Předběžný návrh zadání SW projektu

Základní informace

Jméno projektu	**Nástroj pro analýzu sociálních sítí**
Zkratka	*SNanalyzer*
Vedoucí	*Doc. RNDr. Irena Holubová, Ph.D. (holubova@ksi.mff.cuni.cz)* *RNDr. Martin Svoboda, Ph.D. (svoboda@ksi.mff.cuni.cz)*
Konzultanti	*Mgr. Petr Paščenko (petr.pascenko@profinit.eu)* *Mgr. Jan Hučín (jan.hucin@profinit.eu)*
Anotace	*Cílem projektu je vytvořit rozšiřitelný nástroj, který umožní uživatelsky přívětivou analýzu sociální sítě. Vstupem projektu bude sociální síť klientů banky odvozená na základě znalosti (anonymizované) reálné historie bankovních transakcí a dalších informací. Výsledný nástroj bude umožňovat vhodnou vizualizaci sítě, filtrování/dotazování/prohledávání dat a základní analýzy grafových dat vhodné pro danou doménu. Nástroj bude navržen tak, aby jej bylo možné bez většího úsilí rozšířit i pro sociální sítě v jiných doménách, např. sítě odvozené z dat mobilních operátorů atp.*

Motivace

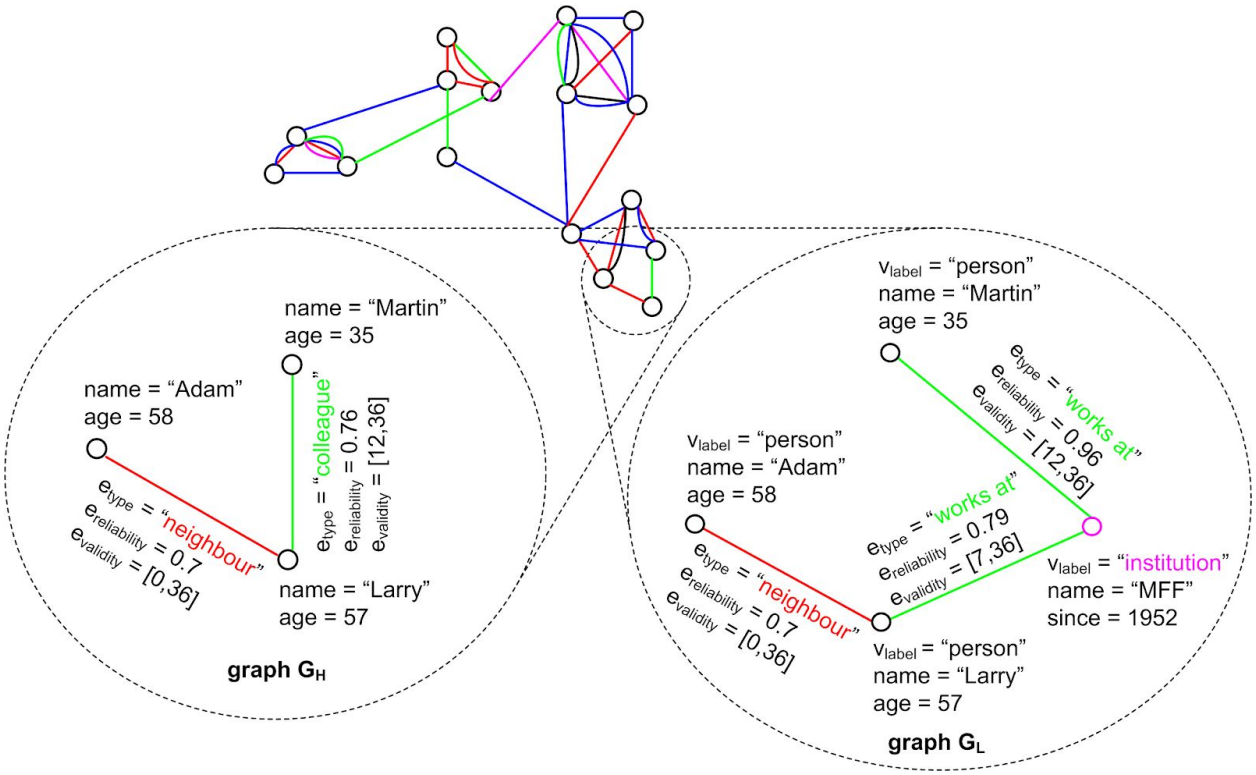
Sociální sítě v současné době představují velmi zajímavý zdroj informací, který umožňuje studovat aktuální a odhadovat budoucí vlastnosti a chování uzlů grafu, jako jsou uživatelé, instituce, zájmové skupiny apod. Přístup k takovému zdroji dat je ovšem z více důvodů značně omezen. Ve spolupráci s firmou Profinit probíhá v současné době na KSI realizace projektu zaměřeného na tzv. ***odvozené sociální sítě****, tj. sítě, které nejsou vytvářeny lidmi, ale které jsou odvozeny nepřímo na základě nejrůznějších informací o jejich chování, vzájemné interakci nebo dalších vlastnostech. Primárně se projekt zaměřuje na bankovní klienty, a síť je tudíž odvozena ze (značně anonymizovaných) dat reprezentujících reálnou historii bankovních operací jednotlivých klientů (jako např. platby kartou, výběry z bankomatů, trvalé příkazy apod.).*

Prvotním nástrojem pro analýzy sociálních sítí je vizualizace grafu, která musí počítat s jejich specifickými vlastnostmi (multigraf, velká data, různé typy uzlů a hran, jejich vlastnosti, časově omezená platnost dat, dynamický vývoj sítě apod.). Dále je možné analyzovat strukturu sítě

prostřednictvím přístupů převzatých z oblasti teorie grafů, jako je např. analýza hustoty, centrality, děr, komunit apod. Také můžeme analyzovat nejrůznější charakteristiky grafu, které vycházejí ze sociologických studií. Např. pojem homofilie vychází z klasického přísloví "vrána k vráně sedá", tedy z pozorování, že lidé mají tendenci vytvářet vztahy s podobnými lidmi. Vybraný uzel tedy můžeme hodnotit na základě vlastností jeho sousedů. Stejně tak můžeme analyzovat např. význam jednotlivého uzlu v rámci celé sítě nebo vybrané části sítě (např. určité komunity), sledovat šíření vlivu, analyzovat chování celé skupiny uzlů apod.

Odvozená sociální síť

Odvozenou sociální síť je možné popsat jako graf, jehož vrcholy reprezentují jednotlivé klienty banky a hrany vztahy mezi nimi. Každý vrchol (klient) má množinu vlastností, z nichž každá má název (např. věk), hodnotu (např. 40), spolehlivost (např. 0,95, tj. danou informaci víme s 95% jistotou) a časovou platnost (např. [0, 47] reprezentující interval prvních 47 týdnů sledovaného období). Každá hrana má typ (např. kolegové) a opět spolehlivost a časovou platnost informace. Takový pohled na klienty banky můžeme označit jako vysokoúrovňový *(viz obrázek 1 vlevo).* Nízkoúrovňový *pohled na danou sociální síť zahrnuje také uzly dalších typů (např. instituce), které je tudíž třeba u vrcholů navíc specifikovat (viz obrázek 1 vpravo).*



Obrázek 1. Vysokoúrovňový (graf G_H) a nízkoúrovňový (G_L) pohled na sociální síť klientů.

Předpokládaná základní analytická práce s grafem zahrnuje následující funkcionalitu:

- *Vizualizace (např. konkrétního klienta a jeho okolí do určité hloubky)*
    - *Důraz na možnost vlastního nastavení barev, tlouštěk, tvarů apod.*
    - *Výběr z layoutů*
    - *Interaktivita (např. "rozkliknutí" informací o uzlu)*
- *Navigace - přesun na další uzly, uzly daného typu apod.*
- *Filtrování (např. vybrané uzly, hrany, časové období, spolehlivost)*
- *Dotazování (např. uzly/hrany/komunity daných vlastností)*
- *Podobnost (např. uzlů/komunit), s možností nastavit parametry jako např. které informace zahrnout, váhy, thresholdy apod.*

Popis SW projektu

Cílem SW projektu je realizovat aplikaci, která umožní uživatelsky přívětivou analýzu dané odvozené sociální sítě bankovních klientů. Aplikace umožní vizualizaci sítě, filtrování/dotazování/prohledávání dat a základní analýzy grafových dat vhodné pro danou doménu. Nástroj nicméně bude (v rámci možností) navržen tak, aby jej bylo možné pohodlně rozšířit pro sociální sítě odvozené z jiných dat (např. mobilních operátorů). Rozšiřitelnost bude upřesněna v rámci podrobné specifikace SW projektu.

Vlastní práce na projektu bude zahrnovat:

- *Seznámení se s problematikou grafových databází, vizualizace a analýz sociálních sítí a samotné bankovní domény.*
- *Návrh cílové aplikace zahrnující*
    - *celkový návrh architektury, příslušných rozhraní a datových struktur,*
    - *volbu konkrétní grafové nebo jiné vhodné databáze pro ukládání dat (předpokládá se některý z osvědčených systémů jako např. Neo4j, OrientDB, ArangoDB apod.),*
    - *volbu konkrétní knihovny pro vizualizaci dat a specifikaci jejího rozšíření/modifikace pro odvozenou sociální síť bankovních klientů,*
    - *volbu konkrétních knihoven pro základní analýzy grafů a specifikaci konkrétní sady analytických operací pro odvozenou sociální síť bankovních klientů.*
- *Implementaci aplikace pro doménu bankovních klientů včetně testování na reálných datech.*

Přesný rozsah vizualizačních a analytických operací bude v rámci podrobné specifikace projektu zvolen tak, aby celková složitost systému odpovídala obecným požadavkům na SW projekty. V rámci navazujících diplomových prací bude možné projekt dále rozšiřovat, tj.

jednotlivé moduly optimalizovat, nahrazovat nebo doplňovat. Možnými tématy mohou být např. další netriviální analytické operace nad grafy, doplňování grafu na základě dalších znalostí apod.

Platforma, technologie

Konkrétní technologie budou stanoveny v rámci podrobné specifikace projektu s ohledem na zkušenosti řešitelského týmu a nalezené vhodné existující knihovny a moduly.

Odhad náročnosti

Pro realizaci aplikace se předpokládá 4-členný řešitelský tým:

- *1 řešitel: načítání, ukládání, dotazování dat*
- *1-2 řešitelé: vizualizace grafu*
- *1-2 řešitelé: analýzy grafu*

V případě většího množství členů může být projekt vhodně rozšířen např. v oblasti realizovaných analytických operací.

Předpokládaný plán prací:

- *Měsíc 1: Seznámení se s cílovou problematikou, podobnými nástroji, grafovými databázemi.*
- *Měsíc 2: Příprava a odevzdání podrobné specifikace projektu, tj. přesné vymezení konkrétních implementovaných funkcí. Rozdělení zodpovědností v rámci týmu, příprava časového plánu.*
- *Měsíc 3-7: Implementace návrhu dle podrobné specifikace.*
- *Měsíc 8: Testování a ladění nad menší sadou reálných dat.*
- *Měsíc 9: Experimenty s kompletními reálnými daty, finalizace, dokumentace.*

Vymezení projektu

Diskrétní modely a algoritmy	
	diskrétní matematika a algoritmy
	geometrie a matematické struktury v informatice
	Optimalizace
Teoretická informatika	
	Teoretická informatika
Softwarové a datové inženýrství	
x	softwarové inženýrství
x	vývoj software
	webové inženýrství
x	databázové systémy
x	analýza a zpracování rozsáhlých dat
Softwarové systémy	
	systémové programování
	spolehlivé systémy
	výkonné systémy
Matematická lingvistika	
	počítačová a formální lingvistika
	statistické metody a strojové učení v počítačové lingvistice
Umělá inteligence	
	inteligentní agenti
	strojové učení
	robotika
Počítačová grafika a vývoj počítačových her	
	počítačová grafika
	vývoj počítačových her