Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra informatiky a výpočetní techniky

Bakalářská práce

Analýza propojení firem využitím PageRanku

Plzeň, 2014 **Václav Suda**

Originální zadání

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 25. června 2014.

Václav Suda

Poděkování

Na tomto místě bych rád poděkoval Ing. Michalu Nyklovi za cenné rady, ochotné vedení práce a čas, který mi věnoval při konzultacích.

Abstract

An Analysis of Companies Interlinking Using PageRank

The presented bachelor thesis is focused on the analysis of companies interlinking using algorithm PageRank. In this thesis there is a description of algorithm PageRank and the description of centrality measures methods. Afterwards, the thesis presents a plan of web crawler which was used for obtaining data from the web and the concept of the database in which, using this crawler, was data saved. During the process of obtaining data from web, there were more than 700,000 recordings saved by crawler. The thesis contains the layout of several graphs which can be created by using obtained data and consequently evaluated. Graphs are created by companies and persons concerned in these companies. Obtained data are evaluated by using the algorithm PageRank and centrality measures methods, and at the end of the thesis there is a discussion of obtained results.

Analýza propojení firem využitím PageRanku

Předkládaná bakalářská práce se zaměřuje na analýzu propojení firem využitím algoritmu PageRank. Naleznete zde popis algoritmu PageRank a metod mír centrality. Dále je zde uveden návrh webového robota, který byl použit pro získání dat z webu, a návrh databáze, do které byla data pomocí tohoto robota uložena. Při získávání dat z webu bylo webovým robotem uloženo do databáze více než 700 000 záznamů. Práce obsahuje návrh několika typů grafů, které lze ze získaných dat vytvořit a dále vyhodnocovat. Grafy jsou tvořeny firmami a osobami do firem zainteresovanými. Využitím algoritmu PageRank a metod mír centrality jsou získaná data vyhodnocována a v závěru práce probíhá diskuze získaných výsledků.

Obsah

1 Úvod

Cílem práce je analýza propojení firem využitím PageRanku. Pro analýzu je nutné nalézt vazby mezi firmami v České Republice. Pomocí provázání firem budeme určovat význam a vliv firem na ostatní firmy. Analýza sítě firem bude prováděna využitím algoritmu PageRank a mírami centrality, jenž graf z těchto získaných dat vyhodnotí. Graf bude tvořen firmami, které budou představovat vrcholy grafu, a osobami do firem zainteresovanými, využitím kterých budou vytvářeny hrany či vazby mezi vrcholy grafu. Dále je možné vytvořit graf, kde osoby budou představovat vrcholy a pomocí firmy budou určovány vazby mezi vrcholy.

V první kapitole se seznámíme se základními metodami pro analýzu grafu. Konkrétně si blíže popíšeme algoritmus PageRank a nejpoužívanější metody pro výpočet mír centrality. Těmi jsou degree centrality, closeness centrality a betweenes centrality. Dále si představíme základní grafové metriky jako je poloměr grafu, hustota grafu a koeficient shlukování.

V další kapitole se zaměříme na výběr vhodného webu, ze kterého bude možné data pro vytvoření sítě firem získat. Firmy musí být na vybraném webu nějaký způsobem propojeny, abychom mohli pomocí těchto spojení vytvořit zmíněnou síť.

V následující kapitole se zamyslíme nad typy grafů, které bude možné ze získaných dat vytvořit. Poté, co si určíme, jaké typy grafů budeme ze získaných dat tvořit, si více specifikujeme typy grafů, které budeme dále chtít vyhodnocovat zmíněným algoritmem PageRank a mírami centrality.

V kapitole 5 se zaměříme na výběr a úpravu webového robota, využitím kterého získáme námi požadovaná data. Bude potřeba vybrat vhodnou implementaci některého z robotů, který bude splňovat naše požadavky na funkčnost a nastavení. Dále robota nastavíme, tak aby stáhnul námi vybraná data. Bude dobré zamyslet se nad otázkou jak daná data uložit, aby s nimi poté nebyla náročná práce.

V následující kapitole zmíníme vytvoření aplikace, pomocí které získaná data vyhodnotíme. Ze získaných dat budou vytvořeny všechny typy navržených grafů, které budou v kapitole 7 vyhodnoceny PageRankem a mírami centrality.

V předposlední kapitole se zaměříme přímo na vyhodnocení navržených typů grafů mírami centrality.

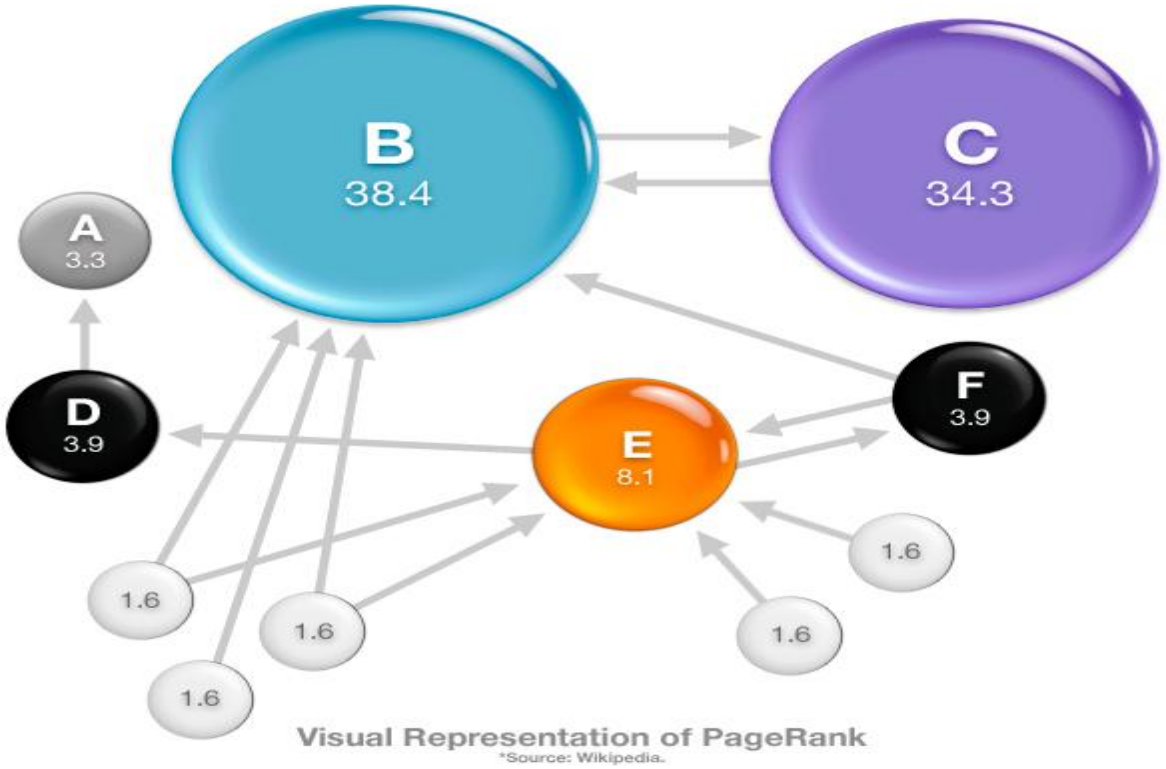
V poslední kapitole porovnáme výsledky vyhodnocení grafů algoritmem PageRank a mírami centrality. Následně pak toto porovnání a výsledky vyhodnocení budeme diskutovat.

2 Metody pro analýzu grafu

Tato kapitola popisuje metody pro analýzu grafu. Konkrétně je zde popsán algoritmus PageRank a míry centrality.

2.1 Algoritmus PageRank

Text v této části vychází ze zdroje [1]. Algoritmus PageRank byl jako první použit pro vylepšení full-textových vyhledávačů, které hodnotí webové stránky s ohledem na jejich hypertextové odkazy. PageRank je iterativní algoritmus a jeho koncept vychází z citační analýzy. Využívá ho například Google.com k řazení stránek při vyhledávání. Při řazení webových stránek je jedním z faktorů řazení počet odkazů na danou webovou stránku. Dalším faktorem je významnost webové stránky, ze které se odkazuje na danou webovou stránku. Webová stránka, na kterou odkazuje jiná významná stránka, je významnější než stránka, na kterou odkazuje např. 10 méně významných stránek. Tím je myšleno, že pokud na danou webovou stránku odkazuje stránka, která není významná, tak tento odkaz zvyšuje významnost dané webové stránky o malé procento. V opačném případě, kdy na danou webovou stránku odkazuje stránka, která významná je, pak tento odkaz zvyšuje významnost dané webové stránky více. Každý vrchol grafu má svou hodnotu, ohodnocení hran se řeší přidělením vah. V původní interpretaci PageRanku vrcholy představují webové stránky a hrany představují hypertextové odkazy mezi webovými stránkami (viz Obrázek 1). PageRank ve vyhledávačích nepoužívá při hodnocení webové stránky interní odkazy (samocitace). Interním odkazem je myšleno, že hypertextový odkaz vede na stránku, na které se nachází.



Obrázek 1 - Ukázka funkce PageRanku.

2.1.1 Matematický zápis

Pro snadnější pochopení a implementaci lze algoritmus PageRank popsat zápisem výpočtu pro jeden prvek. Pro popis lze dále využít maticový zápis, který je vhodný pro matematické zkoumání algoritmu.

V práci [1] autoři uvedli první vzorec PageRanku, který nebyl příliš použitelný. Byl jím vzorec (1), kde $PR_x(A)$ je skóre PageRanku vrcholu A v iteraci x, U je množina všech vrcholů odkazujících na vrchol A a N_u je počet výstupních hran vrcholu u.

$$PR_{x+1}(A) = \sum_{u \in U} \frac{PR_x(u)}{N_u} \tag{1}$$

Výsledné skóre PageRanku poté udává pravděpodobnost, s jakou se osoba, pohybující se po webu pomocí hypertextových odkazů, dostane na danou webovou stránku. Součet skóre PageRanku všech vrcholů nacházejících se v grafu je roven 1, resp. součet pravděpodobností přístupu na webové stránky je 100%.

Jak již bylo zmíněno výše, vzorec (1) nebyl příliš použitelný a potýkal se s různými problémy. Prvním problémem jsou tzv. *dangling nodes*[1], nebo-li vrcholů, které nemají výstupní hrany. V případě těchto vrcholů dochází ke ztrátě hodnoty PageRanku, čímž je zaviněno, že součet PageRanku všech vrcholů grafu přestává být roven 1. Obvykle se pro řešení toho problému používá rovnoměrné rozdělení, protože jiné způsoby řešení nejsou vhodné nebo nejsou spravedlivé. Rovnoměrné rozdělení znamená, že každému dangling nodes přidáme výstupní hranu na každý vrchol grafu (i na sebe sama).

Druhým problémem je *Rank Sink*[2], který vzniká, když je skupina vrcholů odkazována z vnější a vrcholy ve skupině odkazují pouze sami na sebe, ale už neodkazují vně skupiny (viz Obrázek 2).

Obrázek 1 - Rank Sink

Jako řešení Rank Sink byl navržen *Random surfar model* [2]. Při sledování chování reálných uživatelů internetu autoři [2] zjistili, že každý uživatel jednou za 7 kroků využívá tzv.

[1] Dangling nodes – volně přeloženo – problém visících uzlů/vrcholů
[2] Rank Sink – volně přeloženo – klesání hodnoceni

teleportu[3]. Proto do svého algoritmu zahrnuli toto užití teleportu s pravděpodobností 15%. Do algoritmu Page Rank byl Random surfar model přidán konstantou *d*, zvanou *Damping factor*[4]. Uživatel internetu tedy při procházení webových stránek s pravděpodobností *1-d* využije teleportu a s pravděpodobností *d* následuje hypertextové odkazy. Obvykle se tedy Damping factor nastavuje na hodnotu 0,85.

Po vyřešení těchto problémů je zde poslední nepřesnost ve vzorci (1), která se v prostředí internetu obvykle neprojeví. Tato nepřesnost vzorce (1) spočívá v tom, že váha každého hypertextového odkazu je stejná. Tedy pravděpodobnost užití každého z nich je stejná. Pro zvýhodnění některého z odkazů bychom museli do vzorce přidat váhy hran.

Tímto jsou vyřešeny nesrovnalosti a problémy, které obsahoval vzorec (1). Vzorec (2) je tedy kompletní vzorec PageRanku bez problémů s Dangling nodes, Rank Sink a s váhami hypertextových odkazů [3]. Nyní tedy máme výsledný vzorec PageRanku, viz vzorec (2), kde *d* je Damping factor, $|V|$ je suma všech vrcholů grafu, w_{utoA} je váha hrany vedoucí z vrcholu u do vrcholu *A*, w_{utout} je součet vah všech výstupních hran z vrcholu u a *D* je množina dangling nodes .

$$PR_{x+1}(A) = \frac{(1-d)}{|V|} + d \cdot (\sum_{u \in U} \frac{PR_x(u) \cdot w_{\mathrm{utoA}}}{w_{u\mathrm{out}}} + \frac{1}{|V|} \sum_{s \in D} PR_x(s)) \qquad (2)$$

2.1.2 Vzorec použitý v práci

Konkrétně pro můj program či výpočet jsem vybral ještě jednu úpravu vzorce PageRanku. Jedná se o *personalizaci*, která umožňuje, aby algoritmus PageRank v průběhu výpočtu některé vrcholy zvýhodnil. Pojem personalizace zavedli autoři [1], aby mohli do svých výpočtů zahrnout různé vlastnosti a potřeby uživatelů internetu. Výsledný vzorec, doplněný o personalizaci, je vzorec (3), kde P_a je hodnota personalizace pro vrchol A a *P* je vektor personalizací.

$$PR_{x+1}(A) = \frac{(1-d) \cdot P_A}{\sum_{p \in P} p} + d \cdot (\sum_{u \in U} \frac{PR_x(u) \cdot w_{\mathrm{utoA}}}{w_{u\mathrm{out}}} + \frac{1}{|V|} \sum_{s \in D} PR_x(s)) \qquad (3)$$

2.2 Míry centrality

Míry centrality (*Centrality measures*) jsou kolekcí metod, které umožňují měřit centralitu jednotlivých vrcholů grafu, která určuje, jak jsou jednotlivé vrcholy významné v rámci celého grafu. Vrcholy, jejichž centralita je vysoká, mají v grafu výhodnější pozici oproti vrcholům, jejichž centralita je nízká. Výhodnější pozicí je myšleno například to, že vrcholy s vysokou

[3] Teleport – uživatel internetu přejde na webovou stránku tak, že zadá přímo do prohlížeče URL adresu dané webové stránky

[4] Damping factor – volně přeloženo – faktor tlumení

centralitou mají možnost kontrolovat tok informací v grafu nebo mají možnost ovlivňovat ostatní vrcholy v grafu. V následujících podkapitolách si blíže popíšeme hlavní a zároveň nejpoužívanější metody mír centrality jak jsou popsány v [4, 11 a 12].

2.2.1 Degree centrality

Hodnota *degree centrality*[5] vyjadřuje počet přímých vazeb jednoho vrcholu k ostatním vrcholům v síti. V zásadě se jedná o měření aktivity jednotlivých vrcholů. V [4] je uvedeno, že centrální aktéři musí být nejaktivnější, protože mají nejvíce vazeb na ostatní aktéry v dané síti nebo grafu.

Obecný vzorec degree centrality pro jednotlivé vrcholy, viz vzorec (4), kde $C_D(i)$ je hodnota degree centrality vrcholu *i*, $d(i)$ je stupeň vrcholu *i* a *N* je celkový počet vrcholů v síti. Obecný vzorec degree centrality pro výpočet degree centrality celého grafu viz vzorec (4), kde C_D je hodnota degree centrality celého grafu, $C_D(i*)$ je maximální hodnota degree centrality v grafu, $C_D(i)$ je hodnota degree centrality vrcholu *i* a písmeno *N* ve jmenovateli je počet vrcholů v grafu. Jmenovatel zlomku ve vzorci (5) je normalizace pro případ, že by se v grafu vyskytoval jeden centrální vrchol, který by byl napojený na všechny ostatní vrcholy grafu.

$$C_D(\mathrm{i}) = \frac{\mathrm{d}(i)}{\mathrm{N}-1} \tag{4}$$

$$C_D = \frac{\sum_{\mathrm{i}=1}[C_D(i*) - C_D(i)]}{[(\mathrm{N}-1)(\mathrm{N}-2)]} \tag{5}$$

2.2.2 Closeness centrality

Hodnota *closeness centrality*[6] je nejvyšší, jestliže z vrcholu lze dosáhnout ke všem dalším vrcholům v síti přímou vazbou. Je to hodnota součtu nejmenších vzdáleností k ostatním vrcholům. Vrcholy, které mají vysokou hodnotu closeness centrality, mají velký vliv na to, co se v síti odehrává, protože mají nejrychlejší přístup k celé síti. V [4] se uvádí, že je to vzdálenost mezi jednotlivými aktéry, resp. jak blízko je jeden aktér k ostatním aktérům v síti. Obecný vzorec closeness centrality je vzorec (6), kde $C_C(i)$ hodnota closeness centrality vrcholu *i* a $d(i,j)$ je vzdálenost mezi vrcholy *i* a *j*.

$$C_C(\mathrm{i}) = \frac{1}{[\sum_{\mathrm{j}=1} \mathrm{d}(\mathrm{i},\mathrm{j})]} \tag{6}$$

[5] degree centrality - česky - centralita měřená stupněm uzlu

[6] closeness centrality - česky -centralita měřená blízkostí vrcholu ke všem vrcholům grafu

2.2.3 Betweeness centrality

Hodnota *betweenness centrality*[7] je nejvyšší pokud vazby mezi libovolnými dvěma vrcholy sítě vždy procházejí tímto vrcholem. Body s vysokou hodnotou betweeness centrality tak kontrolují tok informací v síti. V [4] je tento typ míry centrality popsán jako míra interakce mezi dvěma nesousedícími aktéry. Interakce mezi těmito aktéry závisí na ostatních aktérech v síti. Obecný vzorec betweeness centrality je vzorec (7), kde C_B je hodnota betweeness centrality, g_{jk} je celkový počet nejkratších cest z vrcholu j do vrcholu k a $g_{jk}(i)$ je počet těchto cest, které procházejí vrcholem i.

$$C_B(i) = \sum_{j<k} \frac{\mathrm{g}_{jk}(i)}{\mathrm{g}_{jk}} \tag{7}$$

2.2.4 Rozšíření mír centrality

Článek [5] popisuje rozšíření standardních mír centrality - Degree, Closeness, Betweenness - aby mohly být aplikovány na skupiny a třídy, stejně jako na jednotlivce. Skupinová míra centrality umožňuje vědcům odpovídat na různé otázky (např.: Jsou mezi středními manažery více centrální muži nebo ženy?). Díky těmto mírám lze řešit také opačné problémy (např.: Jak zformovat tým, který bude maximálně centrální s ohledem na vazby mezi členy organizace?). Dále je také možné formalizovat míru efektivity centrality skupiny, která ukazuje do jaké míry je centralita skupiny ovlivněna počtem svých členů.

2.3 Základní grafové metriky

2.3.1 Poloměr

Poloměr grafu, značený písmenem *D*, je obvykle roven ohodnocení pouze jedné cesty a to té nejdelší cesty z *cest nejkratších*[8] mezi všemi dvojicemi vrcholů v grafu. Pokud je graf rozdělen do více *komponent*[9], tak je vhodné uvažovat poloměr největší komponenty, protože poloměr takto rozděleného grafu je roven nekonečnu, tj. $D = \infty$.

2.3.2 Hustota

Hustota grafu je poměr existujících hran v grafu a všech možných hran v grafu. Vzorce pro její výpočet se liší podle toho, jestli se jedná o orientovaný graf (viz vzorec 8) nebo graf neorientovaný (viz vzorec 9),

$$\Delta = \frac{2\mathrm{m}}{\mathrm{n}(\mathrm{n}-1)} \tag{8}$$

[7] betweenness centrality - česky - centralita měřená počtem nejkratších cest, na kterých vrchol leží

[8] Nejkratší cesta – cesta mezi dvěma vrcholy, která má ze všech existujících cest mezi těmito vrcholy nejmenší ohodnocení

[9] Komponenta grafu – podgraf původního grafu (podmnožina vrcholů a hran původního grafu)

$$\Delta = \frac{m}{n(n-1)} \tag{9}$$

Kde *m* je počet hran v grafu a *n* počet vrcholů v grafu. Hodnota hustoty grafu se pohybuje v intervalu <0,1>.

2.3.3 Koeficient shlukování

Koeficient shlukování je hodnota průměru hustot skupin tvořenými vrcholy grafu. Skupinou vrcholů je v tomto případě myšlena podmnožina vrcholů grafu, které mezi sebou mají malou vzdálenost, ale naopak velký počet hran [6]. Popišme si to na příkladě. Mějme např. graf tvořený sedmi vrcholy. Vrcholy A, B, C a D tvoří jednu pomyslnou skupinu a vrcholy E, F a G tvoří druhou pomyslnou skupinu. Všechny vrcholy v první pomyslné skupině jsou mezi sebou navzájem spojené hranami a taktéž všechny vrcholy ve druhé pomyslné skupině. Skupiny jsou poté navzájem propojeny například pouze jednou či dvěma hranami. Tím se v jednom grafu, tvořeném sedmi vrcholu, utvoří dvě pomyslné skupiny vrcholů, jejichž vrcholy jsou shlukovány k sobě.

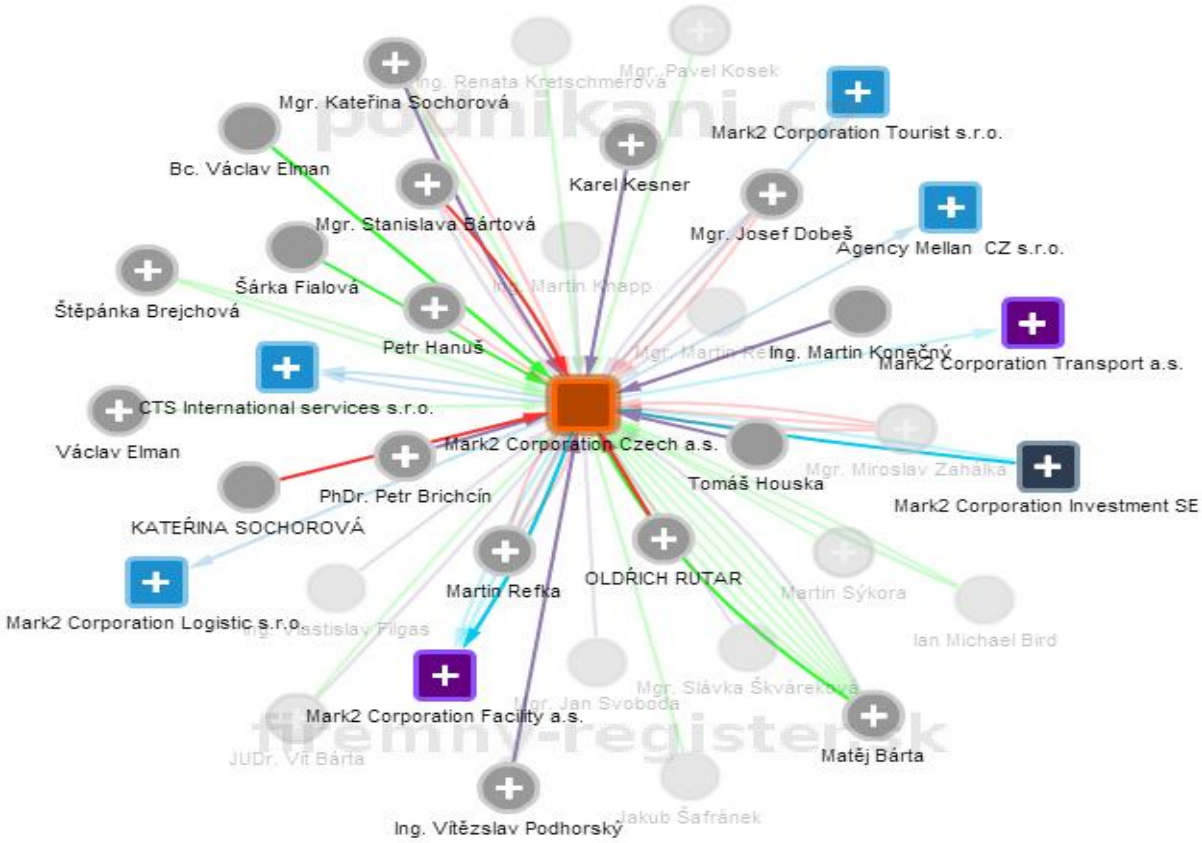
3 Vhodný web k získání informací pro tvorbu grafů

Při hledání vhodného webu, ze kterého bych mohl stahovat informace o firmách, které jsou základem pro tvorbu mé bakalářské práce, jsem navštívil několik webových stránek. Snažil jsem se najít webovou stránku, kde by byla data ve vhodné formě, aby bylo možné je získat využitím webového robota. Firmy a osoby by zde měli být uspořádány takovým způsobem, aby se robot mohl z jedné firmy či osoby dostat do dalších firem nebo k dalším osobám a získat o těchto subjektech informace.

Doporučena mi byla webová stránka http://obchodni-rejstrik.podnikani.cz. Vizualizace vztahů jednotlivých subjektů (firma, osoba) na této stránce je opravdu dobře prezentována (viz Obrázek 2). Při prvním prohlížení této stránky, jsem se příliš neorientoval v tom, jak jsou zde prezentovány vazby mezi jednotlivými subjekty (osoba nebo firma). Tak jsem se rozhodl se ještě poohlédnout po jiném vhodnějším webu. Našel jsem další dvě webové stránky, které se jevily jako vhodný zdroj dat.

Jedním z těchto webů byl http://www.databaze-firem.net. Tento web jsem nakonec nevybral z důvodu, že data je možné získat pouze za finanční poplatky. Druhým kandidátem byl https://or.justice.cz/ias/ui/rejstrik, ale ani tuto stránku jsem nakonec nevybral. Důvodem bylo to, že se z jednoho subjektu nedalo dostat na žádný další subjekt. Nebyl zde žádný hypertextový ani jiný odkaz.

Kvůli výše uvedeným nedostatkům jsem se rozhodl vrátit k původnímu webu http://obchodni-rejstrik.podnikani.cz. Po důkladnějším zkoumání jsem se s tímto prostředím seznámil a díky tomu jsem narazil na webovou stránku, která splňovala moje požadavky. Jako výchozí webovou stránku jsem vybral http://rejstrik-firem.kurzy.cz. Data budou z webu získávána postupně, a to využitím seznamu osob, které mají nejvíce zápisů v obchodním rejstříku. Procházením jednotlivých subjektů by se mělo docílit získání největšího možného množství dat.



Obrázek 2 - Ukázka sítě vztahů Mark2 Corporation Czech a.s..[10]

[10] http://obchodni-rejstrik.podnikani.cz/25719751/mark2-corporation-czech-as [cit. 26.1.2014]

4 Typy grafů

Při získávání dat z vybraného webu jsem uvažoval nad návrhem, co největší počtu typů grafů, které by se daly vytvořit z nabytých informací. Z informací na webu jsem vypozoroval, že je možné vytvořit dva základní typy grafů, a to:

- 1. typ grafu nese označení Graf firem. Tento graf vypadá následovně. Vrcholy grafu reprezentují firmy a podle určitého postavení osob v jednotlivých firmách se tvoří hrany grafu mezi těmito vrcholy. Firma A je tedy spojena s firmou B přes osobu, která je v obou těchto firmách zainteresovaná.
- 2. typ grafu jsem označil jako Graf osob. Graf je také tvořen za pomoci subjektů firma a soba. Vrcholy grafu jsou v tomto případě osoby a hrany se tvoří podle určitého postavení osob v jednotlivých firmách. 1. osoba je spojena s 2. osobou přes firmu, kde je 1. i 2. zainteresovaná.

Tyto dva typy grafů jsou pouze základem pro tvorbu dalších, specifičtějších typů grafů. Pro každý ze dvou základních typů grafů jsem dále navrhl tři vážené grafy. Tyto grafy jsem označil takto: *Graf A, Graf B a Graf C.* Z výše uvedených informací vyplývá, že vznikne 6 různých grafů (pro každý základní typ grafu vzniknou 3 grafy). Rozdělení typů grafů je provedeno podle toho, jestli chci, aby nejlépe ohodnocený vrchol grafu byl nejvíce ovlivňující (Graf A) nebo nejvíce ovlivnitelný (Graf B). Toho docílíme přidělováním vah jednotlivým hranám grafu spojující vrcholy grafu (viz část 4.3). Graf C slouží pro srovnání Grafu A a Grafu B, jelikož u tohoto typu grafu neuvažujeme váhy hran, respektive každá hrana v Grafu C má přidělenou stejnou váhu, a to váhu 1. Nyní si více popíše navržené typy grafů v základních typech grafů.

4.1 Graf firem

V tomto typu grafu firmy představují vrcholy grafu a využitím osob se vytvářejí hrany grafu, které firmy vzájemně propojují.

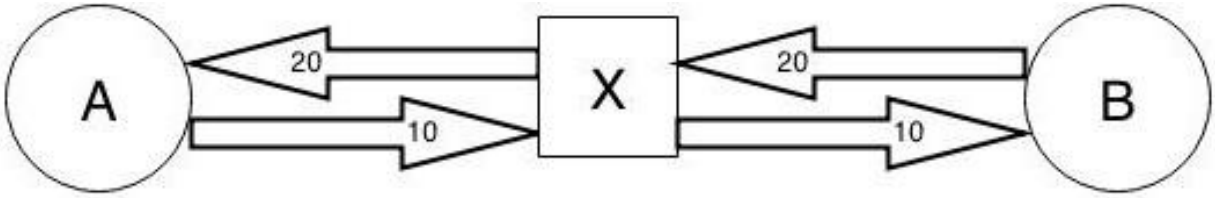
4.1.1 Graf A

Hodnoty vah **vstupních** hran do jednotlivých vrcholů (firmy) budou určeny podle postavení osob v daných firmách. Hodnota váhy vstupní hrany do vrcholu grafu odpovídá postavení osoby ve firmě (viz část 4.3).

Největší hodnotu PageRanku v tomto grafu bude mít vrchol, respektive firma, která nejvíce ovlivňuje ostatní firmy v tomto grafu. V algoritmu PageRank toho docílíme ohodnocením hran, kde váhu hrany určuje postavení jednotlivých osob ve firmách. Z tohoto důvodu bylo nutné získat z webu informace o postavení osob ve firmách. Podrobně si tento typ sítě rozebereme na následujícím příkladu.

Představme si modelovou situaci, kde máme dvě firmy A, B spojené přes jednu osobu X. Osoba X má ve firmě A postavení člena představenstva a ve firmě B zaujímá pozici člena dozorčí rady. V tomto příkladě bude mít hrana z firmy A do firmy B váhu 10 a hrana z firmy B do firmy A váhu 20 (viz Obrázek 3). Důvodem je to, že osoba X zastává ve firmě A vyšší

pozici než ve firmě B. Pokud budou mít oba vrcholy více výstupních hran, tak algoritmus PageRank zvýhodní firmu A.



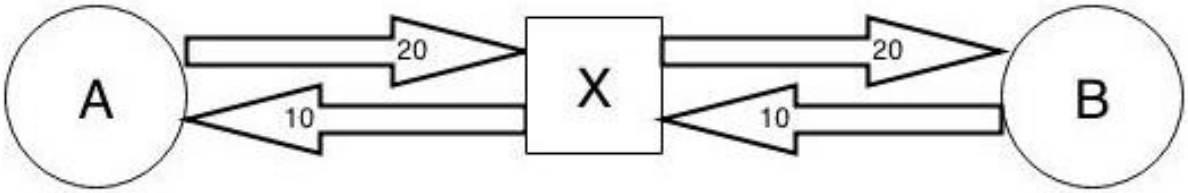
Obrázek 3 - Graf firem A - ohodnocení hran a ukázka propojení

4.1.2 Graf B

Hodnoty vah **výstupních** hran z jednotlivých vrcholů (firmy), budou určeny podle postavení osob v daných firmách. Hodnota váhy výstupní hrany z vrcholu grafu odpovídá postavení osoby ve firmě (viz část 4.3).

V tomto grafu bude mít největší hodnotu PageRanku ten vrchol, respektive firma, která bude nejvíce ovlivnitelná z pohledu ostatních firem v tomto grafu. Obdobně jako v grafu A toho docílíme ohodnocením hran grafu. Váhy hran budou opět určovány podle postavení osob ve firmách. Znovu si to rozebereme na příkladu.

Představíme si stejnou modelovou situaci jako v předešlém typu grafu A. Máme tedy firmy A a B spojené osobou X. Postavení osoby X zůstává stejné, ve firmě A je tedy členem představenstva a ve firmě B členem dozorčí rady. Nyní ovšem bude mít hrana z firmy A do firmy B váhu 20 a naopak hrana z firmy B do firmy A váhu 10 (viz Obrázek 4). Osoba X sice zastává stejné pozice jako v předešlém příkladě, ale tentokrát pomocí algoritmu PageRank chceme zvýhodnit tu firmu, která by měla být více ovlivnitelná jinými firmami v grafu.



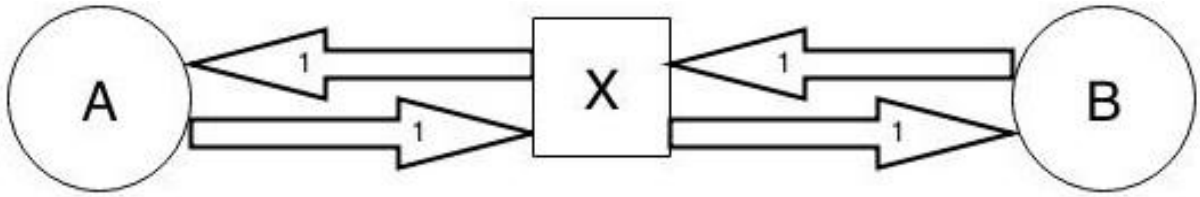
Obrázek 4 - Graf firem B - ohodnocení hrana a ukázka propojení

4.1.3 Graf C

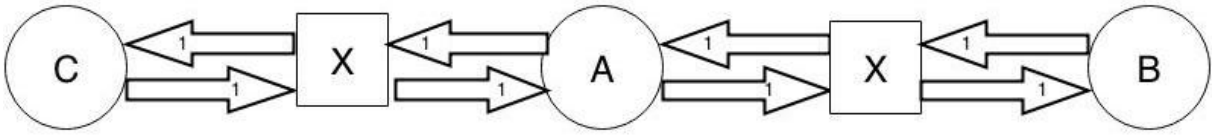
V tomto typu grafu budou mít jednotlivé hrany stejné ohodnocení, takže už nebude záležet na postavení osob ve firmách. Hodnoty vah hran nebudou ovlivňovat výpočet PageRanku pro daný vrchol, tak jak jej ovlivňovali v grafu A a v grafu B. Pokud tedy uvažujeme stejné ohodnocení hran, tak není určení vrcholu, který získá největší hodnotu PageRanku, jednoznačně předvídatelné. Modelový příklad bude vypadat takto.

Jako již v předešlých grafech si představme firmu A a firmu B spojené osobou X. Nyní tedy nezáleží na postavení osoby X v jednotlivých firmách, protože váhy hran jsou v tomto typu grafu stejné, tj. rovné 1. Hrana z firmy A do firmy B bude mít váhu 1 a stejně tomu bude

naopak. Hrana z firmy B do firmy A bude mít také váhu 1 (viz Obrázek 5). To znamená, že v tomto případě nebude zvýhodněna ani jedna z firem. Ke zvýhodnění by došlo v případě spojení jedné z firem s firmou další (novou). Firma A by byla spojena s firmou C buďto přes osobu X nebo přes osobu Y. V tomto případě, by byla firma A zvýhodněna oproti firmám B a C (viz Obrázek 6).

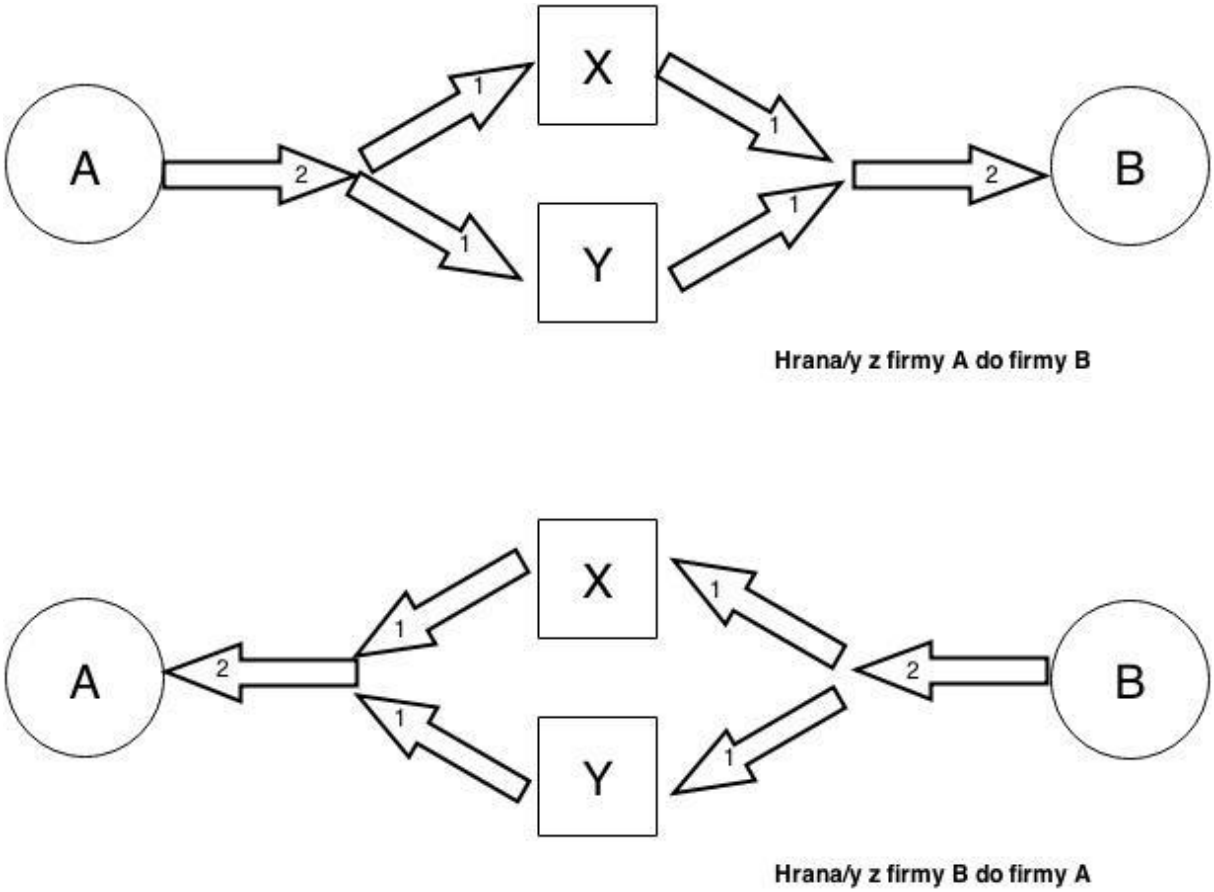


Obrázek 5 - Graf firem C - ohodnocení hran a ukázka propojení



Obrázek 6 - Graf firem C - Přidána firma C, ohodnocení hran a ukázka propojení

V grafu C bylo nutné vyřešit otázku toho, jak bude graf vypadat v případě, že např. firmy A a B budou spojovat dvě hrany, respektive budou existovat osoby X a Y, které budou zainteresovány ve stejných firmách. To znamená, že by mezi firmami A a B vznikly dvě hrany, respektive firma A by měla dvě výstupní hrany do firmy B a firma B by měla dvě výstupní hrany do firmy A. Tato otázka se dala řešit více způsoby. Jedním ze způsobů řešení této otázky je spojení těchto „duplicitních“ hran v jednu s váhou 1. V grafu by, místo několika stejných hran, byla hrana pouze jedna. Já jsem si ale vybral jiný způsob. Rozhodl jsem se, že tyto hrany spojovat v jednu nebudu, ale ponechám je v grafu zachovány všechny. To znamená, že pokud je mezi dvěma firmami více hran, pak tyto hrany prakticky reprezentují hranu jednu, ale už nikoli s váhou 1, ale s váhou přímo úměrnou počtu těchto hran (viz Obrázek 7).



Obrázek 7 - Graf firem C - Osoby X, Y současné ve firmách A, B

4.2 Graf osob

Tento typ grafu je navržen tak, že vrcholy grafu reprezentují osoby a hrany grafu jsou tvořeny podle určitého postavení osob ve firmách. Osoby jsou vzájemně propojeny, pouze pokud zastávají nějaké pozice ve stejné firmě. Modelové příklady sítí v této části, jsou analogické jako v části 4.1. Jedinou změnou je, že osoby jsou nyní dvě X, Y a firma jedna A.

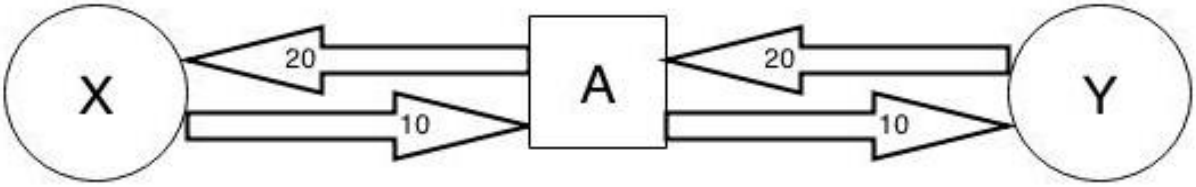
4.2.1 Graf A

Hodnoty vah **vstupních** hran do jednotlivých vrcholů (osoby), budou určeny podle postavení osob v daných firmách. Hodnota váhy vstupní hrany do vrcholu grafu odpovídá postavení osoby ve firmě (viz část 4.3).

Největší hodnotu PageRanku získá v tomto typu grafu ta osoba, která bude nejvíce ovlivňovat ostatní osoby. Ohodnocení hran bude docíleno znovu pomocí postavení jednotlivých osob ve firmách.

Jako modelový příklad si představme, že existují osoby X a Y, které jsou spojeny firmou A, ve které jsou obě osoby zainteresovány, respektive ve firmě A zaujímají nějaké postavení. Pokud je osoba X členem představenstva a osoba Y členem dozorčí rady, tak hrana od osoby X k osobě Y bude mít váhu 10. Tudíž naopak hraně od osoby Y k osobě X přidělíme váhu 20, protože osoba X zaujímá vyšší či významnější pozici než osoba Y (viz Obrázek 8). Pokud

mají obě osoby více výstupních hran, tak PageRank zvýhodňuje osobu X, tedy osobu která více ovlivňuje osobu jinou.



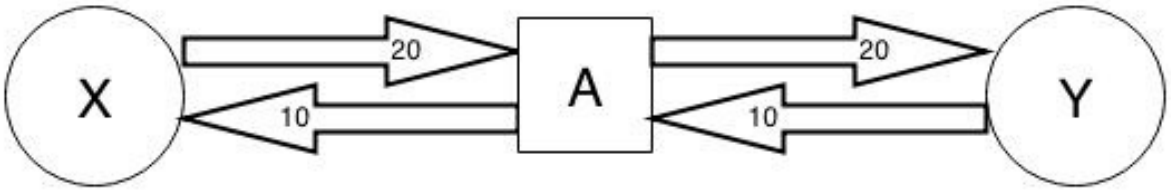
Obrázek 8 - Graf osob A - ohodnocení hran a ukázka propojení

4.2.2 Graf B

Hodnoty vah **výstupních** hran z jednotlivých vrcholů (osoby), budou určeny podle postavení osob v daných firmách. Hodnota váhy výstupní hrany z vrcholu grafu odpovídá postavení osoby ve firmě (viz část 4.3).

Vrchol s největší hodnotou PageRanku bude vrchol, respektive osoba, která bude nejvíce ovlivnitelná z pohledu ostatních osob. Hrany budou opět ohodnoceny podle postavení jednotlivých osob ve firmách.

Pro vysvětlení si uvede následující příklad. Máme dvě osoby X a Y, které jsou zainteresovány ve firmě A a jsou tedy přes tuto firmu spojeny. Osoba X zastává firmě A pozici člena představenstva a osoba Y člena dozorčí rady. V tomto typu grafu, tedy hrana od osoby X k osobě Y dostane váhu 20 a naopak hrana od osoby Y k osobě X bude mít váhu 10 (viz Obrázek 9). V případě, že z osob X a Y povede více výstupních hran, tak pomocí algoritmu PageRank zvýhodníme osobu Y, která bude více ovlivnitelná osobou X.



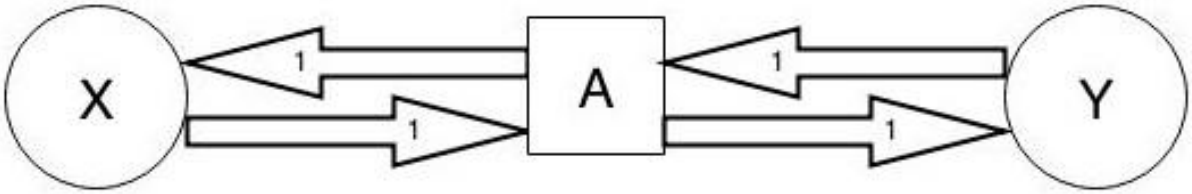
Obrázek 9 - Graf osob B - ohodnocení hran a ukázka propojení

4.2.3 Graf C

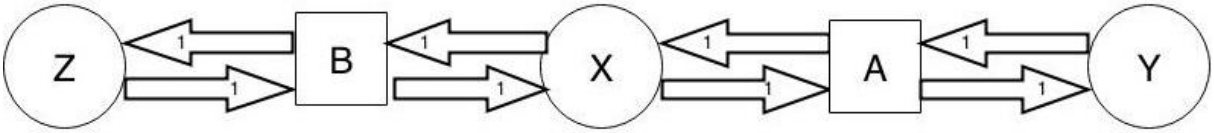
Všechny hrany v grafu budou mít stejné ohodnocení, takže váha každé hrany bude rovna 1. Nezáleží na postavení osob ve firmách. Hodnoty vah hran nebudou ovlivňovat výpočet PageRanku pro daný vrchol, tak jak jej ovlivňovali v grafu A a v grafu B. Při stejném ohodnocení hran není určení vrcholu, který získá největší hodnotu PageRanku, jednoznačně předvídatelné.

Opět si tento graf popíšeme na modelovém příkladě. Mějme dvě osoby X a Y zastávající určité pozice ve firmě A. Nezáleží na tom jaké pozice osoby ve firmě A zastávají. Hrany mezi těmito osobami budou mít stejnou váhu, a to váhu 1 (viz Obrázek 10). PageRank v tomto

případě žádnou z osob nezvýhodňuje, respektive zvýhodňuje obě osoby stejně. Rozdíl by nastal při situaci, že jedna z osob X nebo Y by byla spojena s další osobou, např. s osobou Z (viz Obrázek 11). Pak by tato osoba byla zvýhodněna proti zbylým dvěma osobám.

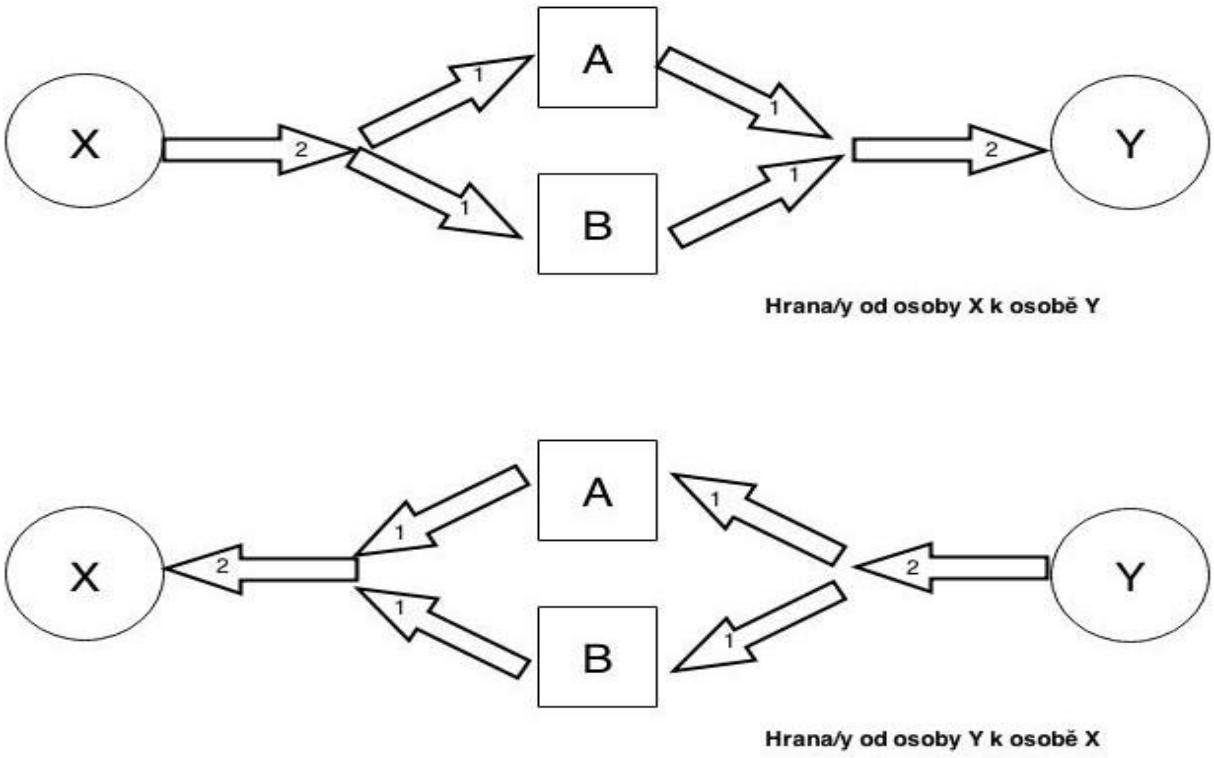


Obrázek 10 - Graf osob C - ohodnocení hran a ukázka propojení



Obrázek 11 – Graf osob C – Přidána osoba Z, ohodnocení hran a ukázka propojení

V grafu C bylo nutné vyřešit otázku toho, jak bude graf vypadat v případě, že např. osoba X a osoba Y budou obě současně zainteresovány ve firmě A a zároveň ve firmě B. To znamená, že by mezi osobami X a Y vznikly dvě hrany, respektive osoba X by měla dvě výstupní hrany k osobě Y a osoba Y by měla dvě výstupní hrany k osobě X. Tato otázka se dala řešit více způsoby. Jedním ze způsobů řešení této otázky je spojení těchto „duplicitních“ hran v jednu s váhou 1. V grafu by, místo několika stejných hran, byla pouze hrana jedna. Já jsem si ale vybral jiný způsob. Rozhodl jsem se, že tyto hrany spojovat v jednu nebudu, ale ponechám je v grafu zachovány všechny. To znamená, že pokud je mezi dvěma osobami více hran, pak tyto hrany prakticky reprezentují hranu jednu, ale už nikoli s váhou 1, ale s váhou přímo úměrnou počtu těchto hran (viz Obrázek 12).



Obrázek 12 - Graf C - Osoby X, Y současné ve firmách A, B

4.3 Hodnoty vah hran v grafech

Hodnoty vah hran v grafech jsou určeny podle postavení jednotlivých osob ve firmách. Největší váhu má osoba, která zaujímá ve firmě největší postavení z hlediska manažerského rozhodování. Osoby, které se starají o dozor nad chodem firmy, mají přidělenou menší váhu než osoby v manažerských pozicích. Nejnižší váhu mají přidělenou osoby na pozici likvidátor. Likvidátor je totiž jmenován až poté, co se firma dostane do stavu likvidace. Kompletní výčet hodnot vah, které se později přidělují hranám grafu, naleznete níže (viz Tabulka 1).

Tabulka 1 - Tabulka hodnot přidělených vah jednotlivým pozicím

Pozice, kterou osoba zastává ve firmě	Ohodnocení pozice
Ředitel	23
Člen představenstva	20
Akcionář, Komplementář, Komanditista, Společník s vkladem, Zakladatel	15
Správní rada	12
Prokurista, Odpovědný zástupce	11
Člen dozorčí rady, Revizor, Zřizovatel nadace	10
Vedoucí oddělení	8
Likvidátor	5

5 Úprava personalizačního vektoru PageRanku

Úprava personalizačního vektoru PageRanku by měla sloužit k vyhodnocení grafu firem či zainteresovaných osob. Pro každý ze dvou typů grafů jsem navrhl jinou úpravu personalizačního vektoru. Zde si uvedeme úpravy pro oba typy grafů.

5.1 Úprava pro Graf firem

Pro tento typ grafu se nabízela snad jediná úprava personalizačního vektoru PageRanku. K úpravě personalizačního vektoru byl použit *kapitál*[11] jednotlivých firem. Většina firem uvádí výši svého kapitálu na webu, ze kterého jsme získávali data pro tvorbu grafů. Nebylo složité rozšířit počet informací o jednotlivých firmách a získat tak jejich výši kapitálu. Menším problémem bylo, že některé firmy výši kapitálu neuvádějí. V tomto případě, jsem se rozhodl těmto firmám přidělit kapitál ve výši 100 000 Kč. K tomuto rozhodnutí mě vedla skutečnost, že pokud jde o jednotlivé *právní formy podnikání*[12], tak každá z nich musí, ke dni založení, složit základní kapitál v rozdílných výších. Nechtěl jsem poškodit ty firmy, které informaci o výši kapitálu bez problému uvádějí. Proto jsem se snažil přidělit firmám, které výši svého kapitálů neuvádějí, průměrnou výši kapitálu, kterou musí jednotlivé právní formy podnikání skládat ke dni založení.

Jako úpravu personalizačního vektoru PageRanku v tomto typu grafu jsem se rozhodl použit kapitál firem. Tato možnost úpravy se nabízela jako nevhodnější a prakticky jediná možnost, jelikož další dostupné informace na webu nebyly takto vhodné.

K úpravě personalizačního vektoru PageRanku využitím výše kapitálu firem mě vedla představa, že firmy disponující vysokou výší kapitálu budou mít vetší vliv na ostatní firmy. V případě grafu B bude výše kapitálu jednotlivých firem spíše snižovat míru ovlivnitelnosti dané firmy ostatními firmami.

5.2 Úprava pro Graf osob

Pro graf osob nebylo jednoduché vymyslet úpravu personalizačního vektoru PageRanku. Po úvaze nad tím, co by bylo vhodné použít, ze získaných dat z webu, jako úpravu personalizačního vektoru v případě grafu osob, se nabízela jediná možnost. Touto možností bylo využití ohodnocení pozic osob v jednotlivých firmách (viz část 4.3). V praxi to znamená, že pokud je osoba (vrchol grafu) zainteresována ve dvou firmách, např. ve firmě A a ve firmě B, pak by součet ohodnocení pozic, na kterých osoba působí ve firmě A a ve firmě B, byl úpravou personalizačního vektoru PageRanku pro tento typ graf. Blíže si vysvětlíme na modelovém příkladě.

Osoba X zaujímá ve firmě A pozici člena představenstva (ohodnocení pozice je 20) a ve firmě B pozici člena dozorčí rady (ohodnocení pozice je 10). V tomto případě by personalizace

[11] Kapitál - značné nashromážděné jmění nebo velké množství hotových peněz, bohatství

[12] Právní formy podnikání – společnost s ručením omezeným (s.r.o.), komanditní společnost (k.s.), akciová společnost (a.s.) a veřejná obchodní společnost (v.o.s.), družstvo, sdružení podnikatelů, nadace, státní podnik (s.p.) a osoba samostatně výdělečně činná (OSVČ)

vrcholu (osoba X) byla upravena na hodnotu 30, protože součet pozic, na kterých působí osoba X, je 30.

Tato úprava personalizačního vektoru PageRanku se nabízela jako jediné řešení použitelné ze získaných dat. Myslím si, že tato úprava je pro graf osob vhodná jelikož, osoby, které působí na lépe ohodnocených pozicích (viz část 4.3), budou mít větší vliv na ostatní osoby v grafu. Naopak osoby, které působí na pozicích s horším ohodnocením, budou mít menší vliv na ostatní osoby v grafu, ale budou více ovlivnitelné.

6 Stahování obsahu webových stránek

Tato kapitola obsahuje základní informace o webových robotech a informace o získávání a ukládání dat z webu pomocí webového robota. Dále se zmíním o tom, podle čeho jsem webového robota vybíral, respektive jaké jsem měl požadavky při výběru.

6.1 Webový robot

V [8] je popsán *crawler*[13] takto, cituji: *„Základní algoritmus, na kterém crawler pracuje, je jednoduchý. Stáhnou se všechny stránky na známých URL, z nich se vyextrahují další URL, a tento postup se může libovolně-krát opakovat“*. Web se neustále mění a rozšiřuje, proto stáhnout všechny známé URL je dost obtížné.

Dále musí crawler plnit určitá pravidla pro vyhledávání, např. jak často danou stránku navštěvovat a jak často z ní obnovovat data. Nástroje využívané k získávání dat mají určitá omezení např. *bandwidth*[14] a vlastní paměť nástroje.

Základní informace o tom, co to webový robot je a jak s webovými roboty pracovat jsou dostupné v [10].

6.2 Vhodný webový robot

Při výběru vhodného webové robota mi byl doporučen robot Scrapy. Scrapy se používá k procházení webových stránek a k extrahování strukturovaných dat z těchto stránek. Scrapy je naprogramován v jazyce Python. Po stažení, konfiguraci PC pro jazyk Python a instalaci robota Scrapy je, s pomocí vlastního zdrojového kódu, možné jej libovolně upravit a nastavit pravidla pro stahování a následné ukládání dat. Po seznámení se s tímto již implementovaných řešením, jsem se rozhodl, že robot Scrapy bude vhodná volba, protože splňoval mnou požadované vlastnosti:

- Možnost ukládat data do databáze.
- Nastavení robota vlastním zdrojovým kódem.
- Přítomnost DOM[15] parseru nebo jiné možnosti výběru části HTML stránky, která bude ukládána.
- Možnost přerušit stahování a při opětovném spuštění pokračovat tam, kde se skončilo.
- Nastavení doby prodlevy mezi stahováním jednotlivých stránek.

6.3 Tvorba webového robota

Základem pro tvorbu mého webového robota byl framework Scrapy (v. 0.20.1) [9].

6.3.1 Databázová struktura a ukládání dat

Pro ukládání dat jsem se rozhodl využít relační databázi. Vedly mě k tomu tyto důvody:

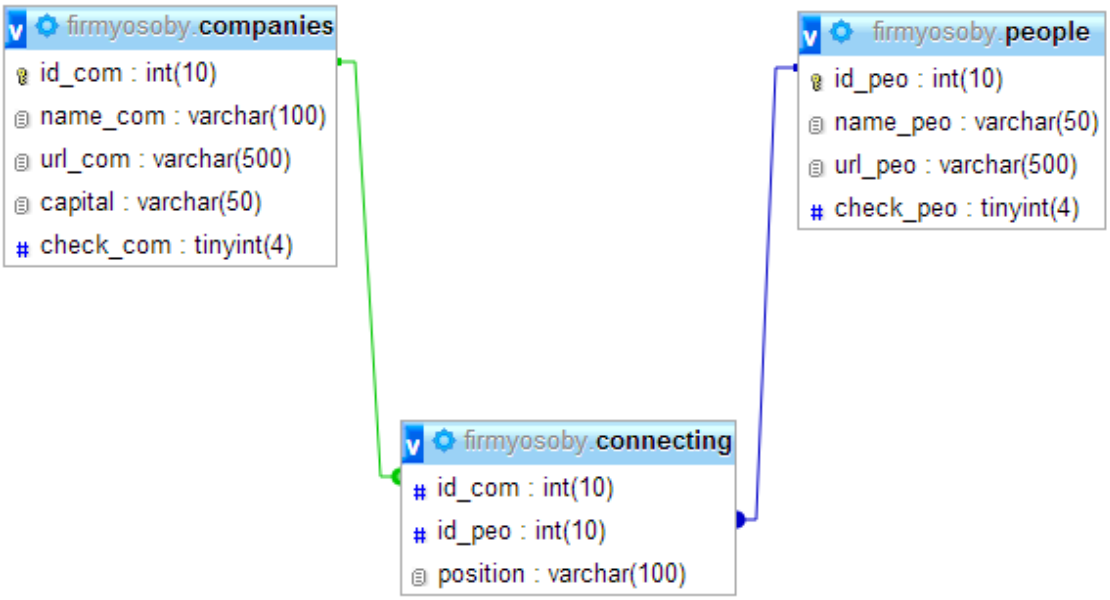
[13] Crawler – označení pro webové roboty
[14] Šířka pásma
[15] Document Object Model (DOM) – pomocí stromové struktury umožňuje přistupovat k datům uloženým v XML nebo HTML souborech, viz http://www.w3.org/DOM.

- Objem dat je, oproti XML souborům, menší.
- Nehrozí ztráta dat, protože jsou ukládána postupně.
- Poskytuje dostatečnou rychlost práce s daty.
- Lze snadno a rychle vyhledávat potřebná data.

Cílovou databází jsem zvolil MySQL databázi, která je používána v rámci programové kolekce EasyPHP[16] (v. 12.1). Tato databáze bude provozována na lokálním stroji. EasyPHP je šířeno pod GPL[17] licencí a nabízí dostatečnou výkonnost. Díky využití nástroje phpMyAdmin nabízí příjemné uživatelské rozhraní. Robot byl na databázi napojen pomocí knihovny MySQLdb[18].

V databázi jsem navrhl tři tabulky (viz Obrázek 13). Kardinalitu mezi tabulkami *people* a *companies* jsem zvolil N:M, protože člověk může zastávat určité pozice ve více firmách a v rámci jedné firmy může na více pozicích působit více lidí. Spojovací tabulka *connecting* uchová spojení mezi jednotlivými firmami a osobami. Třetím sloupcem v tabulce connecting je pozice, na které daná osoba působí v dané firmě.

Tabulky *people* a *companies* mají jako poslední atribut *příznak zpracování*, který říká, jestli už byla daná osoba nebo firma zpracována či nikoli. Pokud má tento příznak hodnotu 0, tak osoba nebo firma ještě zpracována nebyla a pokud mám hodnotu 1, tak daný záznam v tabulce již zpracován byl.



Obrázek 13 - E-R-A model databáze

[16] http://www.easyphp.org/
[17] GNU General Public License – lincence pro svobodný software (viz http://www.gnu.org)
[18] http://sourceforge.net/projects/mysql-python/

6.3.2 Riziko odpojení

Kvůli hrozbě odpojení robota od portálu, na který přistupuje, v důsledku nadměrného počtu přístupů na webové stránky portálu, jsem zvolil prodlevu mezi stahováním jednotlivých stránek jednu vteřinu. Toto opatření by mělo zvýšit pravděpodobnost, že robot nebude odpojen. V případě odpojení robota, by tato prodleva mezi stahováním měla zajistit, že stažených dat bude dostatek.

Robota je možné v libovolnou chvíli pozastavit a při dalším spuštění zajistit, aby pokračoval tam, kde skončil. Případně můžeme u robota nastavit, po kolika záznamech se bude síť firem a osob prohledávat. Pokud tuto proměnnou nastavíme např. na 1000 záznamů, tak si robot načte těchto 1000 záznamů, respektive maximálně 1000 záznamů s příznakem zpracování 0, do paměti a poté zpracovává a prohledává dané webové stránky.

6.3.3 Získávání potřebných dat

Všechny profilové[19] stránky osob mají stejnou strukturu, taktéž profilové stránky firem. Jediná odlišnost struktury stránky mezi jednotlivými firmami je v tom, že některé firmy neuvádějí na svých profilových stránkách výši kapitálu. Buď to jsou neziskové organizace[20] nebo se tyto firmy rozhodly informaci o výši kapitálu nepodávat.

Jelikož se v případě těchto stránek jedná o relativně malý HTML dokument, tak bude vhodné pro procházení jednotlivých stránek použít DOM parser. Potřebná data poté získáme využitím *XPath*[21] cest. Správnou XPath cestu nebylo jednoduché určit, ale nakonec se mi to podařilo i bez použití nástrojů pro určení XPath cest.

6.3.4 Strategie procházení

Jako výchozí body, ze kterých začne robot celou síť osob a firem procházet, jsem určil osoby, které mají nejvíce zápisů v obchodním rejstříku. Webový portál, ze kterého se získávala data, umožňuje zobrazit seznam osob, které mají více než 50 zápisů v obchodním rejstříku. Jinak řečeno, každá osoba v tomto seznamu je zainteresována v minimálně 50 firmách.

Díky tomuto řešení by se měly získat informace o co možná největším počtu firem, osob a samozřejmě spojení mezi nimi. Postupným procházením profilových stránek firem a osob získáme další užitečné informace o firmách, jako je dříve zmíněný kapitál, v případě osob pak pozice, kterou zastávají v jednotlivých firmách.

6.3.5 Které stránky procházet?

Profily osob

Každá fyzická osoba, která má alespoň jeden zápis v obchodním rejstříku, dostane na portále http://rejstrik-firem.kurzy.cz unikátní profilovou stránku na adrese **http://rejstrik-firem.kurzy.cz/osoby/jmenoOsoby/jmenoOsoby+IDOsoby**. Na této adrese najdeme informace o tom, v jakých firmách je tato osoba zainteresována a jaké pozice v nich zastává.

[19] Profil – webová stránka o určité osobě nebo firmě
[20] Nezisková organizace – společnost (právnická osoba), jejímž účelem není vytvářet zisk
[21] Viz http://www.w3school.com/xpath

Profily firem

Každá právnická osoba či firma, která je zapsána do obchodního rejstříku, dostane na již zmíněném portále http://rejstrik-firem.kurzy.cz unikátní profil na adrese **http://rejstrik-firem.kurzy.cz/IDFirmy/jmenoFirmy**. Na tomto profilu jsou základní informace o dané firmě. Důležitější stránkou pro získání informací je stránka **http://rejstrik-firem.kurzy.cz/IDFirmy/jmenoFirmy/vztahy**, kde jak už URL adresa napovídá, jsou k nalezení veškeré vztahy dané firmy s osobami, které jsou v ní zainteresovány.

6.3.6 Rozdělení robotů a popis algoritmů

Pro procházení sítě firem a do nich zainteresovaných osob vznikli celkem tři roboti.

- Start-Bot – sloužící pouze pro načtení výše zmíněného seznamu osob s více než 50 zápisy v obchodním rejstříku do databáze.
- Osoby-Bot – sloužící pro procházení profilových stránek osob. Řeší načtení firem, ve kterých jsou osoby zainteresovány, do databáze (tabulka companies) a vytvoření spojení mezi nově načtenými firmami a osobami (tabulka connecting).
- Firmy-Bot – sloužící pro procházení profilových stránek firem. Řeší zjištění výše kapitálu jednotlivých firem a uložení nových osob do databáze (tabulka people).

Robot Start-Bot vyextrahuje určitá data z výchozích stránek a uloží tato data do databáze. Data představují seznam osob, které mají více než 50 zápisů v obchodním rejstříku.

Algoritmus nejdůležitějších akcí robota Osoby-Bot:

1. Načte do pole z databáze ID osoby a URL adresy osoby těch záznamů (osob), které mají příznak zpracování nastaven na 0. Můžeme nastavit maximální načtený počet těchto záznamů v paměti.
2. Vybere první záznam v poli (ID osoby, URL adresa osoby).
3. Stáhne daný web (URL adresa osoby) a vyextrahuje z něj požadovaná data.
4. Uloží do databáze spojení vybrané osoby s firmami a případně uloží do databáze nové firmy z profilové stránky vybrané osoby.
5. Odstraň zpracovávaný záznam z pole.
6. Pokud je pole prázdné, tak jdi na bod 7. Pokud prázdné není, tak jdi na bod 2.
7. Ukonči program.

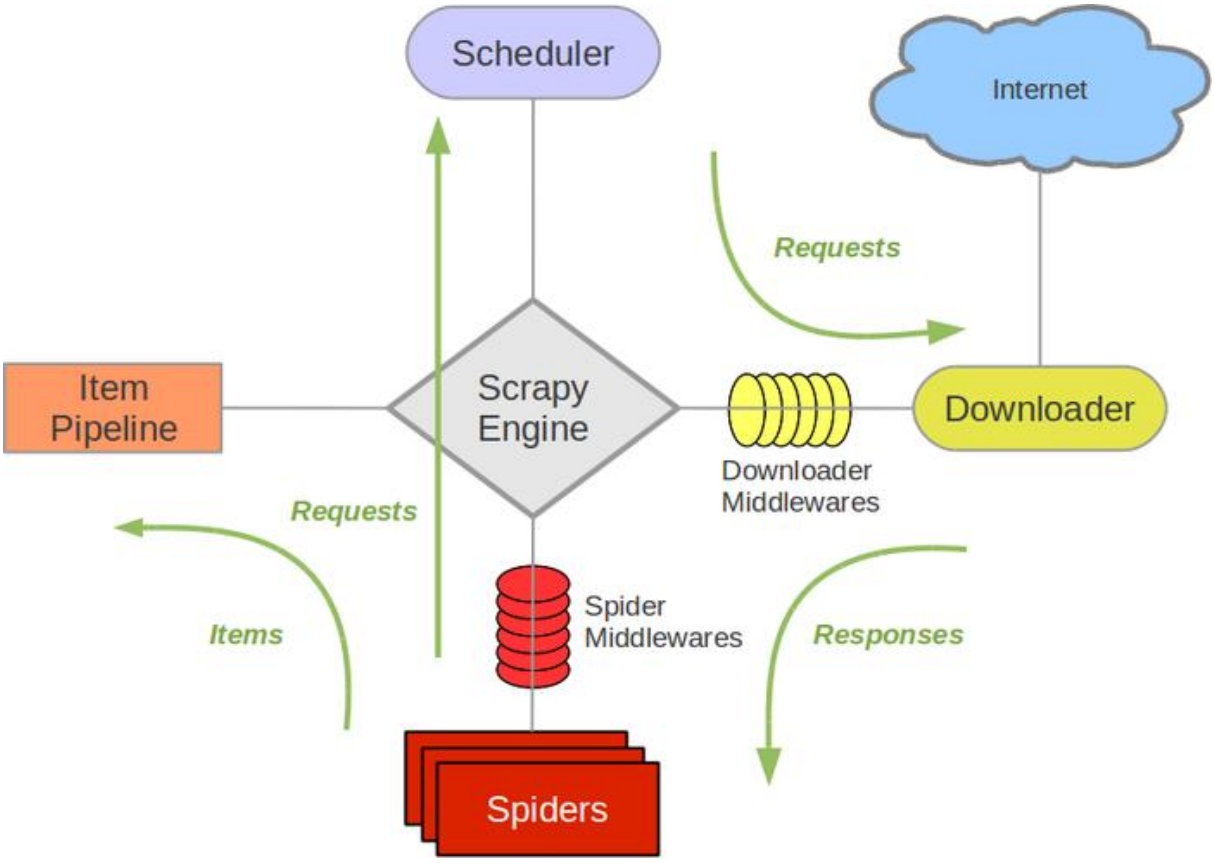
Algoritmus nejdůležitějších akcí robota Firmy-Bot:

1. Načte do pole z databáze ID firmy a URL adresu firmy těch záznamů (firem), které mají příznak zpracování nastaven na 0. Můžeme nastavit maximální načtený počet těchto záznamů v paměti (poli).
2. Vybere první záznam v poli (ID firmy, URL adresa firmy).
3. Stáhne daný web (URL adresa firmy) a vyextrahuje z něj požadovaná data.
4. Uloží do databáze informaci o výši kapitálu právě procházené firmy a případně uloží do databáze nové osoby z profilové stránky procházené firmy.
5. Odstraň zpracovávaný záznam z pole.
6. Pokud je pole prázdné, tak jdi na bod 7. Pokud prázdné není, tak jdi na bod 2.

7. Ukonči program.

6.3.7 Krátké poznámky z implementace frameworku Scrapy

V této části práce si uvedeme několik specifických konstrukcí frameworku Scrapy. Úplné podrobnosti o implementaci jsou k nalezení v oficiální dokumentaci projektu Scrapy, viz [3].



Obrázek 14 - Architektura webových robotů na frameworku Scrapy[22]. Obrázek převzat z [9].

U robotů je nám umožněno nastavit pravidla, díky kterým lze obsah stránek zpracovávat a přiřazovat stránkám různé akce. Nastavit pravidla pro roboty je možné kódem:

```
rules = (
    Rule(SgmlLinkExtractor(allow = "http://rejstrik-firem.kurzy.cz/osoby/(.+)/(.+)"), callback = 'parse_people', follow = True),
)
```

Řetězec přiřazený proměnné **allow** je regulárním výrazem. Pokud bude odkaz odpovídat alespoň jednomu regulárnímu výrazu, pak se na něj uplatní metoda, která má název jako řetězec přiřazený proměnné **callback**. Odkazy, na které nepůjde uplatnit žádné z pravidel, budou ignorovány.

[22] Spiders (roboti), Request (požadavky), Response (odpovědi), Sheduler (plánovač), Item Pipeline (úložiště)

Výchozí či startovní adresy se určují využitím příkazu **start_urls = [”www.prvniadresa.cz”, “www.druhaadresa.cz“ …]**. Pokud jsou tyto startovní adresy předány robotovi v této podobě, pak nebudou zpracovány, protože na ně nejsou uplatňována pravidla. V případě, že chceme, aby byly startovní URL adresy zpracovány uplatněním pravidel, tak musíme tyto adresy předat robotovi v lokálním souboru příkazem **start_urls = [“file:start.html“]**

Po získání HTML dokumentu lze data vybírat využitím XPath cest. Konkrétní příkazy vypadají takto:

```
xpath = HtmlXPathSelector(response)
xpath.select("//table[@id='orsmallinfotab']/tr[3]/td[2]")
```

Ve výše uvedených příkazech je **response** získaný HTML dokument, řetězec vstupující do metody **select** je XPath cesta, pomocí které vybíráme požadovaná data v HTML dokumentu.

V případě robotů je důležité nastavovat prodlevu mezi stahováním jednotlivých stránek (viz kapitola 6.3.2). Tuto prodlevu nastavíme přidělením hodnoty do proměnné **download_delay**.

6.4 Získaná data

Robot pro sběr dat byl několikrát pozastaven, a to nejčastěji po cca 8 hodinách běhu. To znamená, že sběr dat potřebných pro vytvoření grafu propojení firem a následně osob do těchto firem zainteresovaných neprobíhal kontinuálně. Data byla získávána z webu přibližné po dobu 2 měsíců a podařilo se nashromáždit 187 952 záznamů do tabulky people (osoby), 156 941 záznamů do tabulky companies (firmy). Z tabulek people a companies budou vytvořeny vrcholy výsledných grafů, což znamená, že vrcholů se podařilo nashromáždit 344 893, přičemž je 156 941 pro graf firem a 187 952 pro graf osob. Celkový počet nashromážděných záznamů pro tvorbu hran je 404 439. Všechna data pro tvorbu grafů osob a firem byla uložena do již zmíněné databáze na lokálním stroji (viz kapitola 6.3.1).

Při získávání záznamů do tabulky companies (firmy) došlo k zapsání firem, které jsou v *likvidaci*[23]. Tyto firmy stále figurují v obchodním rejstříku, protože likvidace je dlouhodobý proces, který končí zánikem firmy. Během likvidace dochází k úhradě závazků firmy a vyřízení pohledávek firmy. To znamená, že firma, která je v likvidaci, se může objevit například ve výsledcích vyhodnocení grafu osob algoritmem PageRank.

[23] Likvidace je zákonem řízený postup, přičemž dochází k mimosoudnímu vyrovnání majetkových vztahů zanikající právnické osoby, podniku. (viz http://www.finance-management.cz/080vypisPojmu.php?IdPojPass=37&X=Zruseni+likvidace+zanik+spolecnosti)

7 Popis aplikace

V této kapitole si popíšeme, jak vypadají vstupní a výstupní soubory vytvořené aplikace a dále také jak aplikace funguje.

7.1 Struktura vstupního souboru

Vstupní soubor je vytvořený z dat uložených v relační databázi. Pro každý typ grafu (graf osob a graf firem) je vytvořen jeden vstupní soubor. Pro graf firem je vytvořen vstupní soubor s názvem *sitFirmaOsobaFirma.csv* a pro graf osob je vytvořen soubor, který je pojmenován *sitOsobaFirmaOsoba.csv*. Oba soubory jsou ve formátu CSV[24] a mají stejnou strukturu, kterou si nyní popíšeme.

Prvním řádkem obou vstupních souborů je hlavička souboru. Konkrétně první řádka vypadá takto:

- **IdVrcholu;JmenoVrcholu**

To znamená, že z následujících řádků souboru budou tvořeny vrcholy grafu. V každé řádce bude ID vrcholu a jméno vrcholu Tato struktura je stejná až po řádek s klíčovým slovem *Hrany*. Po tomto řádku následuje nová hlavička následujících řádků. Tato hlavička je pro záznamy, ze kterých se budou tvořit hrany grafu. Hlavička těchto řádků vypadá následovně:

- **VRCHOL;personalizace;vahaVystup;HRANA;vahaVstup;personalizace;VRCHOL**

Takováto struktura je až do konce každého vstupního souboru. To znamená, že řádka vypadá následovně: ID zdrojového vrcholu, personalizace zdrojového vrcholu, váha výstupní hrany ze zdrojového vrcholu, ID hrany mezi zdrojovým a cílovým vrcholem, váha vstupní hrany do zdrojového vrcholu, personalizace cílového vrcholu a ID cílového vrcholu.

7.2 Struktura výstupních soborů

7.2.1 PageRank

Výstupní sobory se od souborů vstupních liší hlavně tím, že neobsahují, žádné informace o hranách grafu. Výstupní soubory jsou uloženy také ve formátu CSV, což znamená, že jednotlivá data v řádcích jsou rozdělena středníkem. Ve výstupních souborech najdeme v první řádce hlavičku:

- **Jmeno vrcholu;PageRank vrcholu**

Po této hlavičce následuje výpis vrcholů podle hodnoty PageRanku jednotlivých vrcholů od největší hodnoty po hodnotu nejmenší. V prvních buňkách řádků nalezneme jména vrcholů grafu (firmy, osoby) a ve druhých buňkách hodnoty PageRanku pro daný vrchol. Výstupních souborů s hodnotami s PageRankem vrcholů vznikne dohromady 6.

[24] CSV - Comma-separated values – je to jednoduchý souborový formát určený pro výměnu tabulkových dat. Skládá se z řádků, ve kterých jsou jednotlivé položky (sloupce) odděleny středníkem (**;**)

Pro graf firem vzniknou 3 výstupní soubory, protože máme tři typy grafů pro základní graf firem. Soubory s výsledky jednotlivých typů grafů (viz část 4.1) jsou pojmenovány takto: *vysledkyPRSitAFirmy.csv, vysledkyPRSitBFirmy.csv* a *vysledkyPRSitCFirmy.csv.*

Pro graf osob vzniknou 3 výstupní soubory s názvy: *vysledkyPRSitAOsoby.csv, vysledkyPRSitBOsoby.csv* a *vysledkyPRSitCOsoby.csv*. Soubory nám poskytují výsledky vyhodnocení jednotlivých typů grafů osob (viz část 4.2).

Pro výpočet PageRanku jsem vytvořil ještě výstupní sobory, které obsahují součty hodnot PageRanků za jednotlivé iterace (viz část 7.3). Tyto soubory jsou opět pro všechny typy grafů, což znamená, že jich je 6. V prvních řádcích těchto souborů je opět hlavička:

- **TypSit;PageRank;Iterace**

V dalších řádcích nalezneme typ sítě, pro kterou se PageRank počítal, součet hodnot PageRanku vrcholů za danou iteraci a v poslední buňce číslo aktuální iterace (1 až 25). Tyto výsledky se nacházejí v souborech: *vyslednyPRSitAFirmy.csv vyslednyPRSitBFirmy.csv a vyslednyPRSitCFirmy.csv* a analogicky pro grafy osob.

7.2.2 Míry centrality

Pro každý typ základních grafů (graf osob a graf firem) vzniknou další tři soubory, a to soubory obsahující výpočty mír centrality pro jednotlivé grafy. Struktura těchto souborů je obdobná jako struktura souborů s výsledky pro algoritmus PageRank. V prvním řádku souboru tedy nalezneme hlavičku:

- **Vrchol;Mira centrality**

To znamená, že v první buňce dalších řádků souboru nalezneme názvy jednotlivých vrcholů grafu (firmy, osoby) a v druhé buňce míru centrality jednotlivých vrcholů, kde použitá metoda výpočtu je dána názvem souboru.

Z výše uvedených informací vyplývá, pro graf firem vzniknou další tři výstupní soubory: *degreeCentralityFirma.csv, betweenessCentralityFirma.csv* a *closenessCentralityFirma.csv*.

Pro graf osob také vzniknou další tři výstupní soubory s názvy analogickými jako v případě grafu firem, tj. *degreeCentralityOsoba.csv, betweenessCentralityOsoba.csv* a *closenessCentralityOsoba.csv*.

7.3 Aplikace

Samotná aplikace funguje tak, že při spouštění zadáme soubor, ze kterého chceme graf vytvořit. Dalším vstupním parametrem je výběr typu grafu, v tomto případě se jedná o typy: Graf A, Graf B a Graf C (viz části 4.1 a 4.2). Aplikace načítá řádky ze vstupního soboru a dokud nenarazí na klíčové slovo *Hrany* (viz část 7.1), tak vytváří objekty datového typu Vrchol. Až se dostane na řádek, kde se klíčové slovo *Hrany* nachází, tak přestane vytvářet vrcholy grafu a začne mezi vrcholy grafu vytvářet hrany. Až se přečte celý vstupní soubor a tím se vytvoří graf vrcholů spojených hranami, tak přichází na řadu výpočet PageRanku.

Je volána metoda, která vypočítá hodnoty PageRanku pro jednotlivé vrcholy grafu. Výpočet PageRanku se provádí ve 25 iteracích, aby se docílilo co nejpřesnějších hodnot. Vrcholy se seřadí sestupně podle jejich hodnot PageRanku a vypíšou se do příslušných souborů. Součty hodnot PageRanku za jednotlivé iterace jsou taktéž vypsány do určených výstupních souborů (viz část 7.2.1).

Po naprogramování aplikace byly vytvořeny všechny typy navržených grafů a vyhodnoceny algoritmem PageRanku. Výsledky byly uloženy do příslušných výstupních souborů.

8 Programová knihovna JUNG a míry centrality

8.1 Knihovna JUNG

Pro výpočet byla použita již naprogramovaná knihovna *JUNG*[25] , která obsahuje metody pro vytvoření grafu, přidaní vrcholů a hran a konečně i metody pro výpočet výše zmíněných mír centrality. Bylo nutné si projít dokumentaci této knihovny a seznámit s jejími metodami pro tvorbu grafu a následný výpočet mír centrality.

8.1.1 Problémy s knihovnou JUNG

Knihovna JUNG je bohužel dosti pomalá, pokud vyhodnocuje graf s velkým počtem vrcholů. Týká se to především metody pro výpočet closeness centrality, jelikož její výpočet je časově nejsložitější.

Při vytvoření grafu firem, kde se nachází 156 941 vrcholů, trvá výpočet všech tří mír centrality cca 32 hodin. V případě grafu osob, který je tvořen 187 952 vrcholy, trvá výpočet cca 35 hodin. Oba výpočty byly prováděny na stroji s následujícími parametry:

- Procesor: Intel(R) Pentium(R) CPU B970 @ 2.40GHz (2 CPUs)
- Naistalovaná pamět (RAM): 4,00GB (použitelné 3,46GB)
- Systém: Win 7 Professional 64bit

Navíc metoda closeness centrality potřebuje k výpočtu jednotlivých hodnot velkou část paměti a je tedy nutné mít stroj, který bude programu poskytovat dostatek paměti. Pro tuto metodu je dále nutností nastavit velikost haldy *JVM*[26], protože metoda, jak již bylo zmíněno, potřebuje pro svůj chod dostatečné množství paměti. Vypočítané míry centrality jednotlivých vrcholů se vypíšou sestupně do příslušných výstupních souborů (viz část 7.2.2).

V mém případě nejsou výsledky metody closeness centrality úplné, důvodem je velký objem dat na slabé parametry počítačů, které jsem pro veškeré výpočty použil. Výpočty degree a betwees proběhly celé, i když trvali dlouhou dobu (viz výše). Pro metodu closeness centrality se vypočítaly (na výše zmíněném stroji) hodnoty pouze pro cca 140 vrcholů. Důvodem bylo překročení vymezené paměti pro daný proces. Při optimalizaci PC a také jazyka Java na stroji používaném (v. 1.7.0_09) se podařilo získat hodnoty pro cca 180 vrcholů, přičemž výpočet trval cca 4 minuty. Za důležité považuji připomenout, že graf firem tvoří cca 156 000 vrcholů a graf osob dokonce cca 187 000 vrcholů. Vypočtené hodnoty jsou sice správné, ale neúplné. Můžeme hodnotit pouze počet vrcholů, pro které byly hodnoty získány.

Snažil jsem se tento problém odstranit dlouho dobu, nakonec jsem použil dvakrát silnější stroj než stroj zmíněný výše. Bohužel se výsledky pro všechny vrcholy nedopočítaly. Na dvakrát silnějším stroji se sice vypočítaly hodnoty pro více vrcholů, konkrétně pro cca 900 vrcholů, ale i tento počet vrcholů je zanedbatelný. Na silnějším stroji trvaly výpočty cca 35 minut, a to

[25] http://jung.sourceforge.net/

[26] Java Virtual Machine je sada počítačových programů a datových struktur, která využívá modul virtuálního stroje ke spuštění dalších počítačových programů a skriptů vytvořených v jazyce Java (viz http://cs.wikipedia.org/wiki/Java_Virtual_Machine).

pouze pro výše zmíněných cca 900 vrcholů grafu. Můžeme si tedy dopočítat, jak dlouho by trval výpočet closeness centrality pro cca 156 000 vrcholů. Dostali bychom se k číslu 6 066 minut, což je přibližně 101 hodin.

Z výše uvedených informací usuzuji, že je výpočet closeness centrality náročný z hlediska přidělené paměti a velikosti obejmu dat. Ve výsledných souborech pro closeness centrality je tedy pouze zlomek celkového žebříčku získaných osob.

Problém bude pravděpodobně v řeší metody pro výpočet closeness centrality, protože v obou další mírách se výpočty provedly všechny.

Pro výpočet mír centrality bych doporučil použit *datové centrum*[27] nebo opravdu silný stroj, jinak bude výpočet opravdu dlouhý, viz výše. Výběr naprogramované knihovny JUNG se příliš nezdařil a pravděpodobně by bylo vhodné použít jinou implementaci pro výpočet mír centrality nebo naprogramovat knihovnu vlastní.

8.1.2 Tvorba grafu a následný výpočet

Graf se tvoří pomocí vstupních souborů (viz část 7.2.1), kde se postupně do grafu přidají vrcholy a poté hrany. Po vytvoření grafu (doba trvání cca 1 minuta) se volají metody pro výpočet jednotlivých mír centrality. Po výpočtu těchto hodnot se vrcholy a jejich výsledné hodnoty pro jednotlivé míry centrality zapisují do příslušných výstupních souborů.

[27]Datové centrum je místo, ve kterém je možno v téměř ideálních podmínkách, s vysokou mírou fyzické bezpečnosti a velmi dobrou konektivitou do Internetu provozovat servery, datová úložiště a další prvky ICT infrastruktury.

9 Získané výsledky vyhodnocení

V této kapitole si ukážeme výstupní soubory jednotlivých metod vyhodnocení a popíšeme si jednotlivé výsledky. Dále budu výsledky diskutovat a porovnávat.

9.1 PageRank

Výsledky získané pomocí algoritmu PageRank jsou uloženy v příslušných výstupních souborech.

9.1.1 Graf firem

Pro graf firem byly navrženy tři typy grafů, které byly vyhodnoceny algoritmem PageRank. Dále si jednotlivé výsledky ukážeme a popíšeme.

Graf A

V tomto typu grafu získá nejvyšší hodnotu PageRanku ten vrchol, který bude nejvíce ovlivňovat ostatní vrcholy v tomto grafu. Jedná se o graf firem, tudíž by firma s nejvyšším hodnocením měla mít největší vliv na ostatní firmy v síti. Top 10 nejlépe hodnocených firem v grafu A naleznete na Obrázku 15.

	Jmeno vrcholu	PageRank vrcholu
2	AMBI, a.s.	0.0021073284075892344
3	AGR BRNO spol. s r.o., v likvidaci	0.0017932663645339802
4	Hutní montáže, a.s.	0.0016599890668183568
5	AERO Vodochody a.s.	0.0015705602967277204
6	ELMALS s.r.o.	0.0015305180988997385
7	Sport Fit Praha s.r.o.	0.001530407169307188
8	Fügnerova Real s.r.o.	0.0015303469486780254
9	EUROMONT - SM s.r.o.	0.0015303469486780254
10	Toyota Peugeot Citroën Automobile Czech, s.r.o.	0.0015303469486780254
11	KLEOS stavební s.r.o.	0.0014965130486763072

Obrázek 15 - Graf A - Výsledky algoritmu PageRank - TOP 10 vrcholů grafu

Graf B

V grafu B získá nejvyšší hodnocení vrchol, který bude nejvíce ovlivnitelný z pohledu ostatních vrcholů v grafu. Nejvyšší hodnotu PageRanku získá ta firma, která bude nejvíce ovlivnitelná ostatními firmami v grafu. 10 nejovlivnitelnějších firem můžete vidět na Obrázku 16.

1	Jmeno vrcholu	PageRank vrcholu
2	AGR BRNO spol. s r.o., v likvidaci	0.001948343936982254
3	KINO PLUS s.r.o. - v likvidaci	0.0016640440680787349
4	AMBI, a.s.	0.001550487996256979
5	ELMALS s.r.o.	0.0015305180988997385
6	Sport Fit Praha s.r.o.	0.001530407169307188
7	Fügnerova Real s.r.o.	0.0015303469486780254
8	EUROMONT - SM s.r.o.	0.0015303469486780254
9	Toyota Peugeot Citroën Automobile Czech, s.r.o.	0.0015303469486780254
10	KLEOS stavební s.r.o.	0.001496513048676307
11	TG MEDICAL s.r.o.	0.001438841336001959

Obrázek 16 - Graf B - Výsledky algoritmu PageRank - TOP 10 vrcholů grafu

Graf C

Graf C slouží pro srovnání výsledků z grafu A a grafu B, jelikož se v tomto grafu neuvažovaly váhy hran. 10 nejlépe hodnocených vrcholů (firem) grafu je možné vidět na Obrázku 17.

1	Jmeno vrcholu	PageRank vrcholu
2	AGR BRNO spol. s r.o., v likvidaci	0.001995180209864831
3	AMBI, a.s.	0.0016498940968505337
4	KINO PLUS s.r.o. - v likvidaci	0.001647986328244413
5	ELMALS s.r.o.	0.0015305180988997383
6	Sport Fit Praha s.r.o.	0.001530407169307188
7	Fügnerova Real s.r.o.	0.0015303469486780254
8	EUROMONT - SM s.r.o.	0.0015303469486780254
9	Toyota Peugeot Citroën Automobile Czech, s.r.o.	0.0015303469486780254
10	KLEOS stavební s.r.o.	0.0014965130486763072
11	BSH Holice a.s.	0.0013939733768147707

Obrázek 17 - Graf C - Výsledky algoritmu PageRank - TOP 10 vrcholů grafu

9.1.2 Graf osob

Pro graf osob byly, stejně jako pro graf firem, navrženy tři typy grafů, které byly vyhodnoceny algoritmem PageRank. Ukážeme si a popíšeme jednotlivé výsledky vyhodnocení. Popis jednotlivých typů grafů je analogický jako v části 9.1.1, jen s tím rozdílem, že vrcholy reprezentují osoby.

Graf A

V tomto grafu získá nejvyšší hodnocení vrchol, který nejvíce ovlivňuje ostatní vrcholy. Osoba, která získá v tomto typu grafu nejvyšší hodnotu PageRanku, je nejvlivnější osobou v grafu - na ostatní osoby má největší vliv. 10 nejvlivnějších osob grafu můžete vidět na Obrázku 18.

	Jmeno vrcholu	PageRank vrcholu
2	Mykhaylo Shcherbyak	9.2655304816660 17E-4
3	Vladislav Tuček	9.161481319568773E-4
4	Heinz Walter Schott	8.253396063963795E-4
5	Alena Vašková	8.130757423421721E-4
6	Jaromír Pospíšil, Prostějov - Krasice	8.082850389201151E-4
7	Ing. Ladislav Verner, Lanškroun	6.53682482411298E-4
8	Vasyl Berets	6.358470223094593E-4
9	Ing. Adéla Pelechová	5.147976990099754E-4
10	EVA KUCHAŘOVÁ	5.03415328673976E-4
11	Jaroslav Pařízek, Lomnice u Tišnova	4.5894150826521153E-4

Obrázek 18 - Graf A - Výsledky algoritmu PageRank - TOP 10 vrcholů grafu

Graf B

Zde získá největší hodnotu PageRanku ten vrchol, který bude nejvíce ovlivnitelný ostatními vrcholy v grafu. Osoba s nejvyšším hodnocením je ta osobou, která je nejvíce ovlivnitelná osobami jinými. Top 10 nejovlivnitelnějších osob v tomto grafu zobrazuje Obrázek 19.

	Jmeno vrcholu	PageRank vrcholu
2	Alena Vašková	0.0011958379906551835
3	Mykhaylo Shcherbyak	0.0011462156279260093
4	Heinz Walter Schott	9.7010649024421 29E-4
5	Vladislav Tuček	8.243216419129865E-4
6	Ing. Ladislav Verner, Lanškroun	7.591201530104288E-4
7	Jaromír Pospíšil, Prostějov - Krasice	7.231698756170224E-4
8	Ing. Adéla Pelechová	7.024776433558579E-4
9	Vasyl Berets	6.70326884561718E-4
10	Pavel Andrýsek	5.988922726617428E-4
11	Jaroslav Pařízek, Lomnice u Tišnova	5.592399565222647E-4

Obrázek 19 - Graf B - Výsledky algoritmu PageRank - TOP 10 vrcholů grafu

Graf C

V tomto grafu neuvažujeme váhy hran, respektive váhy všech hran jsou v tomto grafu rovné 1. V případě vyššího počtu hran mezi dvěma vrcholy je hodnota váhy hrany rovna součtu hodnot vah jednotlivých hran. Pokud jsou mezi dvěma vrcholy dvě hrany, pak výsledná hodnota váhy hrany je rovna 2. Výsledky grafu C mohou posloužit k porovnání výsledků z grafu A a grafu B. 10 nejlépe ohodnocených osob v grafu C je vidět na Obrázku 20.

1	Jmeno vrcholu	PageRank vrcholu
2	Alena Vašková	9.317949718165182E-4
3	Mykhaylo Shcherbyak	8.787281053559071E-4
4	Ing. Ladislav Verner, Lanškroun	8.68615286900304E-4
5	Heinz Walter Schott	6.947389617628406E-4
6	ing. Vladimír Marek, Dlouhá Třebová	6.380297495312657E-4
7	Marek Hanák	6.367980879998807E-4
8	Pavel Andrýsek	6.122825287167515E-4
9	Ing. Adéla Pelechová	5.707014147322488E-4
10	Vasyl Berets	4.5991903734676624E-4
11	ADAM BRYCHTA	4.4569744679396993E-4

Obrázek 20 - Graf C - Výsledky algoritmu PageRank - TOP 10 vrcholů grafu

9.1.3 Porovnání výsledků a diskuze

Graf firem

Porovnání výsledků pro jednotlivé grafy vytvořené ze základního grafu firem můžeme vidět na Obrázku 15, 16 a 17. Na těchto obrázcích je vždy zobrazeno 10 nejlépe hodnocených firem pro jednotlivé typy grafu A, B a C.

Pokud se podíváme na grafy A a B, tak je celkem překvapivé, že firma *AMBI, a.s.* je v grafu A na prvním místě a v grafu B na místě třetím. Podle mého předpokladu by to vypadalo tak, že pokud by firma *AMBI,* a.s. byla pro graf A nejlépe hodnocena, tak by pro graf B měla hodnocení nejhorší. Důvodem jsou výsledky, které oba grafy vracejí. Pro graf A je nejlépe hodnocena firma, která má největší vliv na ostatní firmy v grafu. Proto jsou tyto výsledky trochu neočekávané. V tomto případě se nejedná pouze o firmu *AMBI, a.s.*, protože takovýchto případů je ve vyhodnocení pro typy grafů A a B hned 7. Pořadí těchto firem se mění vždy pouze o pár míst, respektive zůstávají nejlépe hodnocenými firmami jak pro graf A, tak pro graf B.

Srovnání výsledných žebříčků grafů A a B s posledním „srovnávacím“ typem grafu C je podobné jako v případě grafů A a B. Ve všech třech výsledných hodnoceních se objevuje 7 stejných firem. Znamená to tedy, že tyto firmy mají největší vliv na ostatní firmy v grafu A, ale zároveň jsou jedny z nejovlivnitelnějších firem v grafu B. Graf C umožňuje srovnání grafu

A a B. Můžeme vidět, že firma *AGR BRNO spol. s.r.o*. má v grafu A druhé nejlepší hodnocení ze všech firem. V grafu B je dokonce tato firma hodnocena nejlépe. Poté při náhledu do výsledků získaných z grafu C zjišťujeme, že je tato firma opět nejlépe hodnocena. Tudíž můžeme graf C brát jako srovnávací graf výsledků grafu A a grafu B (viz Tabulka 2).

Tabulka 2 - Srovnání 10 nejlépe hodnocených firem využitím PageRanku (Graf A, Graf B, Graf C)

Graf A	Graf B	Graf C
AMBI, a.s.	AGR BRNO spol. s.r.o.	AGR BRNO spol. s r.o.
AGR BRNO spol. s.r.o.	KINO PLUS s.r.o.	AMBI, a.s.
Hutní montáže, a.s.	AMBI, a.s.	KINO PLUS s.r.o.
AERO Vodochody a.s.	ELMALS s.r.o.	ELMALS s.r.o.
ELMALS s.r.o.	Sport Fit Praha s.r.o.	Sport Fit Praha s.r.o.
Sport Fit Praha s.r.o.	Fugnerova Real s.r.o.	Fügnerova Real s.r.o.
Fugnerova Real s.r.o	EUROMONT – SM s.r.o	EUROMONT - SM s.r.o.
EUROMONT – SM s.r.o.	Toyota Peugeot Citroen Automobile Czech, s.r.o.	Toyota Peugeot Citroën Automobile Czech, s.r.o.
Toyota Peugeot Citroen Automobile Czech, s.r.o.	KLEOS stavební s.r.o.	KLEOS stavební s.r.o.
KLEOS stavební s.r.o.	TG MEDICAL s.r.o.	BSH Holice a.s.

Graf osob

Výsledky získané z grafu osob pro jednotlivé typy grafů A, B a C je možné porovnávat z Obrázků 18, 19 a 20. Z těchto obrázků můžeme porovnat vždy 10 nejlépe ohodnocených osob z navržených a vytvořených typů grafů pro graf osob.

Při porovnání výsledků získaných z grafů A a B je podobnost výsledných žebříčků dokonce větší než v případě grafu firem. Když se podíváme na výsledné žebříčky, tak zjistíme, že se liší pouze jedno jméno. Z toho lze vyvodit, že osoby, které mají největší vliv na ostatní osoby v grafu, jsou zároveň nejovlivnitelnějšími osobami v grafu.

Stejně jako v případě grafu firem bylo mým předpokladem, že osoba s nejlepším hodnocením v grafu A získá nejhorší hodnocení v grafu B. Ze získaných výsledů ohodnocení vidíme, že tomu tak rozhodně není.

Graf C znovu poslouží pro srovnání výsledků získaných z grafu A a grafu B. Pro graf osob se výsledky ze „srovnávacího“ grafu C trochu rozcházejí. Zatímco ve výsledcích grafů A a B bylo mezí 10 nejlepšími osobami 8 stejných osob. Při srovnání výsledků ze všech tří grafů, zjišťujeme, že pro graf osob je mezi 10 nejlepšími osobami ve všech grafech pouze 6 stejných osob (viz Tabulka 3).

Tabulka 3 - Srovnání 10 nejlépe hodnocených osob využitím PageRanku (Graf A, Graf B, Graf C)

Graf A	Graf B	Graf C
Mykhaylo Shcherbyak	Alena Vašková	Alena Vašková
Vladislav Tuček	Mykhaylo Shcherbyak	Mykhaylo Shcherbyak
Heinz Walter Schott	Heinz Walter Schott	Ing. Ladislav Verner, Lanškroun
Alena Vašková	Vladislav Tuček	Heinz Walter Schott
Jaromír Pospíšil, Prostějov – Krasice	Ing. Ladislav Verner, Lanškroun	ing. Vladimír Marek, Dlouhá Třebová
Ing. Ladislav Verner, Lanškroun	Jaromír Pospíšil, Prostějov - Krasice	Marek Hanák
Vasyl Berets	Ing. Adéla Pelechová	Pavel Andrýsek
Ing. Adéla Pelechová	Vasyl Berets	Ing. Adéla Pelechová
EVA KUCHAŘOVÁ	Pavel Andrýsek	Vasyl Berets
Jaroslav Pařízek, Lomnice u Tišnova	Jaroslav Pařízek, Lomnice u Tišnova	ADAM BRYCHTA

9.2 Míry centrality

Výsledky výpočtů jednotlivých metod mír centrality jsou uloženy v příslušných výstupních souborech. Výpočty mír centrality byly prováděny na neorientovaných a neohodnocených grafech. Kdy se v prvním případě jednalo o graf firem a v případě druhém o graf osob (viz kapitola 4).

9.2.1 Degree centrality

Hodnota degree centrality určuje počet přímých vazeb vrcholu k ostatním vrcholům v grafu. Vrchol s největší hodnotou degree centrality je ten, který má největší počet přímých vazeb na ostatní vrcholy v grafu.

Graf firem

Firma s největší hodnotou degree centrality je ta, která je přímo spojena s největším počtem firem v grafu. Spojení mezi jednotlivými firmami v grafu je utvářeno pomocí osob v nich zainteresovaných. 10 firem s nejvyšší hodnotou degree centrality je možné vidět na Obrázku 21.

1	Vrchol	Mira centrality
2	ABRASMONT s.r.o.	1295.0
3	4Support Investment, a.s.	1042.0
4	1. CZECH TRANS COMPANY s.r.o.	816.0
5	3D CHEMOPRAG a.s.	741.0
6	1. Realitní a stavební s.r.o., v likvidaci	631.0
7	1. ENERGO CZ, a.s.	602.0
8	ABLON s.r.o.	503.0
9	1. PEKAŘSKÁ a.s.	482.0
10	4 DVORY a.s.	466.0
11	AGR BRNO spol. s r.o., v likvidaci	463.0

Obrázek 21 - Vyhodnocení grafu firem metodou degree centrality

Graf osob

Osoba s největším počtem přímých vazeb na ostatní osoby v grafu má nejvyšší přidělenou hodnotu degree centrality. Top 10 nejlépe hodnocených osob metodou degree centrality, viz Obrázek 22.

1	Vrchol	Mira centrality
2	ing. Vladimír Marek, Dlouhá Třebová	643.0
3	Marek Hanák	642.0
4	Alena Vašková	637.0
5	Mykhaylo Shcherbyak	591.0
6	Ing. Adéla Pelechová	544.0
7	Heinz Walter Schott	472.0
8	Ing. Ladislav Verner, Lanškroun	441.0
9	Pavel Andrýsek	392.0
10	Fanisa Stefanovich	346.0
11	Vasyl Berets	344.0

Obrázek 22 - Vyhodnocení grafu osob metodou degree centrality

9.2.2 Betweeness centrality

Největší hodnotu betweeness centrality má ten vrchol, který leží na největším počtu nejkratší cest mezi všemi dvojicemi vrcholů grafu.

Graf firem

Firma, která má v rámci grafu největší hodnotu betweeness centrality, je firma ležící na největším počtu nejkratších cest mezi dvojicemi firem v grafu. Největší hodnotu betweeness centrality má firma, přes kterou je spojeno nejvíce dvojic firem grafu. Top 10 nejlépe hodnocených firem metodou betweeness centrality je k náhledu v obrázku 23.

1	Vrchol	Mira centrality
2	1. ENERGO CZ, a.s.	2.1105843598144362E9
3	3D CHEMOPRAG a.s.	2.0508816566141593E9
4	1. PEKAŘSKÁ a.s.	1.6732067632064297E9
5	1. CZECH TRANS COMPANY s.r.o.	6.940493459586923E8
6	1. Realitní a stavební s.r.o., v likvidaci	5.866183784042944E8
7	1.Zdiměřická stavební s.r.o.	3.30697557735429 17E8
8	Agrární komora České republiky	3.0733349677666 91E8
9	ADAST a.s. v likvidaci	2.928249009006894E8
10	Hospodářská komora České republiky	2.5214259682464966E8
11	ANCO Besitz spol. s r.o. v likvidaci	2.4481668637971 07E8

Obrázek 23 - Vyhodnocení grafu firem metodou betweeness centrality

Graf osob

Osoba s největší hodnotou betweenes centrality zprostředkovává nejvíce spojení mezi dvěma libovolnými osobami v grafu. Nejvyšší hodnotu betweeness centrality má osoba ležící na největším počtu nejkratších cest mezi všemi dvojicemi vrcholů grafu. 10 nejlépe hodnocených osob metodou betweeness centrality můžeme vidět na Obrázku 24.

1	Vrchol	Mira centrality
2	Mykhaylo Shcherbyak	2.532617861484381E9
3	Heinz Walter Schott	2.157978050916606E9
4	Alena Vašková	2.046092118782298E9
5	Jaroslav Ďurčovič	7.60746552283388E8
6	Pavel Andrýsek	6.929741984642 54E8
7	Ing. Ladislav Verner, Lanškroun	3.61119792070911 05E8
8	Vasyl Berets	3.51551852301930 67E8
9	Petr Švábek	3.4313962593837E8
10	Ing. Miloslav Pitro	2.28960110909871 28E8
11	Zdeněk Hruška	2.25754935103849E8

Obrázek 24 - Vyhodnocení grafu osob metodou betweeness centrality

9.2.3 Closeness centrality

Vrchol s největší hodnotou closeness centrality je vrchol, ze kterého lze dosáhnout na všechny další vrcholy grafu přímou vazbou. Výpočet této metody je nejsložitější. Výsledné žebříčky na obrázcích (viz Obrázek 25 a Obrázek 26) jsou poněkud zkreslené, jelikož výsledné žebříčky pro closeness centrality nejsou úplné (viz část 8.1.1)

Graf firem

Firma, jejíž hodnota closeness centrality nabývá nejvyšší hodnoty, je firma, která má nejkratší spojení na všechny ostatní firmy v grafu. 10 nejlépe hodnocených firem metodou closeness centrality je k náhledu na Obrázku 25.

1	Vrchol	Mira centrality
2	VUMS, a.s.v likvidaci	1.0
3	Továrny textilních potřeb - ELITEX, a.s.	1.0
4	GRANDHOTEL ZLATÝ LEV, LIBEREC, S.P.	1.0
5	Podblanicko Louňovice v likvidaci	1.0
6	Auto Praha, spol. s r.o.	1.0
7	HITEC, spol. s r.o. v likvidaci	1.0
8	Plynoservis Jílek s.r.o.	1.0
9	Liarinbel s.r.o.	1.0
10	Jezdecká stáj Bílichov s.r.o.	1.0
11	ASIMINA, s.r.o.	1.0

Obrázek 25 - Vyhodnocení grafu firem metodou closeness centrality

Graf osob

Největší hodnotu closeness centrality má ta osoba, ze které vedou nejkratší hrany ke všem dalším osobám v grafu. 10 nejlépe hodnocených osob metodou closeness centrality je možné vidět na Obrázku 26.

1	Vrchol	Mira centrality
2	Jiří Plhal, Drnovice	1.0
3	Ing. Stiva Jokeš	1.0
4	Pavel Flajšhans, Plzeň	1.0
5	MUDr. Hana Holubová, Bílina - Teplické předměstí	1.0
6	Eva Novotná, Kutná Hora	1.0
7	Ing. Vladimír Bureš, Hradec Králové	1.0
8	Alena Květová	1.0
9	Ing. Pavel Soukup, Praha	1.0
10	PhDr. Jana Mertlová, Trstěnice u Litomyšle - Trstě	1.0
11	Dalibor Štěpán, Jihlava	1.0

Obrázek 26 - Vyhodnocení grafu osob metodou closeness centrality

9.2.4 Porovnání výsledků a diskuze

Graf firem

Výsledky získané metodami mír centrality pro graf firem je možné porovnávat na Obrázcích 21, 23 a 25. Na obrázcích vždy vidíme 10 firem, které mají nejvyšší hodny jednotlivých mír centrality.

Při porovnání jednotlivých žebříčků pro graf firem nenacházíme velkou shodu jako v případě výpočtu hodnot PageRanku všech navržených typů grafů. Při porovnání prvních 10 firem nacházíme v jednotlivých žebříčcích (Degree a Betweeness centrality) 5 stejných firem, což je menší shoda než při porovnávání výsledků PageRanku.

Výsledkem je tedy fakt, že polovina firem s největší hodnotou degree centrality nemá zároveň jednu z největších hodnot betweeness centrality (viz Tabulka 4). Z toho je možné vyvodit závěry, že firmy, které mají vysoký počet přímých vazeb (degree centrality) na ostatní firmy v grafu, nemusí nutně kontrolovat tok informací v grafu (betweeness centrality). Analogicky firmy kontrolující tok informací v grafu nemají největší počet přímých vazeb na ostatní firmy v grafu.

Najdou se však i výjimky, tedy firmy, které mají velký počet přímých vazeb na ostatní firmy v grafu, zároveň kontrolují tok informací v grafu. Takovou firmou je například *3D CHEMOPRAG a.s.*. Tato firma figuruje jak mezi firmami s vysokou hodnotou degree centrality, tak mezi firmami s vysokou hodnotou betweeness centrality.

Při srovnávání výsledků vyhodnocených metodami degree a betweeness centrality s výsledky vyhodnocenými metodou closeness nenacházíme shodu žádnou (viz část 8.1.1). Zcela je to

způsobeno neúplnými výsledky v případě této metody a z části tím, že i při neúplných výsledcích hodnocení spousta vrcholů získává využitím metody closeness centrality hodnocení nejvyšší.

Tabulka 4- Porovnání 10 nejlépe hodnocených firem metodami mír centrality

Degree centrality	Betweeness centrality	Closeness centrality
ABRASMONT s.r.o.	1. ENERGO CZ, a.s.	VUMS, a.s.v likvidaci
4Support Investment, a.s.	3D CHEMOPRAG a.s.	Továrny textilních potřeb - ELITEX, a.s.
1. CZECH TRANS COMPANY s.r.o.	1. PEKAŘSKÁ a.s.	GRANDHOTEL ZLATÝ LEV, LIBEREC, S.P.
3D CHEMOPRAG a.s.	1. CZECH TRANS COMPANY s.r.o.	Podblanicko Louňovice v likvidaci
1. Realitní a stavební s.r.o., v likvidaci	1. Realitní a stavební s.r.o., v likvidaci	Auto Praha, spol. s r.o.
1. ENERGO CZ, a.s.	1.Zdiměřická stavební s.r.o.	HITEC, spol. s r.o. v likvidaci
ABLON s.r.o.	Agrární komora České republiky	Plynoservis Jílek s.r.o.
1. PEKAŘSKÁ a.s.	ADAST a.s. v likvidaci	Liarinbel s.r.o.
4 DVORY a.s.	Hospodářská komora České republiky	Jezdecká stáj Bílichov s.r.o.
AGR BRNO spol. s r.o., v likvidaci	ANCO Besitz spol. s r.o. v likvidaci	ASIMINA, s.r.o.

Graf osob

Míry centrality grafu osob lze porovnávat na základě Obrázků 22, 24 a 26. Na každém ze tří obrázků je zobrazeno 10 osob, které získaly největší hodnoty jednotlivých mír centrality.

Stejně jako v případě grafu firem, ani zde není shoda ve výsledných žebříčcích, jako tomu bylo u PageRanku. Míry centrality pro graf osob mají ve výsledných žebříčcích (Degree centrality, Betweeness centrality) 6 shodných záznamů (viz Tabulka 5).

Vyhodnocení získaných výsledků využitím mír centrality bude stejné jako v případě grafu osob. Při srovnání najdeme určitou shodu mezi 10 osobami s vysokou hodnotou jednotlivých mír centrality. Tato shoda je ovšem v grafu osob větší než 50%, což znamená, že osoba s největším počtem přímých vazeb na ostatní osoby v grafu nebude se 40% pravděpodobností kontrolovat tok informací v grafu. Naopak existuje osoba, která figuruje mezi 10 osobami s nejlepším ohodnocením betweeness centrality (kontrolující tok dat v grafu), která bude s 60% pravděpodobností figurovat mezi 10 osobami s vysokou hodnotou degree centrality.

Tabulka 5 - Porovnání 10 nejlépe hodnocených osob metodami mír centrality

Degree centrality	Betweeness centrality	Closeness centrality
ing. Vladimír Marek, Dlouhá Třebová	Mykhaylo Shcherbyak	Jiří Plhal, Drnovice
Marek Hanák	Heinz Walter Schott	Ing. Stiva Jokeš
Alena Vašková	Alena Vašková	Pavel Flajšhans, Plzeň
Mykhaylo Shcherbyak	Jaroslav Ďurčovič	MUDr. Hana Holubová, Bílina - Teplické předměstí
Ing. Adéla Pelechová	Pavel Andrýsek	Eva Novotná, Kutná Hora
Heinz Walter Schott	Ing. Ladislav Verner, Lanškroun	Ing. Vladimír Bureš, Hradec Králové
Ing. Ladislav Verner, Lanškroun	Vasyl Berets	Alena Květová
Pavel Andrýsek	Petr Švábek	Ing. Pavel Soukup, Praha
Fanisa Stefanovich	Ing. Miloslav Pitro	PhDr. Jana Mertlová, Trstěnice u Litomyšle - Trstě
Vasyl Berets	Zdeněk Hruška	Dalibor Štěpán, Jihlava

9.3 Porovnání výsledků PageRanku a mír centrality

V této části si stručně porovnáme výsledné hodnocení grafu firem a grafu osob, které jsme získali použitými metodami pro analýzu grafů (PageRank, míry centrality).

V obou typech grafů nejdeme alespoň jednu firmu či osobu, která získala jedno z nejlepších hodnocení, jak využitím algoritmu PageRank, tak mírami centrality s výjimkou metody closeness centrality (viz část 8.1.1).

Pro graf osob tou jsou například Alena Vašková a Mykhaylo Shcherbyak a pro graf firem je možné zmínit AGR BRNO spol. s.r.o., které sice nefiguruje v žebříčku pro výpočet betweeness centrality, ale v ostatních žebříčcích má své místo.

Závěrem můžeme říci, že existují firmy, které mají velký vliv na ostatní firmy, zároveň jsou ovlivnitelné z pohledu ostatních firem, ve výsledném grafu kontrolují tok informací a současně mají jeden z největších počtů přímých vazeb na ostatní firmy v grafu.

Stejné závěrečné slovo k vyhodnoceným výsledkům lze říci o osobách. Tedy existuje osoba, která má velký vliv na ostatní osoby v grafu a zároveň je ovlivnitelná z pohledu ostatních osob. Současně je jedním z nejvýznamnějších vrcholů grafu, protože kontroluje tok informací grafem a má jeden z největší počtů přímých vazeb na ostatní osoby v grafu.

10 Závěr

Práce si kladla za úkol analyzovat síť firem v České Republice. Analýza byla provedena využitím algoritmu PageRank a mírami centrality. Grafy pro analýzu byly tvořeny firmami a osobami v těchto firmách zainteresovanými. Rozsáhlá databáze byla vytvořena pomocí dat získaných z webu **rejstirk-firem.kurzy.cz** (viz kapitola 3). Provozovatel webu se nebrání automatickému procházení, a proto se podařilo získat dostatečné množství dat k vytvoření grafů firem a osob. Databáze obsahuje spojení mezi firmami a osobami a základní informace o firmách a osobách.

V teoretické části byly představeny základní metody pro analýzu grafů a základní grafové metriky. Jako první byl představen algoritmus PageRank a metody pro výpočet mír centrality. Dále jsme se seznámili s poloměrem grafu, hustotou grafu a koeficientem shlukování, což jsou základní grafové metriky.

V další části jsem popsal navržené typy grafů a předpoklady pro jejich vyhodnocení algoritmem PageRank.

V praktické části jsem se nejprve zaměřil na vytvoření databáze, do které budou data uložena. Po vytvoření databáze jsem se zaměřil na získání dostatečného množství dat využitím webového robota. Webového robota jsem upravil a nastavil u něj pravidla pro prohledávání obsahu určitých webových stránek. Dále jsem díky robotovi získal data pro pozdější tvorbu grafů a jejich vyhodnocení.

Po získání dostatečného množství dat jsem vytvořil aplikaci pro vyhodnocení těchto nabytých dat. Implementoval jsem algoritmus PageRank, který jsem použil pro vyhodnocení všech typů grafů. Dalé jsem použil knihovnu JUNG pro výpočty daných mír centrality, díky které jsem získal výsledky hodnocení jednotlivých typů grafů uvedenými metodami mír centrality.

V poslední části práce jsem porovnal a diskutoval získané výsledky hodnocení jednotlivých grafů. Nejprve jsem porovnával výsledky hodnocení všech typů grafů algoritmem PageRank a poté jsem porovnával výsledky hodnocení všech typů grafů získané metodami pro výpočet mír centrality.

Při porovnání výsledků získaných algoritmem PageRank jsem zjistil, která z firem či osob nejvíce ovlivňuje ostatní firmy či osoby a také jaká firma či osoba je naopak nejvíce ovlivnitelná z pohledu ostatních firem či osob. Žebříčky firem a osob s nejvyšším hodnocením PageRanku byly porovnány jednotlivě pro grafy firem a grafy osob.

Při porovnání bylo zjištěno, že firma či osoba, která nejvíce ovlivňuje ostatní firmy či osoby, může být zároveň i jedna z nejvíce ovlivnitelných firem či osob. Stejně tak nejvíce ovlivnitelná firma či osoba z pohledu ostatních firem či osob může získat nejvyšší hodnotu PageRanku v případě, kdy se vyhodnocuje graf jejímž výsledkem jsou firmy či osoby, které nejvíce ovlivňují ostatní osoby či firmy v grafu

Míry centrality umožnily určit firmy a osoby, které mají v daných grafech klíčovou roli. Výsledné žebříčky klíčových firem a osob pro jednotlivé typy mír centrality mezi sebou byly porovnány a následně diskutovány.

Literatura

[1] Page, L. - Brin, S. - Motwani, R. - Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 66, *Stanford InfoLab*, *Stanford*, *CA*, 1999.

[2] Brin, S., - Page, L. The anatomy of a large-scale hypertextual Web search engine. (S. University, Ed.) *Computer Networks and ISDN Systems*, 30(1-7), 107–117. Elsevier Science Publishers B. V. Amsterdam, The Netherlands, 1998.

[3] Nykl, M. Určování významnosti vrcholů grafu: PageRank a jeho modifikace. *Technical report* No. DCSE/TR-2013-09, University of West Bohemia, 2013.

[4] Wasserman, S. *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge, UK, 1994.

[5] Everett, M. G. – Borgatti, S. P. *The Centrality of Groups and Classes.* Journal of Mathematical Sociology 23(3), 1999.

[6] HANNEMAN, R. A. – RIDDLE, M. *Introduction to social network methods*. University of California Riverside, 2005.

[7] BRANDES, U. *A faster algorithm for betweenness centrality* (PDF). The Journal of Mathematical Sociology. vol. 25, issue 2, s. 163-177, 2001. Dostupné z: http://www.tandfonline.com/doi/abs/10.1080/0022250X.2001.9990249.

[8] Šaloun, P. – Genči, J. *Pokroky databázových a znalostních technologií 1* 2012 s. 57.

[9] *Scrapy* [online]. [cit. 18.6.2014]. Dostupné z: http://scrapy.org.

[10] *The Web Robots Pages* [online]. [cit. 18.6.2014]. Dostupné z: http://robotstxt.org.

[11] Brandes, U. *A Faster Algorithm for Betweenness Centrality.* Journal of Mathematical Sociology 25(2), 2001.

[12] Carrington, P. J. – Scott, J. – Wasserman, S. *Models and Methods in Social Network Analysis.* Cambridge University Press, Cambridge, UK, 2005.

Obsah DVD

Struktura a obsah přiloženého DVD je následující:

- **Vstupní data**
  - sitFirmaOsobaFirma.csv – síť firem ve formátu .csv
  - sitOsobaFirmaOsoba.csv – síť osob ve formátu .csv
  - FirmyOsoby_Data.sql – SQL Dump databáze se získanými daty
  - FirmyOsoby_Struktura.sql – SQL Dump struktury databáze

- **Roboti Scrapy**
  - Firmy-Bot – robot pro zpracování firem (viz část 6.3.6)
  - Osoby-Bot – robot pro zpracování osob (viz část 6.3.6)
  - Scrapy.pdf – manuál Scrapy

- **Výsledky analýzy**
  - PageRank – složka obsahující soubory s výsledky PageRanku všech grafů ve formátu .csv
  - Míry centrality - složka obsahující soubory s výsledky hodnocení všech grafů získanými mírami centrality ve formátu .csv

- **Analyza_grafu.jar** – program umožňující vytvoření všech typů grafů je jich následné vyhodnocení (PageRank nebo míry centrality) (viz příloha A)

- BP_sudav.pdf – elektronická verze BP

- BP_sudav.rar – zdrojové dokumenty elektronické verze BP

Přílohy

A Uživatelské příručky

A.1 Webový robot

Pro spuštění webového robota je potřeba nainstalovat Python (v. 2.6 nebo 2.7) a knihovnu Scrapy, kterou lze získat v [9]. V oficiální dokumentaci nalezneme postup či průběh instalace v části *Installation guide*. Pro správný chod webového robota je nutné vytvořit databázi s předepsanou strukturou (viz část 6.3.1). Testování a vývoj robotů bylo řešeno na operačním systému Windows 7. Roboty by mělo být možné spouštět i na jiných platformách, protože všechny použité části jsou platformě nezávislé. Před samotným spuštěním webového robota je potřeba ještě v souboru scrapy.bat nastavit cestu ke složce, ve které je Python nainstalován.

Před samotným stahováním je nutné mít spuštěný program EasyPHP (viz část 6.3.1), který zajistí databázi pro ukládání získávaných dat. Pro samotné spuštění webového robota, respektive zahájení stahování, spusťte příkazový řádek a přepněte se do složky, v níž se robot nachází. Robota následně spustíme příkazem:

scrapy crawl jmeno_robota

Kde jmeno_robota je jméno webového robota, kterého chceme spustit. Weboví roboti, které je možné výše uvedeným příkazem spustit jsou: Start-Bot – uloží do databáze startovní záznamy pro robota Osoby-Bot (musí být spuštěn pouze na úplném začátku stahování), Firmy-Bot – zpracovává záznamy firem uložených v databázi (viz část 6.3.6) a Osoby-Bot – zpracovává záznamy osob uložených v databázi (viz část 6.3.6).

Nedoporučuji spouštět více robotů najednou, protože se může stát, že budou odpojeny nebo nebudou správně fungovat. Robota ukončíme stisknutím kombinace ctrl + c v příkazové řádce. Robot se ukončí po krátké době. Pokud chceme vynutit ukončení robota, tak je nutné stisknout výše uvedenou kombinaci dvakrát. V tomto případě však není zaručeno korektní zpracování všech stránek. Pro navázání na předchozí stahování opět zadáme v příkazové řádce výše uvedený příkaz.

A.2 Analyza_grafu.jar

Tento nástroj obsluhuje celkový chod aplikace. Uživatel se pouze rozhoduje, co přesně chce vyhodnotit, což je řešeno vstupními parametry. Aplikace vytvoří ze vstupního soboru určený typ graf, který bude vyhodnocen buď PageRankem nebo jednou z mír centrality (viz část 7.3). Získané výsledky budou zapsány do příslušných výstupních souborů (viz část 7.2). Pro správnou funkčnost je třeba, aby vstupní soubory měly danou strukturu a pojmenování (viz část 7.1). Dalším předpokladem je, že vstupní soubory budou ve stejné složce jako aplikace, respektive soubor Analyza_grafu.jar.

Pro spuštění aplikace spustíme příkazový řádek a přepne se do adresáře, v němž se nachází soubor Analyza_grafu.jar a příslušené vstupní soubory. Příkaz ke spuštění aplikace vypadá následovně:

java –jar Analyza_grafu.jar –p/c –f/o –a/b/c

Kde první parametr určuje, jestli chceme vypočítat hodnoty PageRanku (-p) nebo některé míry centrality (-c). Druhý parametr to jestli se bude jedna o graf firem (-f) nebo o graf osob (-o) (viz kapitola 4). Třetí parametr určuje přímo daný výpočet. V případě že jde o výpočet PageRanku, tak třetí parametr určuje typ grafu: Graf A -> -a, Graf B -> -b, Graf C -> -c. Pro výpočet míry centrality třetí parametr určuje metodu výpočtu: Degree centrality -> -a, Betweeness centrality -> -b, Closeness centrality -> -c

Při zadání chybného počtu vstupních parametrů se vypíše na monitor hláška o tom, že byl zadán chybný počet vstupních argumentů a aplikace se ukončí. Při chybném zadání některého z parametrů se na monitor vypíše hláška o tom, že vstupní parametry nebyly zadány korektně a aplikace se ukončí.