Masarykova univerzita
Přírodovědecká fakulta

Data mining I

Projekt – Internet

Tomancová Nikola, 484549
Finanční a pojistná matematika
22.05.2021

Úvod

Tento semestrální projekt pracuje s datovým souborem Internet, který vznikl na základě výzkumu populace České republiky. Průzkum proběhl přibližně na přelomu 20. a 21. století a zaměřil se na okolnosti, které by mohly ovlivnit například množství času stráveného na internetu nebo způsoby, jakým obvykle dotázaní uživatelé tyto chvíle trávili, tedy jaké internetové služby nejčastěji využívali.

Mým úkolem bylo na základě těchto dostupných dat zjistit, jaká byla situace v ČR z hlediska využívání internetu právě na přelomu 20. a 21. století. Dále se můj projekt věnuje hledání v tehdejší době nejčastěji využívaných služeb na internetu, a nakonec zjišťování služeb, které byly obvykle využívány současně.

1 Příprava dat

Seznámení s daty

Data Internet obsahují celkem 32 proměnných a 2 889 případů, jedná se o reprezentativní vzorek občanů České republiky, kteří byli v té době v produktivním věku (tj. 15 – 65 let). Proměnné můžeme rozdělit do dvou skupin, první jsou proměnné demografické, které se týkají místa bydliště respondentů, jejich pohlaví, věku, vzdělání nebo například jejich finanční situace. Podrobněji se popisu jednotlivých demografických proměnných věnuje Tabulka 1.

Ve druhé skupině figurují především dummy proměnné, tedy proměnné přiřazující daným pozorováním pouze hodnoty 1 nebo 0. Většina otázek se týkala využívání základních internetových služeb, a proto možné odpovědi na takové otázky byly jen „ano“ (zakódované v datech jako hodnota 1) nebo „ne“ (hodnota 0 v našich datech). Za dummy proměnnými pak následují ještě tři proměnné týkající se času stráveného na internetu. Nejprve je to ordinální proměnná s kategoriemi 1 až 6 a za ní následují dvě numerické proměnné udávající množství času v minutách. Tuto skupinu proměnných budeme v tomto projektu nazývat „internetové proměnné“ a jsou shrnuty v Tabulce 2.

Vidíme, že proměnné v našem datovém souboru jsou převážně kvalitativní, jediné kvantitativní jsou proměnné *id*, *hmem*, *q75_1* a *q78_1*. Vzhledem k tomu, že ne všechny proměnné byly v datovém souboru přesně specifikovány, bylo nutné si popisy některých proměnných ještě podrobněji dodefinovat. Především se to týkalo proměnných *q75_1* a *q78_1* označující počet minut strávených na internetu. Budeme předpokládat, že v případě proměnné *q75_1* měli respondenti sdělit počet minut strávených na internetu průměrně ve všední dny a naopak u proměnné *q78_1* udávali průměrný počet minut strávených na internetu během víkendového dne.

Vzhledem k tomu, že ordinální proměnné *agecat* a *q40_23* (označující věkové kategorie a množství času stráveného surfováním po internetu) budeme v následujících kapitolách ještě podrobněji rozebírat, bude vhodné zde uvést ještě tabulky vysvětlující význam jednotlivých kategorií těchto proměnných, viz Tabulky 3 a 4.

Název	Popis	Typ
id	Pořadí respondenta	Numerická
reg	Kraj (Středočeský apod.)	Nominální
vb	Velikost města	Ordinální
hinc	Celkový čistý příjem v domácnosti	Ordinální
hmem	Počet členů v domácnosti	Numerická
hint	Jak často jsou připojeni k internetu	Ordinální
fixline	Pevná linka v domácnosti	Nominální
hcar	Počet aut v domácnosti	Ordinální
sex	Pohlaví	Nominální
agecat	Věkové kategorie	Ordinální
marits	Rodinný stav	Nominální
red	Vzdělání	Ordinální
empst	Zaměstnanecké postavení	Nominální
hhead	Hlava domácnosti	Nominální
pinc	Celkový čistý příjem na osobu	Ordinální
sim	Přítomnost SIM karet v domácnosti	Nominální
seg_ls	Typ osobnosti	Nominální

Tabulka 1: Demografické proměnné

Čištění dat, transformace proměnných

Ještě před samotnou analýzou dat bylo nutné data zkontrolovat a ujistit se, že proměnné neobsahují žádné zřejmě chybné hodnoty. V případě kategoriálních proměnných nebyl z tohoto hlediska nalezen žádný problém. Jak je znázorněno v tabulce na Obrázku 1, po vypsání základních popisných statistik těchto proměnných se ukázalo, že všechny kategoriální proměnné se pohybují ve svém oboru hodnot. Jediný problém u kategoriálních veličin by mohl být v proměnné *pinc* znázorňující osobní měsíční příjem, jelikož obsahuje velmi vysoký počet chybějících pozorování. Ovšem většina respondentů, kteří na otázku týkající se osobního příjmu neodpověděli, patří do věkové kategorie 1, případně 2, tzn. jsou ve věku od 15 do 24 let. Je proto velmi pravděpodobné, že se jedná o studenty bez vlastního příjmu a bylo by možné tyto chybějící hodnoty nahradit nulou.

Tabulka dále ukázala podezřelé hodnoty u numerických proměnných *q75_1* a *q78_1*. Maximum těchto proměnných je totiž v obou případech 1 440 minut, což se rovná v přepočtu 24 hodinám. Jedná se tudíž evidentně o nesprávné hodnoty, které by ale mohly náš další výzkum významně ovlivnit. Z toho důvodu byla tato pozorování vyjmuta z datového souboru, stejně jako pozorování s dalšími vysokými hodnotami, které se objevily v tabulkách na

Název	Popis	Typ
q74_1	Elektronická pošta (e-mail)	Nominální
q74_2	Hledání informací, kontaktů (obecně)	Nominální
q74_3	Vyhledávání, objednání nebo nákup zboží či služeb	Nominální
q74_4	Hraní her na internetu (on-line)	Nominální
q74_5	Obsluha bankovního konta z domova	Nominální
q74_6	Telefonování přes internet	Nominální
q74_7	Chat, diskusní skupiny	Nominální
q74_8	Využívání služeb okamžité komunikace (ICQ apod.)	Nominální
q74_9	Poslouchání hudby nebo sledování hud. klipů přes internet	Nominální
q65_13	Internetové připojení	Nominální
q65_17	Stolní počítač	Nominální
q65_18	Notebook	Nominální
q40_23	Surfování na internetu	Ordinální
q75_1	Čas na internetu – denně	Numerická
q78_1	Čas na internetu – o víkendu	Numerická

Tabulka 2: Proměnné - internet

Obrázku 2. Celkem bylo odstraněno 20 pozorování.

S ohledem na analýzu v následující kapitole byla vhodná ještě jednoduchá transformace části „internetových“ proměnných. Byla vytvořena nová proměnná *inernet*, která vznikla jako součet hodnot u proměnných *q74_1* až *q74_9*. Jinými slovy, proměnná *internet* označuje počet internetových služeb, které daný respondent využívá, jedná se tedy o ordinální proměnnou, která nabývá hodnot od 0 do 9. Relativní četnosti této proměnné znázorňuje sloupcový graf na Obrázku 3.

2 Češi a internet

Nyní se zaměříme na některé zajímavé vztahy mezi proměnnými. V této kapitole nás bude zajímat vztah některých demografických proměnných k využívání internetových služeb a množství času stráveného na internetu.

Nechali jsme si vypsat k jednotlivým proměnným vždy tři další nejvíce korelované proměnné. Z výsledků bylo patrné, že s „internetovými“ proměnnými je ve vyšší míře korelovaná nejčastěji demografická proměnná *agecat*, největší závislost má věk na poslech hudby, chatování, surfování po internetu a trávení času na internetu během víkendu. Nicméně takový výsledek, že způsob a množství využívání internetu závisí na věku uživatele, byl vcelku

Proměnná *agecat*	
Kategorie	**Věk od - do**
1	15 - 20
2	21 - 24
3	25 - 30
4	31 - 40
5	41 - 50
6	51 - 65

Tabulka 3: Proměnná *agecat*

Proměnná *q40_23*	
Kategorie	**Frekvence surfování**
1	denně nebo téměř denně
2	přibližně dvakrát až třikrát týdně
3	přibližně jednou týdně
4	přibližně jednou až třikrát měsíčně
5	méně často
6	nikdy nebo téměř nikdy

Tabulka 4: Proměnná *q40_23*

očekávatelný. Naopak poměrně zajímavá je korelace mezi rodinným stavem a poslechem hudby na internetu nebo například mezi vzděláním a využíváním internetového bankovnictví, vyhledávání informací na internetu nebo vlastnictvím notebooku. Ostatní demografické proměnné už s těmi „internetovými“ příliš velkou korelaci neměly.

Jako názorná ukázka korelace mezi věkem a využíváním internetu slouží tabulka na Obrázku 4. Lze pozorovat, že s přibývajícím věkem klesá množství využívaných služeb, množství času stráveného surfováním na internetu i čas strávený na internetu během víkendů.

Následující tabulka (Obrázek 5) pak ještě přibližuje vztah proměnných *agecat* a *internet*. Můžeme vidět, že skupina lidí nevyužívající žádné z nabízených internetových služeb patří nejčastěji do nejstarší věkové kategorie, jsou tedy ve věku 51 až 65 let. Naopak všechny služby, ze kterých bylo možné vybírat, využívá nejvíce druhá věková kategorie (věk 21 až 24 let). Jsou to totiž lidé, kteří jsou sice dost mladí na to, aby na internetu chatovali, poslouchali hudbu či hráli hry, ale zároveň i dostatečně staří na to, aby již byli samostatní a měli tedy například i své vlastní internetové bankovnictví (na rozdíl od respondentů ve věku 15 až 20 let).

Podrobněji se na korelace mezi některými proměnnými zaměřuje ještě korelogram na Obrázku 6. Korelogram se dělí na 2 části – horní a dolní trojúhelník. V horním trojúhelníku se nacházejí korelační koeficienty proměnných (některé z nich jsou vypsané už i v tabulce na Obrázku 4), vypsané modrou barvou v případě kladné korelace a červenou v případě korelace záporné. Míra korelace je pak naznačena i v dolním trojúhelníku korelogramu pomocí

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum
id		2889	0	1.0000000	2889.00
reg		2889	0	11.0000000	81.0000000
vb		2889	0	1.0000000	6.0000000
hinc		2887	2	11.0000000	31.0000000
hmem		2885	4	1.0000000	11.0000000
hint		2864	25	1.0000000	5.0000000
fixline		2887	2	1.0000000	2.0000000
hcar		2887	2	1.0000000	3.0000000
sex		2889	0	1.0000000	2.0000000
agecat		2889	0	1.0000000	6.0000000
marits		2889	0	1.0000000	5.0000000
red		2889	0	1.0000000	4.0000000
empst		2889	0	1.0000000	9.0000000
hhead		2889	0	1.0000000	2.0000000
pinc		2403	486	11.0000000	31.0000000
sim		2889	0	1.0000000	2.0000000
seg_ls		2889	0	1.0000000	8.0000000
q74_1	q74#1	2889	0	0	1.0000000
q74_2	q74#2	2889	0	0	1.0000000
q74_3	q74#3	2889	0	0	1.0000000
q74_4	q74#4	2889	0	0	1.0000000
q74_5	q74#5	2889	0	0	1.0000000
q74_6	q74#6	2889	0	0	1.0000000
q74_7	q74#7	2889	0	0	1.0000000
q74_8	q74#8	2889	0	0	1.0000000
q74_9	q74#9	2889	0	0	1.0000000
q65_13	q65#13	2889	0	0	1.0000000
q65_17	q65#17	2889	0	0	1.0000000
q65_18	q65#18	2889	0	0	1.0000000
q40_23		2889	0	1.0000000	6.0000000
q75_1		2889	0	0	1440.00
q78_1		2889	0	0	1440.00

Obrázek 1: Chybná pozorování, minima a maxima

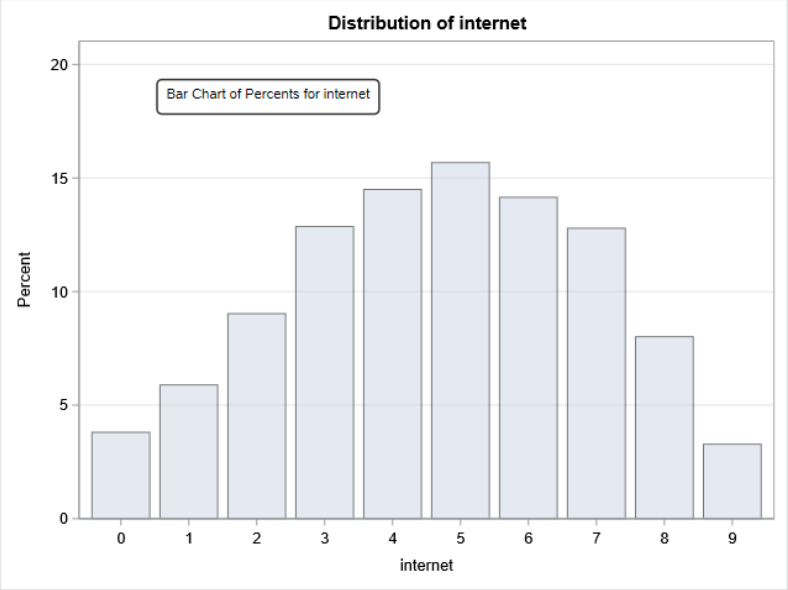
odstínu barev (čím tmavší odstín, tím silnější korelace, tj. tím vyšší korelační koeficient v absolutní hodnotě).

Korelogram potvrdil naše předchozí závěry týkající se proměnné *agecat*. Čím byli respondenti starší, tím méně času trávili u internetu (platí pro všední dny i pro víkendy) a tím méně internetových služeb používali. U proměnné *red* se ukázalo, že úroveň maximálního dosaženého vzdělání nemá příliš velký vliv na celkové využívání internetu. Množství služeb má samozřejmě kladnou korelaci s časem stráveným u internetu. Nakonec jsme se pokusili porovnat čas, který respondenti tráví na internetu ve všední dny a o víkendu, protože by mohlo platit například pravidlo, že ti, kteří tráví na internetu svůj čas v práci během všedních dnů, si pak chtějí o víkendu od počítačů naopak odpočinout a chodí na něj méně. Výsledky korelogramu ovšem tuto hypotézu popírají – platí naopak pravidlo, že respondenti, kteří tráví

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	136	78	1000	1
2	1	8	79	1020	1
3	2	14	80	1200	2
4	3	13	81	1400	1
5	4	7	82	1440	2

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	0	283	68	1000	4
2	1	7	69	1020	1
3	2	16	70	1200	7
4	3	6	71	1400	2
5	4	4	72	1440	5

Obrázek 2: Extrémní hodnoty proměnných *q75_1* a *q78_1*



Obrázek 3: Relativní četnosti proměnné *internet*

více času na internetu v týdnu, pak tráví na internetu více času i o víkendu. Toto pravidlo dokonce platí napříč všemi věkovými kategoriemi, i když bychom vzali jednotlivé věkové kategorie zvlášť, korelační koeficient u proměnných *q75_1* a *q78_1* vychází vždy vyšší než 0,5.

3 Analýza dat

V této kapitole se budeme zabývat dalšími otázkami, které se týkají datového souboru Internet. Budeme zkoumat, jaké služby nejčastěji lidé využívají nebo které služby jsou používány současně. Budeme tedy hledat nějaká nejčastější očekávaná i překvapivá pravidla ve využívání jednotlivých internetových služeb.

Právě pro tyto účely je určená analýza nákupního košíku. Tato metoda se totiž snaží najít

Pearson Correlation Coefficients, N = 2869				
	agecat	internet	q40_23	q78_1
agecat	1.00000	-0.39981	0.42249	-0.32961
internet	-0.39981	1.00000	-0.53388	0.41103
q40_23	0.42249	-0.53388	1.00000	-0.37723
q78_1	-0.32961	0.41103	-0.37723	1.00000

Obrázek 4: Korelace

mezi daty nějaké často se vyskytující vztahy a objevit v nich tímto způsobem zajímavé vzory či pravidla. V případě našich dat to znamená, že se pokusíme naleznout určité vzory chování uživatelů internetu, zjistit, které služby jsou obvykle využívány současně a které naopak nikoliv.

Transakční data

Aby bylo možné provést analýzu nákupního košíku, bude nejprve nutná úprava původních dat. Budeme nyní brát v potaz pouze „internetové" dummy proměnné, vytvoříme proto nejprve tabulku tvořenou pouze vyhovujícími proměnnými.

Ani nově vzniklá tabulka ovšem stále není dostačující. Dále je nutné tato „raw" data převést na transakční. Transakční data obsahují pouze dvě proměnné, jednou z nich je proměnná *ID* označující jednotlivé respondenty, zákazníky apod. Druhá proměnná bývá obvykle pojmenována jako Položka (*Item*) nebo například Transakce (*Transaction*). Jak již název napovídá, do této proměnné uložíme internetové služby, které daný respondent používá.

Četnost služeb

Nově vzniklá data využijeme nejprve k tomu, abychom zjistili, jaké služby jsou využívány nejčastěji. Jak ukazuje graf relativních četností na Obrázku 7, nejrozšířenější internetovou službou byl e-mail, avšak poměrně často byl internet využíván samozřejmě i za účelem vyhledávání informací. Třetí největší zastoupení mezi službami mělo nakupování přes internet. Tabulka relativních četností tedy nepřinesla příliš překvapivé výsledky.

Pro zajímavost ještě uvádíme graf relativních četností pro vlastnictví počítače či notebooku, viz Obrázek 8. Notebooky výrazně převyšují stolní počítače. Lze předpokládat, že vzhledem k technologickému pokroku a potřeby počítačů k práci či výuce z domova by toto porovnání v dnešní době již vyšlo velmi odlišně.

Analýza nákupního košíku

Nyní bude následovat přímo samotná analýza nákupního košíku prováděná v programu SAS Enterprise Miner. Hladinu podpory (*support*) jsme stanovili na 15 % a s mírou spolehlivosti (*confidence*) jsme se pohybovali v hodnotách 80 % a více. Jako maximální počet položek v

The FREQ Procedure

Table of agecat by internet

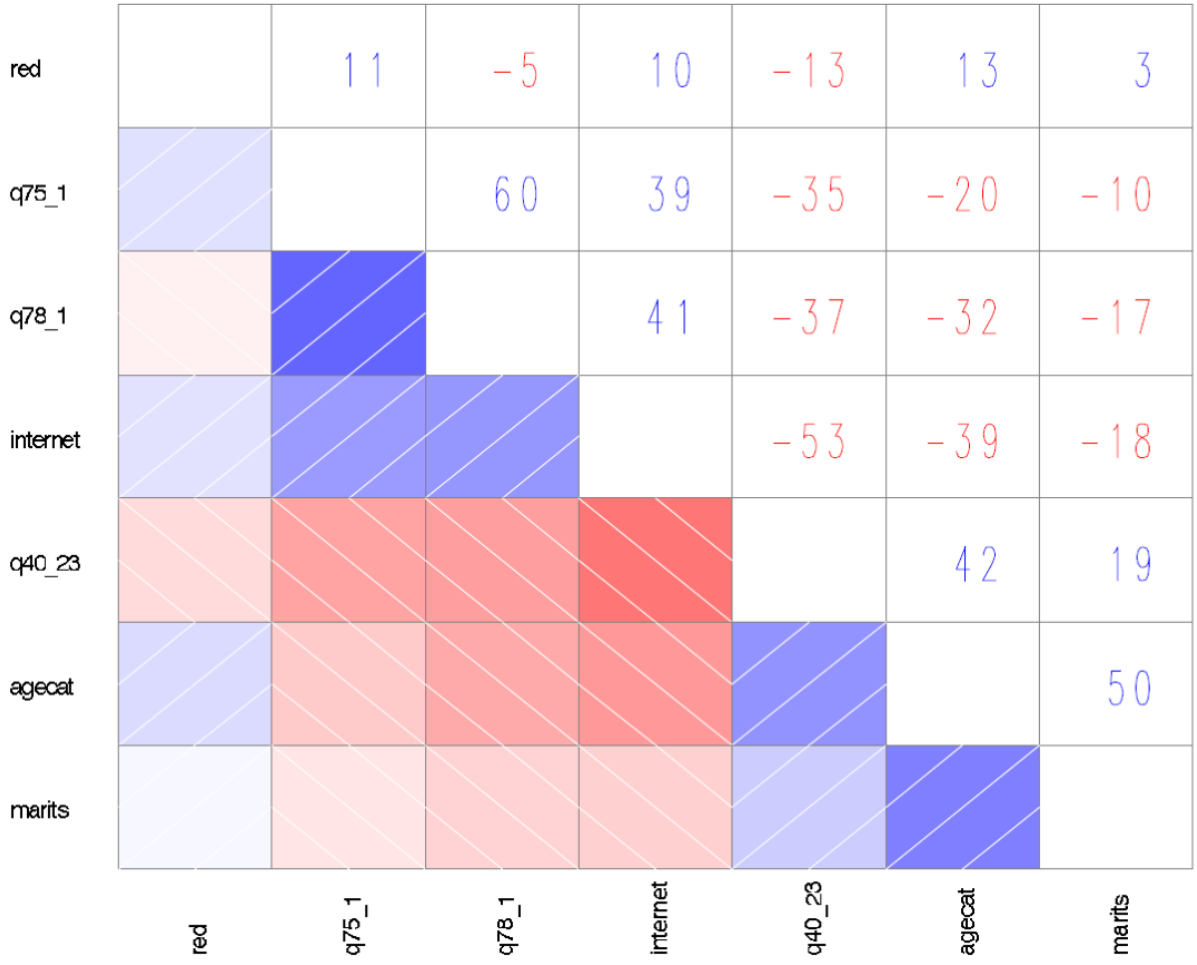
Frequency / Percent / Row Pct / Col Pct	internet										
agecat	0	1	2	3	4	5	6	7	8	9	Total
1	1 0.03 0.26 0.92	11 0.38 2.84 6.51	18 0.63 4.64 6.95	25 0.87 6.44 6.78	42 1.46 10.82 10.10	63 2.20 16.24 14.00	52 1.81 13.40 12.81	96 3.35 24.74 26.16	67 2.34 17.27 29.13	13 0.45 3.35 13.83	388 13.52
2	4 0.14 0.99 3.67	14 0.49 3.45 8.28	16 0.56 3.94 6.18	42 1.46 10.34 11.38	41 1.43 10.10 9.86	57 1.99 14.04 12.67	74 2.58 18.23 18.23	63 2.20 15.52 17.17	55 1.92 13.55 23.91	40 1.39 9.85 42.55	406 14.15
3	2 0.07 0.55 1.83	6 0.21 1.66 3.55	14 0.49 3.88 5.41	44 1.53 12.19 11.92	40 1.39 11.08 9.62	72 2.51 19.94 16.00	68 2.37 18.84 16.75	59 2.06 16.34 16.08	41 1.43 11.36 17.83	15 0.52 4.16 15.96	361 12.58
4	3 0.10 0.58 2.75	17 0.59 3.30 10.06	40 1.39 7.77 15.44	64 2.23 12.43 17.34	96 3.35 18.64 23.08	95 3.31 18.45 21.11	80 2.79 15.53 19.70	76 2.65 14.76 20.71	33 1.15 6.41 14.35	11 0.38 2.14 11.70	515 17.95
5	21 0.73 4.09 19.27	47 1.64 9.14 27.81	76 2.65 14.79 29.34	81 2.82 15.76 21.95	83 2.89 16.15 19.95	71 2.47 13.81 15.78	64 2.23 12.45 15.76	43 1.50 8.37 11.72	18 0.63 3.50 7.83	10 0.35 1.95 10.64	514 17.92
6	78 2.72 11.39 71.56	74 2.58 10.80 43.79	95 3.31 13.87 36.68	113 3.94 16.50 30.62	114 3.97 16.64 27.40	92 3.21 13.43 20.44	68 2.37 9.93 16.75	30 1.05 4.38 8.17	16 0.56 2.34 6.96	5 0.17 0.73 5.32	685 23.88
Total	109 3.80	169 5.89	259 9.03	369 12.86	416 14.50	450 15.68	406 14.15	367 12.79	230 8.02	94 3.28	2869 100.00

Obrázek 5: Internet a věk

jednom pravidle byla nejprve nastavena hodnota 2. Zahrnutí proměnných *q65_13*, *q65_17* a *q65_18* žádné zajímavé závěry nepřineslo, rozhodli jsme se je proto z dat určených k analýze odebrat a nadále ponechat pouze položky týkající se služeb samotných.

Analýza dvouprvkových pravidel především potvrdila důležitost e-mailu a vyhledávání na internetu, protože všechna pravidla nacházející se v grafu spolehlivosti a podpory vpravo (tzn. ty s největší spolehlivostí) byla vytvořena ve formě „nějaká služba ⇒ e-mail“ (resp. „služba ⇒ hledání“ nebo „služba ⇒ nákupy“). Vzhledem k tomu, že e-mail, vyhledávání a nákupy byly nejčastěji používanými službami (viz Obrázek 7), tak není příliš překvapivé, že měla právě pravidla s těmito službami tak velkou spolehlivost. Pro podrobnější výsledky uvádíme na Obrázku 9 několik prvních řádků z Tabulky pravidel (*Rules Table*) seřazených podle spolehlivosti.

Při seřazení tabulky podle sloupce *Lift* se na první příčku dostala dvojice hudba a hraní her (služby se spolu vyskytovali současně 1,65x častěji oproti náhodě). Může to být způsobeno například tím, že hráči online her si často pustí hudbu přímo k hraní jako kulisu. Po seřazení

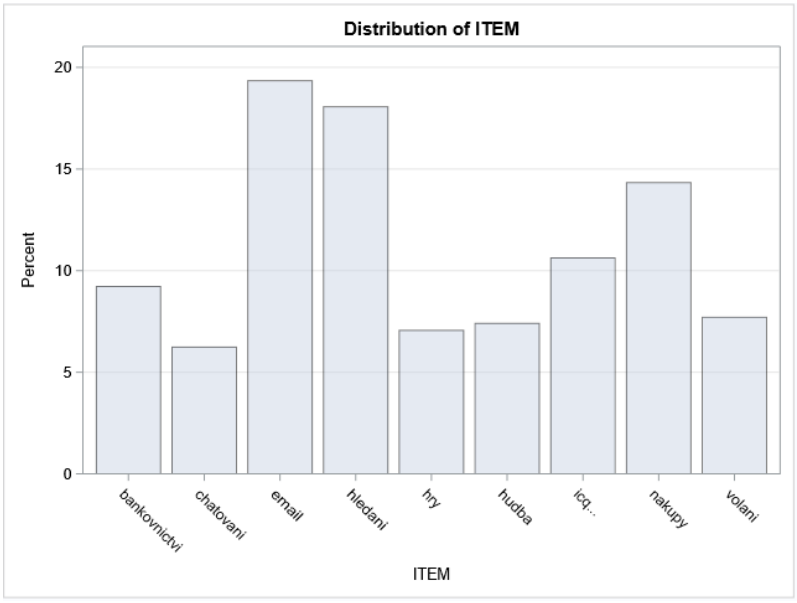


Obrázek 6: Korelogram vybraných proměnných

tabulky podle podpory lze vidět, jaké služby jsou nejčastěji využívány současně. Takovými dvojicemi služeb byly především kombinace služeb s největší relativní četností (např. email a vyhledávání: 83,7 %, nákupy a email 67,54 % apod.).

Pro dokreslení uvádíme ještě Link Graph s orientovanými šipkami (Obrázek 10). I z něj jsou patrná ta nejsilnější pravidla (největší podporu mají ta největší a nejčervenější kolečka, sílu pravidel pak znázorňuje tloušťka a barva šipek).

Dále jsme vyzkoušeli variantu analýzy pro až čtyřprvková pravidla. Výsledky této analýzy zpracované graficky se nachází na Obrázku 12. Čtverečky grafu jsou u vysoké spolehlivosti v naprosté většině ve formě „služba + služba + služba ⇒ e-mail“, například „hudba + nákupy + bankovnictví ⇒ e-mail“ (podpora 17,13 %, spolehlivost 99,79 %). Pokud tedy respondent využívá některou ze zjištěných trojic služeb, pak téměř se stoprocentní jistotou lze říci, že bude mít i e-mail. Po kombinacích se službou e-mail následují obdobná pravidla, ve kterých

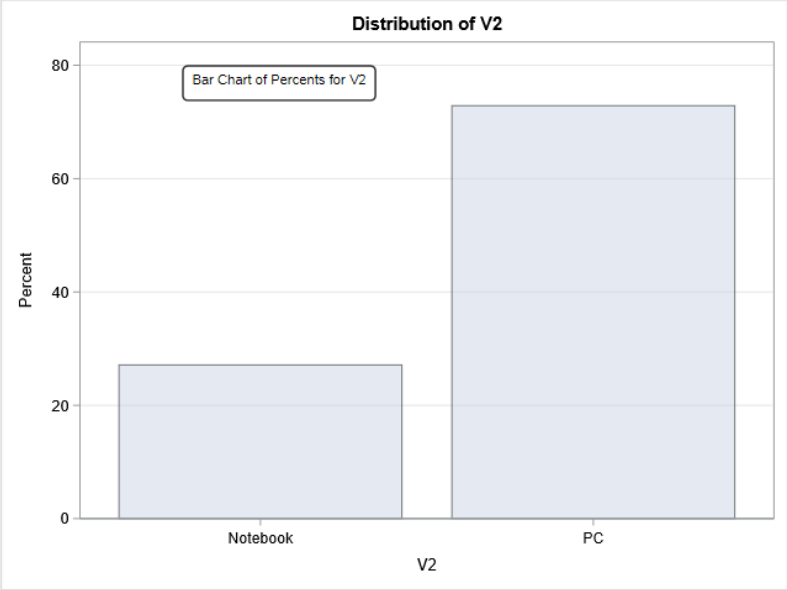


Obrázek 7: Relativní četnosti jednotlivých služeb

figuruje na pravé straně vyhledávání nebo vyhledávání + e-mail (resp. nákupy, nákupy + e-mail nebo nákupy + hledání), poté se na pravou stranu přidávají i služby okamžité komunikace (dále IM podle anglického překladu *Instant Messaging*) atd. V případě IM se už ale spolehlivost pohybuje v nižších hodnotách, mezi 80 až 90 %, například „hudba + hledání + chatování ⇒ IM“ (podpora 14,86 %, spolehlivost 87,5 %), „hudba + chatování ⇒ IM + e-mail“ (podpora 15,11 %, spolehlivost 85,37 %).

Na zajímavý bod, který ukázala matice pravidel (*Rule Matrix*) poukazuje Obrázek 11. Jedná se o pravidlo „nákupy + hledání + e-mail ⇒ bankovnictví“ (spolehlivost 80,51 %), které je svou strukturou velmi odlišně od ostatních okolních pravidel. Vyplývá z něj vztah mezi nakupováním po internetu a vlastnictvím internetového bankovnictví. Tato vazba je poměrně logická, jelikož nákupy přes internet jsou často hrazeny online formou právě prostřednictvím internetového bankovnictví.

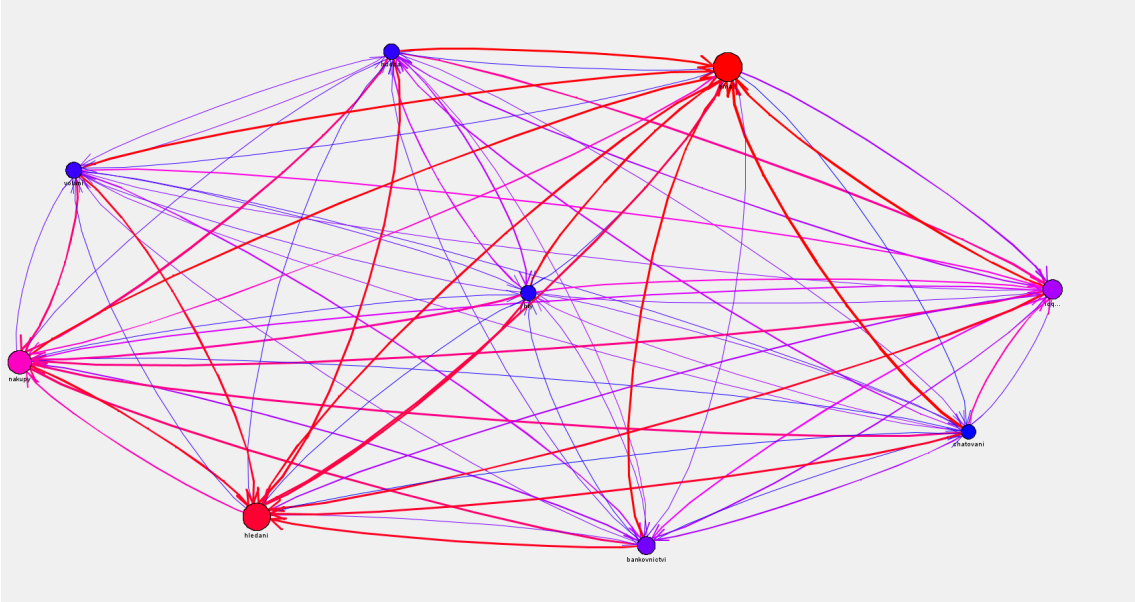
Nakonec jsme pro zajímavost vyzkoušeli přidat do transakčních dat i informaci, do jaké věkové kategorie daní respondenti patří. Nalezená pravidla ovšem opět pouze poukázala na významnost služeb e-mail a internetového vyhledávání. Největší spolehlivost pro využívání e-mailu měla 2. věková kategorie (tj. věk 21 – 24 let), a to ve výši 97 %, takže většina respondentů ve věku 21 až 24 let měla a používala e-mail.



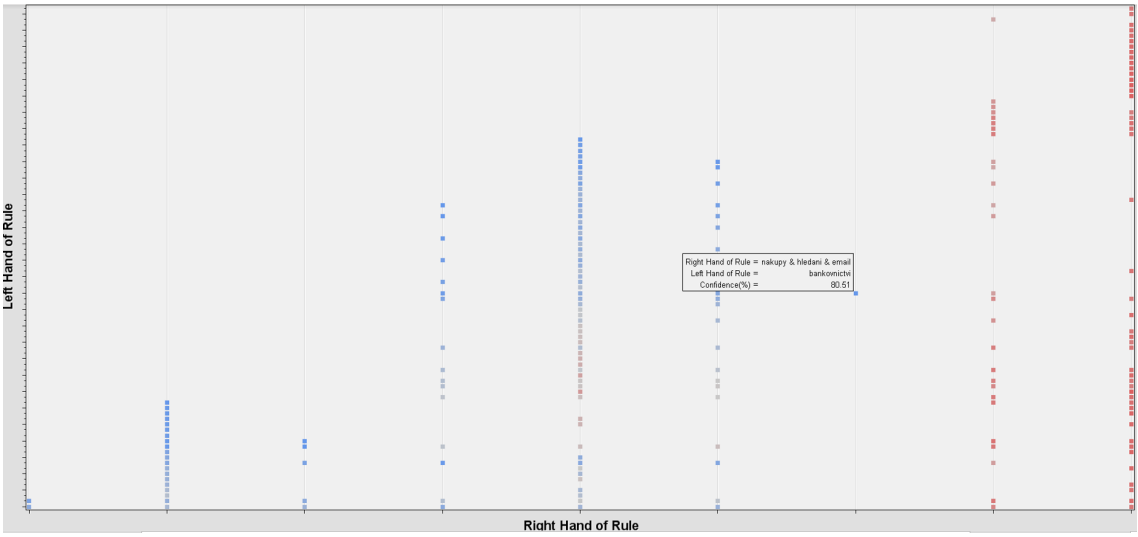
Obrázek 8: Notebooky vs. počítače

Confidence(%) ▼	Support(%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3
98.58	29.90	1.05	831.00	chatovani...	chatovani	email	chatovani	======...	email
98.40	50.88	1.05	1414.0	icq... ==>...	icq...	email	icq...	======...	email
98.00	43.97	1.04	1222.0	bankovni...	bankovni...	email	bankovni...	======...	email
97.90	35.26	1.04	980.00	hudba ==...	hudba	email	hudba	======...	email
97.31	36.49	1.03	1014.0	volani ==...	volani	email	volani	======...	email
96.85	67.54	1.03	1877.0	nakupy =...	nakupy	email	nakupy	======...	email
96.65	33.21	1.03	923.00	hry ==> ...	hry	email	hry	======...	email
95.29	83.77	1.01	2328.0	hledani =...	hledani	email	hledani	======...	email
94.41	34.01	1.07	945.00	hudba ==...	hudba	hledani	hudba	======...	hledani
94.15	42.25	1.07	1174.0	bankovni...	bankovni...	hledani	bankovni...	======...	hledani
93.96	65.53	1.07	1821.0	nakupy =...	nakupy	hledani	nakupy	======...	hledani
93.86	35.19	1.07	978.00	volani ==...	volani	hledani	volani	======...	hledani
93.60	48.40	1.06	1345.0	icq... ==>...	icq...	hledani	icq...	======...	hledani
93.00	28.21	1.06	784.00	chatovani...	chatovani	hledani	chatovani	======...	hledani
90.79	31.20	1.03	867.00	hry ==> ...	hry	hledani	hry	======...	hledani
88.99	83.77	1.01	2328.0	email ==...	email	hledani	email	======...	hledani
84.20	37.78	1.21	1050.0	bankovni...	bankovni...	nakupy	bankovni...	======...	nakupy
83.01	31.13	1.19	865.00	volani ==...	volani	nakupy	volani	======...	nakupy
82.52	29.72	1.18	826.00	hudba ==...	hudba	nakupy	hudba	======...	nakupy
81.35	42.07	1.17	1169.0	icq... ==>...	icq...	nakupy	icq...	======...	nakupy
81.26	24.65	1.17	685.00	chatovani...	chatovani	nakupy	chatovani	======...	nakupy
78.32	28.21	1.51	784.00	hudba ==...	hudba	icq...	hudba	======...	icq...
77.28	26.56	1.11	738.00	hry ==> ...	hry	nakupy	hry	======...	nakupy
75.33	22.85	1.46	635.00	chatovani...	chatovani	icq...	chatovani	======...	icq...

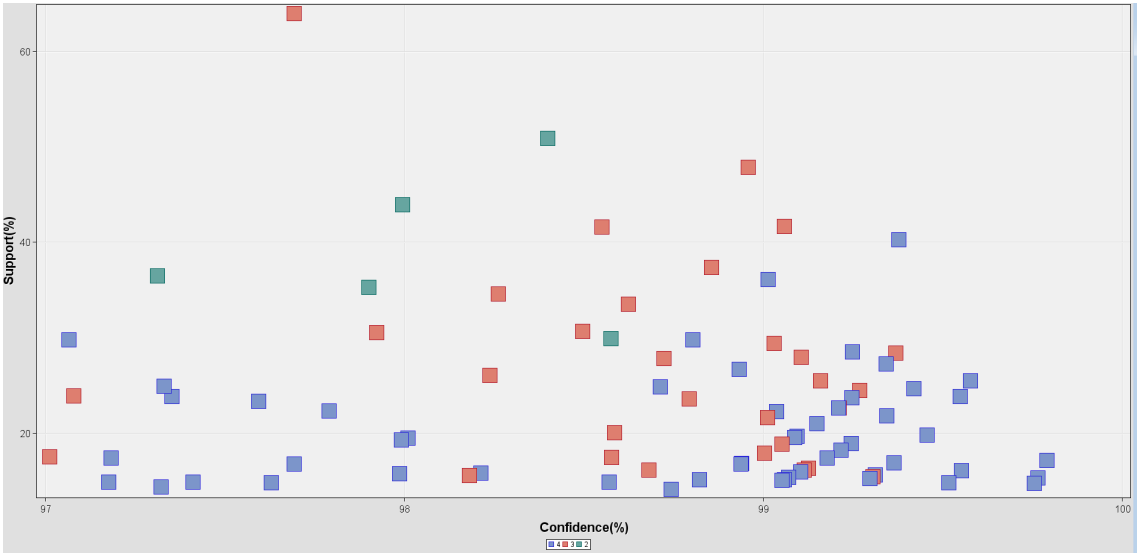
Obrázek 9: Pravidla podle spolehlivosti



Obrázek 10: Link Graph



Obrázek 11: Rule Matrix, max. čtyřprvková pravidla



Obrázek 12: Dvou, tří a čtyřprvková pravidla

Závěr

V tomto projektu jsme hledali odpověď postupně na tři položené otázky. Nejprve jsme se zabývali obecnou otázkou, jaká byla situace v ČR z hlediska využívání internetu. Zjistili jsme, že věk respondenta výrazně ovlivňuje jeho míru využívání internetu. Mladší lidé trávili na internetu více času a používali více služeb než starší ročníky. Pro všechny věkové kategorie je také kladná závislost mezi množstvím času stráveného u internetu během všedního dne a víkendu.

Účelem další části projektu bylo zjistit, jaké internetové služby respondenti využívali nejčastěji. Necelá pětina respondentů udala, že aktivně využívá e-mail. Dalšími hojně se vyskytujícími službami v našich datech bylo vyhledávání na internetu, nákupy přes internet nebo například využívání služeb okamžité komunikace.

Nakonec jsme zjišťovali, jaké služby byly v té době nejčastěji používány současně. Vzhledem k vysoké četnosti výskytu kladné odpovědi u služeb e-mail a vyhledávání byly nejčastěji současně využívané služby právě různé kombinace služeb obsahující e-mail a vyhledávání. Narazili jsme ale i na překvapivější souvislosti, například na vztah mezi hraním her a poslechem hudby nebo nakupováním přes internet a vlastnictvím internetového bankovnictví.

Možným nedostatkem této analýzy by mohl být fakt, že kvůli přílišnému výskytu pravidel „služba $\Rightarrow$ e-mail nebo vyhledávání“ jsme mohli přehlédnout nějaká sice méně spolehlivá, ale zajímavá a méně očekávaná pravidla.

Bylo by nepochybně zajímavé podobnou analýzu provést i na datech pocházejících z dnešní doby a získané výsledky porovnat s výsledky z tohoto projektu. Bylo by možné pozorovat změny četnosti využívání jednotlivých služeb, které by byly jistě podstatně vyšší než na přelomu tisíciletí.