OverNight a změny v NetMonitoru

www.gemius.com

Sběr dat a gemiusPrism

V nové vlně NetMonitoru bude sběr dat probíhat za pomoci rozhraní gemiusPrism na rozdíl od gemiusTraffic. Do gemiusPrism jsou již nyní zmigrovány všechny účty pro měření webových stránek, všechny účty pro měření Streamovaného obsahu a také účty pro měření aplikací.

Stále je možné v aplikaci gemiusTraffic editovat strukturu skriptů, nicméně tato možnost bude vypnuta po přechodu na gemiusPrism. Přepnutím editace stromu do gemiusPrism bude gemiusTraffic v podstatě uzamčen do read-only podoby a další administrace webu bude probíhat už jen přes gemiusPrism.

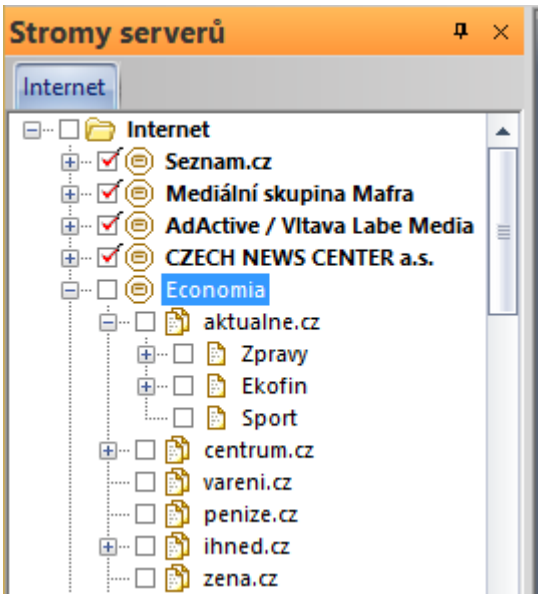
Z rozhraní gemiusPrism se získávají data pro produkci dat OverNight.

Za zmínku také stojí to, že gemiusTraffic je založený na měření Cookies, zatímco gemiusPrism je založený na měření browserID.

Strom médií

Strom médií se v OverNight se neskládá jen z webů a jejich sekcí, jako to bylo v předchozí metodice, ale základním pohledem je pohled přes Provozovatele. Na první úrovni stromu jsou tedy provozovatelé, pod kterými jsou jejich měřené weby včetně sekcí.

Kategorizace a uživatelské balíčky jsou nadále součástí definice agregátů.



Prezentace dat v rozhraní gemiusAudienceRatings (gAR)

Hlavním veřejným rozhraním pro zobrazení výsledků NetMonitoru bude gAR. Rozhraní gAR je primárně jedním z modulů gemiusPrism, nicméně bude zpřístupněný jako zvláštní rozhraní pro veřejnost. Níže je výpis rozdílů mezi OLA a rozhraním gAR:

OLA	gAR
Interface: http://online.netmonitor.cz/	Interface: veřejně dostupný na http://online.netmonitor.cz Navíc Rating tab v rozhraní gemiusPrism.
Pouze pohledy na návštěvnost bez sociodemografie.	Mimo klasické návštěvnosti také základní údaje o základní sociodemografii a možnost sledovat trendy.
Strom v OLA se může lišit od oficiálních výsledků.	Strom v gAR je stejný jako v gem souboru.

Data v OLA se mohou lišit od dat v gem souboru.	Data v gAR jsou stejná jako data v gem souboru.
Data jsou publikována následující den a jsou nezávislá na gem souborech.	Data jsou publikována následující den společně s publikací gem souboru.
Hodnoty RU jsou nahrávány do OLA z produkce dat jednou za měsíc.	Hodnoty RU jsou do gAR nahrávány na denní bázi.
Funguje	Funguje rychleji

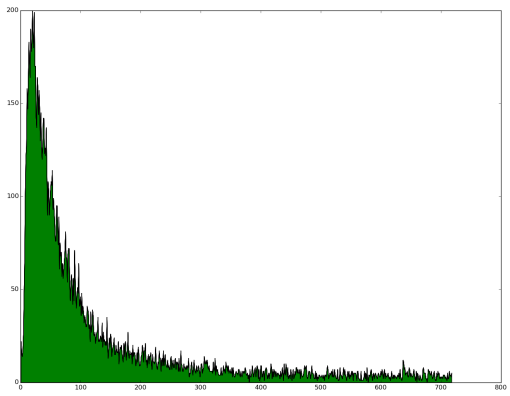
Filtrování trafiku

Filtrování autorefreshe

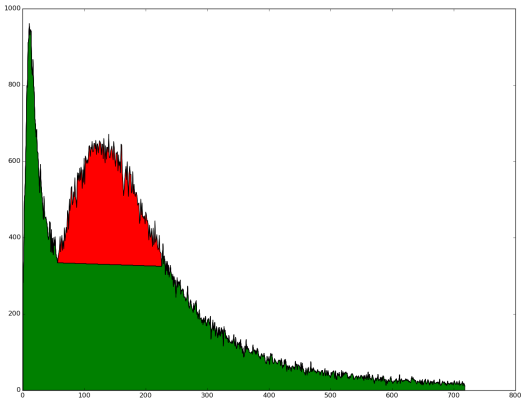
Cílem filtrování autorefreshe je vyfiltrovat z dat zobrazení stránek, které nebyly vykonány akcí uživatele, ale na základě automatického obnovení stránky na straně provozovatele webu.

Hity z autorefreshe jsou identifikovány na základě rozložení zobrazení v čase pro daný uzel. Pokud stránka využívá autorefresh, zpravidla je vidět, že zobrazení stránek má vrchol v určitých pravidelných časových intervalech (např. 180, 300, 420 sekund).

Rozložení zobrazení v čase – Bez autorefreshe:



Rozložení zobrazení v čase - s autorefreshem:



Pokud je zaznamenaný náhlý nárůst zobrazení v určitém časovém úseku, lze usoudit, že se jedná o autorefresh. Analýza rozložení PV v čase pak umožňuje zjistit, kde vrchol začíná a kde končí. Tuto informaci lze pak použít pro výpočet podílu PV pro danou službu, které vznikly na základě autorefreshe. Pokud takový podíl vyjde např. 20 %, pak pro každého uživatele, který byl na stránce určitou dobu, existuje pravděpodobnost, že 20 % jeho aktivity pochází z autorefreshe. Tato nadbytečná aktivita je pak odfiltrována.

Filtrování iFrame

Cílem filtrování iFrame je odfiltrování nevalidních nebo podvodných zobrazení z NetMonitoru. Užití iFrame často znamená, že více zobrazení může být odesláno během jednoho načtení stránky. Může také dojít k tomu, že iframe na stránce, která náleží účtu1, vyvolá zobrazení na stránce náležící účtu2.

Proto jsou z výsledků filtrovány zobrazení, která jsou vyvolána z framu, který náleží jiné doméně, než které patří hlavní stránka (iframe typu 3).

Výjimku z filtrování iFrame mají AMP – Accelerated Mobile Pages, které jsou odlišeny na základě dodatečných parametrů a mohou tak být započítány do návštěvnosti stránek.

Určování denních univerz

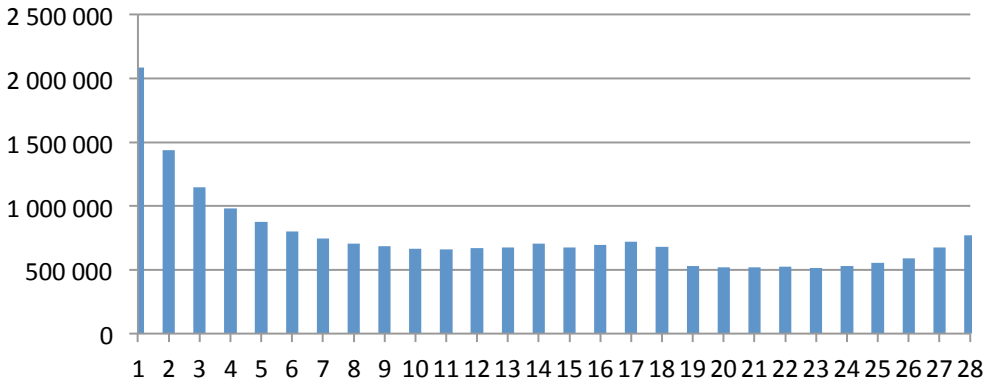
V této kapitole bude popsán algoritmus pro určení aktivní denní populace internetových uživatelů. Využívá panelistů narekrutovaných na jednotlivých platformách a site-centric data pro jednotlivé platformy.

Zdroj dat

Pro určení denní aktivní populace jsou potřeba panely pro jednotlivé analyzované platformy a mimo sociodemografických dat je zohledňována také naměřená aktivita good BID. Tato naměřená aktivita vytváří

vážící proměnnou, která se nazývá Frekvence využívání internetu a je definovaná jako počet dnů aktivity good BID za posledních 28 dní před produkcí dat.

Graf frekvence využívání internetu může vypadat např. následovně:



Dále je využívána struktura a odhad velikosti internetové populace pro daný měsíc z externího strukturálního výzkumu.

Algoritmus určení denní populace

V prvním kroku se sestaví set reportujících panelistů (těch kteří byli zastoupení v panelu před více jak 28 dny). Tento set se váží na strukturu populace, velikost internetové populace a frekvenci využívání internetu. Na základě tohoto procesu získá každý panelista základní (iniciační) váhu. V následujícím kroku je upravena iniciační velikost internetové populace (vychází ze strukturálního výzkumu). Při tom se vychází z předpokladu, že v krátkém časovém období je počet lidí surfujících na internetu a BrowserID stejný. Pokud je suma iniciačních vah rozdílná od počtu browserID, vypočítá se faktor, který by měl upravovat tuto inkonzistenci. Průměrný faktor je následně použit na velikost internetové populace. Proces vážení je opakován, dokud každý panelista nezíská finální váhu. Suma finálních vah aktivních panelistů v daném dnu pak udává velikost denní aktivní populace.

Demografická struktura

Panel výběru je vážen na strukturu populace. V případě měsíčních dat je panel jednoduše vážen na strukturu danou z externího výzkumu. V případě denních dat je situace složitější. V tomto případě je používán Software panel, který registruje aktivitu u obou auditovaných i neauditovaných webových stránek.

Panelisté musí splňovat následující podmínky:

- Vstup do panelu alespoň 28 dnů před dnem, kdy odhadujeme aktivní populaci.
- Aktivní v daný den.
- Nebo v případě neaktivity v daném dni, aktivní během předcházejících 28 dnů (pro vytvoření měsíční populace).

Panel je vážen na měsíční strukturu dat, dále na frekvenci používání internetu (viz určení denní populace), a dále na velikost internetové populace pro daný měsíc.

Poté, co je panel zvážen, vybereme ty panelisty, kteří byli v daný den aktivní. Denní struktura je definována jako distribuce demografických charakteristik těchto panelistů po vážení.

Taková demografická struktura je považována za vstup pro vážení konečných denních dat. Na tuto strukturu je vážen výsledný panel po fúzi Pop-up a Software panelu (BPS) a po fúzi Home & Work.

Výpočet RU

V rámci výpočtu RU došlo k několika metodickým změnám. Cílem této kapitoly je tyto změny představit.

Beast28

Metodika Beast je již dva roky používána v rámci NetMonitoru. V OverNight metodice dochází k jejím drobným modifikacím.

První z modifikací se týká přechodu z Beast15 na Beast28. Čísla za Beast se odkazují k metodikám EC15 resp. EC28. Metodiky EC byly založeny na principu určování Estimate Cookies. Proto, aby bylo možné vypočítat Estimate Cookies, bylo potřeba definovat tzv. Good cookies. Good cookie byla taková cookie, která se objevila před měřeným měsícem, v rámci měřeného měsíce a po skončení měřeného měsíce. Číslo za EC pak říkalo, do kolikátého dne následujícího měsíce se musí cookie objevit na měřené stránce, aby mohla být prohlášena za dobrou. EC15 tedy říkala, že cookie se musí objevit do 15. dne následujícího měsíce, aby s ní bylo počítáno jako s dobrou cookie. V rámci EC28 by se muselo na cookie čekat 28 dnů. EC28, tak není příliš praktická metodika pro výpočet RU, protože by výsledky byly nejdříve až 4 týdny pro ukončení měřeného měsíce. Přesto by metodka EC28 dávala teoreticky nejlepší výsledky pro výpočet RU.

V rámci Beast se EC metodiky (kde se ovšem nepočítají good cookies, ale good browsers) používají pro verifikaci a kalibraci algoritmu pro výpočet Estimate Browsers v rámci Beast. Metodika Beast28 je tedy verifikována proti algoritmu EC28 (počítajícím s browsery). Při přechodu od Beast15 k Beast28 se tak zlepší algoritmus, proti kterému se Beast verifikuje a kalibruje.

Metodika dynamického J-koeficientu

Pro připomenutí je dobré uvést, že J-koeficient je koeficient, kterým se násobí počet Estimate browserů (EB), aby tak vzniklo RU (RU(web) = EB(web) * J). V klasické Beast (nebo EC) metodice je J = Universum * Reach / EB(Internet). J je tak pro všechny weby stejné. Metodika dynamického J-koeficientu však počítá pro každý web J-koeficient zvlášť. Metoda výpočtu je pak založena na následujících předpokladech:

- Pro malé stránky (uživatel navštíví stránku jednou měsíčně) představuje jeden Estimate Browser jednoho uživatele.
- Pro větší stránky je větší pravděpodobnost, že uživatel přistupuje na stránku z více prohlížečů, čili J koeficient se blíží globálnímu J koeficientu.

Platí tedy, že čím je reach web z hlediska návštěvnosti menší, tím více se J koeficient blíží 1.

Při stanovování J koeficientu v rámci této metodiky se přihlíží k:

- Frekvenci návštěv;
- Reachi;
- Konzistenci výsledků při pohledu na uzel a jeho podřízené části;
- Konzistenci výsledků při pohledu přes krátká a dlouhá období.

Tato metodika vykazuje lepší chování v porovnání se statickým J, na vzorku dat získaných ze softwarového panelu.

RU ČR bez ohledu na platformu

Výpočet RU z ČR bez ohledu na platformu vychází z konstrukce panelu na základě BPS. Z fúzovaného panelu pomocí BPS metody se vezmou panelisté, kteří navštívili daný uzel a součet jejich vah dá dohromady celkové RU z ČR pro daný web. Metodika fúze panelu pomocí BPS je popsána dále v tomto dokumentu.

RU ČR+zahraničí bez ohledu na platformu

Pro výpočet celkových RU z ČR a zahraničí bez ohledu na platformu se využívá matematické funkce. Tato funkce je popsána v současné metodice. Zatímco dříve se pomocí matematické funkce počítaly jak RU z ČR, tak z ČR+zahraničí, v metodice OverNight jsou počítány jen RU ze zahraničí (resp. z ČR+zahraničí). RU z ČR jsou počítány na základě BPS.

Behaviorální syntéza Panelů (BPS)

Fúze Pop-up panelu a Softwarového panelu se skládá ze tří kroků:

- Vážení
- Metrické klastrování
- Spojování nejbližších sousedů

Každý krok je následován verifikačním procesem, během kterého je vyprodukován automatický report, na základě kterého se dá ověřit správnost procesu.

Vážení

Nejdříve je každý z panelů vážen podle standardních pravidel vážení NetMonitoru a to jak na strukturální proměnné získané z externího výzkumu, tak na frekvenci používání internetu (viz kapitola Určování denních populací). Na základě tohoto vážení pak oba panely reprezentují stejnou populaci internetových uživatelů a je tedy možné oba panely spojit dohromady.

Metrické klastrování

Po navážení je každý z panelů (Pop-up a Software panel) rozklastrován do podmnožin uživatelů, kteří mají podobné vybrané sociodemografické proměnné a také na základě toho, zda navštěvují jen neoskriptované weby. To znamená, že klastry se skládají z panelistů, kteří mají stejné hodnoty pro vybrané proměnné.

V rámci každého klastru jsou zkombinováni dohromady panelisti z různých panelů. Je tím zajištěno, že dva rozdílní panelisti (z různých klastrů) nebudou spojeni dohromady (např. muž a žena).

Výpočet vzdálenosti

V rámci každého klastru je stále velmi mnoho možností, jak panelisty z různých panelů zkombinovat. Vzhledem k tomu, že sociodemografický profil není jediná proměnná, která ovlivňuje chování uživatelů, existuje v rámci klastrů velké množství vzorců chování. Aby se předešlo fúzování panelistů s výrazně jiných chováním, používá se pro jejich spojování vzdálenostní funkce (distance function).

V rámci každého klastru je tedy pro každé dva panelisty z Pop-up panelu a Software panelu vypočítána jejich vzdálenost.

Spojování nejbližších sousedů

Pop-up panel je definovaný jako příjemce, Software panel je definovaný jako dárce. Aby bylo zaručeno, že oba klastry reprezentují stejný počet uživatelů, je vypočítána suma vah pro každý klastr. Pokud je v sumách rozdíl, jsou oba panely přeškálovány na průměr sum vah z předchozího kroku.

Vzhledem k tomu, že v Pop-up panelu je více panelistů než v Software panelu, je pravděpodobné, že panelista za Software panelu bude napárován s více než jedním Pop-up panelistou. Na algoritmu, který se používá pro párování, je založený na procesu absorbce vah: Software panelisti jsou seřazeni na základě vah pro auditované weby a následně jsou pro každého Software panelistu jeho příslušní Pop-up panelisti seřazeni podle vzdálenosti na základě vzdálenostní funkce. Následně software panelista absorbuje váhy jednotlivách Pop-up panelistů v pořadí, jak byli seřazeni vzdálenostní funkcí, dokud se absorbované váhy nebudou rovnat váze Software panelisty. V okamžiku, kdy dojde k rovnosti, nemusí být použita celá váha Pop-up panelisty. Nicméně tato zbývající váha může být přiřazena dalšímu Software panelistovi.

Po fúzi všech panelistů je pak BPS panel použit v rámci standardního produkčního cyklu. Tedy je znovu vážen na strukturální a behaviorální proměnné a použit pro výpočet výsledků.

PC Home a PC Work

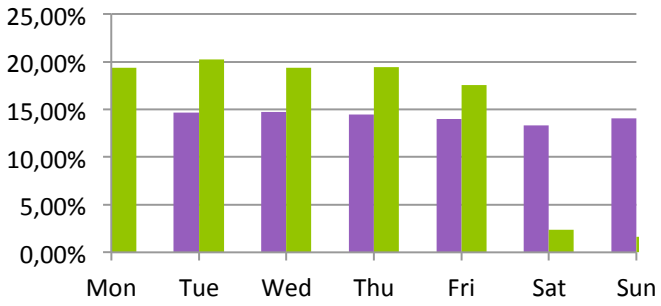
Zavedení platforem PC Home a PC Work nám umožňuje rozdělit prezentovaná data podle lokací používání internetu se zachováním na zřeteli, že někteří uživatelé se připojují z jednoho PC např. jen doma nebo jen v práci. Jiní uživatelé se ale mohou díky laptopu připojovat z jednoho PC jak doma tak v práci. Různé lokace připojení také mohou označovat různé vzorce chování. Např. pro PC home může být typická konzumace video obsahu, zatímco v rámci návštěvnosti z práce může být typické využívání např. korporátních webů. Stejně tak se pro obě platformy může lišit strávený čas s tendencí trávit více času na internetu doma.

Platformy Home and Work jsou vytvořeny na základě toho, že každá IP adresa je přiřazena do jedné z kategorií Home nebo Work a stejně tak je do kategorie Home a Work přiřazen i každý validovaný panelista. Na základě spojení těchto platforem následně vzniká celá platforma PC.

Klasifikace IP adres

První fáze klasifikace IP adres vychází z údajů z dotazníku, kde panelisti vyplňují, kde jsou připojeni v okamžiku vyplňování dotazníku.

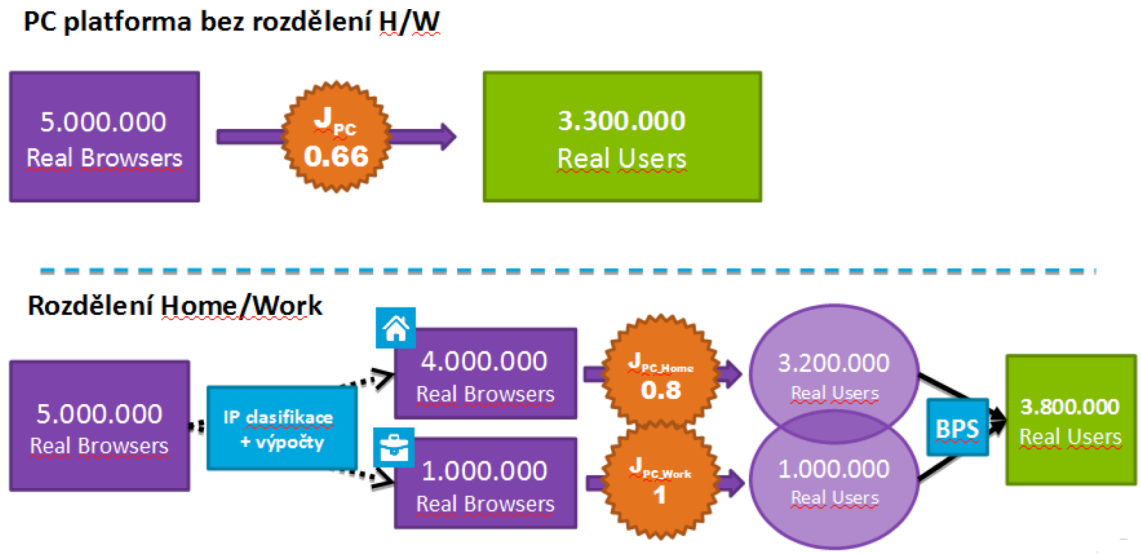
Pro tyto IP adresy se následně vypočítává, kdy jsou aktivní (jaké dny, jaké části dne) a jak dlouho je každá IP adresa aktivní. Cílem je identifikovat vzorce aktivity typické pro platformu Home a pro platformu Work. Tyto informace jsou použity v rámci automatického strojového učení, aby byl vytvořen a validován model pro přiřazování IP adres do kategorie Home nebo Work. Výsledkem je kategorizace všech IP adres z NetMonitoru (všechna site-centric data) do jedné z kategorií Home nebo Work.



V rámci OverNight je tato klasifikace prováděna jednou týdně je platná po dobu jednoho týdne. V rámci měsíční produkce se tak může stát, že jedna IP adresa je jeden týden klasifikována jako Home a další týden jako Work.

Klasifikace panelistů

V dalším kroku jsou panelisté klasifikování jako Home nebo Work. Tato klasifikace je prováděna na základě kategorií IP adres a toto přiřazení je validováno na základě údajů z dotazníku. Zejména se k tomu využívá otázka, z jakých míst se panelista připojoval k internetu za poslední měsíc z právě používaného zařízení. Pouze pokud souhlasí IP klasifikace s odpovědí v dotazníku, rozhodne se o tom, zda bude panelista zařazen do platformy Home nebo Work. Panelista může být na základě své aktivity přiřazen do obou platforem, pokud tomu odpovídá IP klasifikace a odpovědi v dotazníku. Avšak jeho aktivita je vždy započítána do jedné z platorem, podle klasifikace IP adresy.



Ilustrace 1: Vysvětlení vlivu rozdělení Home and Work na RU PC

Behaviorální syntéza (BPS) pro Home a Work

Cílem Behaviorální syntézy panelů je obecně z jednoplatformových panelů získat multiplatformní panel. Při tom se využívají znalosti z kalibračního panelu, ze kterého se získávají vzorce pozitivních duplikací, vzorce negativní exkluzivity a také se usuzuje na důležitost jednotlivých behaviorálních charakteristik. Popis BPS je v jíné části tohoto dokumentu, ale v případě Home and Work platforem jsou hlavní kroky následující:

1) Před samotnou syntézou jsou obě platformy váženy na strukturální data, frekvenci užívání internetu, behaviorální informace (návštěvnost na zapojených webech) a počet internetových uživatelů pro každou platformu (informace získané ze strukturálního výzkumu).

2) Kalibrační panelu se skládá z panelistů, kteří jsou kategorizovaní v obou platformách Home a Work.

3) Obě platformy Home a Work jsou klastrovány na základě pohlaví a věku. Následně v rámci klastrů jsou panelisté z obou platforem propojeni a spojeni za pomoci BPS algoritmu.

Výsledkem je, že platformy Home a Work jsou spojeny do platformy PC. Tato platorma obsahuje tři typy panelistů: Pouze Home, Home a Work, pouze Work.

Gemius, s. r. o.

Českobratská 2778/1

130 00 Praha

contact.cz@gemius.com

www.gemius.com