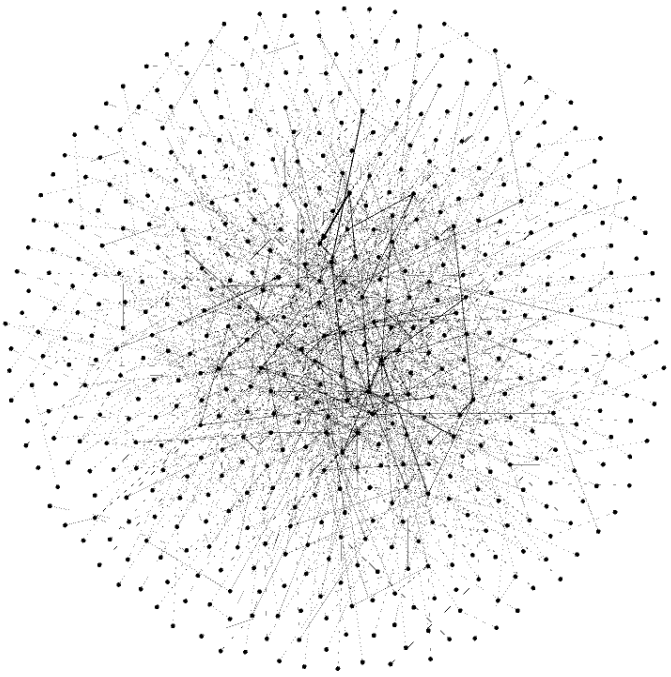
Zhluk 6 – 4 uzly

Zhluk 7 – 49 uzlov

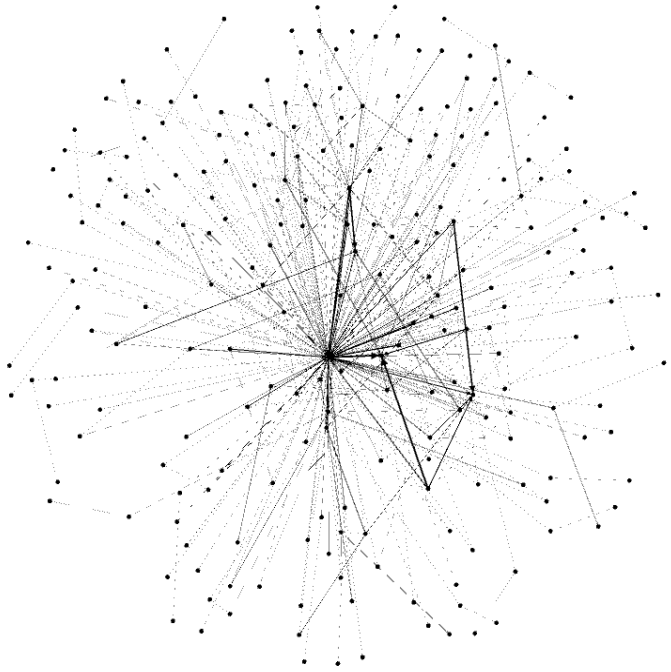
Testovaním selekcie bolo zistené, že najväčší počet uzlov sa nachádza v zhluku č.1 a tvoria ho práve vyššie spomínané uzly, ktoré sú veľmi početné ale málo informatívne. Zhlukovaním boli eliminované a sociálna sieť nadobudla charakter, ktorý ilustruje Obr.4.7.



Obr. 4.5 Prvotná vizualizácia jednomesačnej selekcie



Obr. 4.6 Vizualizácia jednomesačnej selekcie po predspracovaní Fruchterman-Reingoldovým algoritmom.



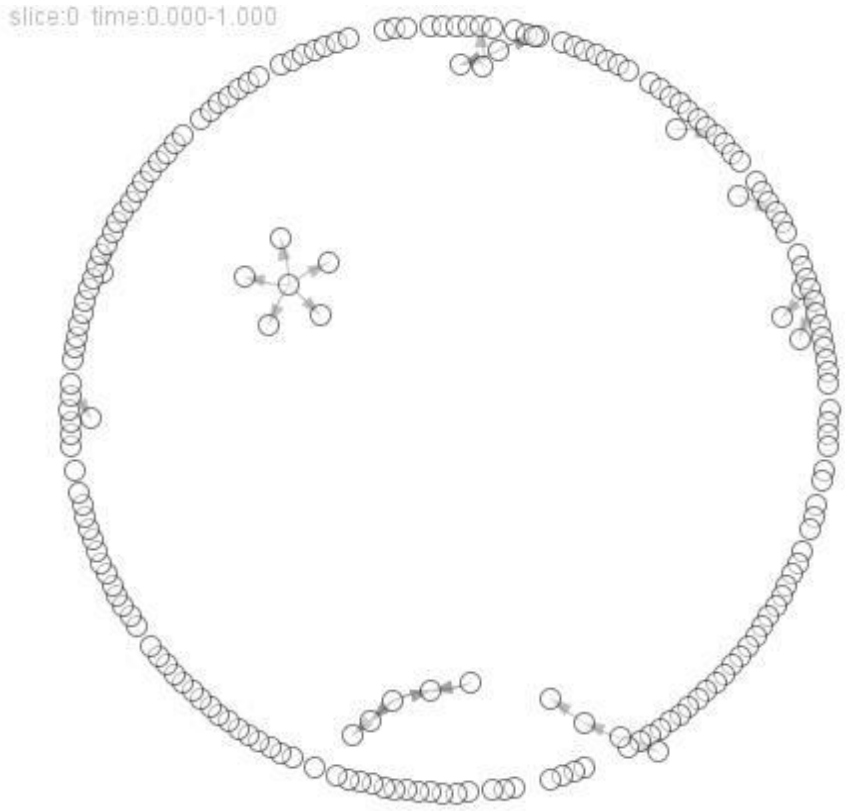
Obr. 4.7 Vizualizácia jednomesačnej selekcie po eliminácii nevýznamových uzlov.

4.4 Vizualizácia dynamiky sociálnej siete

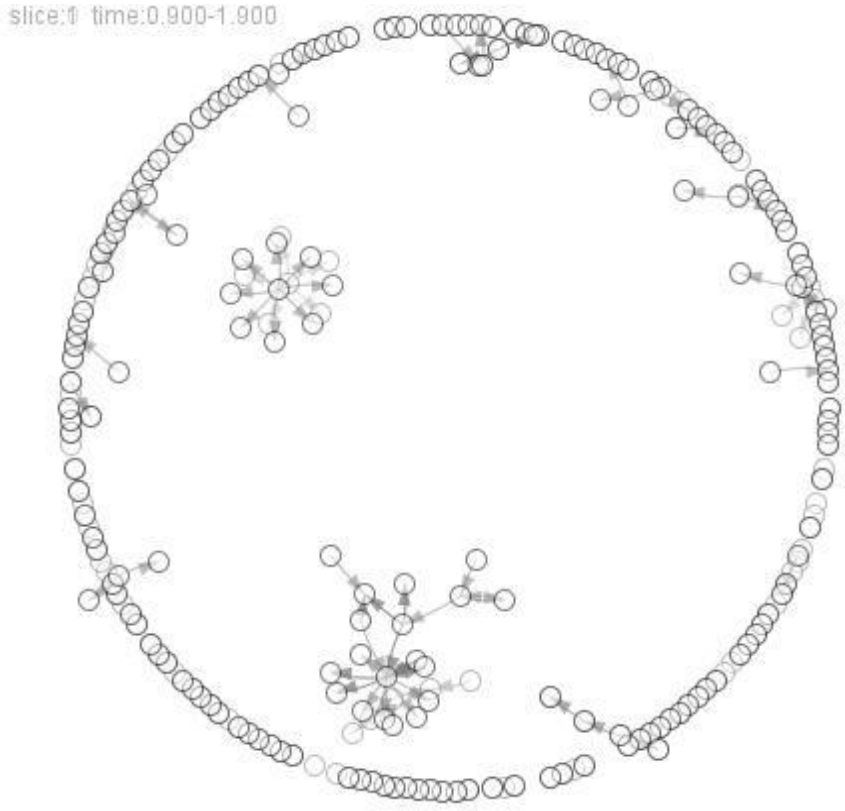
Spracovaním selektovanej jednomesačnej sociálnej siete a jej následným normovaním je možné vizualizovať dynamiku sociálnej siete. Pre tieto účely bola použitá aplikácia SoNIA (Social Network Image Animator) [SoNIA, 2014]. Vhodnou selekciou z databázy a normovaním pomocou parametra gama, ktorý definuje posun v dňoch od štartovacieho (najmenšieho) dátumu v selekcii, bola vytvorená báza interakcií medzi jednotlivými uzlami v čase t_0, t_1, až t_{31}.

```
mysql> select ID_A, ID_B, gama from friends where gama <=31;
+------+------+------+
| ID_A | ID_B | gama |
+------+------+------+
|  406 |  369 |    0 |
| ...  | .... | .... |
| 2434 | 2884 |    7 |
| ...  | .... | .... |
|624286|  349 |   31 |
+------+------+------+
```

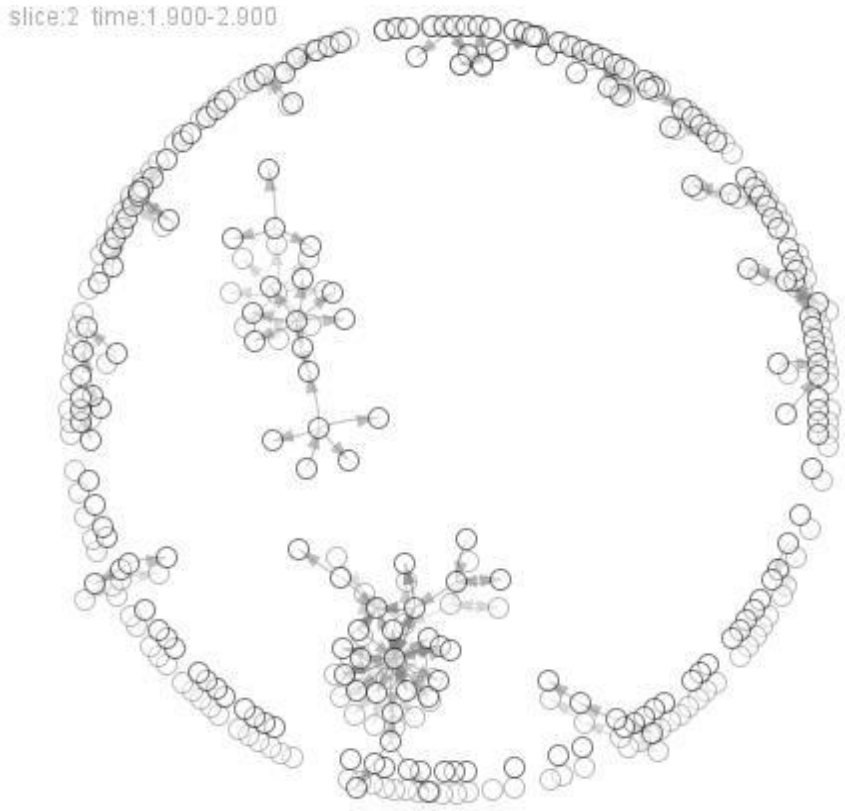
Takto selektované údaje boli konvertované do formátu „.son“, s ktorým pracuje vizualizačný program SoNIA. Výsledné vizualizácie ilustrujú Obr.4.8 až Obr.4.13. Tieto vybraté obrázky ilustrujú budovanie sociálnej siete v čase. Ako je možné vidieť, dochádza k zhlukovaniu sa jednotlivých uzlov, ktoré sú tvorené výberom priateľstiev v danom čase. Ako je možné vidieť z druhej animácie aj v tomto prípade dochádza k zvyšovaniu neprehľadnosti siete a je preto nutné mať na zreteli vhodné eliminovanie informačne slabých uzlov a sústavne sieť vhodne čistiť a zachovať tak maximálnu obrazovú výpovednú hodnotu danej vizualizácie.



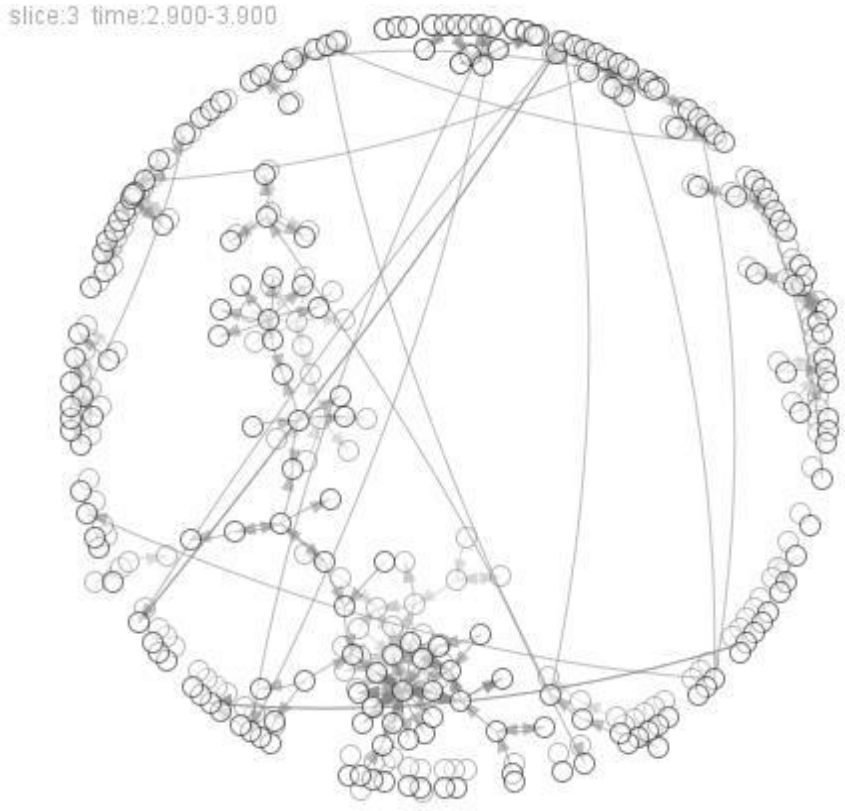
Obr. 4.8 Vizualizácia selekcie v danom čase - 1. Iterácia.



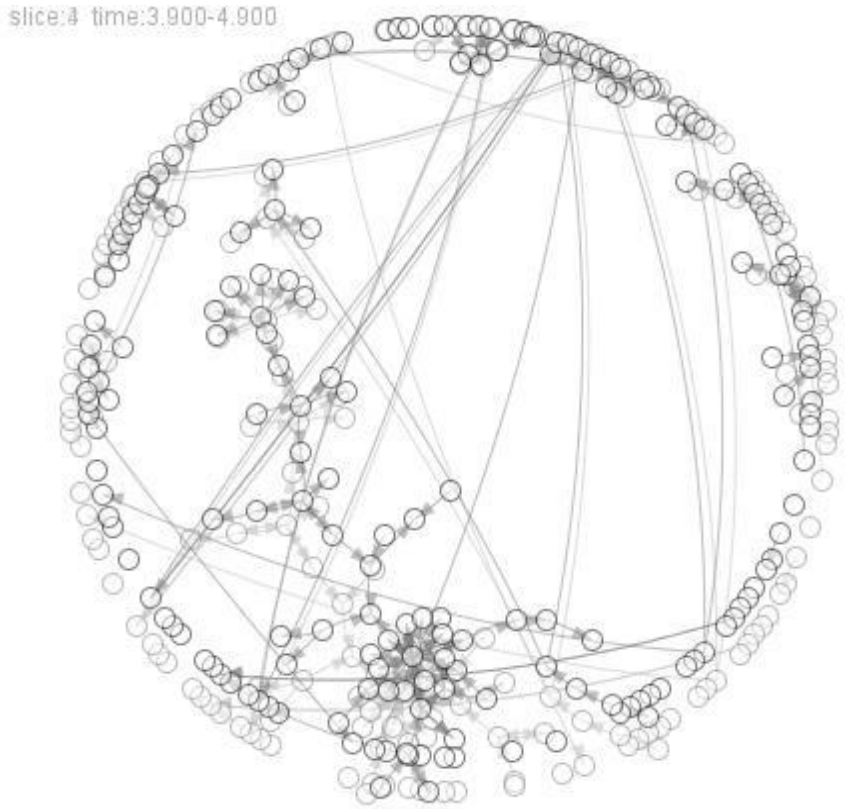
Obr. 4.9 Vizualizácia selekcie v danom čase - 10. Iterácia.



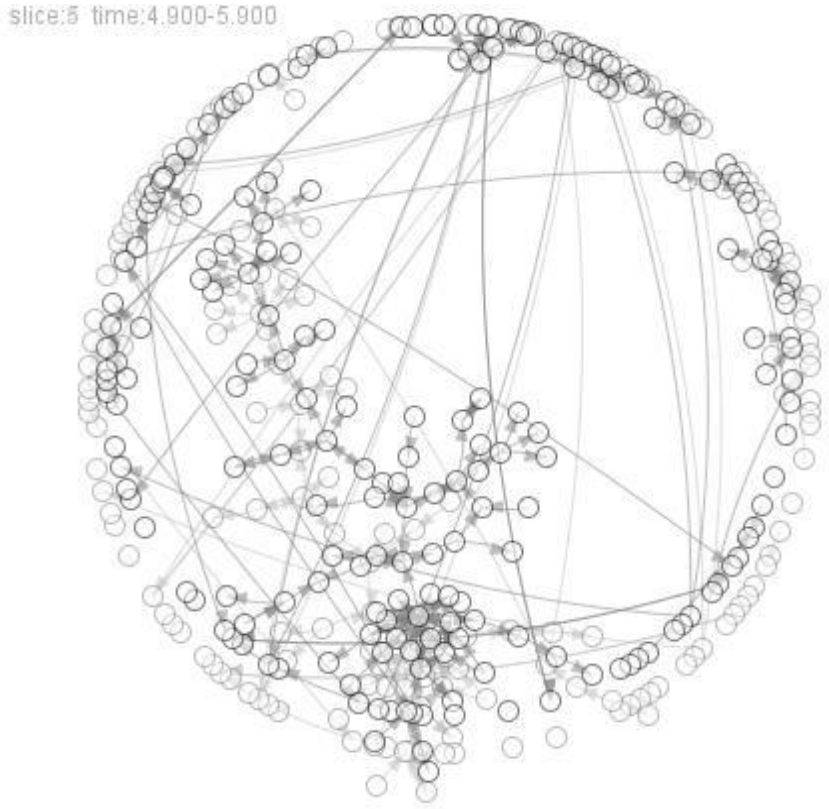
Obr. 4.10 Vizualizácia selekcie v danom čase - 20. Iterácia.



Obr. 4.11 Vizualizácia selekcie v danom čase - 30. Iterácia.



Obr. 4.12 Vizualizácia selekcie v danom čase - 40. Iterácia.



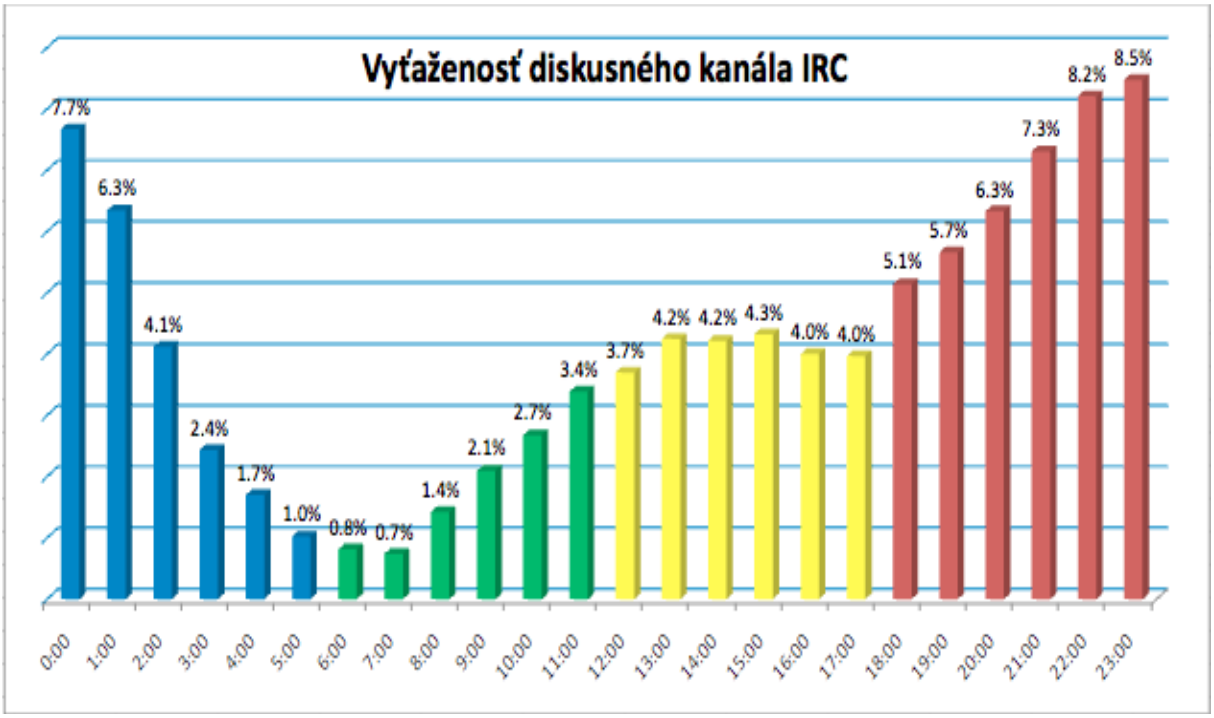
Obr. 4.13 Vizualizácia selekcie v danom čase - 50. Iterácia.

Dynamika v sociálnych sieťach verne ilustruje procesy, vytvárané ľuďmi. Dochádza k uzatváraniu priateľstiev, k ich rušeniu a s rastúcim počtom používateľov aj k rastu siete a tým pádom aj k zrýchleniu samotnej dynamiky siete. Čím je sieť robustnejšia, tým viac zmien je možné pozorovať v stále kratšom a kratšom časovom intervale. Sledovaním a štúdiom dynamiky sociálnej siete je možné zistiť vývoj lokálnych trendov, vývoj popularity a reakcií okolia na mnohé podnety. Zaujímavé je aj porovnanie, akou rýchlosťou vznikajú jednotlivé zhluky a ako rýchlo rastie okolie centrálnych uzlov. Pozorovaním a porovnávaním týchto zhlukov je možné získať užitočné informácie. Takto je možné dospieť k znalostiam o tom, ktorá lokálna sociálna sieť je ako robustná a takisto akou rýchlosťou vznikla. Rýchlejší vznik sociálnej siete predznamenáva jej väčšiu popularitu. Mapovanie dynamiky sociálnej siete má preto isto mnohoraké využitie a dá sa predpokladať, že v budúcnosti sa nájde mnoho foriem uplatnenia dolovania v dátach tohto charakteru.

4.5 Mapovanie dynamiky diskusného kanála IRC

IRC (Internet Relay Chat) je komunikačný kanál, ktorý umožňuje svojím používateľom komunikovať navzájom, či už pri riešení technických problémov, diskutovaní noviniek alebo udržiavať bežnú sociálnu komunikáciu. Výhodou tohto komunikačného kanálu IRC je nepretržitá a dlhotrvajúca komunikácia jeho používateľov. Preto je veľmi vhodný ako zdroj dát pre mapovanie dynamiky sociálnych sietí. Samotné vytváranie sociálnej siete je totiž nepretržitý proces a vždy reflektuje na aktuálnu komunikáciu so zreteľom na jej intenzitu aj do istej historickej hĺbky. Dynamika sociálnej siete je tvorená emergenciou. Takáto sociálna sieť, keďže je tvorená komunikáciou v čase, nesie isté prvky dynamického systému. Grafická vizualizácia sociálnej siete predstavuje potom pohľad na aktuálny stav prebiehajúcich diskusií. Ak komunikácia medzi entitami – používateľmi po čase zanikne, väzby medzi takýmito entitami sú roztrhnuté a sú odstránené z grafu. Každá nová diskusia je okamžite reflektovaná do grafu pridaním nových uzlov a hrán. Kvantita komunikácie medzi jednotlivými entitami v čase je prenesená do váh väzieb. Nepretržitá komunikácia je reprezentovaná hrubnutím väzby – spojenia v grafe. Z výsledného grafu v nejakom konkrétnom čase je možné takto určiť charakter komunikácie a na základe toho aj dynamické parametre sociálnej siete.

V prvom rade bolo potrebné sústrediť sa na vyťaženosť komunikačného kanála IRC v priebehu jednotlivých fáz dňa, čo ilustruje Obr.4.14. Každá fáza reprezentuje konkrétnu hodinu dňa. Spomenutá vyťaženosť je určená počtom interakcií medzi používateľmi.



Obr. 4.14 Vyťaženosť kanála IRC v jednotlivých hodinových fázach dňa.

Z Obr.4.14 vyplýva že najviac interakcií medzi používateľmi prebieha vo večerných a nočných hodinách. Dáta pre túto analýzu boli zbierané počas 211 dní. Tab.4.1 obsahuje absolútny presný počet interakcií medzi používateľmi v daných fázach dna. Počas doby sledovania, používatelia si vymenili celkovo 1 672 263 správ. Dáta boli zbierané pomocou IRC robota *Eggdrop* [Eggdrop, 2014] a následne spracované programom PISG (Perl IRC Statistics Generator) [Pisg, 2014].

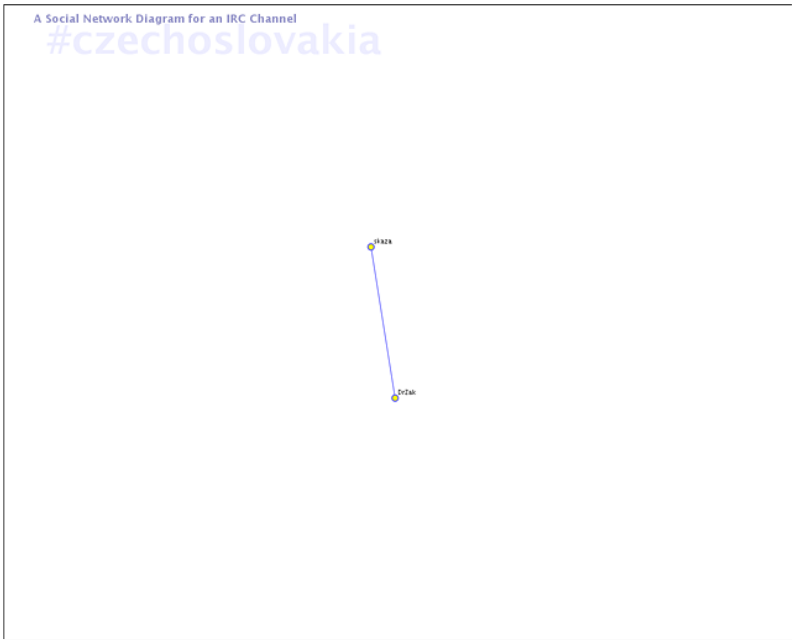
Tab. 4.1 Absolútne hodnoty počtu iterácií v rámci kanála IRC v jednotlivých fázach dňa.

Day hour	00:00-00:59	01:00-01:59	02:00-02:59	03:00-03:59	04:00-04:59	05:00-05:59
Number	128242	106138	68807	40686	28396	17086
Day hour	06:00-06:59	07:00-07:59	08:00-08:59	09:00-09:59	10:00-10:59	11:00-11:59
Number	13563	12391	23755	35083	44776	56598
Day hour	12:00-12:59	13:00-13:59	14:00-14:59	15:00-15:59	16:00-16:59	17:00-17:59
Number	61863	71047	70561	72301	66927	66373
Day hour	18:00-18:59	19:00-19:59	20:00-20:59	21:00-21:59	22:00-22:59	23:00-23:59
Number	85975	94583	105862	122248	137262	141740

4.5.1 Dynamika komunikačného portálu

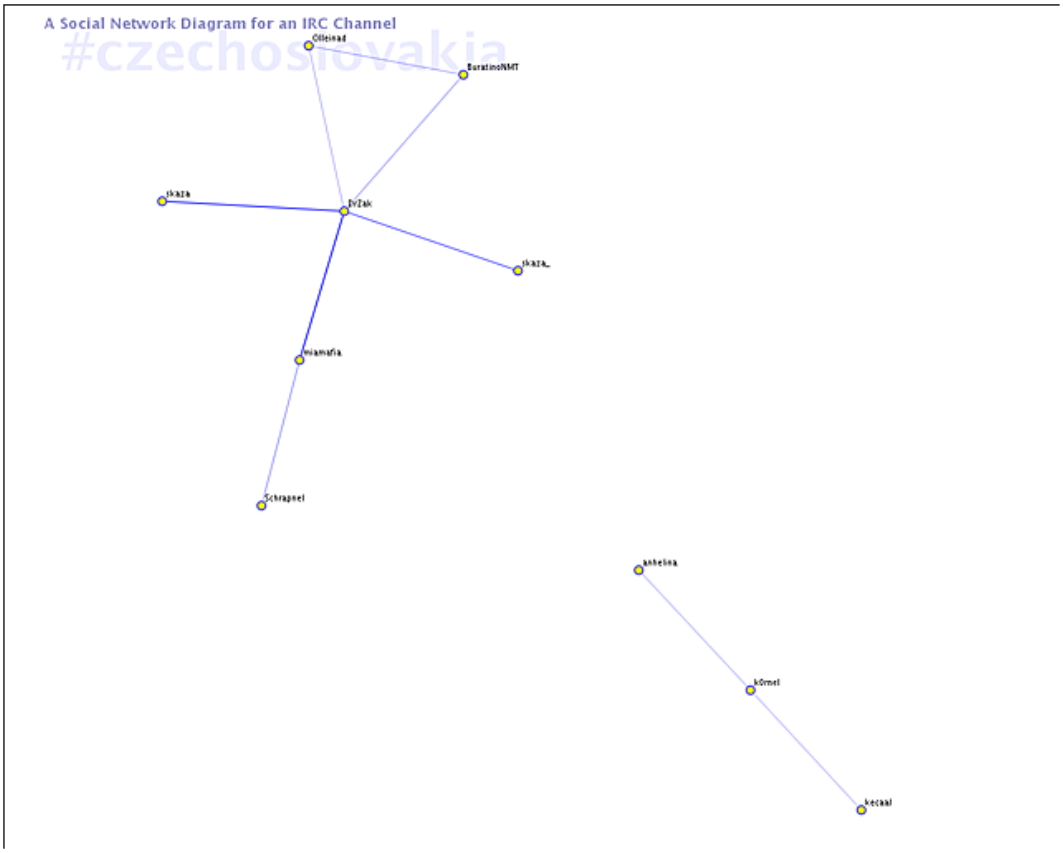
Prvotné pokusy o vizualizáciu dynamiky sociálnej siete boli založené na grafickom znázornení partikulárnej sociálnej siete v konkrétnom čase. Vizualizácia sa uskutočňovala kontinuálne, keďže diskusný kanál bol aj v tom čase v nepretržitej prevádzke. Bolo použité sledovanie formou detekcie zmeny. Každá nová komunikácia bola po zahájení hneď zavedená do grafu. Počas trvania komunikácie sa kontinuálne zvyšovali váhy existujúcich spojení, čo malo za následok vykreslenie hrubšou čiarou medzi entitami - používateľmi. Ak na určitý čas komunikácia ustala, boli spojenia medzi danými entitami pretrhnuté. Ak to bolo jediné spojenie týchto entít, boli z grafu odstránené aj tieto entity. Celková vizualizácia dynamiky sociálnej siete predstavuje sériu 8497 grafických vizualizácií. Na vizualizáciách v Obr.4.15, Obr.4.16 a Obr.4.17 je znázornená evolúciu sociálnej siete od prvotnej inicializácie, až po jej finálnu podobu v čase ukončenia mapovania.

Prvotná inicializácia sociálnej siete. Počas tejto fázy nie sú k dispozícii žiadne informácie o aktéroch sociálnej siete. Mapovací mechanizmus čaká, kedy dôjde k prvej komunikácii medzi aktérmi, Keďže ide o zber údajov v čase detekcie zmeny, nie klasické vzorkovanie v pravidelných časových intervaloch, mechanizmus zaznamenáva iba dichtonomické relácie dvoch aktérov. Po vzniknutí prvej komunikácie medzi aktérmi je táto zaznamenaná ako prvá iterácia (viď Obr.4.15).



Obr. 4.15 Prvotná iterácia vizualizačného procesu.

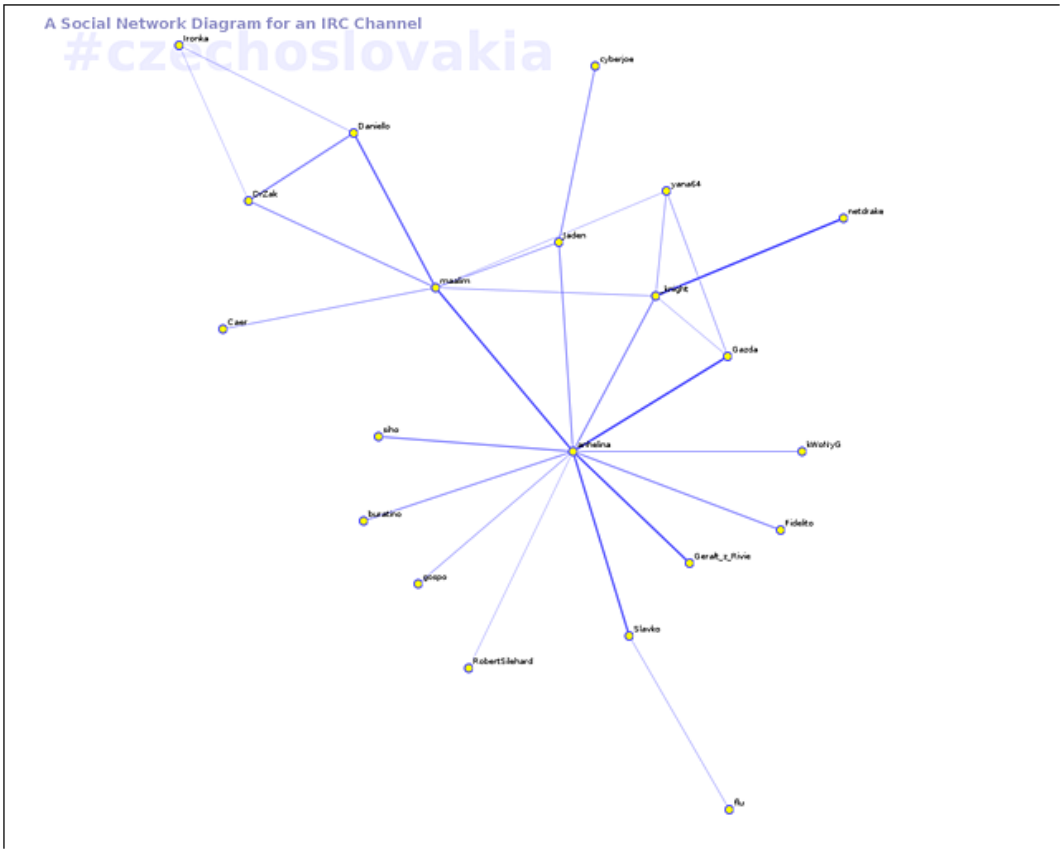
Popis 1778. Iterácie. Do sociálnej siete pristupujú ďalší diskutéri, ktorí ale diskutujú medzi sebou separátne od už vybudovanej sociálnej siete, čo spôsobuje rozbitie grafickej vizualizácie na dva nezávislé fragmenty. Separátni aktéri sa nezapájajú do diskusie iných aktérov a tak vzniká oddelená sociálna sieť, ktorá sa vôbec nemusí pripojiť k už existujúcim reláciám, ale môže (viď Obr.4.16).



Obr. 4.16 V poradí 1778 iterácia vizualizačného procesu.

Popis 8497. iterácie. Posledná vizualizácia je zobrazená na Obr.4.17. Je možné sledovať existenciu autoritatívneho aktéra v sociálnej sieti, ktorý má najviac väzieb.

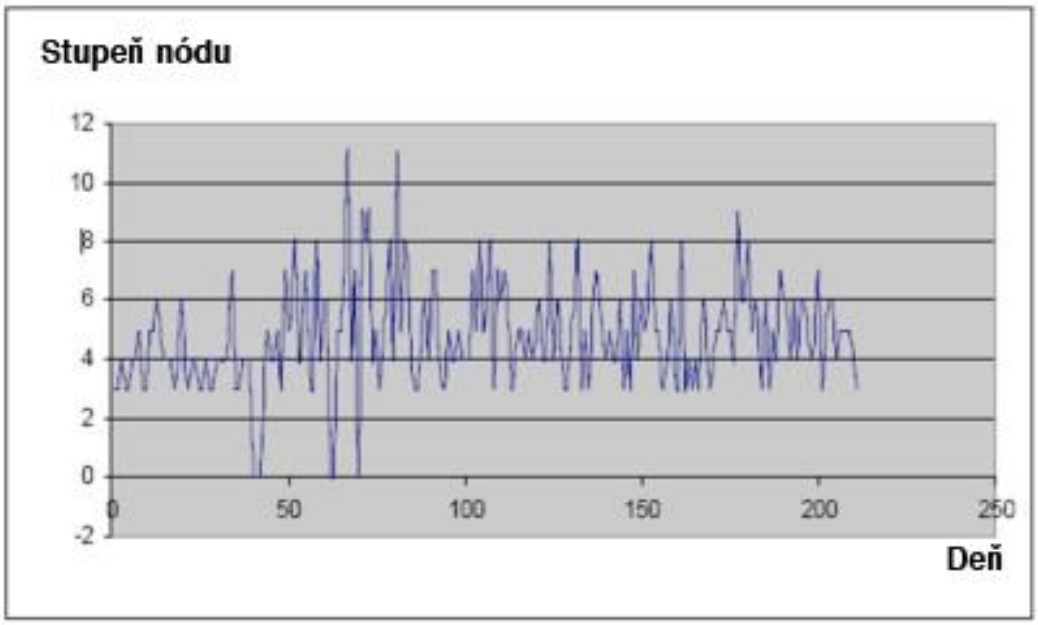
Podľa [Rakuščinec, 2009] je možné dynamiku sociálnej siete rozdeliť do troch hlavných fáz. Prvotná fáza je samotná inicializácia. Následne dochádza k prelínaniu dvoch fáz a to fázy *evolúcie* sociálnej siete a fázy *oscilácie* sociálnej siete. Vo fáze evolúcie sociálnej siete dochádza k zmene počtu aktérov a k zmene poctu väzieb medzi aktérmi. Táto fáza formuje sociálnu sieť do podoby, v akej ju máme možnosť sledovať v danom čase. Fáza oscilácie predstavuje upravovanie váh väzieb na základe detekcie komunikácií medzi aktérmi, ale samotná sociálna sieť vykazuje istú stabilitu.



Obr. 4.17 Posledná v poradí 8497. iterácia vizualizačného procesu.

4.5.2 Dynamické vlastnosti komunikačného portálu

Niektoré statické vlastnosti sociálnych sietí (boli uvedené v kapitole 3.) je možné a účelné sledovať v dynamike. Stupeň nódu je najvhodnejším parametrom pre skúmanie dynamiky, pretože najlepšie charakterizuje vývoj entity používateľa z hľadiska jeho aktívnosti na sociálnej sieti. Stupeň nódu *k* reprezentuje počet najbližších susedov diskutéra v danom čase. Diskusný kanál IRC je možné považovať za uzavretú sociálnu sieť, v ktorej sa uvažujú iba väzby medzi diskutérmi len v rámci tejto siete, čo nevylučuje možnosť diskutéra pôsobiť na viacerých diskusných kanáloch. Z predpokladu uzavretosti siete vyplýva, že priemerný vstupný a výstupný stupeň nódu všetkých diskutérov sa rovnajú. Ide totiž o taký typ uzavretej siete, ktorá komunikuje na základe scenára „otázka – odpoveď“ a predpokladá sa, že počet používateľov, ktorým sa nedostalo odpovede je minimálny a zanedbateľný. Obr.4.18 ilustruje vývoj stupňa nódu jedného diskutéra, ktorého je možné považovať za autoritu, pretože počas celého procesu mapovania vykazoval najvyššie priemerné hodnoty stupňa nódu.



Obr. 4.18 Distribúcia stupňa nódu najviac autoritatívneho aktéra.

V rámci mapovania dynamiky diskusného kanála IRC bola pozornosť sústredená na detekciu zmeny v čase. Každý graf - vizualizácia predstavuje jednu iteráciu v diskrétnom časovom období, kedy došlo k zmene, keďže nešlo o mapovanie vzorkovaním. Toto mapovanie umožňuje sledovať nielen evolúciu sociálnej siete ale aj jej vlastností.

Prvotnou inicializáciou mapovacieho procesu vzniklo jediné spojenie, pretože detekcia komunikácie je najmenšia jednotka, ktorú je možné merať. Následne sa sociálna sieť začala formovať a vznikali známe usporiadania entít ako sú popísané v kapitole 3. a to h*viezda* (Obr.4.16 a Obr.4.17) a č*iara* (Obr.4.16). Taktiež bol zaznamenaný autoritatívny aktér, ktorý svojim pôsobením na sociálnej sieti "priťahoval pozornosť" a mnohí iní používatelia sa s ním chceli radiť.

Uskutočnené mapovanie sociálnej siete reprezentuje aj dve hlavné fázy evolúcie sociálnej siete a to s fázu evolúcie sociálnej siete (kedy je možnosť sledovať vývoj sociálnej siete od inicializačnej fázy až po ukončenie mapovania) a fázou oscilácie sociálnej siete (kedy nedochádza k zmene usporiadania sociálnej siete pribúdaním alebo miznutím entít ale úpravou váh reprezentujúcich početnosť komunikácie. Celkový objem nazbieraných dát predstavoval 181310 grafov a presahoval 5GB diskového priestoru. Táto sociálna sieť bola mapovaná počas relatívne dlhej doby a to od 9. Októbra 2009 po 5. December 2010.

Boli objavené existencie skupín diskutérov, centrálnych uzlov a izolovaných entít. Dynamický proces vývoja siete dokumentuje jej vznik a rast v dôsledku javu emergencie. Znamená to, že používatelia nekomunikujú navzájom preto aby budovali sociálnu sieť. Tá teda nevzniká ako produkt cielenej činnosti žiadneho z diskutérov.

4.6 Analýza komunitného portálu

Pre účely analýzy komunitného portálu bol zvolený portál www.kyberia.sk, ktorý obsahoval dostatočné množstvo dát pre analýzu (6000 používateľov). Tento portál slúži na výmenu informácií a umožňuje diskusie v diskusných fórach. Je orientovaný hlavne na slovenských a českých používateľov, ktorí sa stretávajú aj v reálnom živote, nielen v kyber-priestore. Tento portál umožňuje komunikáciu s mnohými osobnosťami zo sveta umenia, vedy, techniky a pod. Taktiež organizuje podujatia (napr. hudobné), ktoré sa následne prenášajú z kyber-priestoru do reálneho života. Boli zaznamenané aj manželstvá používateľov, ktorí sa zoznámili práve pomocou Kyberie. Portál Kyberia nie je verejne prístupný registráciám a tak každý nový "občan" portálu je prijatý po podaní žiadosti až na základe hlasovania ostatných používateľov „občanov" Kyberie.

Požívatelia majú samozrejme možnosť vytvárať priateľstvá, čím vzniká silno štruktúrovaná sociálna sieť. Vznik priateľstva sa na tomto portáli inicializuje zverejnením krátkej správy na nástenke používateľa, ktorá charakterizuje okolnosti, za akých sa noví priatelia spoznali. Vytvorenie priateľstva je taktiež spätnou väzbou od iného používateľa.

4.6.1 Údaje obsiahnuté v databáze

Zdroj označuje iniciátora priateľstva. Každé zaznamenané priateľstvo má orientáciu od zdroja k používateľovi, ktorý ho prijal. Zdroj predstavuje „donora" priateľstva. *Cieľ* označuje prijímateľa priateľstva, teda jeho „akceptora". *Priateľstvo* je jedinečná identifikácia vzniknutého vzťahu, ktorý sa zakladá uverejnením krátkej správy na nástenku prijímateľa. *Dátum* je reprezentovaný časovým „timestamp" priateľstva.

Na účely analýzy bola Kyberia, ktorá predstavovala bohato štruktúrovanú sociálnu sieť za relatívne dlhé časové obdobie, veľmi vhodná. Na druhej strane bola tak rozsiahla, že pre účely vizualizácie nebolo účelné pracovať s celou štruktúrou. Vytváranie priateľstiev na portáli Kyberia má skôr charakter začleňovania nových jednotlivcov do komunity. V tom sa líši od podstaty priateľstva na Facebook-u. Portál Kyberia umožňuje používateľovi plnú kontrolu prístupových práv k vlastným fóram.

Úplná databáza popisujúca sociálnu sieť komunitného portálu, pred akoukoľvek modifikáciou má nasledovné parametre:

Počet uzlov: 5852

Počet hrán: 150392

Časové rozpätie: 25. 2. 2004 - 16. 3. 2009

Priemerný stupeň siete: 51,39.

Táto databáza obsahuje enormne vysoké množstvo hrán, aby sa dala s úspechom a prehľadne vizualizovať. Preto bolo potrebné zúžiť jej rozsah na jeden rok budovania tejto siete. Bol zvolený práve prvý rok, ktorý obsahoval údaje o vzniku a evolúcii sociálnej siete, teda informačne najvýznamnejšie. Štatistické údaje tejto selekcie sú nasledovné:

Počet uzlov: 233

Počet hrán: 291

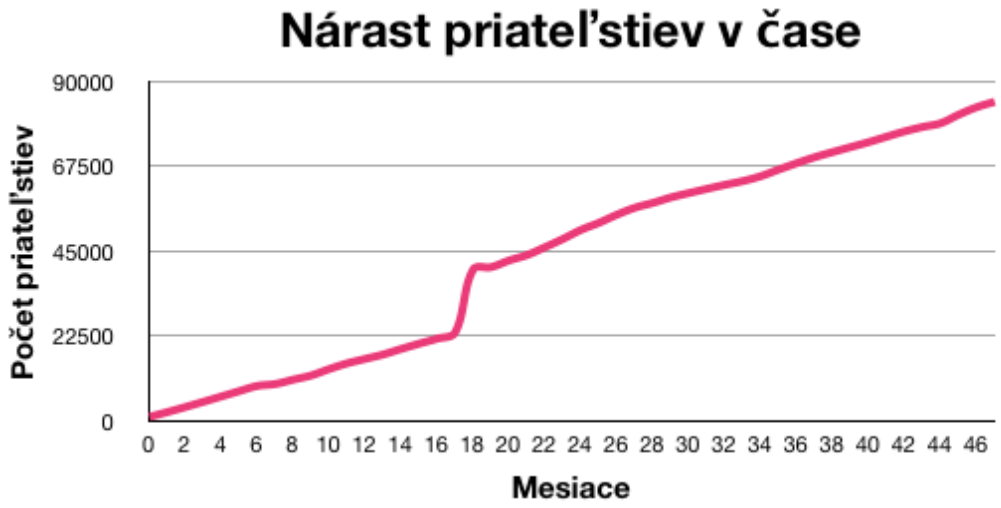
Časové rozpätie: 25. 2. 2004 - 25. 3. 2004

Priemerný stupeň siete: 15,62.

4.6.2 Dynamika komunitného portálu – nárast priateľstiev

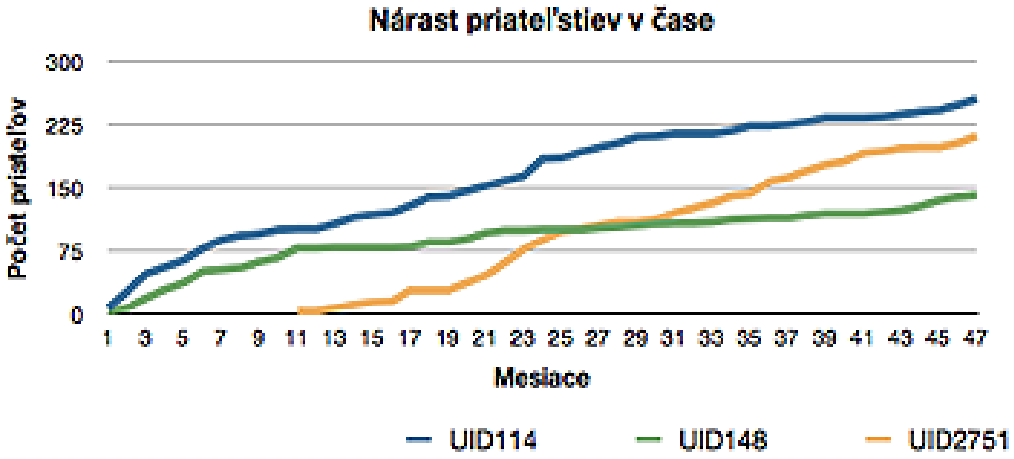
Pre potreby vizualizácie komunitného portálu boli upravené dáta z databázy mysql do tabuľkovej podoby, ktorá bola lepšie prispôsobená vizualizačnému programu Gephi [Bastian, 2014]. Jednotlivé priateľstvá boli reprezentované tabuľkou hrán s údajmi o odosielateľovi, žiadosti o priateľstvo, o prijímateľovi priateľstva a o časových intervaloch, v ktorých bolo skúmané priateľstvo platné. Nárast priateľstiev bol skúmaný opäť pomocou charakteristiky - parametra *stupeň nódu*, ktorý bol definovaný v kapitole 3. Tento parameter reprezentuje entity v sociálnej sieti pomocou počtu väzieb, ktoré vytvárajú s inými entitami.

Obr.4.19 vizualizuje celkový nárast priateľstiev v priebehu 47 mesiacov. Táto analýza ukázala, že približne od 18. po 20. mesiac fungovania tohto portálu prebehol najvýraznejší nárast počtu priateľstiev. Zaujímavosťou je, že celkový nárast nemá exponenciálny priebeh ale skôr lineárny, čo môže byť spôsobené relatívne nízkym počtom nových používateľov.



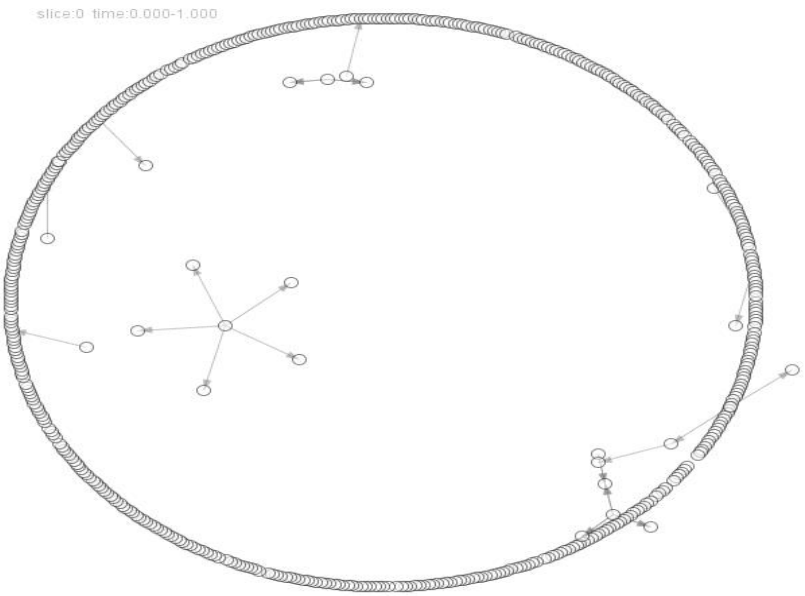
Obr. 4.19 Nárast celkového počtu priateľstiev na portáli Kyberia.

Súčasťou analýzy je sledovanie vývoja rastu (respektíve poklesu priateľstiev troch najobľúbenejších používateľov počas celej existencie sociálnej siete (viď Obr.4.20). Zaujímavý je tretí používateľ (oranžová farba, UID 2751), ktorý vstúpil do Kyberie až v jedenástom mesiaci existencie portálu a približne rok na to začal počet jeho priateľstiev narastať tak rýchlo, až prekonal druhého používateľa (zelená farba, UID 148) a priblížil sa k hodnotám počtu priateľov prvého používateľa (modrá farba, UID 114). Každý používateľ má svoj unikátny identifikátor, ktorý sa prideľuje pri registrácii takým spôsobom, že používateľ, ktorý vstúpil do sociálnej siete neskôr, má svoje unikátne ID vyššie ako ten, ktorý prišiel pred ním. Okrem ID používatelia pri registrácii zadávajú prezývku, ktorá má nižšiu informačnú hodnotu, preto sú v rámci tejto analýzy používatelia identifikovaní podľa UID X (Univerzálny Identifikátor X).



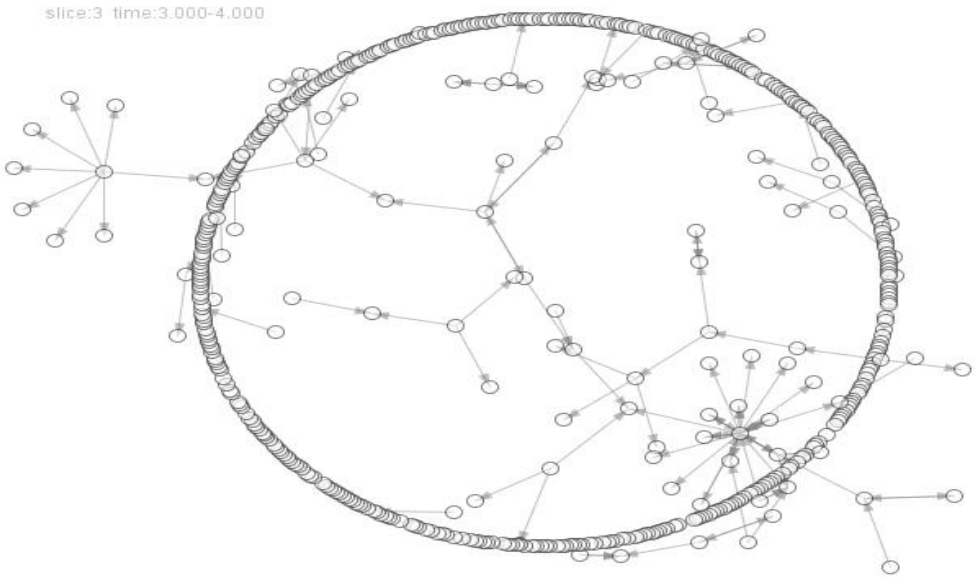
Obr. 4.20 Nárast celkového počtu priateľstiev troch najobľúbenejších používateľov portálu Kyberia.

Ako už bolo spomenuté, pre potreby analýzy dynamiky sociálnej siete bola použitá vzorka dát zozbieraná v prvom roku existencie komunitného portálu. Prvý rok totiž ponúkal možnosť zachytiť formovanie sociálnej siete od jej samotného začiatku. Samozrejme proces mapovania môže byt inicializovaný v ľubovoľnom čase. Pravdou je, že skrátením času sledovania môžeme prísť o možnosť sledovania zaujímavých sociálnych štruktúr. Taktiež spustenie sledovania až vtedy, keď je sieť už vybudovaná, môže viesť k strate niektorých informácií o štruktúrach, charakteristických pre proces vznikania siete. Séria obrázkov (Obr.4.21 až Obr.4.24) ilustruje evolúciu sociálnej siete na portáli Kyberia počas prvého roka jej existencie. Tieto obrázky dokumentujú formovanie siete cez rôzne iterácie. Obr.4.21 obsahuje oddelenú sociálnu štruktúru „hviezda“ už v prvý deň mapovania sociálnej siete.

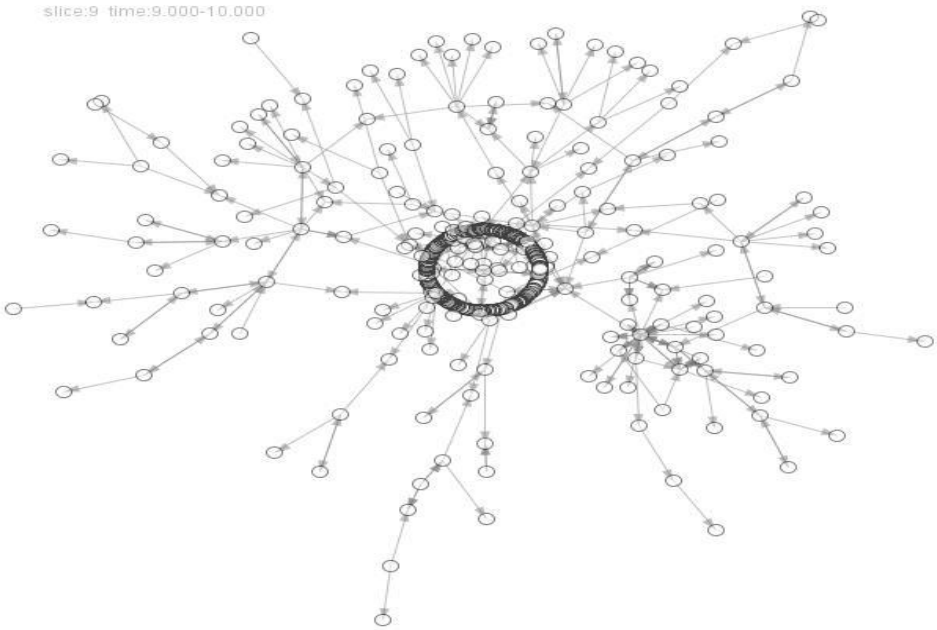


Obr. 4.21 Inicializácia sociálnej siete Kyberia – 1.iterácia.

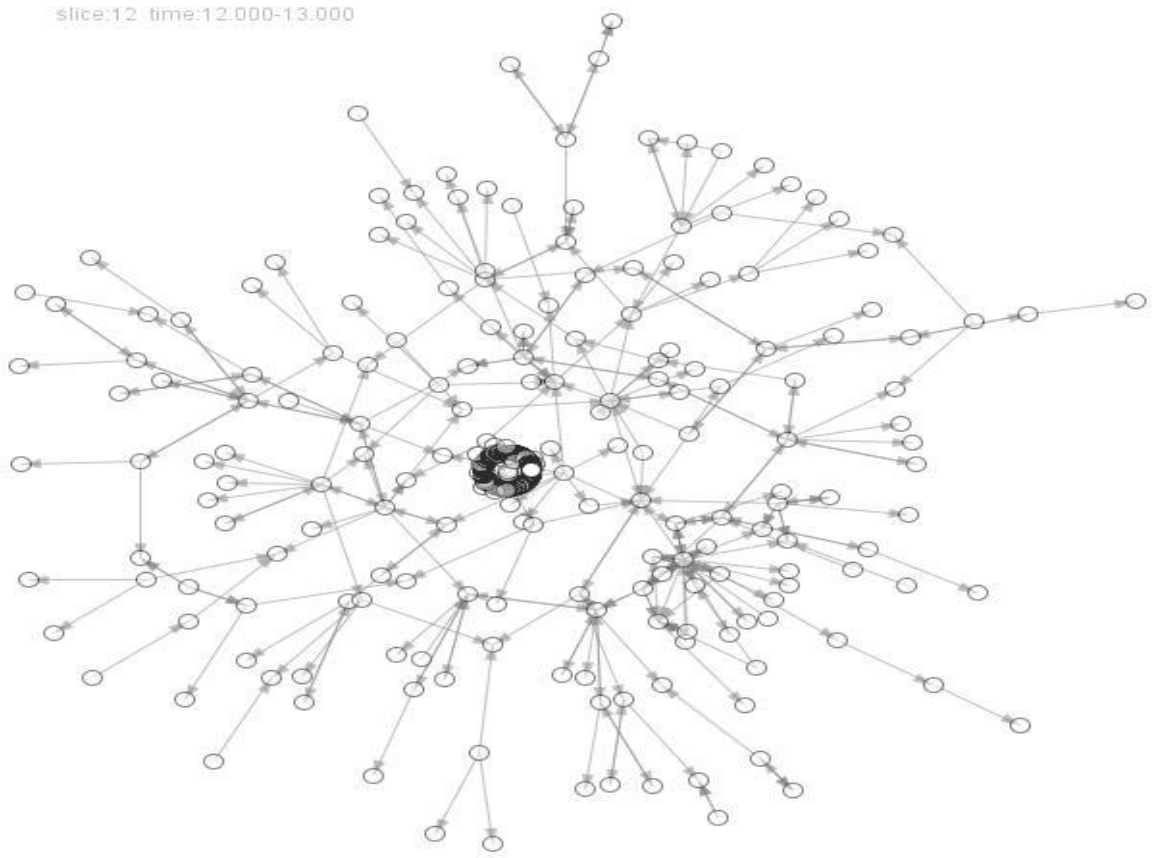
Počas ďalšej evolúcie tejto siete (Obr.4.22 až Obr.4.24) došlo k usporiadaniu entít do zložitých štruktúr.



Obr. 4.22 Inicializácia sociálnej siete Kyberia – 91.iterácia.



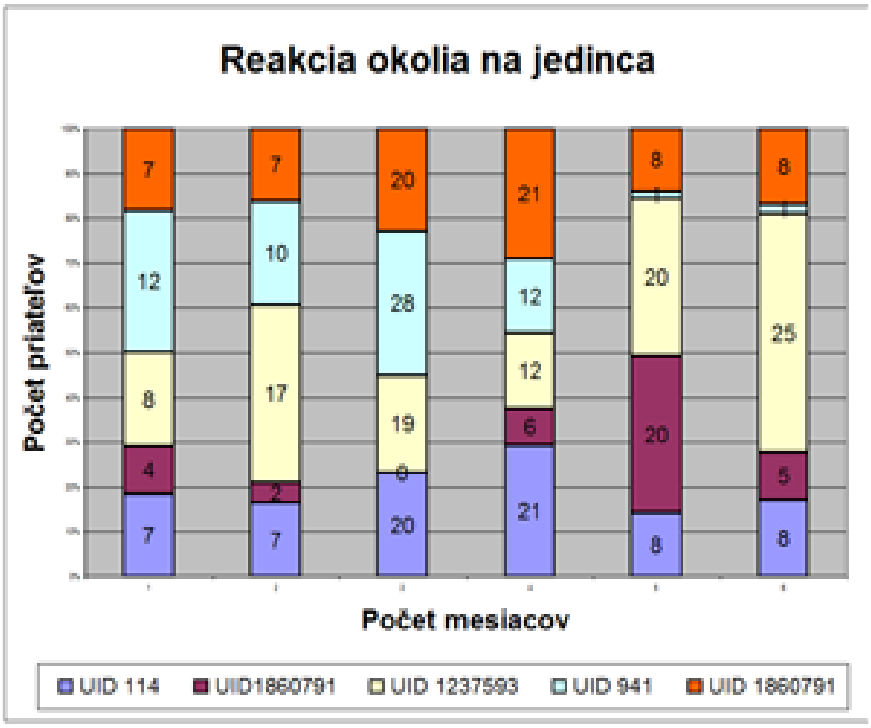
Obr. 4.23 Inicializácia sociálnej siete Kyberia – 271.iterácia.



Obr. 4.24 Inicializácia sociálnej siete Kyberia – 361.iterácia.

4.6.3 Reakcie okolia na entitu

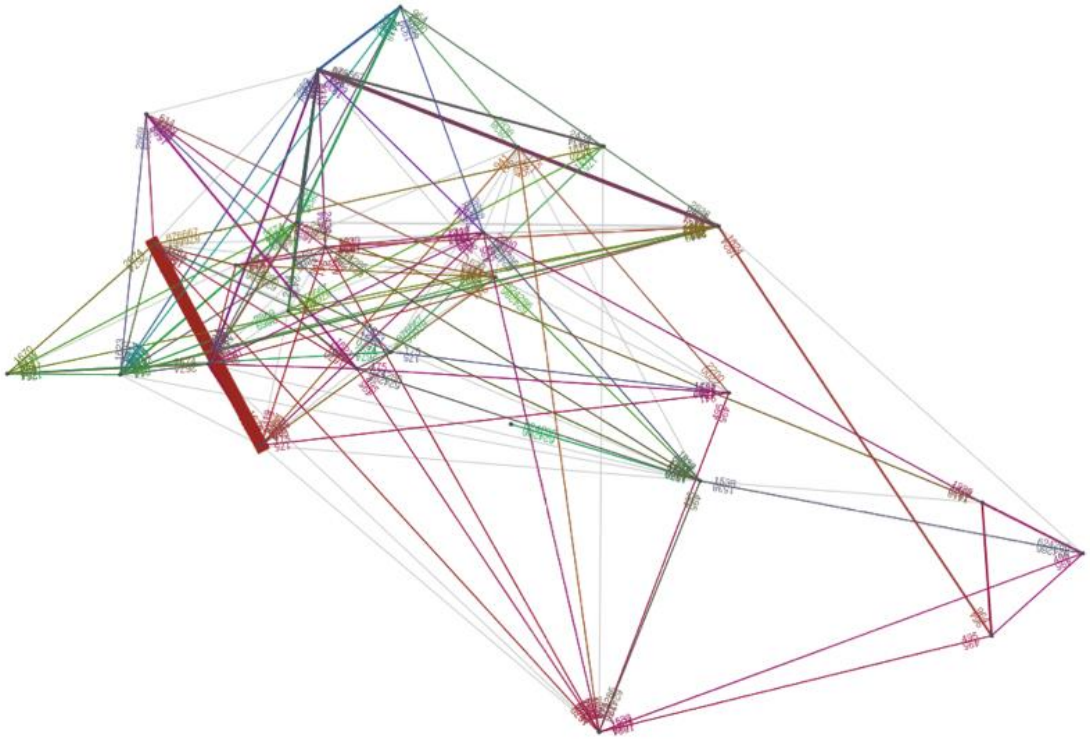
Analýza dynamiky sociálnej siete Kyberia bola uskutočnená aj z pohľadu reakcií okolia na entitu – na aktuálneho používateľa. Je to inverzný proces k sledovaniu prírastku počtu priateľov entity. V rámci tohto procesu bolo zaznamenávané, koľko používateľov si pridalo za priateľa novo vstupujúceho člena do komunity. Mapovanie prebiehalo počas prvých šiestich mesiacov po vstupe každého zo sledovaných používateľov do komunity. Bola skúmaná vzorka pätice používateľov, ktorí patria medzi entity s vysokým počtom priateľov. Obr.4.25 ilustruje vývoj počtu reakcií sociálnej siete na nových používateľov. Zaujímavý je používateľ UID 114, ktorý je má najväčší počet priateľov, ako ilustruje spomenutý Obr.4.25, hoci sám nepatrí medzi “najpopulárnejších” z používateľov. Z tohto obrázku je možné vyčítať, ako u niektorých používateľov došlo v čase k postupnému nárastu počtu priateľov počas celého mesiaca (ako napríklad u používateľa UID 1237593) Na druhej strane u niektorých iných používateľov počet priateľov stúpol a krátko na to začal klesať (ako napríklad u používateľa UID 1860791).



Obr. 4.25 Reakcia sociálnej siete na nových používateľov.

4.6.4 Vek priateľstva

Analýza sociálnej siete z hľadiska veku priateľstva nepredstavuje mapovanie dynamiky sociálnej siete v pravom slova zmysle, keďže nesleduje zmenu parametra v čase. V tomto prístupe bola zaznamenávaná dĺžka priateľstiev u entít s vysokým počtom väzieb. Následne boli jednotlivé väzby medzi priateľmi váhované dĺžkou trvania priateľstva. Zistili sme, že najstaršie spojenia sa nachádzajú u tých entít, ktoré vykazujú vysoký počet väzieb. Bola skúmaná selekcia používateľov, ktorí majú nad 35 väzieb. Vizualizáciu veku priateľstiev ilustruje Obr.4.26. V tomto obrázku hrúbka čiary reprezentuje vek priateľstva v zmysle dĺžky jeho trvania.



Obr. 4.26 Mapovanie veku priateľstiev – najstaršie priateľstvo je reprezentované najhrubšou čiarou..

POUŽITÁ LITERATÚRA

[Albert-Barabási, 2001] Albert, R., Barabási, A. L.: *Statistical mechanics of complex networks*. Cond-mat/0106096 (2001)1, [online]. [cit. 2014-06-28] Dostupné na internete: <http://www.barabasilab.com/pubs/CCNRALB_Publications/200201-30_RevModernPhys-StatisticalMech/200201-30_RevModernPhys-StatisticalMech.pdf>

[Barabási-Albert, 1999] Barabási, A. L., Albert, R.: *Emergence of scaling in random networks*. Science 286, 1999, 509-512.

[Bastian, 2014] Bastian M.: *Gephi, an open source graph visualization and manipulation software*. [online] [cit. 2014-07-20] Dostupné na internete: http://gephi.org/about/.

[Dorogovstev-Mendes, 2003] Dorogovstev S. N., Mendes J. F. F.: *Evolution of networks*. Oxford, 2003.

[Eggdrop, 2014] Eggdrop development and the Eggdev team. [online] [cit. 2014-06-28] Dostupné na internete: http://www.eggheads.org/about/.

[Fruchterman-Reingold, 1991] Fruchterman, T. M. J., & Reingold, E. M.: *Graph Drawing by Force-Directed Placement. Software*. Practice and Experience 21(11), 1991.

[Markošová-Náther, 2010] Markošová M., Náther, p.: *Networks Dynamika sietí a jazyk*. Katedra aplikovanej informatiky, Fakulta matematiky, fyziky a informatiky, Univerzita Komenského, Bratislava.

[Markov, 2014] Markov, A.: *Markov Cluster Algorithm (MCL) - a cluster algorithm for graphs*. [online] [cit. 2014-06-28] Dostupné na internete: http://www.micans.org/mcl/.

[Pisg, 2014] PISG: Perl IRC Statistics Generator. [online]. [cit. 2014-06-28] Dostupné na internete: http://pisg.sourceforge.net/index.php?page=about.

[Rakuščinec, 2009] Rakuščinec, T.: *Vizualizácia sociálnych sietí.* Košice, Technická univerzita v Košiciach, Fakulta elektrotechniky a Informatiky, 2009, 1-30.

[Repka, 2011] Repka, M.: *Analýza určitých typov sociálnych sietí.* Košice, Technická univerzita v Košiciach, Fakulta elektrotechniky a Informatiky, 2011. 1-75.

[SoNIA, 2014] SoNIA: *Social Network Image Animator*. [online] [cit. 2014-06-28] Dostupné na internete: http://www.stanford.edu/group/sonia/.

5 Analýza sentimentu

5.1 Úvod

V súčasnosti web ponúka mnoho webových služieb umožňujúcich nielen efektívne vyhľadávanie informácií, ale aj spoluprácu a komunikáciu medzi používateľmi a to hlavne v rámci sociálneho webu, ktorý (ako už bolo spomenuté) podporuje interakcie medzi používateľmi. Tieto webové služby predpokladajú komunitu používateľov, ktorí ich aktívne využívajú a tak zdieľajú svoje znalosti. Takýmto spôsobom ožívajú pojmy kolektívna inteligencia a múdrosť davu. Avšak modely kolektívnej inteligencie a múdrosti davu prinášajú zo sebou problém so zabezpečením ochrany súkromia používateľov sociálneho webu 2.0. Napriek tomu aplikácie webu 2.0 priniesli nové možnosti, ako znalosti zdieľať a podieľať sa na ich rozširovaní medzi ostatných členov komunity. Príkladom takej aplikácie je aj automatická analýza názorov. V súčasnosti sa stalo kolaboratívne zdieľanie znalostí, názorov a postojov v komunitách ľudí predmetom nielen informačných vied, ale aj ďalších oblastí ako psychológia, sociológia a pod. [Strba-Bielikova, 2013].

Kolektívna inteligencia predstavuje „*zdieľa-nú alebo skupinovú inteligenciu, ktorá vzniká ako dôsledok spolupráce a súťaženia viacerých jednotlivcov a objavuje sa počas procesu hľadania konsenzu*“ alebo „*skupiny jednotlivcov spolupracujúcich na úlohách, ktoré sa zdajú byť inteligentné*“, podľa T.W. Malone, riaditeľa Centra pre kolektívnu inteligenciu, v MIT [Malone, 2012]. Žiadny z členov skupiny nevie všetko, ale keď sa spoja znalostí jednotlivcov môže vzniknúť rozsiahla kolektívna znalosť. Kolektívna inteligencia existovala už dávno pred vznikom informačných technológií (napríklad rodiny, národy alebo armády). Spolu s kolektívnou inteligenciou môže existovať aj kolektívna hlúposť, ak ľudia slepo nasledujú správanie ostatných používateľov. Informačné technológie súčastnosti hlavne internet, umožnili vznik nových foriem kolektívnej inteligencie ako je Wikipédia alebo aplikácie analýzy názorov.

Múdrosť davu predstavuje podľa Surowieckeho proces, ktorého cieľom je sumarizovať anonymne vytvorené údaje. Múdrosť davu je tvorená z odhadov veľkého počtu hodnotení jednotlivcov, ktoré aj keď sú nedokonalé, vďaka fenoménu emergencie môžu priniesť lepší výsledok ako najlepší získaný jednotlivý odhad [Surowiecki, 2004]. V knihe *Wisdom of the Crowd* „ [Surowiecki, 2004] sú definované štyri princípy potrebné pre využitie múdrosti davu:

- *rozmanitosť názorov*, ktorá predpokladá spoluprácu v rámci skupiny, ktorá nie je rovnorodá,
- *nezávislosť hodnotenia* jednotlivca bez ovplyvňovania hodnotením okolia,
- *decentralizácia*, ktorá zabezpečí, že nikto nediktuje davu svoj názor,
- *agregácia*, hodnotení jednotlivcov do kolektívneho rozhodnutia.

Podobne ako model kolektívnej inteligencie, tak aj model múdrosti davu má svoje obmedzenia. Prvým obmedzením je nesplnenie princípu nezávislosti názorov členov komunity. V reálnych podmienkach sa členovia komunity navzájom významne ovplyvňujú a to môže viesť ku skupinovému mysleniu (angl. „groupthink“), ktoré nereprezentuje múdrosť davu, pretože komunite, ktorá ho kreuje, chýba potrebná rôznorodosť. V takej komunite vzniká výsledná znalosť (múdrosť) výrazne iná ako

v prípade dostatočne diverzifikovanej skupiny. Ďalším obmedzením modelu múdrosti davu je, že ho je možné použiť len na riešenie objektívnych problémov.

5.1.1 Charakteristika analýzy sentimentu

Dnes komerčné firmy ponúkajú mediálny monitoring na sledovanie ohlasov v mediach pre:

- veľké firmy a organizácie,
- politické strany,
- subjekty verejnej a štátnej správy,
- rôzne odvetvia hospodárstva.

Monitorovanie mediálnych ohlasov je len prvým krokom. Dôležitejším je určovanie charakterov príspevkov a analýza sentimentu obsiahnutého v týchto príspevkoch, správach a ohlasoch. Prvýkrát použili metódu analýzy sentimentu Bo Pang a Lillian Lee v roku 2004 na analýzu recenzií filmov.

Analýza sentimentu z oficiálnych a neoficiálnych platforiem sociálneho webu (tweeter, blogy, microblogy, chat, chatrooms a rôzne diskusné fóra) môže byť s úspechom použitá na:

- predikovanie vývoja nálad spoločnosti a
- stanovenie miery pre určenie slobody prejavu v tlačených médiách.

Systémy využívajúce analýzu sentimentu určujú pocity a názory vyjadrené v texte formou prirodzeného jazyka. Analýzu textu je potrebné robiť s ohľadom na spracovávanú doménu. Podľa Osgooda rozoznávame tri emočné rozmery písaného textu, ktoré definujú sémantický priestor:

- *hodnotenie* (pozitívne alebo negatívne)
- *účinok, potencia* (závisia od nasledovných faktorov):
    - *vzdialenosť* (vzťah autora k téme, nakoľko je orientovaný v téme a nakoľko ho zaujíma)
    - *špecifickosť* (jasná, vágna formulácia)
    - *určitosť* (istota respektíve pochybnosti autora o jeho názoroch)
- *intenzita* (sila hodnotenia alebo emócie).

Metódy analýzy sentimentu pracujú so slovami, ktoré nevyjadrujú priamo pocity iba hodnotia tému. Sentiment je možné vyjadriť aj pomocou irónie až sarkazmu, ktoré sa veľmi zložito a komplikovane identifikujú. Špeciálnym prostriedkom je používanie emotikonov, ktoré samé o sebe veľmi jednoznačne ukazujú na vyjadrovanú emóciu. Základom analýzy sentimentu je vytvorenie preddefinovaných slovníkov, v ktorých sú slová zoskupené podľa psycho-sociálnych kritérií merania intenzity textu.

Analýza sentimentu využíva podľa autorov Budinská – Balogh – Gatial (UI SAV Bratislava) dva základné prístupy k spracovaniu prirodzeného jazyka:

- *Symbolické metódy*, respektíve slovníkový prístup – text sa spracúva ako súbor slov bez zohľadnenia vzťahu medzi nimi alebo gramatických pravidiel. Spracovanie sentimentu slov je základom k určeniu sentimentu dokumentu, pričom sa využíva agregácia sentimentu jednotlivých slov. Ako slovník alebo ako zdroj pre budovanie slovníka môže byť použitý napríklad WordNet.
- *Metódy strojového učenia* – tieto metódy sa klasicky používajú na dolovanie v textoch a analýza sentimentu sa v drvivej väčšine prípadov zužuje na dolovanie názorov v krátkych textoch, typických pre sociálny web. Hľadaný

sentiment je potrebné rýchlo objavovať, teda v reálnom čase, kým je aktuálny. Často sa nezaobídeme bez štatistického strojového prekladu pomocou kompresovaných polí prípon a stromov prípon.

5.2 Analýza názorov

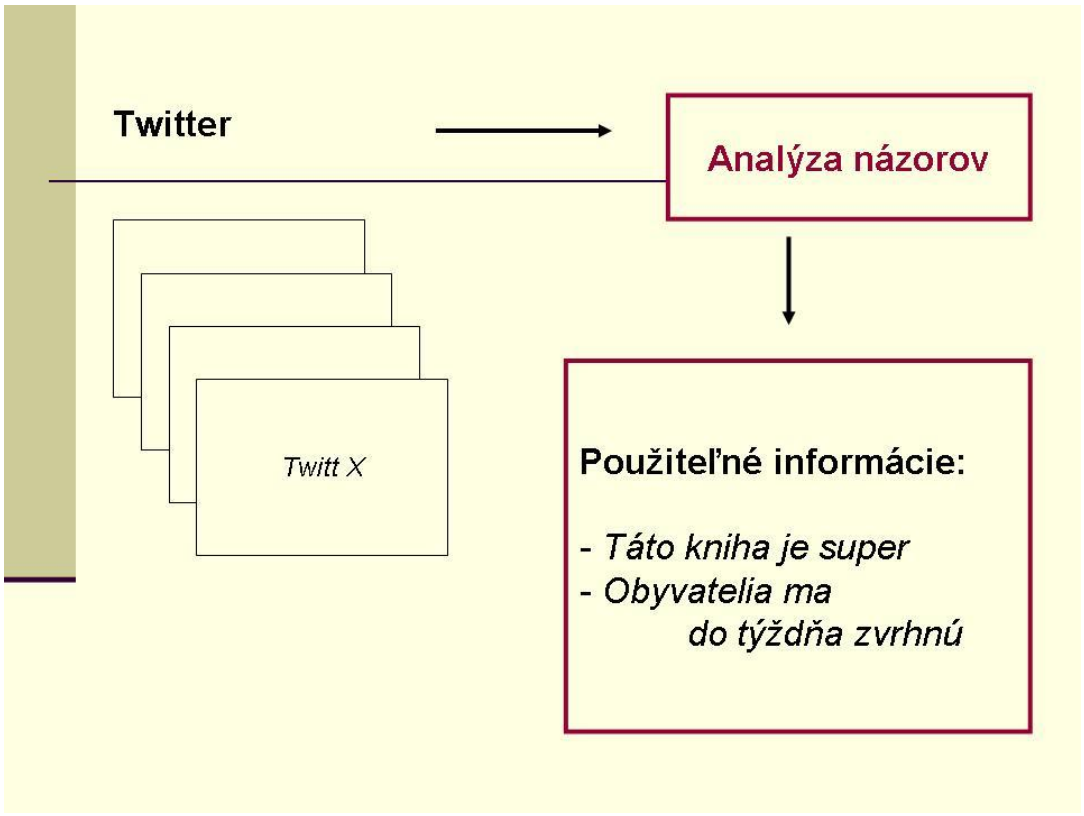
V rámci tejto publikácie sa budeme venovať určitému zúženému výseku problémov sociálneho webu a to skúmaniu typu sentimentu, ktorý je prezentovaný v konverzačnom obsahu webu. Sentiment môže predstavovať rozličné aspekty naladenia používateľa sociálneho webu, ako napríklad emócie, názory, postoje, pocity a podobne. My sa sústredíme na analýzu názorov. Uvedieme rozličné metódy získavania a analýzy názorov z webových diskusných fór. Niekedy sa uvedený problém analýzy názorov označuje aj názvom dolovanie názorov (opinion mining), keďže ide o extrakciu pozitívneho alebo negatívneho postoja participanta ku komentovaným objektom z textových zdrojov. Dolovanie názorov, resp. dojmov môže byť rozšírené z vnímania celkových textov na úroveň vlastností posudzovaných objektov [Ding-Liu-YuA, 2008]. Paralelne s dolovaním názorov sa v literatúre objavuje pojem analýza sentimentu (sentiment analysis) [Pang-Lee, 2008]. Anglické. ekvivalenty, ktoré sa najčastejšie používajú sú: sentiment analysis, opinion analysis, opinion classification, opinion mining, opinion extraction.

5.2.1 Webové diskusné fóra

Webové diskusné fóra predstavujú jeden z mnohých prostriedkov umožňujúcich prispievanie k obsahu webu, a teda prostriedkom, pomocou ktorého aj bežný používateľ, ktorý nie je programátor alebo inak informaticky zdatný, môže pretvárať obsah webu. Tak sa z konzumenta webového obsahu stáva jeho producent. Diskusné fóra vytvárajú takzvaný konverzačný obsah v rámci už spomenutých platforiem ako: blog, mikroblog, chat, IRC a pod. Diskusia v nich môže mať formu „point-to-point“ alebo „multicast“ diskusie. My sa zameriavame na „multicast“ diskusie.

Diskusné fóra sú taktiež zdrojom poznatkov o názoroch, pocitoch a postojoch používateľov. Nárastom množstva dát na diskusných fórach sa stávajú pre človeka ťažko spracovateľné. Preto sa vynára potreba ich automatického spracovania, napríklad pomocou automatickej klasifikácie názorov.

Diskusné fóra reprezentujú rozsiahle databázy týchto názorov, pocitov, postojov a nálad ľudí, ktorí používajú Internet ako spôsob komunikácie. Na rozdiel od klasických databáz neobsahujú štruktúrované dáta, preto vyžadujú špeciálne postupy. Takýto špeciálny postup je navrhnutý a vylepšovaný v rámci klasifikácie názorov. Na klasifikáciu názorov sa pozeráme ako na metódu, systém, ktorý má na vstupe množinu krátkych textov (napríklad twittov) zozbieraných z jednej alebo aj viacerých diskusií na konkrétnu tému. Na výstupe tohto systému sa očakáva sumarizovaná použiteľná informácia typu „Tento výrobok je u používateľov obľúbený“ alebo „Obyvatelia prijímajú túto reformu s veľkou nevôľou“, čo je ilustrované na Obr.5.1.



Obr. 5.1 Vstupy a výstupy systému analýzy názoru.

Táto publikácia sa bude zaoberať špeciálnym druhom analýzy názorov a to klasifikáciou názorov, keďže pôjde o klasifikáciu krátkych textov do preddefinovaných tried a to *pozitívny názor, neutrálny názor a negatívny názor* na základe analýzy obsahu textu pomocou slovníkových metód alebo metód strojového učenia.

5.3 Klasifikácia názorov

5.3.1 Základné problémy klasifikácie názorov

Klasifikácia názorov sa na rozdiel od klasifikácie dokumentov zameriava na subjektívne stránky textu. Musí byť jasné, na čo sa tieto subjektívne informácie vzťahujú, inak povedané aká je *téma diskusie*. Či ide o hodnotenie produktu, hotelu, politickej situácie, osoby, lekára, politika, udalosti, filmu, knihy, alebo pocitov autora k objektu hodnotenia.

Autor názoru resp. držiteľ názoru (opinion holder) je podľa [Ding-Liu-YuA, 2008] osoba, resp. organizácia, ktorá má konkrétny názor na konkrétny objekt. Objekt definuje ako tému, na ktorú sa názor vzťahuje. Názor je potom pohľad, postoj alebo hodnotenie objektu držiteľom názoru.

V rámci klasifikácie názorov sa pozornosť sústreďuje na tie slová, ktoré sú dobrými nositeľmi subjektívnych postojov prispievateľov do diskusného fóra, ako prídavné mená a príslovky, ale aj niektoré slovesá a podstatné mená. Nie je potrebné ani vhodné brať do úvahy všetky slová textu a všetky slovné druhy, keďže niektoré slová nevyjadrujú názor, iba popisujú fakty, preto neznamenajú žiadny prínos ku

klasifikácii názorov. Základné problémy, ktoré klasifikácia názorov musí riešiť sú:

- určenie subjektivity slova
- určenie polarity (orientácie) slova
- určenie intenzity polarity.

Určenie subjektivity lexikálnych jednotiek spracovávaného textu je prvý problém, ktorý je potrebné vyriešiť. Napríklad veta „Klasifikácia názorov ma zaujíma.“ vyjadruje osobný postoj autora k danej veci a teda je prospešná pre klasifikáciu názorov. Na druhej strane veta „Klasifikácia názorov je v tejto prezentácii.“ obsahuje iba informáciu o fakte bez akýchkoľvek postojov a teda je pre extrahovanie postojov bezvýznamná.

Určenie polarity lexikálnych jednotiek predstavuje jej zatriedenie do jednej z troch kategórií: pozitívny názor, neutrálny názor a negatívny názor. Predpokladá sa, že autori príspevkov na niečo reagujú, na otázku, príspevok v rámci diskusie o konkrétnom objekte. Je potrebné určiť polaritu každej lexikálnej jednotky na základe polarity jednotlivých slov.

Nestačí určiť polaritu lexikálnych jednotiek ale je potrebné rozlíšiť *intenzitu polarity*. Hovoríme o určovaní stupňa polarity (polarity degree). Každý príspevok vyjadruje spravidla názor s inou vervou, silou teda intenzitou. Sila polarity sa vyjadruje na škále od veľmi slabej po veľmi silnú. Tak napríklad negatívna polarita môže byť: slabo negatívna, mierne negatívna a silno negatívna. Napríklad veta „Tá kniha je otrasná.“ má silnú negatívnu polaritu, zatiaľ čo veta „Tá kniha nebola až taká dobrá.“ má iba slabú negatívnu polaritu.

Je možné definovať nasledovné kroky postupného získavania polarity slov, viet, príspevkov a napokon celej diskusie:

- part-of-speech analýza (priradenie slovného druhu)
- vytvorenie „seedlistu“ respektíve klasifikačného slovníka (zoznam pozitívnych a negatívnych prídavných mien a prísloviek)
- nájdenie synoným a antoným v „seedliste“
- porovnanie počtu kladne a záporne orientovaných častí príspevku
- otočenie polarity záporom
- spracovanie synoným a antoným.

Klasifikácia ku konkrétnej polarite môže byť realizovaná pomocou jednoduchých pravidiel:

počet (positive) > počet (negative) –> positive
počet (positive) < počet (negative) –> negative
počet (positive) = počet (negative) –> neutral

Toto je dosť striktný prístup, lebo keď máme 41 pozitívnych jednotiek a 40 negatívnych, ťažko môžeme povedať, že ide o pozitívnu polaritu. V tomto prípade ide skôr o neutrálnu polaritu. Preto bolo navrhnuté vyhodnotenie polarity s použitím prahu P, nasledovným spôsobom:

- počet (positive) - počet (negative) > P –> positive
- počet (positive) - počet (negative) < -P –> negative
- |počet (positive) - počet (negative)| ≤ P –> neutral.

5.3.2 Webová služba klasifikácie názorov a motivácia jej vzniku

Klasifikácia názorov má uplatnenie v oblastiach, kde je potrebné agregovať veľké množstvo rôznych názorov do jednej výslednej ucelenej použiteľnej informácie. Zvlášť dobre sa klasifikácia názorov dá využiť pri výrobe, vývoji a predaji produktov. Tieto oblasti sa skúmajú z dvoch pohľadov a to z pohľadu spotrebiteľa, ktorému môže funkčná aplikácia klasifikácie názorov uľahčiť kúpu nejakého drahého produktu, a z pohľadu výrobcu, ktorému môže byť takto uľahčený vývoj a predaj jeho výrobkov.

Z pohľadu spotrebiteľa je Internet veľmi užitočný nástroj. Internet môže byť zdrojom vecných informácií pre rozhodnutie o kúpe produktu ako napríklad užitočné informácie o cene, dizajne, doplnkoch a rozličných funkciách výrobku. Avšak, používateľ potrebuje aj informácie iného druhu o spokojnosti iných používateľov (už vlastníkov) s daným produktom. Takéto znalosti je možné objaviť v rozličných diskusiách na webových portáloch. S tým sú spojené isté problémy, ako je obrovský počet diskusných príspevkov ako aj nehomogénnosť týchto príspevkov. Tieto problémy by mohla vyriešiť webová služba, ktorá automaticky spracuje celú diskusiu (respektíve viac diskusií), klasifikovala by názory v nej obsiahnuté a následne extrahovala sumarizovaný názor.

Z pohľadu výrobcu, na ktorého kladie podnikanie v súčasnosti mimoriadne nároky, je Internet tiež veľmi užitočný nástroj. Internet je pre neho zdrojom informácií o dodávateľoch a konkurencii, ale môže byť pre neho aj zdrojom informácií o potrebách zákazníkov. Doteraz sa takýto prieskum potrieb zákazníkov robil pomocou marketingového prieskumu a bol spravidla vykonávaný prostredníctvom dotazníkov alebo telefónu. Nevýhoda marketingového prieskumu spočíva vo vysokých nákladoch na uskutočnenie dotazníkovej metódy, ktorá vyžaduje zapojiť veľké množstvo ľudí, respektíve v nákladoch na telefónne účty. Je taktiež časovo náročná. Je potrebné podotknúť, že rýchlosť získavania informácií o zákazníkovi je zásadná. Preto z pohľadu výrobcu sú webové diskusné fóra výborným zdrojom informácií pre takzvaný internetový prieskum. Odpadá čas potrebný na zber údajov, keďže sú prístupné na Internete okamžite. Čas získavania informácií z diskusných skupín pomocou webovej služby klasifikácie názoru by bol veľmi krátky. Dotazníky by neboli potrebné a telefónne účty by tiež výrazne poklesli.

V rámci metódy klasifikácie názorov platí, že polarita diskusnej skupiny je tvorená súčtom polarít jej príspevkov, že polarita príspevku je sumárom polarít jeho častí, teda slov, fráz alebo viet. Je potrebné brať do úvahy, že zápor mení polaritu časti textu, že synonyma majú rovnakú polaritu, a že antonyma majú opačnú polaritu.

Je množstvo problémov znižujúcich úspešnosť klasifikácie názorov:

- Prídavné meno s kladnou (respektíve zápornou) polaritou nesie v skutočnosti opačný postoj, ktorý je vyjadrený kontextom: „Rád si prečítam dobrú knihu. Táto taká nebola.“
- Prídavné mená a príslovky majú opačnú polaritu ako sa predpokladalo: „Tento výrobok je dobrá hlúposť.“
- Názor je vyjadrený nepriamo, bez prídavných mien a prísloviek, ktoré by napomohli určeniu polarity: „Na ten film by som už nešiel.“, „Inú značku by som si nekúpil.“

Na druhej strane existujú riešenia, ktoré by mohli zvýšiť úspešnosť klasifikácie názorov, ako napríklad:

- Uvažovanie kontextu.

- ❖ Zahrnutie ďalších slovných druhov.
- ❖ Identifikácia typických znakov hodnotenej veci. Napríklad pri filme alebo knihe uvažovanie zápletky, deja, efektov alebo pri digitálnom fotoaparáte uvažovanie ostrosti, obrazu alebo farby.
- ❖ Tvorba špecializovaných „seedlistov“ klasifikačných slovníkov pre rôzne domény.

5.4 Slovníkový prístup

Nie všetky slová textu sú rovnako dôležité pri odhaľovaní a klasifikácii názoru. Nositeľmi subjektívnych postojov v texte sú hlavne prídavné mena (perfektný), príslovky (katastrofálne), ale tiež podstatné mená (bomba) a slovesá (zničiť). Je preto nutné identifikovať v texte slová, ktoré majú resp. nemajú význam pri klasifikovaní názorov. To je zároveň aj prvý problém, s ktorým sa klasifikácia názorov stretáva. Ide o tzv. určenie subjektivity. Medzi ďalšie základné problémy, ktoré musí klasifikácia riešiť patrí aj určenie polarity slov a určenie sily polarity.

5.4.1 Základné problémy - subjektivita a polarita slova

Určenie subjektivity slova predstavuje klasifikácia slov do dvoch tried, a to do triedy slov so subjektivitou a slov, ktoré nie sú použiteľné pre ďalšiu analýzu sentimentu. Totiž ani nie všetky tie slová, ktoré patria k sledovaným slovným druhom (prídavné mená, príslovky, podstatné mená, slovesá) sú nositeľmi názoru, resp. potrebnej informácie na určenie polarity príspevku. V Tab.5.1 sú uvedené príklady slov vhodných na klasifikáciu názoru (obsahujúcich subjektivitu) ako aj subjektívne neutrálnych slov.

Tab. 5.1 Triedenie slov podľa vhodnosti pre klasifikáciu názorov.

Slová	Obsahujúce subjektivitu	Subjektívne neutrálne
Prídavné mená	perfektný, nekvalitná	žltá, slovenský, trávnatý
Príslovky	výborne, skvelo, otrasne	mokro, vysoko, úzko
Podstatné mená	super, bomba, hlúposť	motorka, jedlo, človek
Slovesá	zničiť, zefektívniť	točiť, plávať, budovať

Rozpoznávame tri základné stupne polarity slov a to pozitívnu, negatívnu a polaritne neutrálnu, pričom polaritne neutrálne slovo nie je to isté ako subjektívne neutrálne slovo. Následne podľa toho môžeme potom klasifikovať aj vety a príspevky do skupín na pozitívne, negatívne a neutrálne (viď Tab.5.2). Teoreticky, ak to problém vyžaduje, môže byť definovaných aj viac kategórií spojených s jemnejším delením.

Pri určovaní polarity nastáva aj ďalší problém a tým je obracanie polarity slov prostredníctvom záporu. Pomerne často sa v diskusiách stretávame s formuláciou viet, kde na začiatku vety figuruje zápor a neskôr sa v nej nachádza kladné slovo.

Príklad: “Nie je to z jeho strany pekné.”

Príklad: “Nebol to kvalitný film.”

Menej často sa vyskytujú vety so záporom na začiatku a negatívnym slovom v ďalšej časti vety. Príklad: “Nie je to zlé.” alebo “Nie som najhorší.”

Tab.5.2 Triedenie slov podľa vhodnosti pre klasifikáciu názorov.

Pozitívne slová	Negatívne slová	Neutrálne slová
perfektný, skvelo, super, páčiť	nekvalitná, otrasne, hlúposť, zničiť	priemerný, efektívne, komunizmus, mazať

Aj na týchto príkladoch môžeme vidieť, že diskusný príspevok nie je možné klasifikovať len na základe polarity jednotlivých slov, ale treba sa naň pozerať v širšom kontexte z pohľadu viet, alebo častí viet.

5.4.2 Základné problémy - intenzita polarity slova

Intenzita polarity predstavuje silu, resp. veľkosť podpory slova, vety či diskusného príspevku k potvrdeniu alebo vyvráteniu názoru (na produkt, politické rozhodnutie, službu...). Každému slovu je možné priradiť hodnotu na základe stupnice intenzity polarity. Pomocou takejto stupnice je možné roztriediť slová do niekoľkých skupín s príslušnými hodnotami. V Tab.5.3 sú uvedené príklady takýchto stupníc.

Tab. 5.3 Symbolické a číselné stupnice polarity.

Počet stupňov	Stupnice polarity	
2	negatívne	pozitívne
6	slabo negatívne, mierne negatívne, silno negatívne	slabo pozitívne, mierne pozitívne, silno pozitívne
8	-4, -3, -2, -1	1, 2, 3, 4

Silu polarity je možné vyjadriť tak slovne, ako aj číselne. Číselné vyjadrenie je vhodnejšie pre ďalšie spracovanie programom. Uvažujme stupnicu od silno pozitívnych slov až po silno negatívne. Potom je možné slová zatriediť do skupín tak, ako je to uvedené v Tab.5.4. Toto triedenie sa týka sily polarity jedného slova.

Tab. 5.4 Triedenie slov podľa príslušnosti k stupňu polarity.

silno pozitívne	perfektný, vynikajúci, božský, úžasný
mierne pozitívne	pekný, chvályhodný, kvalitný, šikovný
slabo pozitívne	vhodný, dobrý, frajerský, fajn
slabo negatívne	slabší, priemerný, nemastný, neslaný
mierne negatívne	zlý, nefunkčný, slabý, nevyhovujúci
silno negatívne	otrasný, katastrofálny, najhorší, úbohý

Pre zvýšenie presnosti klasifikácie je však nutné sa na diskusné príspevky pozerať z pohľadu viet, resp. minimálne z pohľadu dvojice slov. Príspevky obsahujú často kombinácie slov, ktoré svojím významom posúvajú silu polarity do vyššej, alebo nižšej roviny.

Príklad posunu do vyššej roviny: “veľmi pekný, vysoko kvalitný”

Príklad posunu do nižšej roviny: “o dosť slabší, veľmi nekvalitný”

Prvý príklad reprezentuje zmenu sily polarity pôvodného slova z mierne pozitívnej na silno pozitívnu a druhý príklad zo slabo negatívnej na mierne negatívnu polaritu.

5.4.3 Tvorba špecializovaných slovníkov a ich použitie

Pre účely klasifikácie názorov je potrebné automaticky vytvoriť slovník (seedlist) slov, ktoré sú nositeľmi názoru v danej oblasti (doméne). Nazvime ho klasifikačným slovníkom. Veľmi málo respondentov, ktorí prispievajú na diskusné fóra a tým pádom vytvárajú časť obsahu webu, ovláda svoj materinský jazyk a jeho gramatiku perfektne. Preto je nutné, aby sa aplikácia klasifikácie názorov prispôsobila reči, ktorú bežne používajú používatelia webu vo svojich komentároch a názoroch. Stáva sa takmer pravidlom, že hlavne mladá generácia píše svoje príspevky bez diakritiky a tiež v nich môžeme nájsť gramatické chyby. Z tohto dôvodu je tvorba slovníkov skupín (pozitívnych, negatívnych, neutrálnych) slov veľmi náročná a často je nutné do nich zahrnúť aj nespisovné slová (viď Tab.5.5).

Tab. 5.5 Príklady špecializovaných slov v slovníku.

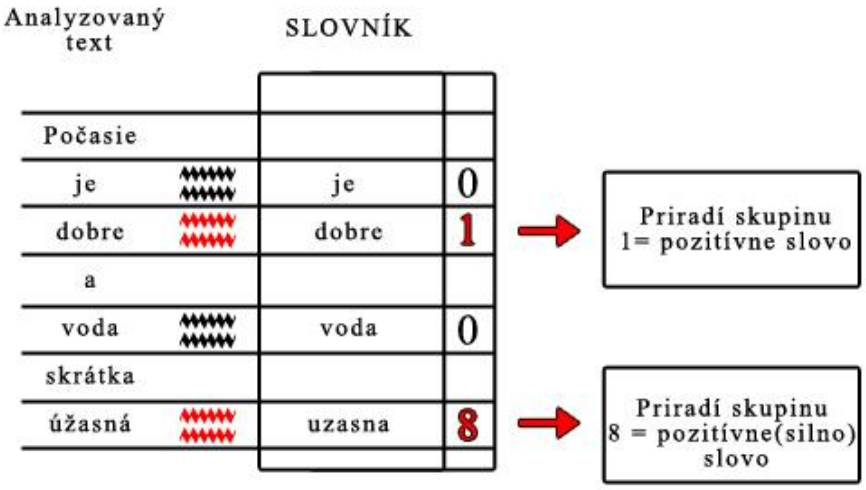
spisovné slová	výborný, pohodlný, slabý
slangové slová	coolový, dzivy
slová bez diakritiky	kvalitny, paci (sa mi to)

Po rozdelení textu na jednotlivé slová sa tieto hľadajú v preddefinovaných slovníkoch, ktoré obsahujú slová, ktoré sú nositeľmi subjektivity. Týmto slovám sa priradí hodnota zo stupnice od 0 do 9. Slovám, ktoré nemajú subjektivitu sa nepriradí žiadna hodnota, ako je to ilustrované na Obr.5.2.

Určenie polarity je ďalším krokom v procese analýzy textu. Nami navrhnutá metóda rieši túto úlohu súčasne s určovaním subjektivity slova. Ak má slovo kladnú polaritu (pozitívnu polaritu), je mu priradená jedna z hodnôt 1 a 8 (viď Tab.5.6). Ak má negatívnu polaritu je mu priradená jedna z hodnôt 2 a 9. Ak má neutrálnu polaritu je mu priradená hodnota 0. Tento proces priraďovania hodnôt slovám podľa slovníka rieši aj určovanie sily polarity, pretože hodnoty 1 a 2 reprezentujú miernu polaritu a hodnoty 8 a 9 silnú polaritu či už pozitívnu alebo negatívnu. Navrhnutá metóda používa pri určovaní sily polarity päťhodnotovú stupnicu (vrátane neutrálnej). Päť stupňov je postačujúci počet, keďže na určovanie sily polarity je tu ešte jeden nástroj a to hodnota 4 pre prípad, že jedno slovo predchádza iné slovo so subjektivitou a zvyšuje jeho polaritu (vysoko kvalitný). Obr.5.2 ilustruje postup určovania sily polarity na príklade vety: „Počasie je dobré a voda skrátka úžasná.“.

Tab.5.6 Klasifikačné stupne pre slová so subjektivitou.

mierne pozitívne a silno pozitívne	1 a 8
mierne negatívne a silno negatívne	2 a 9
neutrálne	0
zápor – obracanie polarity	3
zvyšovanie polarity - intenzifikácia	4
nevyužité hodnoty	5, 6 a 7



Obr.5.2 Určovanie subjektivity a sily polarity slov vo vete (slová: počasie, a, skrátka nemajú subjektivitu).

5.4.4 Ďalšie problémy klasifikácie názorov

Navrhnúť vhodnú metódu klasifikácie názorov nie je jednoduchý proces. Je potrebné zohľadniť všetky časti klasifikácie a implementovať ich v správnom a logickom poradí. Okrem základných problémov klasifikácie názorov uvedených v predošlom (určenie subjektivity slova, určenie polarity slova a určenie intenzity polarity slova) je potrebné vyriešiť nasledovné netriviálne problémy:

- Obracanie polarity záporom
- Určenie sily polarity kombinácie slov
- Určenie dynamického koeficientu
- Selekcia správnych kombinácií slov
- Nahrávanie slov do klasifikačných slovníkov

Prezentovaná metóda klasifikácie názorov zohľadňuje nielen silu polarity jednotlivých slov osobitne ale aj silu polarity kombinácie viacerých slov. Táto metóda je navrhnutá tak, aby analyzovala predložené texty – diskusné príspevky z nejakej domény a pre túto doménu automaticky z daných textov zostaví slovníky slov, ktoré budú hrať kľúčovú úlohu pri samotnej klasifikácii názoru.

Navrhnutá metóda zohľadňuje aj štruktúru a skladbu slovenského jazyka, aby zabezpečila čo najpresnejšiu analýzu textu. Metóda transformuje text na pole slov – termov. Každému slovu je priradená číselná hodnota zo slovníka, ktorá reprezentuje polaritu daného slova. Sú identifikované vety. Keď sa v poli termov vyskytne prvá nenulová hodnota, začne sa vytvárať kombinácia slov. O tom, aká veľká kombinácia slov sa vytvorí, rozhoduje dynamický koeficient *K*. Každej kombinácii slov ako celku sa tiež priradí hodnota. Hodnoty všetkých uvažovaných kombinácií slov sa podieľajú na určovaní polarity celého príspevku a následne z polarít jednotlivých príspevkov sa určuje polarita celej diskusie.

Pri určovaní polarity celého príspevku je rozhodujúce, či v ňom prevažujú slová, resp. kombinácie slov s pozitívnou polaritou (potom je pozitívny) alebo slová s negatívnou polaritou (potom je negatívny). Neutrálny je príspevok vtedy, keď je

počet pozitívnych a negatívnych slov, resp. kombinácií slov rovnaký a teda rozdiel týchto počtov je nulový. Tento prístup k určovaniu neutrality príspevku je striktný. Tento striktný prístup je vhodný pre krátke príspevky, kde aj rozdiel o jednotku medzi počtom pozitívnych a negatívnych lexikálnych jednotiek môže byť rozhodujúci.

Je možné k tomuto problému pristupovať benevolentnejšie a stanoviť pravidlo (14) na určovanie neutrality nasledovne:

$$IF\ |Pocet\_pozit - Pocet\_negat| < H\ THEN\ neutralita. \qquad (14)$$

Šírku pásma neutrality môžeme meniť nastavovaním rôznych hodnôt H, pričom H ≥ 1 a je to celé číslo. Pre veľmi krátke príspevky je vhodnejšie striktné pásmo neutrality s hodnotou H=1 (rozdiel počtu pozitívnych a negatívnych slov je rovný 0), pretože ak taký príspevok pozostávajúci z jednej vety obsahuje iba jedno pozitívne alebo negatívne slovo, širšie pásmo neutrality by ho pohltilo a systém klasifikácie názorov by ho vyhodnotil ako neutrálny príspevok.

Širšie pásmo neutrality je vhodné pre spracovanie dlhších príspevkov. Vo všeobecnosti, pri dlhších príspevkoch je vhodnejšie zaviesť pásmo neutrality H, ktoré zabezpečí, že malý rozdiel medzi pozitívnymi a negatívnymi lexikálnymi jednotkami nebude hrať tak významnú rolu pri veľkom celkovom počte lexikálnych jednotiek.

5.4.5 Obracanie polarity záporom

Obracanie polarity prostredníctvom záporu slúži ku presnejšej klasifikácii analyzovaného textu. Slová reprezentujúce zápor (nie (je), ne...) patria do kategórie s hodnotou 3, ktorá spĺňa svoju úlohu až v kombinácii s inou kategóriou spojenou s jedným zo štyroch stupňov na určovanie polarity, ako je to uvedené v Tab.5.7, pričom mení pozitívnosť na negatívnosť a naopak v rámci daného stupňa.

Tab. 5.7 Obracanie polarity záporom.

3 + 1	3 + 8	3 + 2	3 + 9
zápor + mierna pozitivita = *mierna negativita*	zápor + silná pozitivita = *silná negativita*	zápor + mierna negativita = *mierna pozitivita*	zápor + silná negativita = *silná pozitivita*

Pri tejto metóde obracania polarity predstavuje nehomogénnosť štruktúry vety veľký problém. Rozmanitosť vetných štruktúr v slovenčine je demonštrovaná v Tab.5.8 na príklade troch podôb toho istého tvrdenia. Úlohou implementácie je spracovať tri vstupy vo forme kódu 0301, 0031 a 0013 ako tú istú informáciu s rovnakou silou polarity.

Tab. 5.8 Generovanie kombinačného kódu záporných viet.

Mobil	nie	je	kvalitný
0	3	0	1
Tento	mobil	nebol	kvalitný
0	0	3	1
Tento	mobil	kvalitný	nebol
0	0	1	3

Naviac aj kombinácia s kódom 3000010 reprezentujúca vetu: “Nie je to podľa mňa kvalitný mobil.” má tú istú polaritu aj silu polarity. Tu sa vynára potreba dynamického koeficientu, ktorý by vedel zohľadniť potrebu meniť dĺžku kombinácie slov spracovávaných ako jedna lexikálna jednotka.

5.4.6 Intenzifikácia – určovanie sily polarity

Podobne ako zápor, ktorý sám o sebe nemá polaritu a jeho vplyv na polaritu príspevku je možné vyhodnotiť iba pri analýze kombinácie slov, aj slová zvyšujúce silu polarity (intenzifikátory) má zmysel uvažovať iba v rámci kombinácie slov. Tieto slová sú zaradené do kategórie 4. Sú to väčšinou príslovky, ako napríklad: veľmi, totálne, dosť a pod. Tab.5.9 uvádza dve rôzne vety obsahujúce slová zvyšujúce intenzitu polarity s kódmi 00041 a 4002.

Tab. 5.9 Generovanie klasifikačného kódu viet s intenzifikátormi.

Ten	mobil	je	veľmi	kvalitný
0	0	0	4	1
neutrálne	neutrálne	neutrálne	+ intenzita	mierne pozitívne
Dosť	ma	to	hnevá	
4	0	0	2	
+ intenzita	neutrálne	neutrálne	mierne negatívne	

5.4.7 Dynamický koeficient

Navrhnutá metóda klasifkácie názorov si dáva ambíciu popasovať sa s variabilitou vetných štruktúr v slovenskom jazyku. Používa k tomu dynamický koeficient *K*. V počiatočných fázach návrhu vystupoval tento koeficient ako statický parameter algoritmu (implementácie). Ďalšie návrhy smerovali k tomu, aby sa tento parameter menil dynamicky pri každej spracovávanej lexikálnej jednotke a prispôsoboval sa dĺžke jej kódu (resp. počtu núl v ňom). Hodnota *K* vyjadruje koľko slov v rade má byť podrobených skupinovej analýze, inými slovami určuje dĺžku kombinácie slov. Určuje počet slov, ktoré budú pripočítavané do kombinácie od prvého nenulového prvku, ktorý program v poli slov zachytil. V prípade, že hodnota koeficientu *K* je nastavená tak, že presahuje dĺžku vety, do kombináte slov sa započítajú iba slová z danej vety. Teda koeficient *K* sa dynamicky zmenší. Výsledná kombinácia slov môže byť kratšia aj v prípade, keď pre *K=4* je spracovávaná kombinácia 3011. V tom prípade sa kreujú dve kombinácie a to 301 a 1. Tab.5.10 ilustruje princíp použitia tohto koeficientu.

Tab.5.10 Princíp fungovania dynamického koeficientu *K* (slová, ktoré sa spracovávajú v rámci jednej kombinácie sú zvýraznené boldom a podškrtnutím).

K	Nie	je	to	dobrý	telefón
1	**3**	0	0	**1**	0
2	**3**	**0**	0	**1**	**0**
4	**3**	**0**	**0**	**1**	0

Z Tab.5.10 vyplýva, že *K=1* je nevhodné pre spracovanie vety „Nie je to dobrý telefón“, pretože zápor by bol v inej lexikálnej jednotke ako mierne pozitívne slovo „dobrý“, ku ktorému je tento zápor vzťahovaný. *K=1* predstavuje hraničnú hodnotu

koeficientu, ktorá nepripúšťa spracovanie kombinácií slov, iba spracovanie slov ako izolovaných jednotiek. *K=2* síce pripúšťa spracovanie dvojslovných kombinácií, ale to nerieši problém s izolovaním záporu od slova, ktorého sa týka. Tento problém je uspokojivo riešený pri danej vete až koeficientom *K=4*.

Dynamický koeficient by sa mal stanovovať automaticky podľa charakteru analyzovaného textu (dĺžka viet a pod.). Tri možné spôsoby stanovenia dynamického koeficientu sú:

- Priemerná dĺžka vety
  - Počíta sa aritmetický priemer početností slov každej lexikálnej jednotky analyzovaného textu.
  - Dynamický koeficient je rovnaký pre všetky lexikálne jednotky – vety analyzovaného textu.
- Polovica dĺžky vety
  - Početnosť slov lexikálnej jednotky je delený dvoma so zaokrúhlením na hor.
  - Dynamický koeficient sa nastavuje zvlášť pre každú vetu analyzovaného textu.
- Hybridný prístup
  - Ku dĺžke každej lexikálnej jednotky sa pripočíta priemerná hodnota dĺžok všetkých viet analyzovaného textu a to sa delí konštantou odvodenou od štruktúry textu (napríklad 5).

5.4.8 Typovanie kombinácií slov

Pre účely klasifikácie názorov a určenie polarity textu je potrebné spracovať jednotlivé lexikálne jednotky textu. Pod lexikálnou jednotkou sa myslí skupina slov teda K-tica slov (nemusí to byť iba veta: môže to byť krátka veta, alebo časť vety). Takej K-tici slov budeme hovoriť kombinácia. Teda „kombinácie slov“ sú K-tice, ktorým je priradená polarita ako celku. Táto polarita je celé číslo z intervalu <-3, 3> a každý z uvedených stupňov polarity má svoju interpretáciu, čo je ilustrované v Tab.5.11.

Tab. 5.11 Uvažované stupne polarity kombinácie a ich interpretácia.

SP + I	SP, MP + I	MP	MN	SN, MN + I	SN + I
silná pozitivita + intenzita	silná pozitivita, resp. mierna pozitivita + intenzita	mierna pozitivita	mierna negativita	silná negativita, resp. mierna negativita + intenzita	silná negativita + intenzita
3	**2**	**1**	**-1**	**-2**	**-3**

Napríklad hodnota polarity 2 môže byť interpretovaná ako silná pozitivita alebo mierna pozitivita posilnená intezifikátorom (SP alebo MP + I). Každá z kombinácií reprezentuje práve jednu interpretáciu a je jej priradená práve jedna hodnota polarity. Napríklad kombinácia 4010 reprezentuje interpretáciu SP, MP + I a je jej priradená hodnota pozitívnej polarity +2 (viď Tab.5.12).

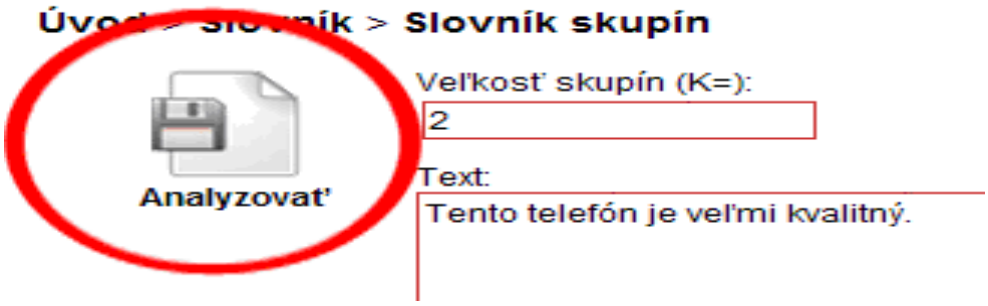
Od hodnoty polarít jednotlivých kombinácií ako stavebných kameňov textu sa odvíja jeho celková polarita. Pozitívny príspevok (resp. diskusia) je taký, v ktorom prevažujú kombinácie (resp. príspevky) s pozitívnou polaritou. Podobne pre negatívnu polaritu.

Tab. 5.12 Uvažované stupne polarity kombinácie a ich interpretácia.

Interpre-tácia	SP + I	SP MP + I	MP	MN	SN MN + I	SN + I
K = 2	48	80, 41	10, 32, 23	20, 31, 13	90, 42	49
K = 3	480,408	800, 410, 401	100, 320, 230, 302, 203	200, 310, 130, 301, 103	900, 420, 402	490, 409
K = 4	4800, 4080, 4008	8000, 4100, 4010, 4001	1000, 3200,2300, 3020,2030, 3002,2003	2000, 3100,1300, 3010,1030, 3001,1003	9000, 4200, 4020, 4002	4900, 4090, 4009
polarita	**+3**	**+2**	**+1**	**-1**	**-2**	**-3**

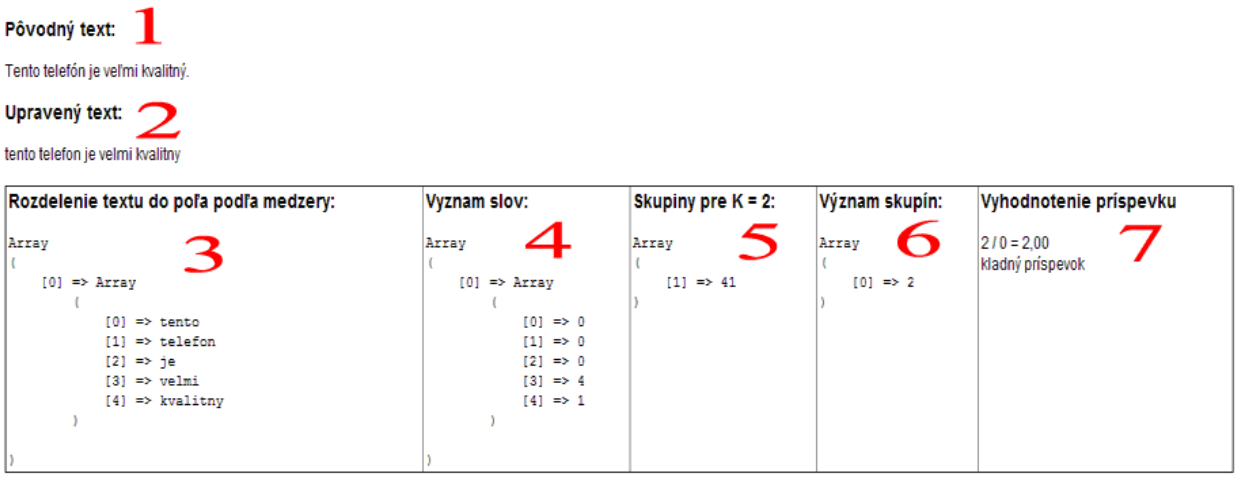
5.4.9 Implementácie

Uvedený prístup ku klasifikácii bol implementovaný. Prvá implementácia bola statická, pretože koeficient určujúci dĺžku kombinácie bol zadávaný ako parameter algoritmu. Táto implementácia bola nazvaná KLAN (systém KLAsifikácie Názoru) a je ilustrovaná na Obr.5.3. Má dve rozhrania „guest“ a „admin“. „Guest“ môže klasifikovať zvolený text a nastavovať koeficient K. „Admin“ môže nahrávať a editovať klasifikačný slovník. Aplikácia mu pri detekcii neznámych slov automaticky ponúka možnosť nahrať ich do slovníka.



Obr. 5.3 Statická implementácia slovníkovej metódy klasifikácie názorov KLAN – úvodná obrazovka.

Nasledovný Obr.5.4 ilustruje výsledok procesu klasifikácie názorov pomocov KLAN.



Obr. 5.4 Statická implementácia slovníkovej metódy klasifikácie názorov KLAN – výsledok spracovania.

Postupnosť krokov spracovania je nasledovná:

1 – pôvodný text (aktuálna veta „Tento telefón je veľmi kvalitný.“)

2 – upravený text (segmentácia slov a odstránenie diakritiky)

3 – transformácia textu do poľa slov

4 – pridelenie polarity slovám

5 - identifikácia kombinácií slov podľa koeficientu *K* (aktuálne *K=2*)

6 - určenie polarity kombinácií

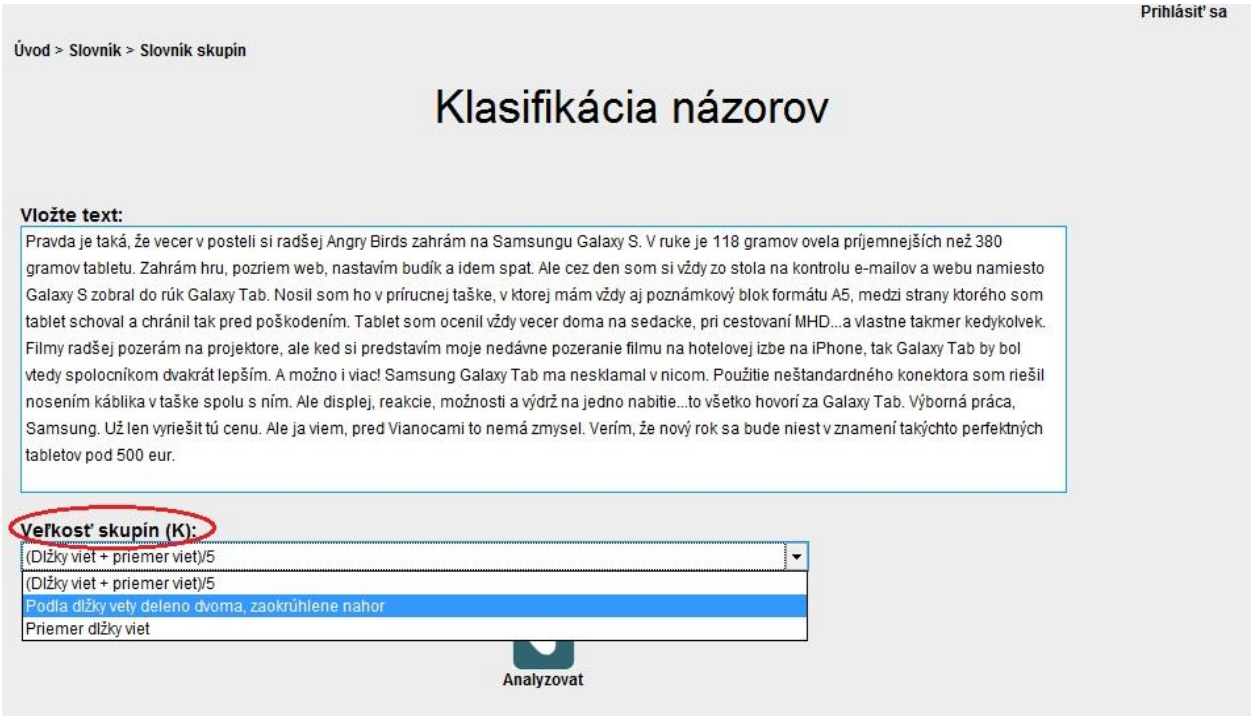
7 – určenie polarity príspevku.

Táto implementácia bola testovaná na diskusných príspevkoch internetového portálu „www.mobilmania.sk“ konkrétne pre vlákno LGKU990. Testovaná bola vzorka: 236 viet, 1558 slov. Klasifikačný slovník obsahoval 27 pozitívnych slov, 27 negatívnych slov, 10 záporných slov a 11slov zvyšujúcich intenzitu, takzvaných intenzifikátorov. Priemerná dosiahnutá presnosť bola 78,2 % ako je vidieť v Tab.5.13.

Tab. 5.13 Výsledky testov statickej implementácia KLAN.

	Názor aplikácie	Názor experta	Percentuálna úspešnosť
pozitívny	29	25	86,2%
negatívny	26	18	69,2%

Následne bol implementovaný dynamický prístup ku klasifikácii názorov, používajúci dynamický koeficient, ktorý nemusí zadávať používateľ, ale je automaticky určovaný na základe štruktúry viet spracovávaného textu. Tento dynamický koeficient môže byť stanovený jedným z uvedených spôsobov: Priemerná dĺžka vety, Polovica dĺžky vety alebo Hybridný prístup. Táto implementácia je ilustrovaná na Obr.5.5. Táto implementácia bola testovaná a výsledky testov sú uvedené v Tab.5.14.



Obr. 5.5 Statická implementácia slovníkovej metódy klasifikácie názorov KLAN – výsledok spracovania.

Tab. 5.14 Výsledky testov dynamickej implementácia Klasifikácie názorov.

Spôsob výpočtu koeficientu K	Presnosť klasifikácie
1. Priemerná dĺžka vety	0.8
2. Polovica dĺžky vety	0.84
3. Hybridný prístup	0.82

Táto implementácia zlyhala pri odhaľovaní skrytej irónie ako napríklad „Veď ešte aj môj starý Sony Ericsson robí lepšie fotky!“ a dvojzmyslov. Problémom bolo správne klasifikovať aj názor vyjadrený nepriamo, keď text obsahoval iba neutrálne slová, ako napríklad „Na tento mobil môžeš pokojne zabudnúť!“ alebo „Iné filmy ani nevyhľadávam.“ Ďalšie problémy znižujúce úspešnosť klasifikácie názorov sú prípady, keď slovo s kladnou resp. zápornou orientáciou nesie opačný postoj alebo je zápor posunutý do inej lexikálnej jednotky ako napríklad „Prišiel som si kúpiť fakt výkonný vysávač. Tento to určite nebude.“ Niekedy majú prídavné mená a príslovky opačnú orientáciu ako sa predpokladalo: „Tento výrobok je celkom slušný paškvil.“.

5.4.10 Použitie n - gramov v klasifikácii názorov

Pri použití dynamického koeficientu je cieľom nastaviť automaticky takú dĺžku tohto koeficientu aby nedošlo k izolácii negácie alebo intenzifikátora od vzťahovaného slova. Ale keďže dynamický koeficient rozdelí text do lexikálnych jednotiek, ktoré sa neprekrývajú, môže sa stať, že negácia alebo intenzifikátor padne do inej lexikálnej jednotky ako slovo, ku ktorému sa vzťahujú. Tento problém je možné riešiť použitím n-gramov. N-gramy riešia uvedený problém izolácie pomocou cyklického posuvu o

jedno slovo. Napríklad majme vetu:

„Naozaj je to pekné a na viac aj veľmi praktické.“

Pri použití 4-gramov, bude táto veta rozdelená do nasledovných lexikálnych jednotiek:

„naozaj je to pekné“ *P = 1 x (1+0,5) = 1,5*

„je to pekné a“ *P = 1*

„to pekné a na“ *P = 1*

„pekné a na viac“ *P = 1*

„a na viac aj“ *P = 0*

„na viac aj veľmi“ *P = 0 x 1 = 0*

„viac aj veľmi praktické.“ *P = 1 x (1+1) = 2*

Každá lexikálna jednotka je nasledovaná výpočtom hodnoty jej polarity.

Tento spôsob predspracovania textu bol implementovaný v slovníkovom prístupe ku klasifikácii názoru. Tento systém používal dva slovníky.

Prvý slovník obsahoval prídavné mená, podstatné mená, slovesá a emotikony. Tento slovník bol určený na riešenie základných problémov prostredníctvom skladania jednoduchých polarít. Tento proces je ilustrovaný prvou sumou vo vzťahu (1). Ukážka časti tohto slovníka je uvedená v Tab.5.15. Tab.5.16 obsahuje úplnú množinu používaných emotikonov.

Tab. 5.15 Ukážka prvého slovníka implementácie založenej na n-gramoch.

Stupeň polarity	Slová a emotikony
3	:D, boží, špičkový
2	:), super, vynikajúci
1	pekný, funkčný, praktický
-1	nepríjemný, slabý
-2	:(, otrasný, chatrný
-3	:((, mizerný, katastrofálny

Tab. 5.16 Úplná množina používaných emotikonov.

Pozitívny	Negatívny
:)	:(
:))	:((
:)))	:(((
:-)	:-(
=)	=(
:D	
=D	

Druhý slovník obsahoval negácie a intenzifikátory (príslovky). Tento slovník bol určený na spracovanie negácie a intenzifikáciu, teda posuvy polarity. Toto spracovanie je ilustrované druhou sumou vo vzťahu (15). Ukážka časti tohto slovníka je uvedená v Tab.5.17.

Tab. 5.17 Ukážka druhého slovníka implementácie založenej na n-gramoch.

Stupeň polarity	Intenzifikátory a negátory
1	veľmi, dokonale, výnimočne
0.5	vhodne, naozaj, fakticky
-0.5	málo, príliš, zbytočne
-2	negácie: nie,nie je, ne, nebol …

$$P = \sum v(w_i^1)[1+\sum v(w_j^2)] \quad (15)$$

Kde:

P … je stupeň polarity analyzovaného textu

$v(w_i^1)$… je hodnota (value) slova w_i v analyzovanom texte nájdené v prvej časti slovníka

$v(w_j^2)$… je hodnota (value) slova w_i v analyzovanom texte nájdené v druhej časti slovníka.

Tento vzťah bol inšpirovaný [Thelwall, M. et al., 2011]. Nasledujú príklady výpočtu polarity podľa vzťahu (1).

- Jednoduché polarity

„Ako samotná myška je pekná, ale spracovanie je mizerné a celkovo je nepodarená.“

pekná(+1) + mizerné(-3) + nepodarená(-1)

P = 1 + (-3) + (-1) = -3

- Negácia

„Nie je to dobré riešenie.“

Násobené: Nie(-2), pripočítané: dobré(+1)

*P = 1 * (1 + (-2)) = 1* (-1) = -1*

- Intenzifikácia

„Celkovo je spracovanie veľmi slušné.“

násobené: veľmi(+1), pripočítané: slušné(+1)

*P = 1 * (1 + 1) = 1 * 2 = 2*

Všetky modifikácie aplikácie klasifikácie názorov boli testované. Na testovanie boli použité nasledovné dáta. Pre aplikáciu so Statickým koeficientom boli použité príspevky z diskusie na stránke http://www.mobilmania.sk, konkrétne diskusné vlákno recenzií k mobilu LGKU990. Pre aplikáciu s Dynamickým koeficientom boli použité príspevky zo stránky http://recenzie.sme.sk. Aplikácia založená na N-gramoch bola testovaná v dvoch nezávislých experimentoch. Prvý na dátach z http://www.mojandroid.sk z diskusného vlákna k mobilom HTC One X a HCT One S

a na dátach z http://www.pocitace.sme.sk z diskusného vlákna k produktom Asus Transformer Prime TF201 and Asus Transformer Pad TF300T. Druhý experiment s aplikáciou založenou na n-gramoch bol vykonaný na dátach z http://tech.sme.sk obsahujúcich recenzie telefónu Samsung Galaxy S4 a na dátach z http://www.mojandroid.sk obsahujúcich recenzie telefónov HTC ONE a Samsung Galaxy S4. Výsledky týchto testov sú uvedené v Tab.5.18.

Tab.5.18 Výsledky testov nad modifikáciami implementácie založenej na statickom koeficiente, dynamickom koeficiente a na n-gramoch.

Version	Positive	Negative	Average precision
Static coefficient	0.86	0.69	0.78
Dynamic coefficient 1	0.76	0.84	0.80
Dynamic coefficient 2	0.80	0.88	0.84
Hybrid	0.80	0.84	0.82
N-grams 1	0.83	0.57	0.70
N-grams 2	0.76	0.42	0.59

Uvedená implikácia používa takzvanú „swich“ negáciu, ktorá spočíva v preklopení polarity na opačnú v tom istom stupni. V budúcnosti by mohla byť nahradená „shift“ negáciou podľa [Choi, Y. - Cardie, C., 2008] a [Taboada, M. et al., 2011].

POUŽITÁ LITERATÚRA

[Choi, Y. - Cardie, C., 2008] Choi, Y., Cardie, C.: Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. Proc. of the EMNLP 2008, Conference on Empirical Methods in Natural Language Processing, 793-801 (2008)

[Ding-Liu-YuA, 2008] Ding, X., Liu, B., YuA, P. *Holistic Lexicon-Based Approach to Opinion Mining*. Proc. of the Int. Conf. on Web Search and Web Data Mining WSDM'2008, New York, NY, USA, 2008, 231-240.

[Malone, 2012] Malone, T.W.: *Collective Intelligence: A Conversation with Thomas W. Malone*. [online]. [cit. 2014-06-28] Dostupné na internete: <http://edge.org/conversation/collective-intelligence>.

[Pang-Lee, 2008] Pang, B., Lee, L.: *Opinion Mining and Sentiment Analysis*. Foundation and Trends in Information Retrieval, *Vol.2, No.1-2*, 2008, 1-135.

[Strba-Bieliková, 2013] Srba, I., Bieliková, M.: *Adaptive Support for Educational Question Answering.* In: Proceedings of the Docto-ral Consortium at the European Conference on Technology Enhanced Learning 2013. Ed. Katherine Maillet & Tomaž Klobučar. Paphos, Cyprus: CEUR, (2013), pp.109–114.

[Surowiecki, 2004] Surowiecki, K.: *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wis-dom Shapes Business*, Economies, Societies and Nations, Doubleday, 2004.

[Taboada, M. et al., 2011] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, Vol. 37, No. 2, 267-307 (2011)

[Thelwall, M. et al., 2011].Thelwall, M., Buckley, K., Paltoglou, G, Cai, D., Kappas, A.: Sentiment Strength Detection in Short Informal Text. Journal of the American Society for Information Science and Technology, Vol. 61, No. 12, 2010, 2544-2558 (2010)

6 Extrakcia informácií z konverzačného obsahu

6.1 Úvod

Konverzačný obsah je to, čo sa vytvára a hromadí v procese využívania sociálneho webu. Sociálny web nielen umožňuje ale aj posilňuje interakcie medzi používateľmi rozličných platforiem tohto internetového média. Tieto interakcie medzi používateľmi sú spojené s ich vzájomným ovplyvňovaním sa v reálnych situáciách, ako je napríklad kúpa drahého produktu, voľba politickej reprezentácie a pod.

Tieto rozhodovacie procesy môžu byť podporované aplikáciami dolovania názorov z konverzačného obsahu. Toto dolovanie môže poskytnúť informácie rôzneho druhu:

- nielen informácie o drahých veciach - môže ísť o nejakú nehnuteľnosť, dovolenkovú destináciu, auto, jachtu a pod.
- kultúrne informácie
- informácie spojené s bezpečnostnými aspektmi, napríklad informácie o webových sídlach, kde sa umiestňuje podozrivý obsah (rasistický, pedofílny a pod.) alebo informácie o autoroch týchto príspevkov.

Čo je to vlastne konverzačný obsah? Konverzačný obsah môžeme vymedziť ako krátke texty, ktoré predstavujú niečo medzi hovoreným písaním a písaným hovorením. Tieto krátke texty sú syntakticky odlišné od klasických dokumentov vyskytujúcich sa na webe, ako sú vedecké články, správy v internetových novinách a pod. Spomenutá syntaktická odlišnosť sa prejavuje hlavne frekvenciou typických slov, interpunkciou, slovosledom, preklepmi často úmyselnými, ktoré odrážajú autorovu osobnosť. Keď už definujeme na tomto mieste konverzačný obsah, je potrebné pripomenúť, kvôli úplnosti fakt už uvedený v tejto publikácii a to, že konverzačný obsah vzniká v rámci rozličných platforiem sociálneho webu, ako napríklad v rámci sociálnych sietí, blogov, microblogov, chatov, chatrooms, twitter, IRC (Internet Relay Chat), diskusných fór, komentárov k článkom, videám a pod.

Teda máme krátke texty k daným témam, ktoré používateľov sociálneho webu práve zaujímajú, pričom téma môže byť v princípe vopred známa ale aj neznáma.

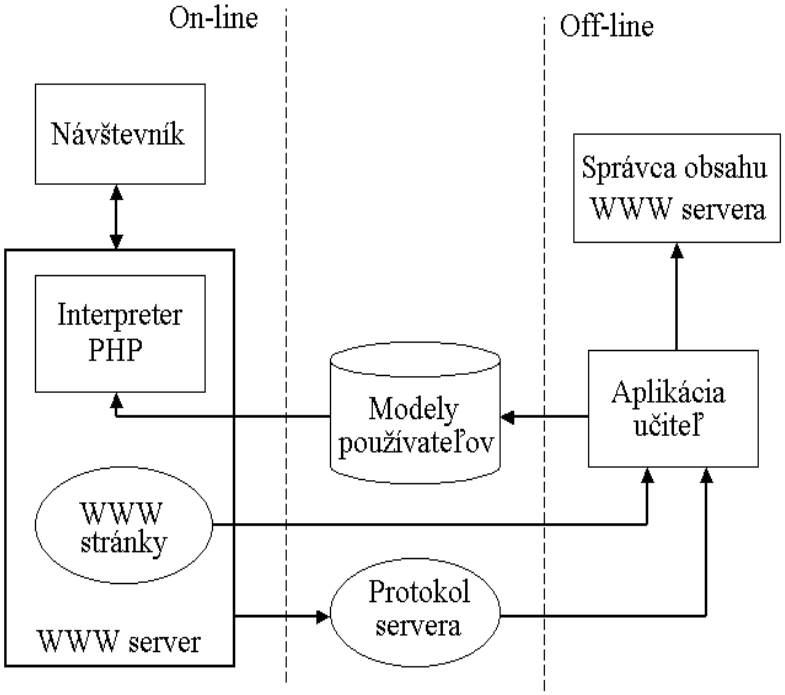
- **Známa téma.** V prípade známej témy, sa pozornosť sústreďuje na hodnotenia objektu, ktorý predstavuje tému diskusie. Môže ísť o drahý produkt, destináciu dovolenky, hotel a jeho služby, politickú situáciu, osobu, lekára, politika, film, knihu, pocity autora, teda čokoľvek, o čom majú používatelia potrebu hovoriť.
- **Neznáma téma.** V prípade neznámej témy je potrebné tému modelovať. Modelovanie témy predstavuje proces spracovania textu, anotovania textu alebo hľadania kľúčových charakteristických slov textu za účelom zistenia témy diskusie. Dôležité je to hlavne vtedy, ak je potrebné zistiť ktoré diskusie nesú pečať podozrivého obsahu.

6.2 Dolovanie v konverzačnom obsahu

Rozpoznávame tri druhy dolovania v dátach (data mining) [Paralič et al., 2010] a to dolovanie z používania, dolovanie zo štruktúry a dolovanie z obsahu.

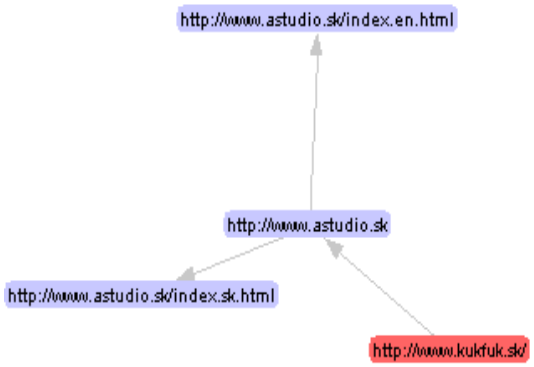
Pri **dolovaní z používania** sa vlastne doluje z log súborov, ktoré obsahujú záznamy o tom, ktoré webové stránky používateľ navštívil (na ktoré linky klikol). Tento druh dolovania vedie k personalizácii webu a navigácii používateľa v bludisku webových

stránok. Použitím metód strojového učenia je možné naučiť model používateľa, ktorý sa následne použije na odporučenie personalizovaného zoznamu nových stránok, o ktorých používateľ nevedel a pri tom môžu byť pre neho zaujímavé. Toto učenie modelov používateľov prebieha v rámci Off-line módu v Aplikácii učiteľ, čo ilustruje Obr.6.1. Učenie prebieha na základe záznamov z protokolu www servera a informácií o www stránkach, ako je to naznačené v dolnej časti obrázka Obr.6.1. On-line mód predpokladá, že už sú naučené modely používateľov a stránok a v rámci tohto módu sa uskutočňuje samotné personalizované odporúčanie.

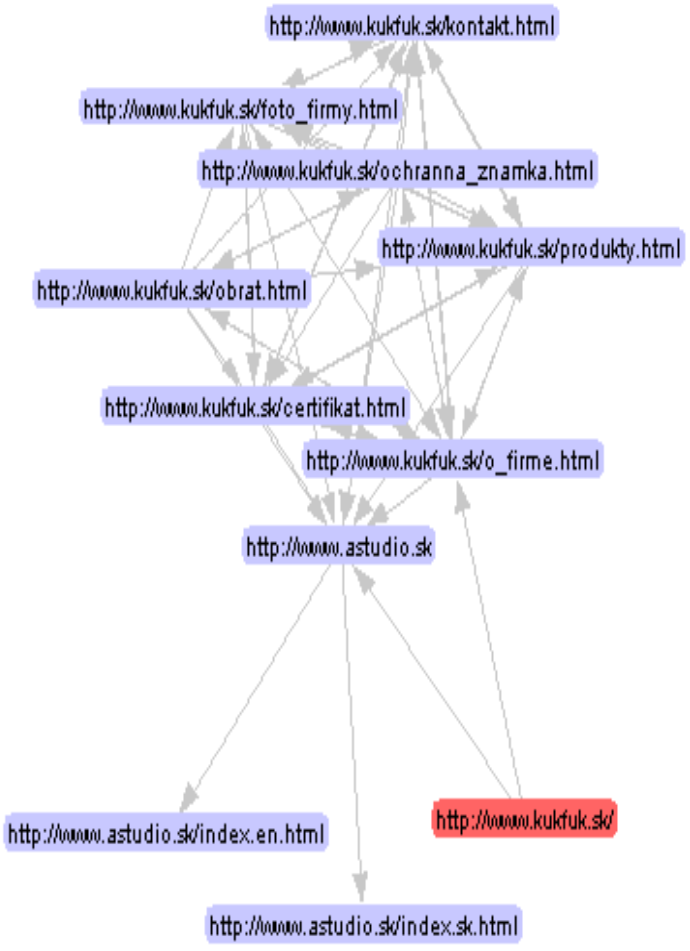


Obr. 6.1 Architektúra systému založeného na dolovaní z používania.

Dolovanie zo štruktúry webu ako takého môže poskytnúť mapu okolia aktuálnej (práve navštívenej) web stránky, inak povedané informácie nápomocné pri navigácii používateľa webu. Takéto dolovanie vedie k parciálnemu mapovaniu okolia aktuálnej web stránky pomocou matice susednosti a matice najkratších vzdialeností. Pritom sa rozlišujú rozličné úrovne vnorenia. Obr.6.2 reprezentuje partikulárnu mapu pri uvažovaní dvoch úrovní vnorenia a Obr.6.3 pri uvažovaní troch úrovní vnorenia. Pri veľkom počte uvažovaných úrovní vnorenia sa aj partikulárna mapa webu stáva neprehľadnou.



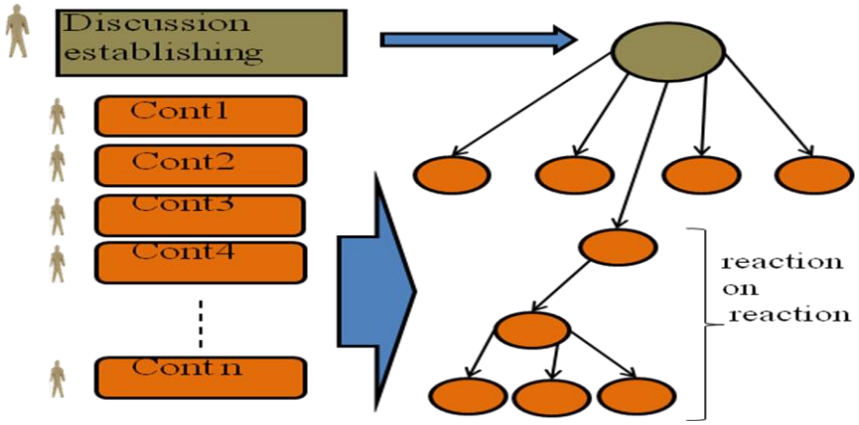
Obr. 6.2 Partikulárna mapa webu pri uvažovaní dvoch úrovní vnorenia.



Obr. 6.3 Partikulárna mapa webu pri uvažovaní dvoch úrovní vnorenia.

Keď sa sústredíme na dolovanie zo štruktúry konverzácie vzniknutej v rámci niektorej platformy sociálneho webu, môžeme získať informácie použiteľné pri

riešení problému identifikácie autorít. Dolovaním štruktúry webovej diskusie, ktorá sa dá reprezentovať pomocou acyklického grafu – stromu, môžeme získať také informácie ako: Počet príspevkov daného prispievateľa, Počet reakcií na jeho príspevky, Počet výskytov na spodnej úrovni (reprezentuje prípad, keď používateľ uzavrel diskusiu) a pod. Webová diskusia reprezentovaná stromom je ilustrovaná na Obr.6.4.



Obr. 6.4 Webová diskusia reprezentovaná stromom.

V rámci dolovania obsahu webu sa sústredíme na **dolovanie z obsahu** konverzácie a budeme sa usilovať o dolovanie, resp. klasifikovanie názorov tak pozitívnych ako aj negatívnych. Je možné taktiež získať informácie umožňujúce analýzu sentimentu (hnev, radosť, znechutenie, nadšenie,...) . Dolovanie z obsahu konverzácie umožňuje riešiť aj iné problémy, ako napríklad:

- identifikácia autorít (Kto je autoritou v tejto diskusii?)
- analýza názorov (Pozitívny, negatívny?)
- vyhľadávanie názorového spamu (Je obsah príspevku informatívny? Alebo sú to vlastne „pokecy“?)
- určovanie užitočnosti názorov (Je tento názor kvalitný, overený?)
- aspektovo orientovaná analýza sentimentu (Aká je názorová polarita v rámci jednotlivých vlastností entity?)
- porovnávacia analýza sentimentu (Ktorý z týchto produktov je lacnejší, komfortnejší, poruchovejší?)
- cielená reklama (Čo má obsahovať, lebo to ľudia oceňujú?)
- detekcia emócií (Čo vyjadruje príspevok: nadšenie, znechutenie?)
- modelovanie témy (O čom sa diskutuje?)
- vyhľadávanie názoru (Kde sa o tom diskutuje?)
- identifikácia autorstva (Kto je autorom príspevku? Aký typ človeka je prispievateľ?).

6.3 Extrakcia dát z webových zdrojov

Extrakcia dát predstavuje proces vyberania špecifikovaných dát z väčšieho množstva. Dáta s ktorými sa pracuje môžu byť zle štruktúrované alebo neštruktúrované. Extrakcia dát je kľúčovým krokom, cieľom ktorého je získať exaktné dáta s čo najmenším šumom. Táto práca sa sústreďuje na získavanie relevantných dát z webový zdrojov, ako napríklad adresa, telefóne číslo, e-mail, cena, produkty, vlastnosti produktu, text príspevku, atď. Podľa typu dát rozpoznávame extrakciu dvoch druhov:

- Extrahovanie vzorov, ako je e-mailová adresa, hypertextová adresa, dátum a pod.
- Dolovanie dát spojené s analýzou stromovej štruktúry HTML dokumentu.

Existujú rôzne prístupy k extrakcii dát z webu:

- **manuálne metódy** - Vo veľkej miere zapájajú do procesu extrakcie človeka a vyznačujú sa vysokou presnosťou a sú časovo náročné.
- **kontrolované učenie** - Zapája metódy strojového učenia, vyžaduje interakciu s človekom pre označkovanie dokumentov, teda získanie pozitívneho alebo negatívneho hodnotenia textu príspevku extrahovaného systémom, aby sa z neho bolo možné strojovo učiť.
- **automatické techniky** - Vyžadujú menej ľudského úsilia ale výsledkom nemusia byť spoľahlivé dáta.

Proces extrakcie dát z webových zdrojov musí zvládnuť nasledovné problémy:

- kvalita spracovania webových zdrojov, ktoré nedodržiavajú štandardy W3C pre webové služby,
- rýchle zmeny webových zdrojov,
- heterogénnosť a komplexnosť dátových typov,
- nekonzistentná sémantika a nekonzistentná štruktúra objektov.

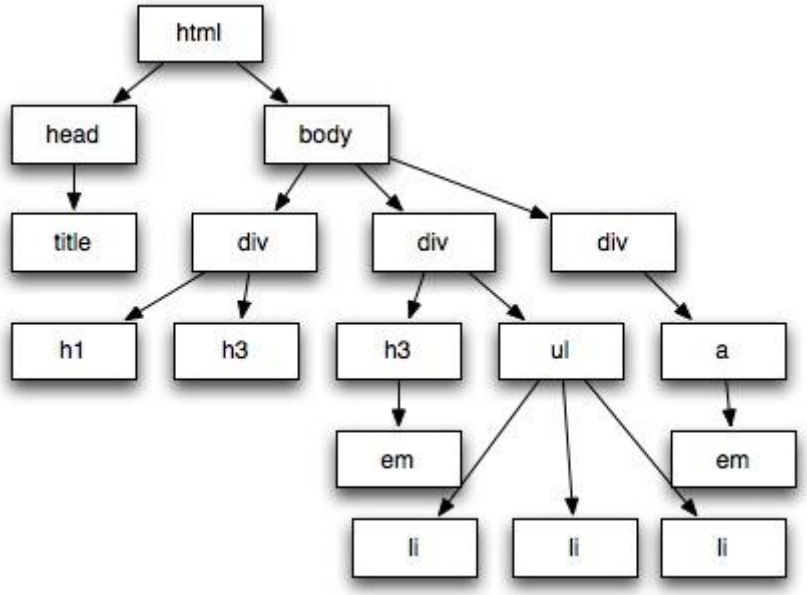
Štruktúra prezentovaných webových stránok je definovaná značkovacím jazykom HTML (XHTML). Úlohou značkovacích jazykov je vytvárať štruktúrované dáta, teda reprezentovať štruktúru webových zdrojov. V súčasnosti je väčšina webových zdrojov generovaná dynamicky. Bližšie špecifikácie o HTML (XHTML) je možné nájsť na stránkach W3C.

Ak chceme na webovej stránke rozpoznať text príspevkov, musíme vedieť rozpoznať niektoré dôležité značky. Značky môžu byť párne a nepárne. Párne majú začiatočný a koncový znak, napr. <*body*> …. <*/body*>. U nepárnych neexistuje koncový znak (ďalej už len tag). Všetky webové dokumenty písané v jazyku HTML obsahujú nasledovnú štruktúru:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
   "http://www.w3.org/TR/html4/strict.dtd">
<HTML>
   <HEAD>
      <TITLE>My first HTML document</TITLE>
   </HEAD>
   <BODY>
      <P>Hello world!
   </BODY>
</HTML>
```

Príklad ilustruje ako môže vyzerať jednoduchý HTML kód. Povinnou súčasťou stránky je tag <HTML></HTML>, tento označuje začiatok a koniec webovej stránky. Extrakcia dát z webových stránok sa uplatňuje nad blokom kódu HTML obsiahnutého medzi tagmi *<body> a </body>.* V tomto bloku kódu sa nachádza celé jadro webovej stránky. Tagy neobsahujú žiadnu explicitnú informáciu o formátovaných dátach. Z toho vyplýva problém ako exaktne získavať relevantné dáta z webových stránok. Štruktúra HTML dokumentu je koncipovaná za účelom zobrazenia vo webovom prehliadači, ktoré bude ľahko a intuitívne pochopiteľné človeku. To ako extrahovať dáta do veľkej miery závisí od charakteru požadovaných dát. Dáta takého druhu ako e-mail, hypertextové odkazy, obrázky a pod. sa dajú jednoducho získať pomocou regulárnych výrazov alebo prostredníctvom jednoduchých JavaScript-ov.

Štruktúru HTML je možné reprezentovať graficky pomocou stromu, čo ilustruje Obr.6.5.



Obr. 6.5 Reprezentácia štruktúry HTML dokumentu pomocou stromu.

Pri takomto pohľade na štruktúru HTML je možné vytvoriť metódy prehľadávajúce strom HTML za účelom automatickej extrakcie dát (textov) z webových zdrojov na určitej vhodnej úrovni vnorenia v strome reprezentujúcej štruktúru HTML dokumentu. Existuje viacero prístupov k tejto extrakcii. V tejto publikácii sa sústredíme na metódu založenú na čiastočnom vyrovnávaní stromu.

6.3.1 Čiastočné vyrovnávanie stromu

Čiastočné vyrovnávanie stromu je metóda vychádzajúca z predpokladu, že webová stránka obsahuje viacero štruktúrovaných dátových záznamov tvoriacich štruktúrované dáta o objektoch. Tieto dáta sú uložené v databázach a na stránke sú zobrazované prostredníctvom pevnej šablóny. Predmetom je segmentovať tieto dátové záznamy a extrahovať dátové položky [Yanhong-Bing, 2010].

Obr.6.6 ilustruje segment webovej stránky obsahujúcej zoznam dvoch produktov. Ide o notebooky. Popis každého produktu predstavuje dátový záznam. V Tab.6.1 je

ilustrovaný segment stránky tak, že každý záznam je reprezentovaný jedným riadkom. Je totiž dôležité reprezentovať údaje zo stránky jednotným spôsobom.

Obr. 6.6 Segment webovej stránky obsahujúcej informácie o dvoch notebookoch.

Tab. 6.1 Tabuľková reprezentácia záznamov segmentu webovej stránky

Years	Persons	(%)	Persons	(%)	Persons	(%)
40-49	51 000	0.1%	80 000	0.2%	131 000	0.3%
50-59	45 000	0.1%	102 000	0.3%	147 000	0.4%
60-69	59 000	0.3%	176 000	0.9%	235 000	1.2%
70-79	134 000	0.8%	471 000	3.0%	605 000	3.8%
>80	648 000	7.0%	1 532 000	16,7%	2 180 000	23,7%

Extrakcia záznamov pozostáva z dvoch krokov:

1. automatické identifikovanie individuálnych dátových záznamov na stránke a
2. automatické zarovnanie a extrahovanie dátových položiek z identifikovaných dátových záznamov.

6.3.2 Extrahovanie dátového záznamu

V prvom kroku sa segmentuje webová stránka a identifikujú sa jednotlivé záznamy. V tomto kroku sa nezarovnávajú ani neextrahujú dátové položky v dátových záznamoch. To sa vykoná až v ďalšom kroku.

Na extrahovanie dátových záznamov bol použitý algoritmus, ktorý je modifikáciou algoritmu MDR (Mining Data Records). Základná myšlienka MDR algoritmu je

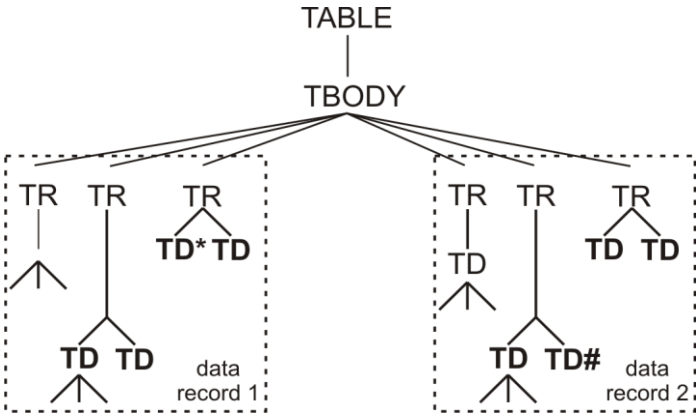
založená na dvoch veciach:

1. na dátových záznamoch na webových stránkach a
2. na editovaní založenom na meraní vzdialenosti zhodných reťazcov znakov.

Skupina dátových záznamov obsahuje popis množiny podobných objektov, väčšinou prezentovaných v spojitej oblasti stránky, ktoré sú formátované podobnými HTML značkami. Takúto oblasť nazývame *dátová oblasť*, ktorá je ilustrovaná na Obr.6.7. Dátové záznamy sú formátované pomocou podobnej postupnosti HTML tagov. Ak by sme reprezentovali HTML značky stránky dlhým reťazcom, bolo by možné použiť zhodu reťazcov na porovnávanie odlišných pod-reťazcov za účelom nájdenia týchto podobných jednotiek, ktoré môžu reprezentovať podobné dátové záznamy. S týmto prístupom je spojený problém, pretože dátový záznam môže začať od každého tag-u a skončiť na každom nasledujúcom tag-u. Množiny dátových záznamov zvyčajne nemajú rovnakú dĺžku vo vzťahu k ich tag-om, pretože nemusia obsahovať presné rovnaké kúsky informácií [Yanhong-Bing, 2010].

Vnorená štruktúra HTML značiek vo webovej stránke prirodzene formuje strom značiek. Na Obr.6.7 je zobrazený príklad reprezentácie HTML pomocou stromovej štruktúry. V tomto strome je každý dátový záznam zabalený v troch TR uzloch s ich pod-stromami pod rovnakým rodičovským uzlom TBODY. Tieto dva dátové záznamy sú obsiahnuté v dvoch čiarkovaných boxoch.

Ďalším užitočným poznatkom je to, že množiny podobných dátových záznamov sú formátované pomocou rovnakého stromu reprezentujúceho potomkov a majú spoločný rodičovský uzol [Yanhong-Bing, 2010].



Obr. 6.7 Príklad stromovej reprezentácie segmentov stránky

Je nepravdepodobné aby dátový záznam začal uprostred stromu potomka a končil uprostred stromu iného potomka. Miesto toho začína na začiatku pod-stromu potomka a končí na konci rovnakého potomka alebo na pod-strome posledného potomka. Napríklad, je veľmi nepravdepodobné, aby dátový záznam začal z TD* a končil v TD* (Viď Obr.6.7). Tento poznatok umožnil vytvoriť veľmi výkonný algoritmus založený na editácii vzdialenosti porovnávaných reťazcov k identifikácii dátových záznamov, pretože tieto sú limitované tagmi, od ktorých dátový záznam začína a končí v hierarchickom strome tagov.

Experimentálne sa podarilo overiť platnosť vyššie spomenutých faktov. V žiadnom prípade nepredpokladáme, že webové stránky obsahujú iba jednu dátovú oblasť obsahujúcu dátové záznamy. Rôzne regióny môžu obsahovať odlišné dátové záznamy. Modifikovaný MDR algoritmus spracúva webové stránky v troch krokoch:

1. Vytvorenie HTML hierarchického stromu tagov stránky.
2. Dolovanie dátových regiónov. K dolovaniu dátových regiónov sa používa stromová hierarchia stránky. Dátová oblasť je oblasť stránky obsahujúca zoznam podobných dátových záznamov. Namiesto dolovania dátových záznamov priamo, čo je náročné, MDR najprv doluje dátové regióny a až následne sa v nich pokúša identifikovať dátové záznamy. Napr. na Obr.6.7 najprv nájdeme jedinú dátovú oblasť pod uzlom TBODY.
3. Identifikovanie dátových záznamov z každej dátovej oblasti. Napr. na Obr.6.7 v rámci tohto kroku nájdeme dátový záznam 1 a dátový záznam 2 v dátovej oblasti pod uzlom TBODY [Yanhong-Bing, 2010].

6.3.3 Extrahovanie konverzačného obsahu

Cieľovými dátami sú texty diskusných príspevkov vo webových diskusiách. Zdrojom webových diskusií môže byť webové fórum, webová stránka umožňujúca pridávanie svojich názorov vo forme komentárov, príspevkov. Diskusné príspevky môžu byť obsiahnuté v rôznych častiach webovej stránky a taktiež ich reprezentácia v HTML kóde môže byť rôzna. Situovanie webových príspevkov v rámci webovej stránky je možné rozdeliť na dva typy. Na webové diskusné fóra a na webové stránky kde nosnú časť tvorí informačný text, po ktorom nasleduje diskusia.

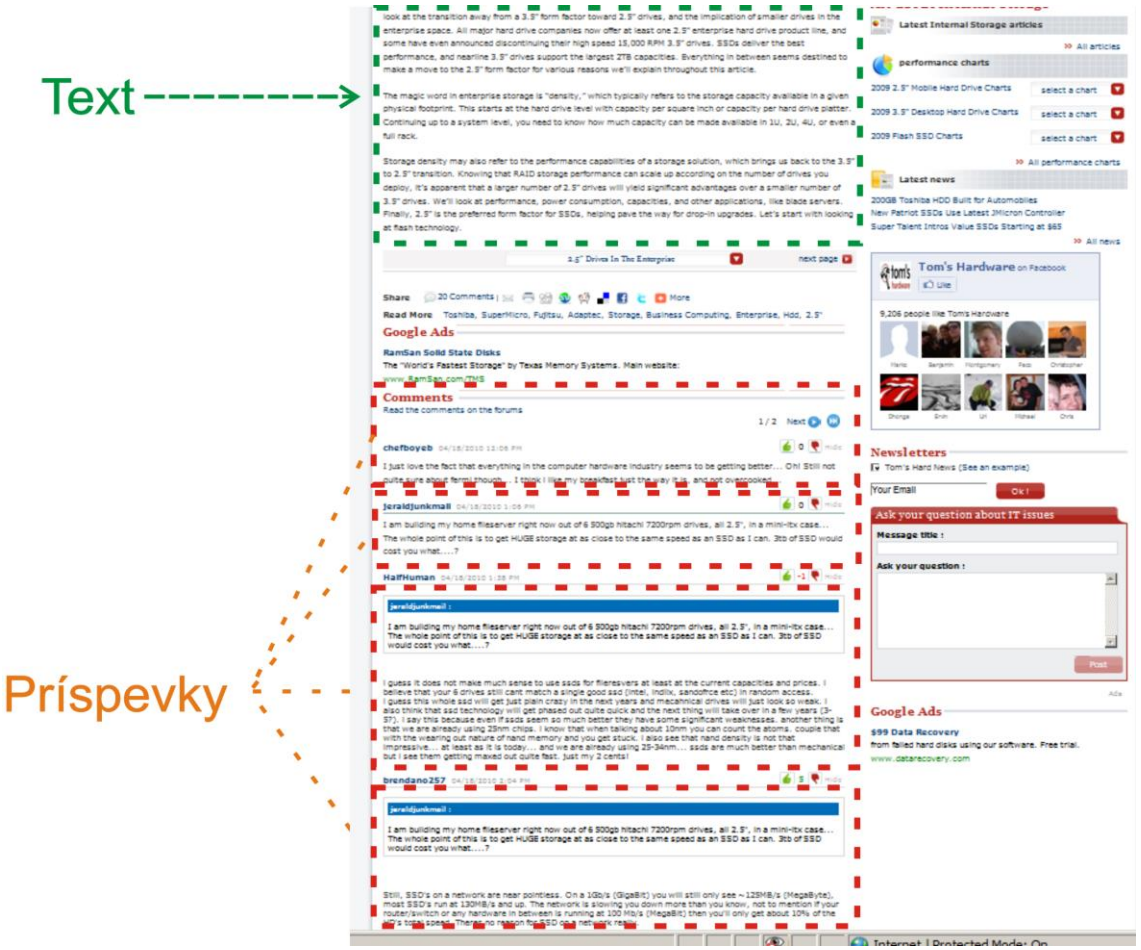
U **webových diskusných** fór jednotlivé príspevky zaberajú celú webovú stránku. Taká stránka väčšinou neobsahuje hypertextové menu. Štruktúra takýchto webových fór je často tvorená pomocou HTML tagov, akými sú TABLE, alebo DIV. Pri riešení prostredníctvom tabuľky, riadky tvoria jednotlivé príspevky. Tieto riadky obsahujú rôzne údaje o používateľoch fóra. Tieto údaje je potrebne identifikovať a odstrániť, lebo z pohľadu analýzy sentimentu sa jedná o šum. Medzi takéto doplnkové údaje patrí aj dátum pridania príspevku, ktorý na druhej strane môže uľahčiť proces extrahovania informácií. Extrahovanie dátumu je jednoduché, nakoľko sa jedná o vzor ľahko opísateľný regulárnymi výrazmi. Dátum je možné použiť ako pomocný parameter pri extrakcii príspevkov. Problémom zostáva formát dátumu alebo prípadná absencia tejto informácie. Príkladom takejto webovej stránky je napr. *http://www.politicalforum.com/*, viď Obr.6.8.

Ak máme spracovať webové stránky kde nosnú časť tvorí **informačný text doplnený diskusiou** (recenzia produkty, správy, atď.), extrakcia textu príspevkov je odlišná. Tento prípad je ilustrovaný na Obr.6.9. Medzi takéto stránky patria stránky predajcov produktov, informačné kanály (CNN), atď. Na týchto stránkach je potrebné odlíšiť popis oblasti webovej diskusie od popisu oblasti informačného textu.

Obr. 6.8 Príklad webového diskusného fóra

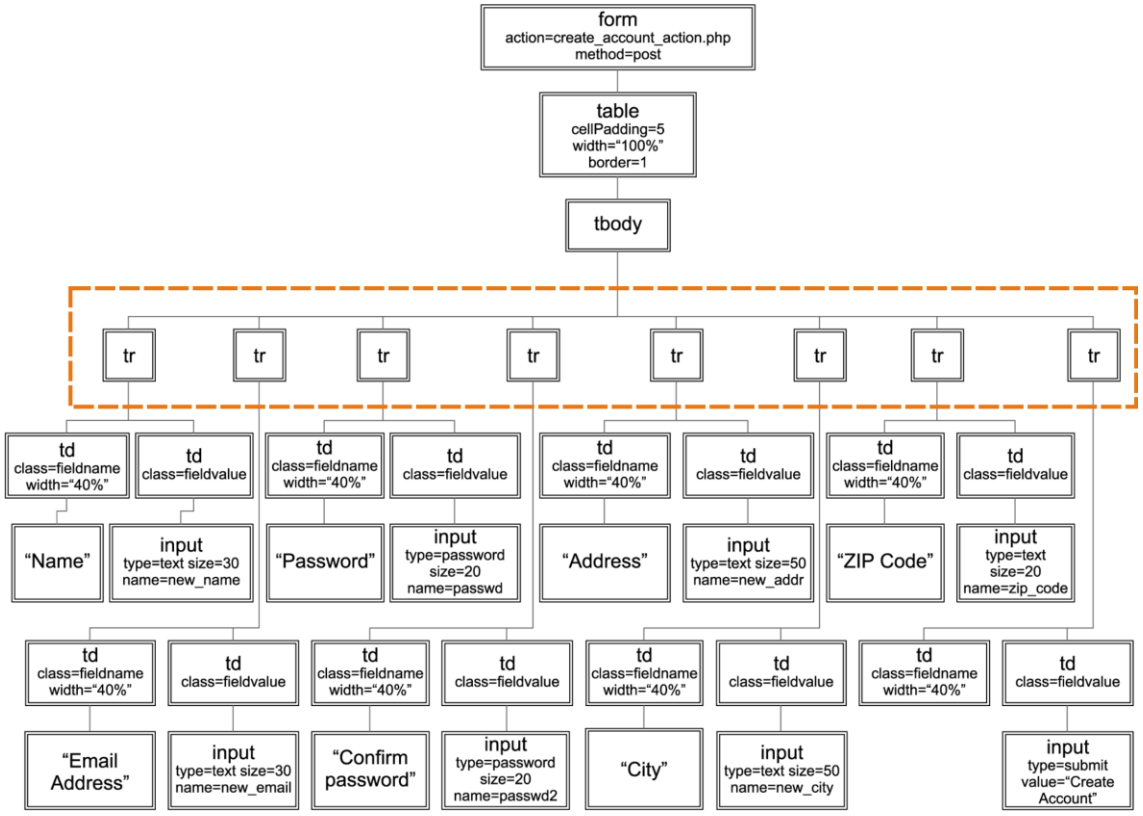
Informačný text býva tvorený postupnosťou textových blokov (….), kde postupnosť týchto blokov predstavuje súvislú oblasť, ktorej je potrebné sa vyhnúť. Tento typ stránok obsahuje väčšie množstvo šumu. Šum predstavujú reklamy začlenené do textu, alebo menu tvorené hypertextovými odkazmi, a ďalšie bloky u ktorých sa vyskytuje súvislá oblasť rovnakých tagov. Je potrebné vhodne popísať oblasti obsahujúce diskusné príspevky pre potreby zefektívnenia procesu extrakcie. Inak povedané, je potrebné vyhnúť sa oblastiam neobsahujúcich diskusné príspevky. Tým, že sa používatelia môžu podieľať na tvorbe obsahu webových stránok, je im dovolené zasahovať do štruktúry HTML kódu. Kľúčovým momentom v kroku extrakcie je vhodné popísanie oblasti obsahujúcej diskusné príspevky, ktoré budú prehľadávané na úrovní HTML kódu.

Na extrakciu dát z webových stránok je potrebné nejakým spôsobom reprezentovať štruktúru web stránok, za účelom ich prehľadávania. Pre tento účel je na prehľadávanie možné použiť tzv. DOM (Document Object Model) HTML dokumentu, pre viac podrobností viď web stránky W3C. Prostredníctvom DOM je možné prechádzať štruktúru HTML ako strom a postupne sa vnárať do nižšej vrstvy HTML kódu. Najbežnejším spôsobom ako identifikovať dátové oblasti v HTML je nájdenie súvislej oblasti opakujúcich sa HTML tagov (viď Obr.6.10).



Obr. 6.9 Informačné webové stránky s následnou diskusiou.

Otázkou ostáva do akej hĺbky je potrebné sa vnoriť aby sme získali texty príspevkov. Pri výbere vhodného uzla na expandovanie je potrebné sa vyhnúť blokom HTML kódu, tvorených súvislou opakujúcou sa oblasťou kódu, ktorá neobsahuje cieľové dáta, napr. menu hypertextových odkazov. V úvode bolo spomenuté, že cieľom je vytvoriť univerzálny prístup k extrakcii príspevkov z webových stránok. Na Obr.6.9 je ilustrovaná cieľová oblasť vo forme červeno orámovaných blokov textu. Cieľová oblasť obsahuje texty príspevkov ako aj šumy (meno užívateľa, dátum, hypertextové odkazy, atď.). Informačný text, na ktorý príspevky reagujú je na Obr.6.9 orámovaný zelenou farbou. Na získanie textu príspevku je potrebné extrahovanie na úrovni cieľovej oblasti. Univerzálny spôsob pre čistenie bloku obsahujúceho príspevok, môže byť odstránenie duplicitných textov obsiahnutých v každom bloku.



Obr. 6.10 Príklad súvislej oblasti v HTML kóde.

POUŽITÁ LITERATÚRA

[Paralič, et al., 2010] Paralic, J. et al.: Dolovanie znalostí z textov 1. vydanie. Košice : Equilibria, 2010, 182 s. ISBN 978-80-89284-62-7

[Pénzeš, 2010] Pénzeš, T.: Karma a sentiment vo webových diskusiách. Technická univerzita Košice, 2010. 62 s.

[Yanhong-Bing, 2010] Yanhong, Z., Bing, L.: Web Data Extraction Based on Partial Tree Alignment. ACM New York, USA, 2005, 76-85, ISBN:1-59593-046-9.

7 Identifikácia autorít

7.1 Úvod

Pojem analýza sociálnych sietí zahŕňa viacero rozličných problémov a rozličné metódy ich riešenia. Väčšina metód sa zaoberá vzormi v priamych a nepriamych väzbách medzi aktérmi so zameraním na vyšetrovanie centralít aktérov, teda autorít ako aj na odhalenie schopnosti aktérov združovať sa do kohéznych zoskupení a ovplyvňovať sa medzi sebou [Scott, 2000]. Analýzou zaujímavých pozícií v rámci siete sa zaoberá pozičná analýza a analýza rolí. V poslednom čase sa orientujú metódy pre analýzu rolí hlavne na on-line dáta, ktoré sú stále viac dostupné. Metódy pre identifikáciu rolí v on-line dátach by mohli byť užitočné tak pre používateľov ako pre organizátorov on-line fór v procese budovania reputačných systémov na identifikáciu užitočných alebo škodlivých používateľov, odhaľovanie podvodov ale aj takzvaných „príživníkov“ v sieti. Taktiež sa môžu použiť pri odmeňovaní a podporovaní cenných prispievateľov alebo pri vytváraní lepších algoritmov pre zber dát v sieti. Najcennejšia je pomoc týchto metód pri rozpoznávaní sociálnych rolí, hlavne autorít v sieti [Welser, 2007].

Sociálne role a sociálne pozície sa klasicky skúmali v sociológii. Idea identifikácie sociálnej role je založená na predpoklade, že akákoľvek rola je kombináciou súborov behaviorálnych a štrukturálnych atribútov. Napríklad sociálna rola otca zahŕňa schému toho, čo otec je, a spája sa s vhodnými interakciami, očakávaným chovaním sa a zastávaním osobitných sociálnych štrukturálnych vzťahov. Je možné ilustrovať to na príklade skúmania interakcií na detskom ihrisku s cieľom identifikácie role otca. Napríklad, ak dospelý muž utešuje malé dieťa objímajúc ho, mohli by sme tieto behaviorálne a relačné dáta považovať za dôkaz toho, že muž je otcom dieťaťa. Čím rozsiahlejšie sú dáta a čím obsiahlejšie sú naše informácie o vzájomnom pôsobení a štrukturálnych vzťahoch, tým lepšie sme schopní identifikovať role jednotlivých aktérov. Údaje získané on-line sú pre štúdium rolí ideálne, pretože poskytujú k analýze tak informácie o štruktúre siete, vzory správania ako aj sémantiku interakcií sprostredkovanú analýzou obsahu interakcií. To všetko vedie k presnejšej identifikácii rolí. Rozvoj metód, ktoré presne rozlišujú sociálnu rolu bez nutnosti rozsiahlej analýzy obsahu, má veľký potenciál byť prínosom pre organizácie, ktoré potrebujú udržiavať konzistenciu on-line siete ako aj pre rozvoj efektívnych stratégií vyhľadávania.

Teoretické definície role sa takmer vždy zakladajú na vlastnostiach (atribútoch) aktérov alebo skupín aktérov. Použite takejto definície je v tvare: „skúmaný aktér je v roli autority v skúmanej skupine aktérov“. V práci [Homans, 1961] je rola definovaná ako „*chovanie očakávané od aktéra, ktorý zastáva určitú sociálnu pozíciu*“. Na rozdiel od sociálnej pozície, ktorá vymedzuje kolekciu aktérov, koncept sociálnej role označuje spôsob ako sa aktéri v určitých sociálnych pozíciách vzťahujú k aktérom v ostatných pozíciách. Preto je potrebné brať do úvahy jednotlivých aktérov, resp. kolekciu aktérov ako sociálnu pozíciu. Potom relácie medzi aktérmi, resp. kolekciami aktérov predstavujú sociálne role. Rozdiely medzi týmito pojmami boli definované v [Goodenough, 1969]. Boli definované nasledovné pojmy:

- postavenie (status),
- pozícia (position),
- rola (role).

Sieťová analýza rolí môže byť aplikovaná v troch úrovniach siete a to buď na celú

skupinu aktérov, na nejakú konkrétnu podskupinu aktérov alebo na jedného aktéra.

Podľa tejto špecifikácie rozdeľujeme role na:

- globálne role,
- lokálne role,
- individuálne role alebo tzv. ego role [2].

V ďalšom sa budeme zameriavať na identifikáciu rolí so zameraním na ego role, konkrétne ide o rolu autorita. Budeme sa zaoberať identifikáciou autorít v rámci danej webovej diskusie. Tento problém je možné riešiť primárne dolovaním štruktúry webovej diskusie ale okrajovo je možné použiť aj dolovanie konverzačného obsahu. V rámci tejto kapitoly sa taktiež budeme venovať problému identifikácie autorít vo vedeckých komunitách so spoločnými profesijnými záujmami. Takto ponímaná identifikácia autorít vedie ku dolovaniu obsahu digitálnych knižníc vedeckých publikácii.

Autorita býva spravidla overená v reálnych situáciách, čo môžu byť aj rozličné webové diskusie. Vo všeobecnosti rozpoznávame dva typy autorít a to formálnu a neformálnu.

- **Formálna autorita** vyplýva z nejakej pozície v organizácii, z dosiahnutého titulu, funkcie v organizácii. Tento status podlieha zmene ak skúmaná osoba bola povýšená alebo degradovaná, respektíve zmenila pôsobisko, filiálku, detašované pracovisko alebo sa zmenili jeho vlastnícke vzťahy v organizácii. Formálne autorita môže vyžadovať poslušnosť, podriadenosť alebo až submisívny postoj od svojich podriadených, čo môže viesť k nechuti podriadených nechať sa riadiť takouto autoritou.
- **Neformálna autorita** je prirodzenou autoritou, založenou na schopnostiach, primeranom sebavedomí, osobnom profile a sociálnych aktivitách skúmaného človeka. Prirodzená autorita je posilňovaná rešpektom vedených ľudí. Je založená na takých vlastnostiach ako: čestnosť, statočnosť, rozhodnosť, schopnosť predvídať vývoj situácie, schopnosť odhadnúť ľudí – spolupracovníkov ako aj konkurentov a pod.

Formálna a prirodzená autorita môžu byť totožné a stelesnené v jednom konkrétnom človeku, čo je ten najšťastnejší prípad hlavne pre ľudí v okolí takéhoto človeka. Formálna autorita sa môže meniť na neformálnu - prirodzenú a naopak. V takom prípade je vhodné sledovať dynamickú zmenu autority, ktorá podlieha zmene statusu. My sa budeme zaujímať o identifikáciu neformálnych – prirodzených autorít.

Autority môžeme z iného pohľadu deliť na:

- **Priateľov,** ktorých veľké množstvo aktérov v rámci sociálneho webu označilo za priateľa. Teda ide o autoritu podporovanú vzťahmi.
- **Šíriteľov vplyvu (influencerov)**, ktorí sú často citovaní, pričom iní aktéri sa odvolávajú na jeho autoritu. Často iných aktérov prekvapí ba dokonca až ohromí. Teda ide o autoritu podporovanú názormi a vedomosťami o objekte diskusie.

My sa budeme zaujímať o identifikáciu autority typu šíriteľ vplyvu. Teda v sumáre, uvedieme dva prístupy k identifikácii neformálnych (prirodzených) autorít typu šíriteľ vplyvu v rámci dvoch podstatne rozličných oblastí a to v definovanej vedeckej oblasti a vo webových diskusiách.

- *Autority vo vede* sa môžu identifikovať v rámci vedeckých článkov na osobných stránkach, v profiloch vedcov, vo webových stránkach rôznych vedeckých inštitúcií, v digitálnych knižniciach a pod.
- *Autority vo webových diskusiách* k rozličným témam z oblasti spoločenského života ale aj z oblasti kultúry ako aj k vlastnostiam rozličných produktov. Zameriame sa na dáta hromadené v konverzačnom obsahu vznikajúcom v rámci sociálnych sietí rôzneho druhu.

Predtým, než budú popísané zmienené prístupy k identifikácii autorít, popíšeme podrobnejšie problematiku pozícií a rolí v sociálnych sieťach.

7.2 Pozície a role v sociálnych sieťach

Identifikáciou pozícií a rolí v sociálnych sieťach sa zaoberá aj štrukturálna analýza sociálnych sietí. Táto štrukturálna analýza sa v prvom rade zaoberá hľadaním vzorov v štruktúre siete. Ďalej sa zaoberá aj skúmaním pod - štruktúr sieťových dát a zoskupovaním aktérov podľa vzdialenosti. Príkladmi takýchto zoskupení sú kliky, bloky, hviezdy a mosty, ktoré definujú spôsob rozdelenia aktérov do podskupín na základe vzorov ich vzájomných vzťahov.

Zložitejším ale aj abstraktnejším prístupom je analýza vzťahov medzi aktérmi použitím tried rovnocennosti. Pre definovanie rovnocennosti aktérov je potrebné vykonať isté zovšeobecnenia v sociálnom správaní a sociálnej štruktúre. Totiž tento prístup uvažuje aktéra nie ako indivíduum, ale ako príklad určitej kategórie, skupiny aktérov v určitej definovanej rovnocennosti.

Abstraktné kategórie opisujúce sociálne postavenie (napríklad: stredná trieda, vyššia trieda) bežne používa sociológia. Takéto kategórie môžu byť používané v sociologickej teórii na opis „sociálnej role“ alebo „sociálnej pozície“ typických členov danej skupiny. Mnoho kategórií je založených na podobnosti hodnôt niektorých „atribútov“ jednotlivých aktérov kategórie. Napríklad by to mohli byť atribúty ako: pohlavie, pôvod, vek, príjem atď. Popis „muž Európan vo veku 40 – 60 rokov s relatívne vysokým príjmom“ , môže identifikovať skupinu – kategóriu ľudí. Pravdou je, že štrukturálna analýza nedokáže veľmi efektívne pracovať s kategóriami, ktoré sú založené na popise podobnosti jednotlivých aktérov. Preto sa uchyľuje k identifikácii členov kategórií z hľadiska podobnosti vzorov vzťahov priamo medzi jednotlivými aktérmi, nie medzi atribútmi – vlastnosťami týchto aktérov. Z toho vyplýva, že definícia kategórie ako „sociálna rola“ alebo „spoločenské postavenie“ závisí od vzťahov k ďalším kategóriám. Tento pohľad na sociálne role a pozície je relačným pohľadom zo strany štrukturálnej analýzy. Avšak nič nebráni tomu, aby okrem štrukturálnych dát boli využité aj iné dáta ako konverzačný obsah, kompozičné a temporálne dáta [Repka, 2011].

7.2.1 Pozície v sociálnej sieti a sociálne role

Na základe predpokladu, že pozície, role alebo sociálne kategórie sociálnych sietí sú definované primárne vo vzťahoch medzi jednotlivými aktérmi – používateľmi sociálnych sietí, je možné identifikovať sociálne pozície pomocou sieťových dát. Intuitívne by sme vedeli povedať, že dvaja aktéri majú rovnakú pozíciu alebo rolu do tej miery, pokiaľ profil ich vzťahov s inými aktérmi je rovnaký. Avšak, v tomto prístupe je niekoľko problémov.

V prvom rade nie je triviálne zistiť, ktoré vzťahy je potrebné vziať do úvahy, lebo sú podstatné pre identifikáciu autoritatívnosti, podobnosti či rovnocennosti. Neexistuje

ani jednoduchý spôsob ako určiť relevantnú množinu aktérov. To všetko závisí od účelu vykonávanej analýzy, použitého teoretického pohľadu a povahy celkovej množiny aktérov. Pri zodpovedaní týchto otázok nie sú klasické analytické metódy veľmi nápomocné.

Druhým problémom je samotná intuitívna definícia role alebo pozície. Funguje iba za predpokladu, že máme kolekciu aktérov a množiny vzťahov, ktoré majú zmysel pre ľudské vnímanie konkrétneho problému. Avšak pre účely identifikácie autorít je potrebné presne definovať rovnocennosť vzhľadom na pozíciu, či podobnosť ako istú mieru rovnocennosti.

Teda vymedzenie rovnocennosti alebo podobnosti je zložité ale potrebné. Metódy analýzy sietí najčastejšie definujú dva uzly (alebo aj zložitejšie štruktúry) za podobné, ak spadajú do rovnakej „triedy rovnocennosti". Zvyčajne sa vymedzí prečo dvaja aktéri (alebo iné štruktúry) sú členmi triedy, ktorá sa líši od ostatných tried, ktoré vlastnosti pozície aktéra sú podstatné pre umiestnenie aktéra do triedy s nejakými inými aktérmi a v akom smere sú rovnocenní. Vo všeobecnosti existuje mnoho spôsobov, ako definovať rovnocennosť aktérov založenú na ich vzťahoch s ostatnými. Napríklad by sme mohli vytvoriť dve triedy rovnocennosti aktérov s výstupným stupňom nula (respektíve jedna – stále by šlo o skupinu izolovaných aktérov) a aktérov s výstupným stupňom viac ako nula (aktívni aktéri). Dnes je k dispozícii veľké množstvo algoritmov, ktoré analyzujú kolekcie aktérov pomocou kategórií, alebo určujú pozície na základe určitých zhôd ich pozície v grafických reprezentáciách sociálnych sietí [Repka, 2011].

7.2.2 Podobnosť a rovnocennosť v sociálnej sieti

Podobnosť môže byť mierou rovnocennosti. Napríklad dvaja nejakým určitým spôsobom podobní aktéri môžu spadať do tej istej definície rovnocennosti. V analýzach sociálnych sietí založených na použití teórie grafov, ktoré by mali napomôcť k pochopeniu sociálnej úlohy a štrukturálnej pozície autoritatívneho aktéra sa obzvlášť osvedčili tri špecifické definície rovnocennosti:

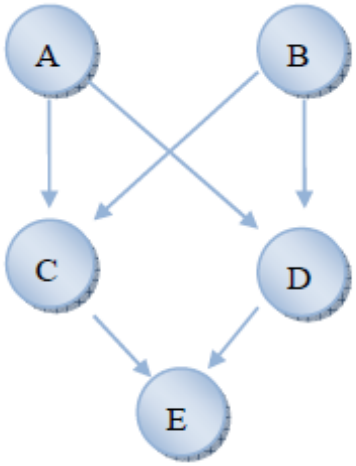
1) *Štrukturálna rovnocennosť* (Structural Equivalence) je vzťah medzi rovnocennými aktérmi, ktorí majú rovnakú alebo podobnú schému nalinkovania sa ku rovnakým susedom a sú vzájomne zameniteľní.
2) *Automorfná rovnocennosť* (Automorphic Equivalence) znamená, že rovnocenní aktéri majú takú schému zapojenia, že istou permutáciou aktérov je možné vytvoriť izomorfnú sieť. Automorfná rovnocennosť je v praxi málo používaná [Hanneman, 2005].
3) *Regulárna rovnocennosť* (Regular Equivalence) znamená, že rovnocenní aktéri majú rovnakú alebo podobnú schému zapojenia k rôznym susedom, preto hrajú rovnakú úlohu v sieti.

Definíciu rovnocennosti špecifikuje „relácia rovnocennosti", ktorá musí byť symetrická, reflexívna a tranzitívna. Na základe relácie rovnocennosti je možné definovať tzv. triedy rovnocennosti. Trieda rovnocennosti (Equivalence Class) alebo pozícia označuje kolekciu rovnocenných (respektíve približne rovnocenných) aktérov.

Štrukturálna rovnocennosť. Platí, že dvaja aktéri sú štrukturálne rovnocenní, ak majú rovnaký vzťah ku všetkým ostatným aktérom a musia byť presne nahraditeľní aby mohli byť považovaní za štrukturálne rovnocenných [Hanneman, 2005].

Na Obr.7.1 sú ilustrované tri triedy štrukturálnej rovnocennosti. Prvá z nich obsahuje iba aktéra *E*, ktorý nenadväzuje žiadne väzby (žiadne výstupné hrany). Keďže nie je

podobný žiadnemu inému aktérovi, je v triede sám. Druhá trieda pozostáva z aktérov *A* a *B*, ktorí majú rovnaké väzby na *C* a *D*. Poslednou triedou je trieda obsahujúca aktérov C a D, ktorí majú totožné vstupné a výstupné väzby a teda sú vzájomne zameniteľní. Aktéri, ktorí sú štrukturálne rovnocenní vytvárajú zhodné „pozície“ v štruktúre diagramu.

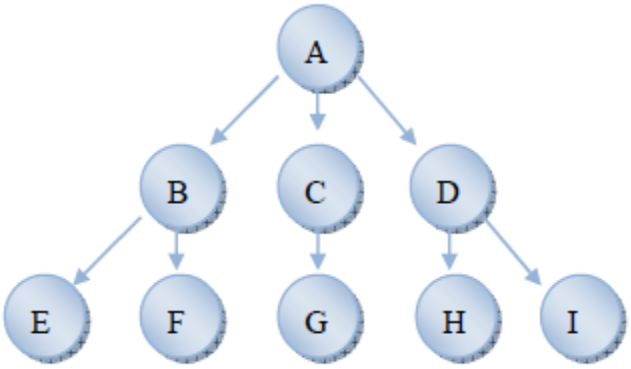


Obr. 7.1 Ilustrácia štrukturálnej rovnocennosti.

Aktéri v triede štrukturálnej rovnocennosti majú v istom zmysle rovnaké postavenie vzhľadom na všetkých ďalších účastníkov. Presné štrukturálne rovnocennosti sú zriedkavé alebo nepravdepodobné, preto sa v reálnych prípadoch väčšinou skúma miera štrukturálnej rovnocennosti. Štrukturálna rovnocennosť je „najsilnejšou“ formou rovnocennosti, keďže dvaja aktéri sú štrukturálne rovnocenní práve vtedy a len vtedy, ak majú identické prepojenia od a k ďalším aktérom v sieti.

Automorfná rovnocennosť predstavuje menej prísnu definíciu rovnocennosti ako štrukturálna rovnocennosť. Formálna definícia automorfnej definície je: dvaja aktéri *A, B* sú automorfne rovnocenní v sieti reprezentovanej značkovaným grafom, ak je možné vytvoriť izomorfný graf vzájomným zamenením aktérov *A, B* a preznačkovaním ostatných aktérov. Automorfná rovnocennosť vychádza z myšlienky izomorfizmu grafov nazývanou automorfizmus. Niekedy sa používa aj názov izomorfná rovnocennosť [Wasserman, 1994]. Automorfne rovnocenní aktéri majú tie isté značkovo-závislé vlastnosti. Permutácia grafu prostredníctvom zámeny rovnocenných aktérov nemá žiadny vplyv na vzdialenosti (najkratšie cesty v grafe) medzi všetkými aktérmi v grafe [Hanneman, 2005].

Graf na obrázku Obr.7.2 zobrazuje hierarchicky štruktúrovanú skupinu aktérov. Aktér *A* je koreňovým uzlom v sieti (napríklad riaditeľ v nejakej spoločnosti), aktéri *B*, *C* a *D* sú tromi medziľahlými aktérmi (napríklad vedúci oddelení) a aktéri *E*, *F, G, H, I* sú listy (napríklad radoví pracovníci). Aktér B a aktér D nie sú štrukturálne rovnocenní, lebo majú síce rovnakého šéfa, ale nemajú rovnakých pracovníkov. Na druhej strane sú rovnocenní v inom zmysle. Obaja vedúci *B* a *D* sú podriadení tomu istému šéfovi a každý z nich má presne dvoch pracovníkov.



Obr. 7.2 Hierarchicky štruktúrovaná skupina aktérov.

Ak by sme vymenili manažérov B a D a ich štyroch zamestnancov, všetky vzdialenosti medzi všetkými aktérmi v grafe by boli úplne totožné. Aktéri *B* a *D* tvoria „automorfnú“ triedu rovnocennosti. Táto menej prísna definícia rovnocennosti znižuje počet tried rovnocennosti.

Regulárna rovnocennosť. Dva uzly sú regulárne rovnocenné, ak majú rovnaký profil väzieb s aktérmi iných skupín, ktoré sú tiež regulárne rovnocenné. Ako príklad regulárnej rovnocennosti môžu byť rodinné vzťahy. Dvaja otcovia sú rovnocenní z hľadiska pozície v rodine, pretože každý z nich má určitý vzor väzby s manželkou, deťmi a ostatnými aktérmi rodiny. Regulárna rovnocennosť najlepšie stanovuje pojem „sociálna rola“, ktorá je základným stavebným blokom všetkých spoločenských inštitúcií.

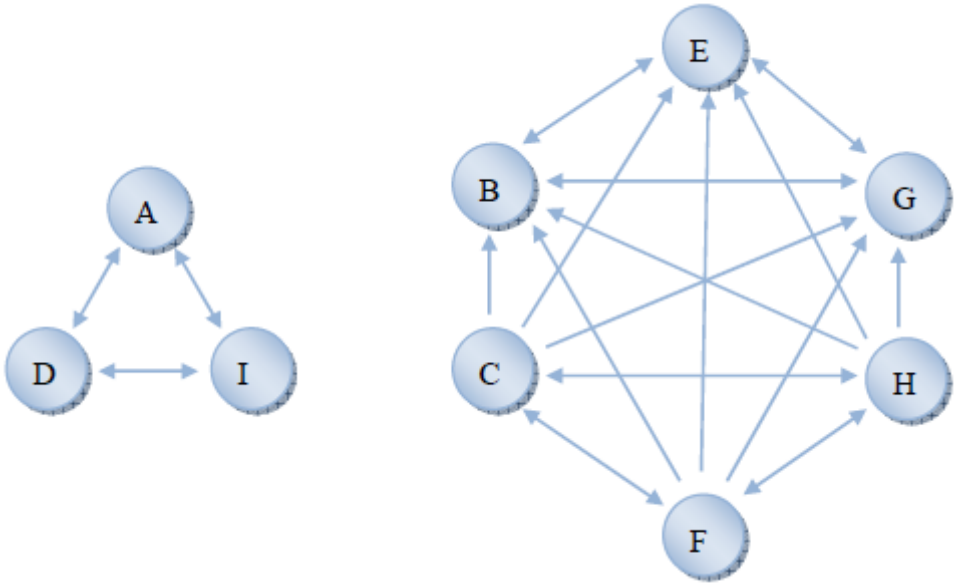
Na Obr.7.2 sú tri triedy regulárnej rovnocennosti. V prvej je aktér A, druhá trieda sa skladá z troch aktérov B, C a D a tretia obsahuje zostávajúcich päť aktérov E, F, G, H a I. Týchto päť aktérov triedy je navzájom regulárne rovnocenných, pretože nemajú väzbu s aktérom v prvej triede a každý z nich má väzbu s aktérom v druhej triede. Aktéri B, C a D tvoria triedu regulárnej rovnocennosti, pretože, každý z nich má väzbu s členom prvej triedy a taktiež každý z nich má väzbu s nejakým členom tretej triedy. Zatiaľ čo aktéri B a D majú väzby s dvomi členmi triedy, aktér C má väzbu iba na jedného člena z triedy (to stačí, lebo vyžaduje sa väzba iba na niektorých členov tretej triedy). Aktér A je osamelý člen v triede, ktorá je definovaná väzbou aspoň na jedného člena druhej triedy.

Rovnako ako pri štrukturálnej a automorfnej rovnocennosti, presná regulárna rovnocennosť môže byť vzácna, hlavne vo veľkej populácii s mnohými triedami rovnocennosti. Preto sa často pracuje s približnou regulárnou rovnocennosťou, ktorá môže mať v reálnych sociálnych sieťach väčší zmysel. Určí aktérov, ktorí budú patriť do jednotlivých sociálnych rolí a ako spoločenské role (nie role osôb) vytvárajú vzťahy medzi sebou. Aj keď sa pracuje s menej prísnymi definíciami, neznamená to, že sú menej presné. Menšia striktnosť predstavuje vyššiu úroveň abstrakcie, ktorá umožní lepšie porozumenie sociálnym sieťam.

V reálnom svete sú však sociálne siete často dosť chaotické, nie sú vyvážené a taktiež môžu byť zle merateľné. Vyhľadať rovnocennosti v reálnych údajoch môže byť značne komplikovaná záležitosť. Tu vyvstáva problém adekvátnej definície rovnocennosti ako aj miery rovnocennosti. A tu môže byť nápomocná pozičná analýza, ktorá zjednodušuje informácie v dátach sociálnej siete.

7.2.3 Pozičná analýza

Jednou z hlavných úloh pozičnej analýzy je zjednodušiť informácie v dátach siete. Tento proces zjednodušenia sa začína reprezentáciou siete pomocou názvov pozícií a následne sa definujú matematické výrazy určujúce vzťahy týchto pozícií medzi sebou. Pozície sú definované väčšinou formou rovnocennosti aktérov. Toto zjednodušenie je ilustrované na Obr.7.3. Príslušná sociometrická reprezentácia je uvedená v Tab.7.1.



Obr. 7.3 Sociálna sieť pred zjednodušením.

Tab. 7.1 Sociometrická reprezentácia sociálnej siete z Obr. 7.3 (dáta pred zjednodušením).

	A	B	C	D	E	F	G	H	I
A	-	0	0	1	0	0	0	0	1
B	0	-	0	0	1	0	1	0	0
C	0	1	-	0	1	1	1	1	0
D	1	0	0	-	0	0	0	0	1
E	0	1	0	0	-	0	1	0	0
F	0	1	1	0	1	-	1	1	0
G	0	1	0	0	1	0	-	0	0
H	0	1	1	0	1	1	1	-	0
I	1	0	0	1	0	0	0	0	-

Zatiaľ nie je možné vidieť voľným okom pozície založené na štruktúre ani v grafe (Obr.7.3) ani v sociometrickej reprezentácii (sociomatici, Tab.7.1). Ale reorganizáciou dát v sociomatici možné vytvoriť bloky, ktoré budú vyplnené jednotkami alebo nulami. Reorganizovaná sociomatica je znázornená v Tab.7.2. Na základe týchto blokov je možné vytvoriť hľadané a skúmané pozície a zobraziť ich v novej sociomatici (Tab.7.3).

Tab. 7.2 Reorganizovaná sociometrická reprezentácia sociálnej siete z Obr.7.3 (dáta pred zjednodušením).

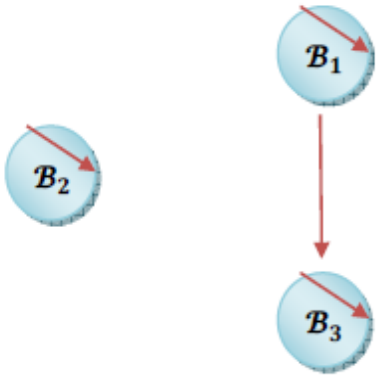
	F	C	H	D	A	I	B	E	G
F	-	1	1	0	0	0	1	1	1
C	1	-	1	0	0	0	1	1	1
H	1	1	-	0	0	0	1	1	1
D	0	0	0	-	1	1	0	0	0
A	0	0	0	1	-	1	0	0	0
I	0	0	0	1	1	-	0	0	0
B	0	0	0	0	0	0	-	1	1
E	0	0	0	0	0	0	1	-	1
G	0	0	0	0	0	0	1	1	-

Tab. 7.3 Sociometrická reprezentácia sociálnej siete po zjednodušení.

	$\mathcal{B}_1$	$\mathcal{B}_2$	$\mathcal{B}_3$
$\mathcal{B}_1$	1	0	1
$\mathcal{B}_2$	0	1	0
$\mathcal{B}_3$	0	0	1

Táto matica je podstatne čitateľnejšia ako pôvodné dáta a vyjadruje taktiež vzťahy medzi jednotlivými pozíciami. Grafickým znázornením týchto zjednodušených dát dostaneme zjednodušený obraz pôvodnej sociálnej siete z Obr.7.3. Táto zjednodušená sociálna sieť je znázornená na Obr.7.4. Podrobnejšie informácie

k pozičnej analýze je možné nájsť v [Repka, 2011].



Obr. 7.4 Redukovaný graf sociálnej siete po zjednodušení.

7.2.4 Metódy identifikácie sociálnych rolí

Sociálne role môžu byť študované porovnaním behaviorálnych a štrukturálnych vzorov interakcií medzi jednotlivými aktérmi (indivíduami) sociálnej siete. Teda to čo nás zaujíma je „štrukturálny profil“. V literatúre zvykne byť tento štrukturálny profil označovaný aj ako „podpis“ (signature). Pojem štrukturálny profil sa vzťahuje k rozlišovacím atribútom, ktoré definujú typy sietí. Pojem „behaviorálny a štrukturálny profil” sa zvykne používať pre vzory správania sa v sociálnej sieti. Je vzťahovaný k rozlišovacím pozičným atribútom, ktoré identifikujú a vymedzujú aktérov ako príslušníkov nejakej sociálnej role. Ak dokážeme identifikovať štrukturálne profily, potom dokážeme identifikovať vzory interakcií indivíduí bez hodnotenia obsahu tejto interakcie, čo výrazne znižuje čas potrebný k rozpoznaniu jednotlivých typov aktérov [Welser, 2007]. V jednej z nasledujúcich podkapitol „Autority webových diskusií“ sa popisuje návrh identifikácie jednej zo sociálnych rolí a to role „autorita“ zo štruktúry webových diskusií, ktorý zjednodušuje a zrýchľuje proces identifikácie autorít práve snahou vylúčiť z procesu analýzu konverzačného obsahu. Pre dedukovanie rolí zo štrukturálnych a behaviorálnych profilov existujú zaujímavé prístupy:

- vizualizácia profilov pre získanie ukazovateľov sociálnych rolí,
- hľadanie rozdielnosti medzi profilmi sociálnych rolí,
- využitie štrukturálnych a behaviorálnych dát pre identifikáciu významných rolí (napríklad autoritatívnych aktérov).

Prístup založený na vizualizácii profilov pre získanie ukazovateľov sociálnych rolí vychádza z predpokladu, že vizualizácia je efektívnym prostriedkom pre rozpoznávanie základných profilov sociálnych rolí. Následná analýza sociálnych rolí môže byť kombináciou viacerých metód. Vizualizácia môže taktiež poskytnúť istú intuitívnu pomoc pre rozoznávanie profilov a vzorov v analýze sociálnych sietí. Táto intuícia môže byť následne formalizovaná do metriky. Táto formalizácia umožní testovanie prostredníctvom systematického porovnávania extrahovaných profilov, vzorov a rolí. Tento proces však nie je ľahko uskutočniteľný, preto sa namiesto metrík častejšie využívajú mnohé známe miery centralít. Mnoho výskumníkov však pracuje na návrhu metrík použiteľných na odhalenie sociálnych procesov a sociálnej štruktúry [Welser, 2007].

Štúdium rolí je založené vo všeobecnosti na predstave, že štrukturálna podobnosť označuje triedu podobných subjektov, ktoré môžu odpovedať nejakej konkrétnej sociálnej roli. Tak napríklad vyššie popísaná skupina metód založená na rovnocennosti sa často používa pri rozdelení populácie aktérov do tried. Následne sa vzťahy medzi týmito triedami používajú na identifikáciu konkrétnej role. Pre algoritmus priraďovania všetkých aktérov do rôznych tried nestačí iba identifikácia všeobecných štrukturálnych čŕt typických pre konkrétnu rolu. Preto sa na identifikáciu štrukturálnych atribútov spojených so sociálnou rolou používajú aj egocentrické sieťové dáta (dáta zamerané na špecifických aktérov, napr. centrality jednotlivých uzlov, autority medzi aktérmi sociálnej siete a pod.). Samozrejme aj v tomto prípade sa používajú vizualizačné techniky.

Existujú štúdie, ktoré porovnávajú relatívne miery kohézie v sieti a štrukturálnu rovnocennosť za účelom predikcie podpory v politických aktivitách. Je pravdou, že použiteľnosť takého modelu a jeho výpovedná hodnota sú obmedzené. Na druhej strane táto štúdia objavila dôležité znalosti týkajúce sa porovnania adekvátnosti jednotlivých druhov rovnocennosti [Mizruchy, 1993].

Ďalšia štúdia štruktúry siete ponúka spôsob použitia lokálnych mier siete na identifikáciu dôležitých pozícií aktérov [Burt, 2004], [Burt, 1992]. Iná štúdia zameraná na identifikáciu prestížneho aktéra, teda autoritu, preukázala relevanciu štrukturálnych charakteristík v identifikácii autorít. V tomto prípade išlo konkrétne o identifikáciu role prestížneho hráča v bridžových tímoch [Erikson, 1984]. Podarilo sa experimentálne dokázať, že štrukturálne atribúty interakcií je možné použiť na identifikáciu role namiesto spoľahlivých ale náročných behaviorálnych mier založených na obsahu týchto interakcií. Z tohto pohľadu je zaujímavá technika porovnania štrukturálnych ukazovateľov a spoľahlivých mier rolí správania sa získaných z analýzy obsahu interakcie [Welser, 2007], teda technika kombinujúca získavanie rolí tak zo štruktúry ako aj z obsahu sociálnych sietí.

Získavanie behaviorálnych profilov sociálnych sietí v špecifických podmienkach je založené na predpoklade, že správanie sa v minulosti bude korelovať so správaním sa v budúcnosti. Z toho vyplýva, že profily môžu byť odvodené aj zo vzorov správania sa v čase. Napríklad, výskumní pracovníci v telefónnych spoločnostiach sa v rámci skúmania bahaviorálnych profilov zameriavajú na odvádzanie profilov podvodných telefónnych účtov. Analyzujú telefónne účty a hľadajú vzory v množstve telefonátov, ich načasovaní, totožnosti adresátov hovorov a totožnosti volajúcich. Je pravda, že v praxi dochádza k častým zmenám v telefónnych číslach a účtovníctve. Napriek tomu by mohli byť podvodné telefónne účty pomerne rýchlo rozpoznané napríklad vtedy, keď novo vytvorený účet začal prijímať hovory od ľudí, ktorí prijímali hovory z účtov, ktoré boli predtým označené ako podvodné. Rovnaký princíp pre rozpoznávanie profilov správania a ukazovateľov základných individuálnych atribútov je možné použiť aj na skúmanie sociálnych rolí v sociálnych sieťach [Cortes, 2004], [Cortes, 2001].

V poslednej dobe sa veľa energie venuje skúmaniu webových diskusných skupín. V rámci toho výskumu sa začali spájať správanie a štrukturálne charakteristiky s rôznymi typmi prispievateľov do on-line webových diskusných skupín. V práci [Viegas, 2004] bol zavedený pojem „authorline“ vizualizácia, ktorá bola použitá na diferenciáciu rôznych typov prispievateľov do on-line diskusií, ako sú spameri, prevažne debatujúci a odpovedajúci používatelia a vyslovené autority. Táto diferenciácia prispievateľov bola založená na analýze počtu a typu správ, ktorými prispievajú do rôznych diskusných vlákien. V práci [Turner, 2005] boli pokusy rozšíriť

túto analýzu správania sa aj na prieskum prispievateľov, lokálnych sieťových atribútov a typov diskusných skupín. Avšak, analýza ich sociálneho prostredia a úloha v on-line komunitách je orientačná a nie je jasné do akej miery vizualizácia dokáže presne vymedziť role [Welser, 2007].

7.3 Autority vo vede

Tento prístup k identifikácii autorít sa zameriaval na identifikáciu často citovaných autorov vedeckých článkov, teda autorov, ktorých názory, myšlienky, návrhy, teórie sú zaujímavé a inšpiratívne pre iných vedcov danej oblasti a ovplyvňujú ich ďalší výskum. Dáta boli získavané z digitálnych knižníc ACM Digital Library a IEEE Database. Predpokladá sa, že daná vedecká oblasť bude definovaná používateľom, ktorý hľadá autoritu v rámci vlastných profesijných záujmov ato vo forme kľúčových slov, resp. kľúčovou frázou. Vyhľadávajú sa všetky vedecké články, v ktorých nadpise alebo abstrakte sa nachádza používateľom zadaná kľúčová fráza. Prístup sa sústreďuje na referencie uvedené v každom vyhľadanom článku. Zisťujú sa početnosti výskytu mien autorov citovaných publikácií. Tab.7.4 ilustruje nájdené autority pre „opinion analysis“ v ACM DL.

Tab. 7.4 Nájdené autority vo vedeckej oblasti „opinion analysis“ v digitálnej knižnici ACM.

Poradie	Slovo	Počet celkových výskytov	Počet dokumentov s výskytom
1	Wiebe	34	9
2	Lee	22	17
3	Pang	19	15
4	Liu	19	9
5	Chen	18	7
6	Cardie	17	8
7	Wilson	17	6
8	Janyce	16	4
9	Zhang	13	8
10	Yu	12	10

Analýza početností mien autorov predstavuje netriviálnu úlohu, ktorú sťažuje variabilita citačných štandardov. Tab.7.5 ilustruje desať najdôležitejších nájdených autorít pre vedeckú oblasť „opinion analysis“ v IEEE Database. Výsledky vyhľadávania autorít vo vedeckej oblasti „opinion analysis“ v digitálnych knižniciach ACM DL a IEEE Database boli vizualizované prostredníctvom TagClouds čo je ilustrované v Obr.7.5. Podrobnejšie informácie je možné nájsť v [Kmeť, 2012].

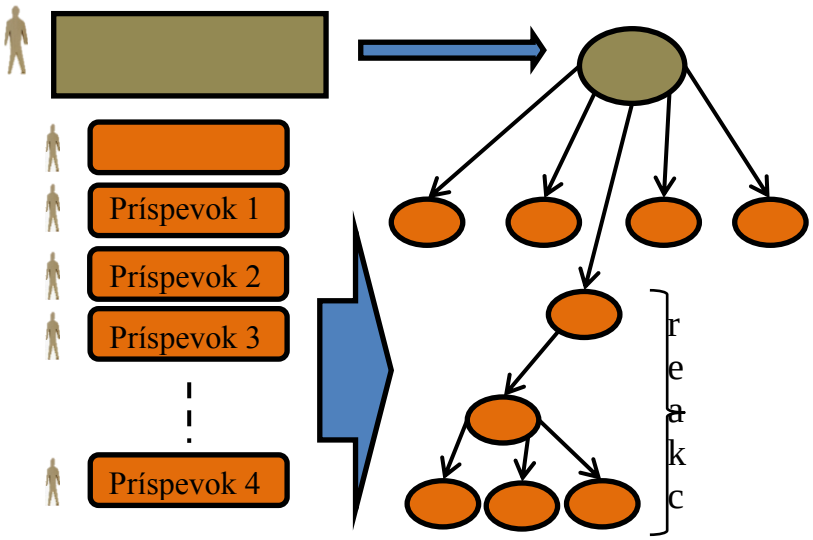
Tab. 7.5 Nájdené najvýznamnejšie autority vo vedeckej oblasti „opinion analysis“ v digitálnej knižnici IEEE Database.

Poradie	Slovo	Počet celkových výskytov	Počet dokumentov s výskytom
1	Liu	26	11
2	Lee	25	13
3	Chen	20	12
4	Pang	18	11
5	Zhang	18	8
6	Li	16	9
7	Kim	16	9
8	Hu	15	8
9	Hovy	14	7
10	Xu	12	7

Obr. 7.5 Výsledky vyhľadávania autorít vo vedeckej oblasti „opinion analysis“ v digitálnych knižniciach ACM DL a IEEE Database vizualizované prostredníctvom TagClouds.

7.4 Autority webových diskusií

Prakticky každý používateľ sémantického webu môže založiť webovú diskusiu k nejakej téme, problému alebo v súvislosti s otázkou na ktorú potrebuje odpoveď. Nasledovné reakcie na založenú otázku – diskusiu vytvárajú webovú diskusiu, ktorú je možné graficky znázorniť. Ak každý príspevok do diskusie bude reprezentovaný uzlom grafu a relácia „reakcia na“ bude reprezentovaná hranou, potom celá diskusia bude reprezentovaná acyklickým grafom, teda stromom, čo je ilustrované na Obr.7.6. Tento acyklický strom dokumentuje štruktúru webovej diskusie, ktorá je základom pre dolovanie informácií v rámci identifikácie autorít webovej diskusie.



Obr. 7.6 Grafická reprezentácia webovej diskusie.

Rôzni používatelia sociálneho webu prispievajú do webovej diskusie z rozličných dôvodov a nie každý je autoritou. Kľúčová otázka je ako rozpoznať, kto je autoritatívny prispievateľ a kto nie je. Je možné formulovať tri hlavné dôvody prispievania do webovej diskusie a teda tri hlavné typy prispievateľov:

- *Hľadanie odpovedí* – Veľká väčšina používateľov hľadá odpovede na svoje otázky, ktoré potrebujú, aby sa lepšie rozhodovali v dôležitých veciach. Takí používatelia očakávajú informované rady od múdrejších prispievateľov a očakávajú samozrejme pravdivé informácie. Títo používatelia nie sú autority aj keď tvoria jadro fóra.
- *Príležitosť prezentovať sa* – Neveľká množina používateľov sociálneho webu vyhľadáva príležitosť prezentovať svoju dôležitosť. Nerozpakuje sa uvádzať aj nepravdivé informácie, niekedy vyvoláva konflikty. To všetko degraduje webovú diskusiu. Preto v prípade moderovanej, riadenej webovej diskusie sú títo problematickí provokatéri vylučovaní z diskusie. Títo používatelia nie sú autority.
- *Príležitosť vyjadriť vedomosti* – Taktiež neveľká skupina prispievateľov do webových diskusií tam prispieva za účelom uistenia sa o správnosti vlastných nápadov a prípadného revidovania názorov. Títo používatelia poskytujú iba pravdivé informácie, seriózne pristupujú k diskusii, prispievajú do diskusie iba keď sa cítia orientovaní v problematike, o ktorej sa diskutuje. Týchto používateľov považujeme za autority a chceme ich identifikovať.

Bol navrhnutý originálny prístup k odhadu autorít, ktorý je založený na spracovaní nasledovných vstupných údajov:

- meno prispievateľa
- polarita príspevku
- dĺžka príspevku
- príspevky - reakcie
- pozícia príspevku v strome.

V procese spracovania, autorita nie je vzťahovaná k príspevkom, ale k prispievateľom. Preto prvým krokom procesu spracovania je predspracovanie, ktoré integruje všetky informácie o každom jednotlivom prispievateľovi. V ďalšom kroku je vykonávaný samotný odhad autority, ktorý je modelovaný vzťahom (16):

$$OA = 4PP^3 + 2PR^3 + 4PKU^2 + ZP + PU + PT \qquad (16)$$

Výsledkom procesu odhadu autority je zostupne usporiadaný rebríček indikujúci špeciálny typ prispievateľov:

- prezentujúcich hlbokú znalosť problematiky,
- vyvolávajúcich mnoho reakcií,
- inicializujúcich najčastejšie prechod na novú tému,

teda autorít. Vyššie uvedený vzorec zahŕňa primárne a sekundárne vplyvy.

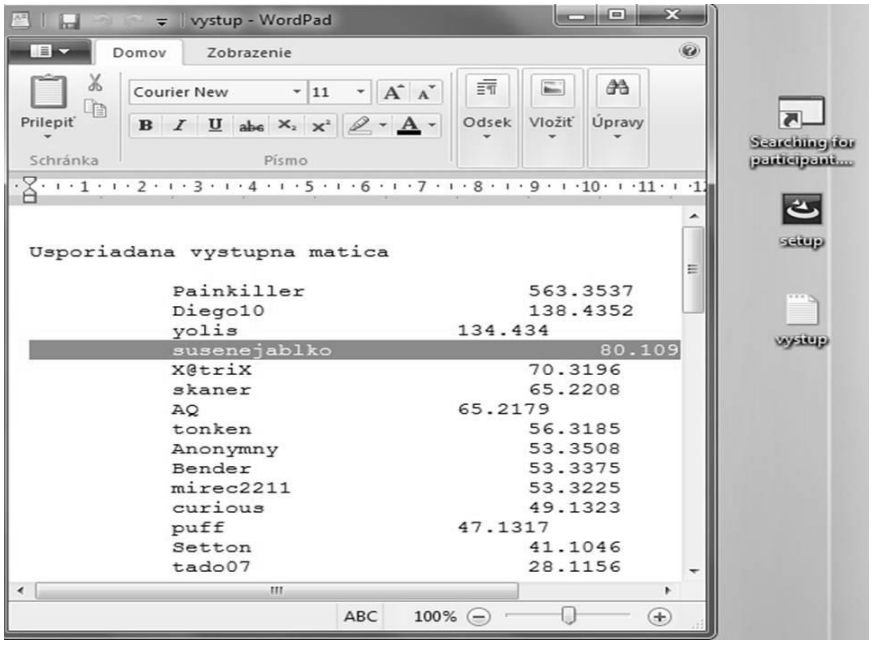
Primárne vplyvy sú reprezentované nasledovnými argumentmi:

- počet príspevkov prispievateľa (PP)
- počet reakcií na príspevky prispievateľa (PR)
- počet výskytov na koncovej úrovni stromu (PKU)

Sekundárne vplyvy sú reprezentované nasledovnými argumentmi:

- zhoda polarity (ZP)
- pozície v strome (počet úrovní - PU)
- počet termov (PT).

Táto metóda odhadu bola implementovaná. Implementácia ponúka po spracovaní zadanej webovej diskusie usporiadaný zoznam prispievateľov od najautoritatívnejšieho po najmenej autoritatívneho, čo je ilustrované na Obr.7.7.



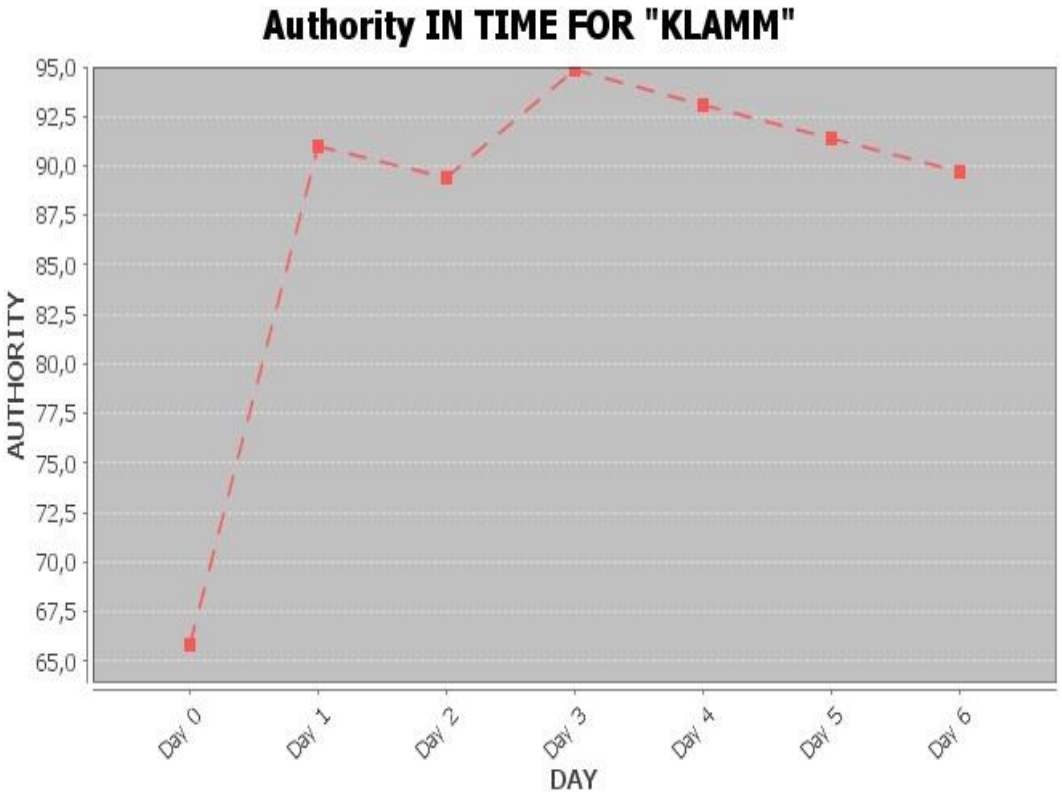
Obr. 7.7 Implementácia metódy odhadu autorít webovej diskusie.

Tab.7.6 obsahuje výsledky testov navrhnutého prístupu na troch diskusiách k trom rôznym témam: Je nejaký vzťah medzi autoritou a počtom liekov?, Slovenskí politici a k téme týkajúcej sa nepokojov na blízkom východe Bomby, letecké útoky a sirény.

Tab. 7.6 Výsledky testov navrhnutého prístupu odhadu autorít v troch rôznych doménach (Autorita a počet liekov, Slovenskí politici, Bomby, letecké útoky a sirény).

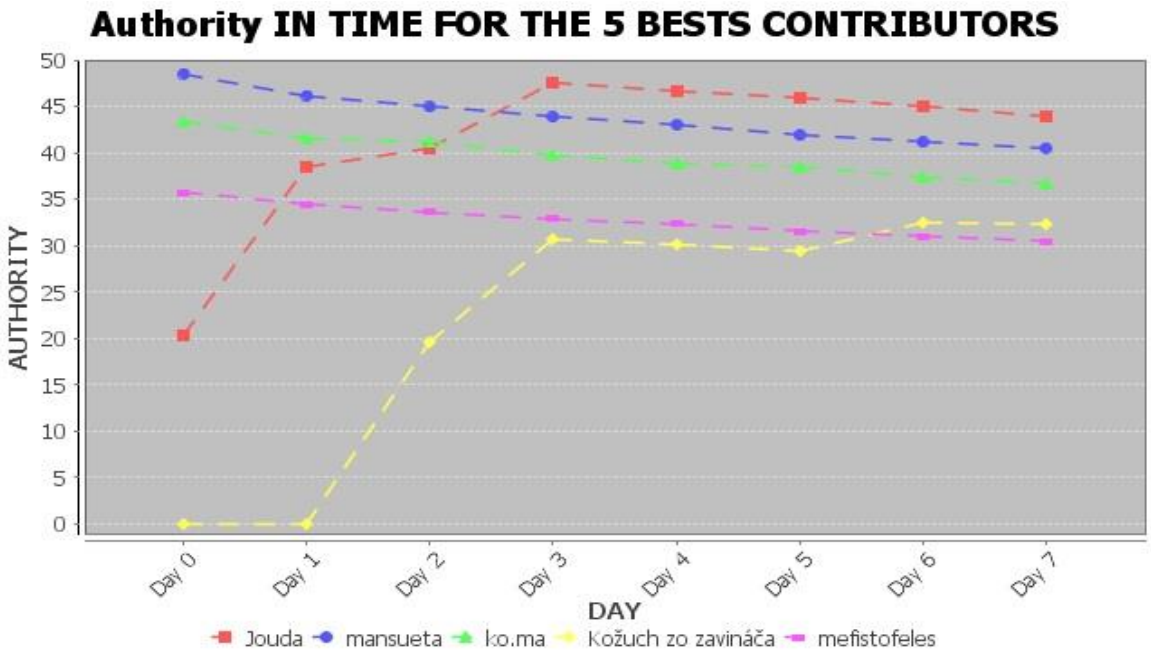
Téma diskusie	**Presnosť**
Autorita a počet “likes”	0.94
Slovenskí politici	0.96
Bomby, letecké útoky a sirény	0.93

Pre používateľa je určite prínosné ak získa informáciu o tom, kto je práve najautoritatívnejší prispievateľ vo webových diskusiách na určitú tému. Avšak oveľa prínosnejšia je informácia, ako dlho si status autority konkrétny prispievateľ udržal. Či neprišiel do diskusie niekto nový kto je v téme lepšie orientovaný. Preto má zmysel sledovanie dynamickej zmeny autority v určitom vopred stanovenom časovom výseku. Obr.7.8 ilustruje dynamickú zmenu autority zvoleného prispievateľa.



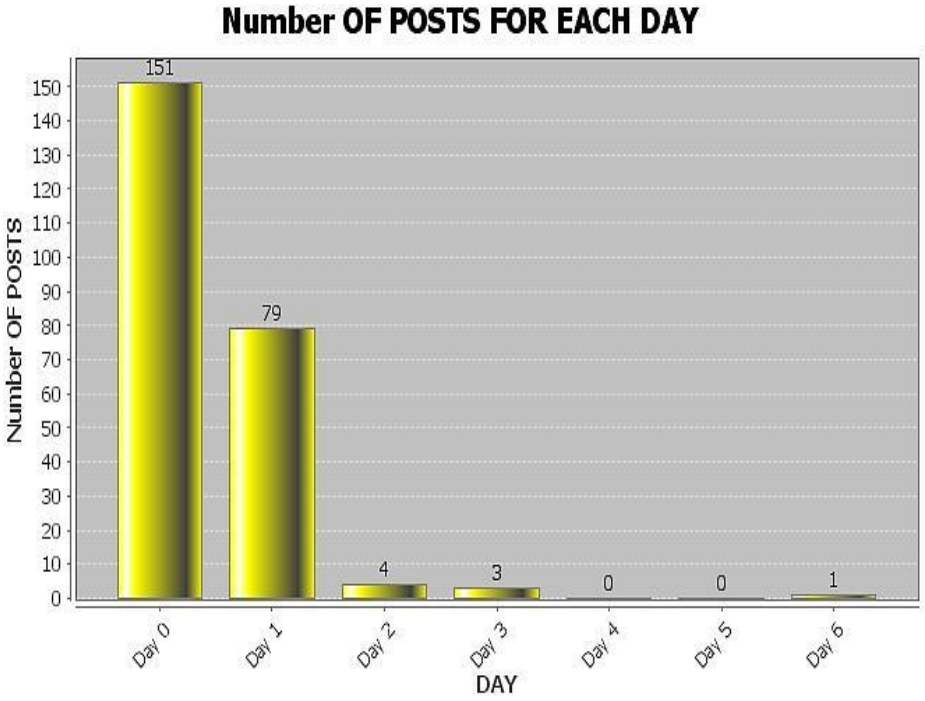
Obr. 7.8 Dynamická zmena autority zvoleného prispievateľa v rámci sledovanej webovej diskusie.

Možno by chcel mať používateľ k dispozícii simuláciu dynamickej zmeny viacerých prispievateľov do danej webovej diskusie aby ich autoritatívnosť a zmenu tejto autoritatívnosti v čase mohol porovnať. Aj to mu uvedená implementácia môže poskytnúť, čo ilustruje Obr.7.9, ktorý obsahuje graficky znázornený vývoj autoritatívnosti v čase pre piatich najvýraznejších prispievateľov do sledovanej webovej on-line diskusie.



Obr. 7.9 Dynamická zmena autority piatich najvýznamnejších prispievateľov do webovej diskusie.

Pri analýze výsledkov simulácie dynamickej zmeny autority sledovaných prispievateľov je potrebné brať do úvahy aj vývoj samotnej diskusie v čase, respektíve je potrebné vedieť ako dlho diskusia trvala a kedy bola najbúrlivejšia. Niekedy môže taká diskusia trvať iba niekoľko dní, čo ilustruje Obr.7.10. Na tomto obrázku je uvedený počet príspevkov, ktoré pribudli do sledovanej on-line diskusie v každom dni jej trvania. Na Obr.7.10 je teda znázornený vývoj webovej diskusie, ktorá kulminovala v nultom (deň založenia diskusie) a prvom dni. Preto je v tomto prípade možné predpokladať, že autorita prispievateľov bude najviac meniť v nultom dni. Tento fakt je podporený najväčším prírastkom komentárov do diskusie spomedzi všetkých dní trvania diskusie. V prvom dni bol pridaný druhý najväčší počet príspevkov. Autorita veľkému percentu prispievateľov bude vo zvyšných dňoch existencie diskusie klesať, pretože počas zvyšných piatich dní bolo pridaných celkovo iba osem príspevkov. Súčasťou návrhu je aj analýza, ktorá umožní určiť najvhodnejšiu dobu pre uzavretie a následnú analýzu dynamického vývoja diskusie. Podrobnejšie informácie je možné nájsť v [Sendek, 2013].



Obr. 7.10 Vizualizácia počtu pribudnuvších príspevkov v jednotlivých dňoch trvania on-line webovej diskusie.

7.4.1 Diskusia k návrhu odhadu autorít

Implementácia metódy odhadu autorít webových diskusií bola testovaná s veľmi dobrými výsledkami na doménach tak z reálneho života ako aj na doméne z technickej oblasti. Tento prístup kombinuje dolovanie zo štruktúry s dolovaním z obsahu. Dolovanie zo štruktúry poskytuje drvivú väčšinu argumentov používaných na odhad miery autority. Dolovanie z obsahu je zastúpené pri odhade miery v akej polarita názorov skúmaného prispievateľa súhlasí s polaritou celej diskusie. Na určenie tejto miery súhlasu, resp. nesúhlasu polarity jednotlivého príspevku s celou diskusiou bola použitá aplikácia klasifikácie názorov. Ale aj naopak, pri klasifikácii názorov (teda odhade polarity príspevku) je možné využiť odhad miery autority. Konkrétne miera autority jednotlivého prispievateľa by sa mohla použiť ako váha, s ktorou sa budú brať do úvahy jeho príspevky. Webová služba, ktorá by implementovala takúto metódu odhadu autorít by mohla byť prospešná aj pre nejakú organizáciu, pre vyhľadávanie schopných zamestnancov na špecifické pozície. Práve autoritám webovej diskusie k problémom, ktoré daná organizácia rieši, by mohli byť ponúknuté pracovné miesta v tejto organizácii. Je možný aj iný scenár a to, že organizácia založí profesionálnu diskusiu k témam, problémom a výzvam, ktoré organizáciu trápia a ktorým čelí. Potom vypíše konkurz, pričom uchádzačov o pracovnú pozíciu požiada, aby sa zúčastnili tejto webovej diskusie, lebo bude bratá ako nulté kolo pohovoru. Po určitom čase budú vyhodnotení autoritatívni prispievatelia do danej profesionálnej diskusie a práve oni budú pozvaní na prvé ostré kolo pohovoru.

POUŽITÁ LITERATÚRA

[Burt, 2004] BURT, R. S. Structural Holes and Good Ideas. American Journal of Sociology 110, 2004, 349-399.

[Burt, 1992] BURT, R. S. Structural Holes: The Social Structure of Competition. Cambridge, Mass.: Harvard University Press, 1992.

[Cortes, 2004] Cortes, C., Fisher, K., Pregibon, D., Rodgers, A., Smith, F.: HANCOCK - A Language for Analyzing Transactional Data Streams. ACM Transactions on Programming Language and Systems, 26, 2004, 301-338.

[Cortes, 2001] Cortes, C., Pregibon, D.: Signature-Based Methods for Data Streams. Data Mining and Knowledge Discovery, 5, 2001, 167-182.

[Erikson, 1984] Erikson, B. H., Nosanchuck, T. A.: The Allocation of Esteem and Disesteem: A Test of Goode's Theory. American Sociological Review, 49(5), 1984, 648-658.

[Goodenough, 1969] Goodenough, W.H.: Rethinking 'status' and 'role:' Toward a general model of the cultural organization of social relationships. In S. A. Tyler(ed.), Cognitive Anthropology, New York: Holt, Rinehart, and Winston, 1969, 311-330.

[Hanneman, 2005] Hanneman, R.A., Riddle, M.: Introduction to social network methods. Riverside, CA, University of California, Riverside. 2005. <http://faculty.ucr.edu/~hanneman/nettext/>

[Homans, 1961] Homans G.C.: Social Behaviour: Its Elementary Forms. New York, Harcout, Brace & World, 1961.

[Kmeť, 2012] Kmet, S., Machová, K.: Web authorities estimation within the given domain. Department of Cybernetics and Artificial Intelligence, Technical University, Košice, 2012, 0-71.

[Mizruchy, 1993] Mizruchi, M. S.Cohesion, Equivalence, and Similarity of Behavior - a Theoretical and Empirical-Assessment. Social Networks 15, 1993, 275-307.

[Repka, 2011] Repka, M.: Analýza určitých typov sociálnych sietí. Košice, Technická univerzita v Košiciach, Fakulta elektrotechniky a Informatiky, 2011. 1-75.

[Sendek, 2013] Sendek, J., Machová, K.: Dynamická zmena autority aktérov sociálneho webu. Technická univerzita v Košiciach, Fakulta elektrotechniky a Informatiky, 2013, 1-80.

[Scott, 2010] Scott, J: Social network analysis: A handbook. Second edition. SAGE Publications, 2000, ISBN 978-0-7619-6339-4.

[Turner, 2005] Turner, T.C. Simth M. Fisher, D. Welser H.T.: Picturing Usenet: Mapping Computer-Mediated Collective Action. Journal of Computer Mediated Communication, 10(4), 2005.

[Viegas, 2004] Viegas, F.B. Smith, M.A.: Newsgroup Crowds and Author Lines: Visualizing the Activity of Individuals in Conversational Cyberspaces, Big Island, Hawaii, IEEE, 2004.

[Wasserman, 1994] Wasserman, S., Faust, K.: Social Network Analysis. Cambridge: Cambridge University Press, 1994, ISBN 9178-0-521-38707-1.

[Welser, 2007] Welser, H. T., Gleave, E., Fisher, D., Smith, M.: Visualizing the Signatures of Social Roles in Online Discussion Groups. Journal of Social Structure, Vol.8, No.2, 2007.