Универзитет у Београду

Математички факултет

Горан Обрадовић

Анализа образаца понашања корисника Википедије

Мастер рад

Београд

2013.

Универзитет у Београду

Математички факултет

Мастер рад

Аутор: **Горан Обрадовић**

Наслов: **Анализа образаца понашања корисника Википедије**

Ментор: **др Саша Малков**

Чланови комисије: **др Филип Марић**

др Владо Филиповић

Садржај

1 Увод

Овај рад представља покушај да се у оквиру софтвера МедијаВики развије алат за упоређивање профила понашања корисника Википедије, на основу јавно доступних података. Циљ је да се са што већом сигурношћу утврди да ли два конкретна одабрана корисничка налога припадају истој особи.

Постоји више мотива због којих злонамерни корисници употребљавају различите налоге. Неки од њих су покушаји да се избегну блокаде[1] или да се оствари већина у гласању, или бројчана премоћ у дискусијама у циљу промовисања неког политичког или другог интереса. Како је Википедија један од најзначајнијих извора информација на Интернету и често једини извор који читаоци консултују, веома је значајно да се деловање злонамерних корисника сведе на најмању могућу меру.

Постојећи алати за проверу да ли два корисничка налога припадају истој особи се заснивају на упоређивању техничких параметара као што су *IP* адресе. Мане таквог приступа су недовољна поузданост, нарушавање приватности и једноставност фалсификовања параметара. Алат који се развија је замишљен као комплемент тренутном приступу како би се повећала поузданост анализа.

Овај рад полази од претпоставке да уређивачке аспекте који зависе од психологије и навика појединца („бихејвиоралне“ параметре), за разлику од техничких параметара, није једноставно изменити. На основу те претпоставке, изразита сличност понашања два корисника може да представља индикацију да би иза оба налога могла да стоји иста особа, док би различито понашање било индикација да иза два налога стоје различите особе. За утврђивање сличности двају корисничких налога су развијене метрике које се базирају на статистичким методама.

1.1 Основни појмови

Вики је тип веб-локације (сајта) чији садржај сваки посетилац може да уређује. Први вики је развио програмер Ворд Канингем (*Ward Cunningham*), и постао је доступан на Интернету 1995. Најпознатији вики је Википедија.

Википедија је слободна Интернет енциклопедија. Одредница „слободна“ има два значења: слободна за коришћење (бесплатна и слободна), и слободна за уређивање (свако може да врши измене у њој). Википедија је доступна на више од 270 језика.

Задужбина Викимедија је непрофитна организација са седиштем у САД, која подржава Википедију и њој сродне пројекте.[1] Циљ Викимедије је да омогући да целокупно људско знање буде слободно доступно свим људима (на њиховом матерњем

[1] У складу са правилима која дефинише заједница, корисници који својим понашањем штете пројекту или нарушавају односе у заједници могу да буду блокирани, то јест да им се привремено или трајно онемогући уређивање Википедије.

језику). Задужбина Викимедија финансира сервере, развој софтвера, правнички тим, и конкретне активности које се тичу стварања, промовисања и ширења слободног знања. Задужбина не води уређивачку политику на Википедији, не бави се уређивањем, и не утиче на садржај, већ су ти задаци остављени заједници.

МедијаВики је софтвер који покреће Википедију. Писан је у програмском језику *PHP* и лиценциран под слободном лиценцом (*GNU GPL*)[2]. Развија га и одржава заједница програмера од којих су неки запослени у Задужбини Викимедија или у неком од локалних огранака, док други долазе из редова волонтера, најчешће Википедијанаца (чланова заједнице људи који активно уређују Википедију) који су стручни и заинтересовани за развој софтвера. МедијаВики се користи на бројним вики веб-локацијама. Примарно је пројектован за *MySQL*, али су подржани и други системи за управљање базама података: *PostgreSQL, SQLite*, а у одређеној мери и *Oracle* и *DB2*[3]. Викимедијини пројекти користе *MySQL*.

Википедију уређују њени корисници. Сваком кориснику је додељена једна или више улога. Следе најзначајније корисничке улоге [4]:

- **Анониман корисник** је сваки корисник који посети/уређује Википедију а нема отворен кориснички налог (или није пријављен). Неки осетљивији чланци (подложнији вандализмима) су закључани за анонимне кориснике и они не могу да их уређују. Када анонимни корисник направи измену, уместо корисничког имена се бележи његова *IP* адреса.

- **Улогован корисник** је сваки корисник који уређује коришћењем свог корисничког налога. Он може да има различите додатне улоге, као на пример „надзирач“ („патролер“) или „корисник коме није потребан надзор“ (енгл. *autopatrolled*). Измене нових корисника које заједница још увек није упознала су испрва обележене као измене којима је потребан надзор тако што у списку измена уз њих стоји знак упозорења да се можда ради о непримереној измени. Када искуснији корисник (надзирач) провери садржај измене и означи да је примерен, знак упозорења нестаје и остали надзирачи знају да не морају и они да прегледају измену.

- **Администратор** може да брише странице, закључава их, блокира кориснике и врши неколико других административних послова.

- **Истражитељ** („чекјузер“, енгл. *checkuser*) је корисник који има приступ техничким подацима других корисника (*IP* адресе, потпис Интернет прегледача, *XFF* атрибути[2]), које користи за проверу да ли сумњиви налози припадају истој особи.

[2] *XFF* (*X-Forwarded-For*) је поље у заглављу *HTTP* захтева које идентификује *IP* адресу клијента који серверу приступа преко неког посредничког сервера. Уколико је ово поље прослеђено, оно може бити од користи приликом истрага, али пошто није обавезно анонимизујући посреднички сервери га сакривају.

- **Бирократа** може да додељује и уклања улоге другим корисницима.

- **Стјуард** је бирократа који има привилегије на свим пројектима (на свим језичким издањима Википедије и других Викимедијиних пројеката).

- **Налози за аутоматске измене** (ботовски налози) су налози преко којих рачунарски програми врше аутоматизоване репетитивне измене на Википедији. Примери оваквих измена су исправљање типичних штампарских и правописних грешака или масовно додавање једнообразних података, као што су резултати пописа у све чланке о насељеним местима. Како ове измене нису контроверзне а има их много, измене корисника са овом улогом се не приказују у списку скорашњих измена.

Проширења (енгл. *extension*) представљају додатке за МедијаВики. Софтвер МедијаВики је дизајниран тако да програмери могу да релативно једноставно у складу са жељама и потребама развију „проширења“ софтвера, која му додају нове опције.

Постоји више десетина проширења за МедијаВики, међу којима су модули за истражитеље, за борбу против нежељених порука (спама), приказ сложених математичких формула, прављење специјализованих извештаја, и тако даље.

1.2 Разлози за уређивање са више налога

Постоји више разлога због којих се неки корисници одлучују да противно прописаним политикама користе више налога за уређивање садржаја:

- корисници који покушавају да произведу привид „већине“ како би остварили своје политичке циљеве;
- блокирани корисници који покушавају да заобиђу блокаду;
- кршење правила са додатног налога како би се избегло блокирање главног налога (провоцирање других корисника, ратови измена)
- класично „троловање“[3]

Један од мотива због кога се злонамерни корисници упуштају у овакво понашање, и главни мотив да се такво понашање онемогући је чињеница да Википедија спада у 10 најпосећенијих веб-локација на Интернету (у тренутку писања, 7. најпосећенији[5]), и за већину корисника прво одредиште у потрази за сваковрсним информацијама.

Када се узме у обзир да Википедију чита 365 милиона људи месечно, и да веб-локација има 12 милијарди посета (учитаних страница) месечно, јасно је да појединац који

[3] Остављање злонамерних коментара чија је једина сврха да изазову свађе, тензије, бес или беспотребно трошење времена других корисника.

успешно утиче на садржај чланака у складу са својом политичком идеологијом, верским опредељењима, историјским схватањима или слично, може осетно да утиче на пропагирање своје идеологије, опредељења или схватања.

1.3 Заштита од злоупотреба више налога

Традиционални приступ заштити од злоупотребе вишеструких налога на Википедији подразумева провере сумњивих корисничких налога од стране истражитеља („чекјузер“, енгл. *checkuser*).

Истражитељи су корисници Википедије са посебним привилегијама, које им омогућавају приступ дневницима активности у којима се чувају технички подаци о корисницима Википедије, као што су:

- *IP* адресе;
- Потписи коришћених Интернет прегледача (*UserAgent* ниске);
- подаци о посредничким (прокси) серверима.

Истражитељи анализирају ове податке и упоређују сумњиве налоге, покушавајући да уоче доказе да два или више налога припадају истој особи.

Како се ради о осетљивим подацима, који задиру у приватност корисника (обично је могуће утврдити место становања а понекад и ближи идентитет корисника), корисници са истражитељским привилегијама морају да буду особе од интегритета и поверења у својој локалној заједници, и да поштују строге политике приватности Задужбине Викимедија. Како би злоупотреба ових података могла да има чак и кривичне импликације, особе које имају истражитељска овлашћења морају да се идентификују Задужбини Викимедија како би се у случају злоупотребе знало ко је за злоупотребу одговоран.

Осим тога, неопходно је да истражитељи поседују довољно техничког знања да би могли да анализирају податке и да доносе исправне закључке.

Основни проблем са којим се истражитељи суочавају је што на основу података којима у својим истрагама располажу често не могу са потпуном сигурношћу да утврде да ли два корисничка налога припадају истој особи.

Злонамерни корисник који има довољно техничког знања може да зна у које податке истражитељи имају увид, па има могућности да прикрије своје трагове (коришћењем различитих прегледача, рачунара или употребом анонимизирајућих прокси сервера).

Поред тога, истражитељи морају да се старају да опасност да неког корисника погрешно оптуже сведу на најмању могућу меру, јер понекад се догоди да многи докази упућују да два корисничка налога припадају истој особи, а ипак се ради само о

случајним поклапањима (Интернет провајдер има посреднички сервер, два корисника су један за другим добили исту динамичку *IP* адресу, или слично).

1.4 Технике и методе

Рад је подељен у следеће целине:

- прављење проширења за МедијаВики
- прикупљање тест-података
- анализа података
- дефинисање индикатора
- приказ података
- дефинисање метрика
- анализа резултата

Прављење проширења подразумева писање кода који омогућава да корисник у формулару унесе параметре жељене анализе, да се анализе покрену и да му се прикажу резултати.

Тест-подаци су неопходни из више разлога: (а) приликом имплементације ради тестирања исправности кода; (б) за потребе анализе природе података и; (в) за анализу резултата, како би се направила што боља оцена квалитета примењених алата.

Циљеви **анализе података** су да се: (а) на основу природе података утврди који статистички методи су најприкладнији за упоређивање понашања корисника и; (б) да се одреди величина узорка који је потребно узети да би резултати били поуздани.

Ради упоређивања понашања два корисничка налога је потребно развити одговарајуће статистичке **индикаторе** како би се добила нумеричка оцена сличности два корисничка профила.

Приказ података подразумева исцртавање дијаграма који кориснику омогућавају да боље разуме резултате индикатора, као и да уочи евентуалне битне детаље које нумерички резултати индикатора не истичу.

Метрике сличности два корисничка налога служе да олакшају интерпретацију резултата то јест да се на једноставан, нумерички начин изрази да ли је сумња да два налога припадају истој особи основана или не.

Анализа резултата је финални део рада, у коме се тестирањем на реалним подацима проверава ефикасност развијених индикатора и метрика сличности.

2 Анализа проблема

Овај рад се заснива на претпоставци да су неке од карактеристика понашања корисника које проистичу из особина и навика особе релативно константне и да их није једноставно променити чак ни свесно, с намером. Стога је неопходно идентификовати што више карактеристика које је могуће анализирати увидом у јавно доступне податке, а које би могле да дају добре резултате приликом упоређивања сумњивих корисничких налога.

На први поглед се намећу карактеристике које се тичу дневног и недељног ритма корисника. Дневни ритам, то јест доба дана када особа обично леже да спава и када устаје, или када јој траје радно време, релативно је фиксан а може значајно да се разликује од особе до особе, тако да је то један од индикатора који би могао да се користи. Такође, висина активности по данима у недељи може да буде значајно условљена радним временом, обавезама и навикама појединца.

Друга група карактеристика које се могу уочити су оне које се тичу начелног понашања особе приликом рада на Википедији. Примери оваквих карактеристика су време које корисник дневно у просеку проводи на Википедији или колико измена врши дневно. Мање очигледна карактеристика би могла да буде у колико одвојених наврата корисник уређује Википедију. Да би се ове карактеристике могле анализирати, неопходно је да се дефинише појам „уређивачке сесије“. Уређивачка сесија је скуп измена које корисник начини у једном наврату, то јест у једној „посети“ Википедији. Сесија као концепт не постоји у МедијаВикију, и подаци о сесији се не бележе у бази, па је стога неопходно на неки начин је вештачки дефинисати на основу података који су доступни. Због овога је узето да нова сесија почиње изменом којој је претходио период од најмање 60 минута без измена, а завршава се изменом након које је наступила пауза од најмање 60 минута.

Трећа група су карактеристике које се тичу неких конкретних аспеката понашања корисника на Википедији, оних које се тичу конкретних измена које корисници врше. Примери оваквих карактеристика су просечна величина измене (број бајтова који је дописан или обрисан у оквиру измене), време између две измене у оквиру једне сесије, као и проценат „исправки“. Исправка је измена која је настала како би била исправљена нека грешка коју је исти корисник направио у претходној измени (на пример лош прелом или штампарска грешка). Као и уређивачка сесија, исправка је појам који није дефинисан у оквиру МедијаВикија и не постоји експлицитно забележен у бази. Стога је исправка дефинисана као узастопна измена на неком чланку у оквиру исте сесије. Узимају се у обзир измене у самим чланцима, а не и на странама за разговор, јер узастопне измене на странама за разговор могу да представљају и реплике другом кориснику у току неке расправе, а не нужно исправке претходно унетог садржаја.

3 Имплементација

3.1 Код проширења

Алат који је предмет овог рада је развијан као проширење („екстензија“) софтвера МедијаВики. Разлог за овакву одлуку је тај што се МедијаВики проширења врло једноставно додају у оригиналну инсталацију софтвера, и затим се још једноставније користе, јер се кориснику приказују као интегрални део апликације.

Системски администратор који управља викијем може врло једноставно да инсталише нова проширења. У општем случају, инсталација проширења се састоји из само два корака:

1. Потребно је да администратор отпакује проширење у директоријум `extensions` и
2. да у датотеци `LocalSettings.php` дода следећу линију:

```
include_once('extensions/NazivEkstenzije/NazivEkstenzije.php');
```

Најзначајнија веб-апликација која користи софтвер МедијаВики је Википедија. Српско језичко издање Википедије, као и издања на свим осталим језицима, се налази на серверима којима управља Задужбина Викимедија и стога су системски администратори Задужбине Викимедија надлежни за инсталацију нових проширења на Википедији.

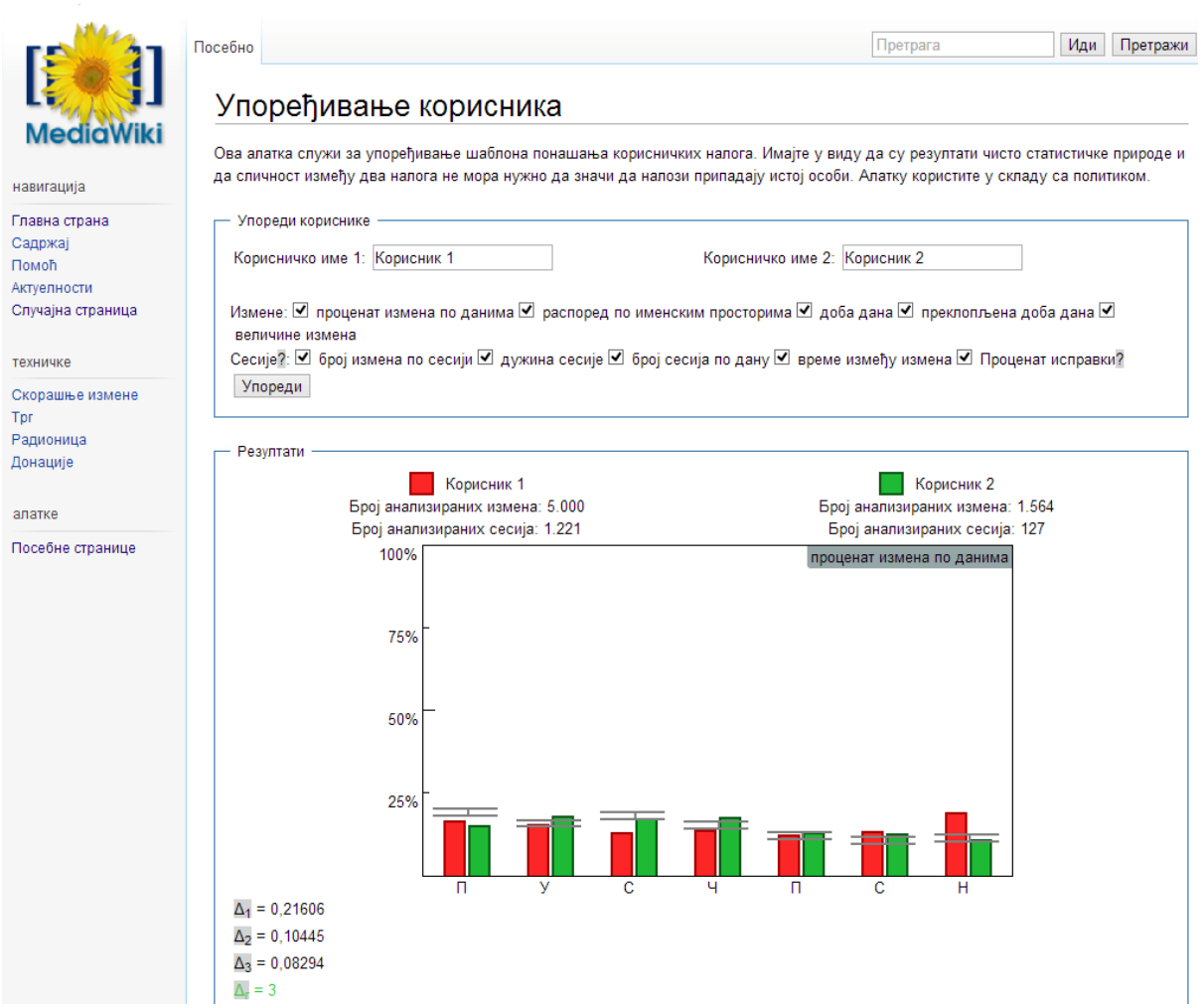
Чињеница да је одабран развој новог проширења за МедијаВики практично намеће избор програмског језика, платформе и технологија у којима се овај рад развија, као и лиценце под којом је доступан (*PHP, MySQL, GNU GPL*). Осим тога, то значи да је неопходно пратити све конвенције и смернице које су прихваћене као стандард приликом развоја проширења за МедијаВики.

Најзначајнији делови кода проширења за МедијаВики су распоређени у следеће три датотеке:

- `SpecialUserDiff.php` је датотека која садржи код задужен за прикупљање улазних параметара и за приказ излазних параметара.

  У овој датотеци/страници се налази формулар који служи да корисник унесе корисничка имена сумњивих налога, и да одабере које анализе је потребно извршити. Након извршених анализа, на истој страници се приказује извештај о сличности понашања два корисничка налога. Такође, ова страница је задужена за контролу приступа, односно за омогућавање приступа извештајима искључиво корисницима са прописаним привилегијама.

На слици 1 је приказан изглед главне странице проширења. У горњем оквиру се налази формулар у коме корисник уноси имена два налога која жели да упореди, и одабира које анализе жели да спроведе. У доњем оквиру су приказани подаци о броју анализираних измена и сесија за оба корисника, графици са резултатима тражених анализа и резултати метрика за те анализе.



Слика 1 — Изглед основне странице проширења

- `UserDiffStatistics.php` садржи *PHP* класе које су задужене за израчунавање захтеваних анализа и за графички приказ резултата, као и неколико помоћних класа.

- `UserDiff.i18n.php` садржи ниске за интернационализацију. Како се ради о вишејезичној апликацији, поруке које се приказују на страницама не смеју да буду тврдо кодиране, већ се смештају у посебне *PHP* низове у датотеци `UserDiff.i18n.php`. За сваки подржани језик постоји по један низ. Формат ових низова је:

```
$messages['sr-ec'] = array(
		'userdiff-user-nonexistent' => 'Корисник „$1“ не
постоји.', 'userdiff-results' => 'Резултати');

$messages['en'] = array(
		'userdiff-user-nonexistent' => 'User "$1" does not
exist.', 'userdiff-results' => 'Results');
```

Ове поруке се у апликацији приказују помоћу функције `wfMsgHtml` на следећи начин:

```
wfMsgHtml( 'userdiff-results' );
```

Проширење је имплементирано на енглеском и на српском (ћирилица и латиница), а заједници је остављено да, као и код других проширења, организује превођење на друге језике.

3.2 Класе

Написане су помоћне класе које представљају појмове измене и сесије, то јест њихове атрибуте и методе који су значајни за анализу која се врши.

Својства измене која су релевантна за анализу и која се прикупљају и обрађују су:

- **време измене** (*timestamp*), које представља време настанка измене (време бележења измене у базу)
- **именски простор**, који означава тип странице која је измењена. Постоји око 20 именских простора, а неки од њих су: главни именски простор, страна за разговор, шаблон, помоћ, и тако даље.[6]
- **наслов странице**
- **величина измене**, која се дефинише као дужина странице уколико се ради о првој измени (прављењу странице), а у супротном као разлика између нове дужине странице и дужине странице у претходној верзији. Величина измене се мери у бајтовима, и може бити и негативна. Ваља имати у виду да како се ради о бајтовима а не о карактерима, карактери који не припадају скупу *ASCII* карактера[7] имају дужину већу од 1 (слова српске ћирилице имају дужину 2). На свим језичким издањима је подразумевана употреба *Unicode* кодирања.

Класа која представља сесију складишти низ измена које чине сесију, и има помоћне методе који враћају број измена у сесији, прву и последњу измену, као и дужину сесије у минутима.

3.2.1 Време измене

Рад са временима у програмирању представља комплексан изазов. Детаљи о којима је неопходно водити рачуна приликом утврђивања временске удаљености два догађаја су:

- 24 стандардне часовне зоне
- 2 посебне часовне зоне (*GMT* и *UTC*)
- преступне године
- преступне секунде
- ере (хришћанска, јуникс, ...)
- летње рачунање времена
- чињеница да су неке државе у неком периоду користиле летње рачунање времена, а потом су га укинуле

Као високо локализован софтвер, МедијаВики омогућава сваком кориснику да у својим подешавањима одреди којој часовној зони припада, како би му времена била приказивана у његовој локалној часовној зони.

У контексту упоредне анализе два корисничка налога фигурише пет потенцијално различитих часовних зона:

- часовна зона у којој се подаци о времену чувају у бази (*UTC*)
- подразумевана часовна зона на серверу (*CET/CEST* у случају `sr.wikipedia.org`)
- часовна зона једног корисничког налога
- часовна зона другог корисничког налога
- часовна зона корисника који упоређује сумњиве корисничке налоге

Као први одговор на питање коју часовну зону користити приликом анализирања података се намеће *UTC*, јер се ради о подразумеваној „основној“ часовној зони.

Међутим, како *UTC* не подлеже променама услед летњег рачунања времена, уколико се корисник чије се понашање анализира налази у држави која упражњава летње рачунање времена, доћи ће до искривљења података који се тичу времена вршења измена (током месеци летњег рачунања времена, корисникове измене ће бити рачунате као да су се догодиле сат времена раније него што заиста јесу).

Часовне зоне корисничких налога који се испитују се не могу узимати у обзир јер се овде ради о корисничким подешавањима, и злонамерни корисник би могао намерно да постави различите часовне зоне код различитих налога како би анализа показала значајна одступања у периодима активности између тих налога.

Стога се као меродавна узима подразумевана часовна зона сервера. У случају Википедије на српском језику, то је централноевропско време, и централноевропско летње време током летњих месеци. Разлог за ово је што се за подразумевану часовну зону сервера поставља управо часовна зона којој припада већина уређивача неке Википедије. Овакав избор осигурава да часовна зона по којој се тестови рачунају одговара часовној зони којој припада већина уређивача Википедије, као и да се

приликом анализе измене оба корисника рачунају по истој часовној зони. Наравно, уколико се један или оба сумњива налога јављају из неке друге часовне зоне, периоди њихових активности ће бити транслирани, али на конзистентан начин (уколико се јављају из исте часовне зоне, периоди њихових активности ће бити транслирани за исти број сати).

4 Подаци

Пожељно је да се приликом израде рада користе реални тест-подаци, јер то омогућава да се током рада на имплементацији стекне увид у исправност планираних решења, да се препознају нови могући правци за развој и да се лакше уоче евентуалне грубе грешке приликом имплементације. Аутор је дугогодишњи корисник и истражитељ на Википедији на српском језику, и стога је најбоље упознат са примерима корисничких налога који су погодни за тестирање на Википедији на српском језику.

База података Википедије на српском језику је доступна за преузимање са веба[8]. Међутим, услед величине базе (извезена база у `xml` формату заузима 70 гигабајта) њен увоз на развојни рачунар представља изазов.

Како би било могуће једноставно тестирање, а и услед чињенице да приликом почетка овог рада није био доступан сервер на који би било могуће извршити увоз ове базе, било је неопходно да се омогући рад са мањим скупом података припремљених у текстуалним датотекама.

Проширење је програмирано тако да у току развоја омогућава преузимање података из два извора: из локалне базе МедијаВикија, или из текстуалних датотека.

4.1 Подаци из текстуалних датотека

Текстуалне датотеке са мањим скуповима тест-података су преузете са Тулсервера (енгл. *Toolserver*)[9]. Тулсервер представља скуп сервера којима управља Викимедија Немачке уз подршку Задужбине Викимедија и неколико локалних Викимедијиних огранака.

Ови сервери омогућавају програмерима приступ живој копији база Викимедијиних пројеката. Корисници Тулсервера могу да извршавају *SQL* упите над овим базама.

Узорци података неопходни за тестирање су преузети са Тулсервера *MySQL* упитима извршеним помоћу *python* скрипта, и сачувани у текстуалним датотекама.

4.2 Увоз базе

Иако је рад са текстуалним датотекама једноставнији и бржи, у неком тренутку је било неопходно да се проширење тестира и над реалним подацима у МедијаВики бази, како би била потврђена исправност свих упита, и како би биле испитане брзинске перформансе кода.

Обезбеђен је тест-сервер са процесором *Intel Core i5 3570, 8GB DDR3* меморије и диском *SATA3 Western Digital Caviar Blue* од 1TB.

За увоз података је коришћен програм *MWDumper*[10], Јава апликација која је направљена за увоз МедијаВики базе из *XML* датотеке са извезеним подацима. *MWDumper* увози податке само у три „главне“ табеле: `revision`, `text` и `page`. Коришћена датотека са извозом базе Википедије на српском језику је имала око 6.300.000 измена и око 580.000 страница у свим именским просторима.

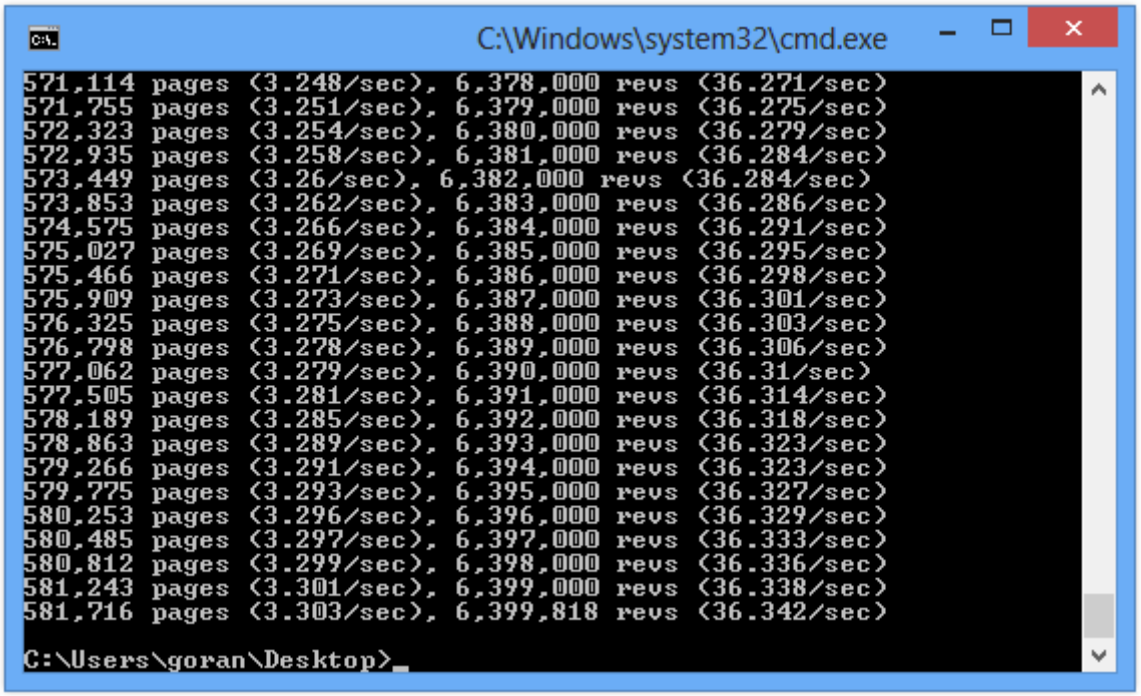
Како увоз података може да траје веома дуго, препоручено[11] је да се спроведу прилагођавања у структури базе података и у систему за управљање базама података да би се поступак убрзао. Спроведене су следеће измене, које су по завршетку увоза поништене:

- уклоњени су сви индекси и ауто-инкремент директиве над три табеле у које се подаци увозе

```
ALTER TABLE revision
       DROP INDEX rev_timestamp;
ALTER TABLE revision
       DROP INDEX page_timestamp;
ALTER TABLE revision
       DROP INDEX user_timestamp;
ALTER TABLE revision
       DROP INDEX usertext_timestamp;
ALTER TABLE revision
       DROP INDEX page_user_timestamp;

ALTER TABLE page
       DROP INDEX page_random;
ALTER TABLE page
       DROP INDEX page_len;
ALTER TABLE page
       DROP INDEX page_redirect_namespace_len;

ALTER TABLE `wiki`.`revision` CHANGE COLUMN `rev_id` `rev_id` INT(10)
UNSIGNED NOT NULL;

ALTER TABLE `wiki`.`page` CHANGE COLUMN `page_id` `page_id` INT(10) UNSIGNED
NOT NULL;

ALTER TABLE `wiki`.`text` CHANGE COLUMN `old_id` `old_id` INT(10) UNSIGNED
NOT NULL;
```

- параметар `innodb_log_file_size` је повећан на 100M
- искључен је бинарни лог.

Упркос свему, увоз података је текао брзином од око 40 измена и око 4 странице по секунди (као што се може видети на слици 2), тако да је трајао око 3 дана. Разлог за оволико трајање увоза је спорост хард диска, која је представљала уско грло.



Слика 2 — Ток увоза базе Википедије на српском језику

Како су осим података из ове три „главне“ табеле потребни још неки подаци, било је неопходно спровести додатни увоз података у још неколико табела. Овај „ручни“ увоз је рађен тако што су подаци преузимани у текстуалну датотеку са Тулсервера помоћу *python* скрипта, а затим помоћу *PHP* скрипта увожени у локалну базу.

Табеле за које је вршен овај додатни увоз су табела `user`, која је неопходна због корисничких имена и табела `user_groups`, која је неопходна због корисничких привилегија (да би се издвојили корисници који нису ботови). Такође, пошто приликом увоза није попуњено поље `rev_len` у табели `revision`, које је неопходно због израчунавања величине измена, и ови подаци су морали да буду „ручно“ попуњени.

Поље `rev_len` је представљало изазов зато што се ради о великој количини података (па и увоз и извоз дуго трају), и зато што су се испољиле извесне неконзистентности између базе извезене у датотеку, и репликоване базе на Тулсерверу.

Због количине података, увоз није рађен одједном, већ су подаци угрубо подељени у десет делова (критеријум поделе је био идентификатор корисника по модулу десет). Увоз свих десет делова је трајао око 5 сати. Након што је увоз завршен, уочено је да од

укупно 6.399.818 редова у табели `revision`, за 1.762 недостаје податак `rev_len` (величина измене у бајтовима), што износи 0,028%. Проверама је утврђено да недостајући подаци нису последица грешака приликом извоза и увоза, већ да подаци нису присутни у реплици базе на Тулсерверу. Неконзистентна репликација базе је пријављена администраторима Тулсервера[12]. Како је удео недостајућих података врло мали (испод једног промила), они не могу значајније да утичу на анализу података, и стога је поступак анализе настављен.

5 Анализа података

Избор најпогоднијих статистичких тестова зависи од природе података, па је неопходно да се утврди расподела вероватноће анализираних параметара. Конкретно, од интереса је да ли подаци одговарају нормалној расподели.

Други циљ анализе података је да се одреде реалне величине узорака за просечне вредности популације и процене потребне величине узорака за два корисничка налога која се упоређују.

Када су у питању просечне вредности популације, циљ је одредити што мањи узорак који је стабилан, то јест у коме параметри имају сличне вредности при сваком израчунавању, како би њихово проналажење било што јевтиније, а опет поуздано, јер се подаци о узорку не складиште већ се рачунају изнова при сваком позивању алата за поређење.[4]

Када су у питању узорци за корисничке налоге који се упоређују, мотив је одредити доњу границу броја измена која даје смислене резултате (налоге са премало измена нема смисла упоређивати). Код корисничких налога је потребно одредити и горњу границу броја измена, јер неки налози имају и по више десетина хиљада измена, и анализа свих измена би била прескупа.

5.1 QQ дијаграми

За испитивање нормалности расподела су коришћени QQ (квантил-квантил) дијаграми[13]. QQ дијаграми представљају једноставну графичку методу за упоређивање две расподеле вероватноће.

За исцртавање QQ дијаграма који служи за упоређивање неке теоријске расподеле са подацима из узорка, неопходно је прво израчунати параметре расподеле који одговарају том узорку. У случају нормалне расподеле, то су аритметичка средина и стандардна девијација. Затим је неопходно одредити скуп квантила који ће фигурисати у дијаграму (број квантила одговара величини узорка). Након тога се израчунавају

[4] Разлог за ово је компликованија имплементација, и мања вероватноћа да ће проширење бити прихваћено, јер изискује промене у структури базе података.

вредности квантила за кумулативну расподелу вероватноће, и ове вредности се упарују са одговарајућим (уређеним по величини) вредностима из узорка. Овако добијени парови су координате тачака на дијаграму. Вредност квантила представља *x* координату тачке док вредност из узорка представља *y* координату тачке на дијаграму. Уколико популација из које је узорак узет одговара теоријској расподели, вредности из узорка ће приближно одговарати вредностима квантила кумулативне расподеле вероватноће, и стога ће тачке бити распоређене приближно дуж праве f(x) = x.

У конкретном случају, QQ дијаграми су направљени на следећи начин:

- За сваки скуп података узети су подаци за неколико стотина корисника.

- У обзир су узети само корисници који имају одговарајући број измена и сесија. За сваки скуп података су направљена по три дијаграма, један где су фигурисали корисници са најмање 50 измена, један где су фигурисали корисници са најмање 100 измена, и један где су фигурисали корисници са најмање 500 измена.

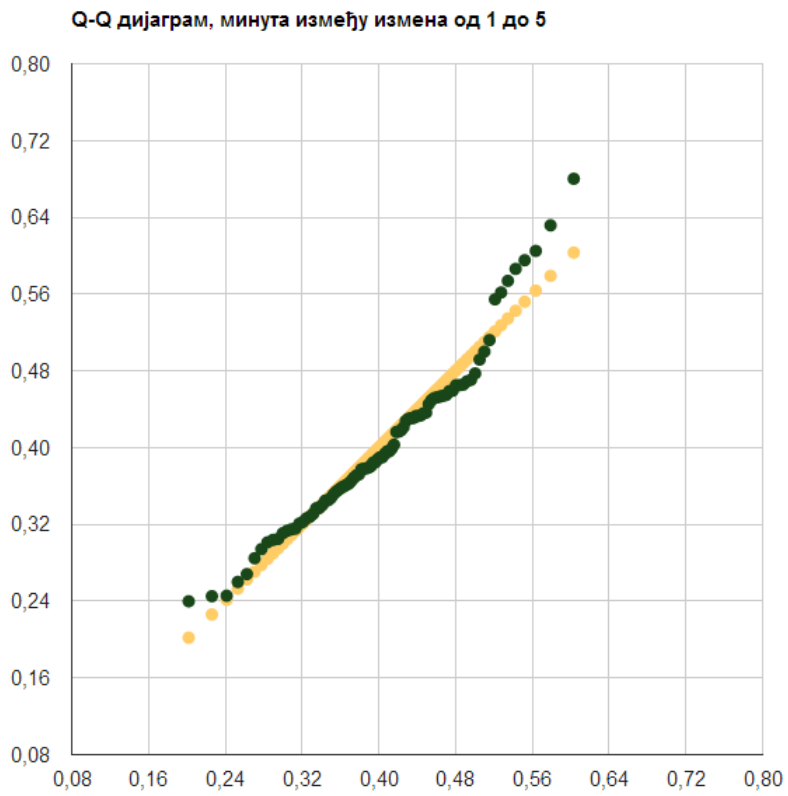
- Израчунате су аритметичке средине за сваког корисника, и вредности су поређане по величини. Тако уређени подаци надаље представљају узорак.

- Израчуната је аритметичка средина и стандардна девијација узорка. Уколико подаци одговарају нормалној расподели, што је полазна претпоставка, ова аритметичка средина и стандардна девијација су параметри који одређују ту нормалну расподелу.

- Величина узорка одређује број квантила. Нађени су одговарајући квантили кумулативне расподеле вероватноће за нормалну расподелу са датим параметрима.

- *i*-ти квантил је упарен са *i*-тим елементом из сортираног низа узорака, и сваки елемент узорка је исцртан као тачка на дијаграму која има за *x*-координату вредност узорка а за *y*-координату одговарајући квантил нормалне расподеле.

На слици 3 је дат пример дијаграма са упоредним приказом расподеле процента измена које су извршене 1 до 5 минута након претходне измене и одговарајуће нормалне расподеле. Из узорка величине 2000 корисника су одабрани корисници који имају најмање 50 измена. Зелене тачке представљају вредности узорка. Свака тачка представља проценат измена које су извршене 1 до 5 минута након претходне измене за по једног од корисника из узорка. Жуте тачке представљају вредности нормалне расподеле са параметрима израчунатим из узорка. То су вредности које би узорак имао кад би саваршено одговарао нормалној расподели.



Слика 3 — QQ дијаграм процента измена које су настале између 1 и 5 минута након претходне корисникове измене код корисника са најмање 50 измена, из узорка од 2000 корисника (укупно 90 корисника)

Мана овог приступа је што се не ради о нумеричком већ о визуелном тесту. Међутим, он може сасвим добро да послужи да се утврди да ли је начелни приступ смислен или не. Са друге стране, предност овог приступа је што поред провере да ли је расподела нормална омогућава да се уоче и евентуалне друге значајне карактеристике анализираних података.

5.1.1 Имплементација

За издвајање података је коришћен код МедијаВики проширења (написана је нова функција `extractMeans` за ту сврху).

Упит је написан коришћењем МедијаВики АПИја за приступ бази:

```
$tableNames = array( 'user', 'r' => 'revision', 'page', 'r1' => 'revision',
'user_groups' );
// Alias in 'r1.rev_len parent_len' format for compatibility with 1.18.0
$fields = array( 'r.rev_timestamp', 'page_namespace', 'page_title',
'r.rev_len', 'r1.rev_len parent_len', 'r.rev_user' );
$conditions = array( 'user_id = r.rev_user', 'ug_group IS NULL', 'r.rev_len
IS NOT NULL', 'r.rev_user IN (SELECT user_id FROM user WHERE user_id > 2000
AND user_id < 4000)' );
// "rev_len is not null" condition is present because in test database we
lack some data in this column

$fname = 'Database::select';
$options = array( 'ORDER BY' => 'rev_timestamp DESC' );
$join = array( 'r' => array( "JOIN", 'r.rev_user = user.user_id' ), 'page'
=> array( "JOIN", 'r.rev_page = page.page_id' ), 'r1' => array( "LEFT JOIN",
"r1.rev_id = r.rev_parent_id" ),
             'user_groups' => array( "LEFT JOIN", "ug_user = user_id AND
ug_group = 'bot'" ) );
```

Преведен у *MySQL*, упит изгледа овако:

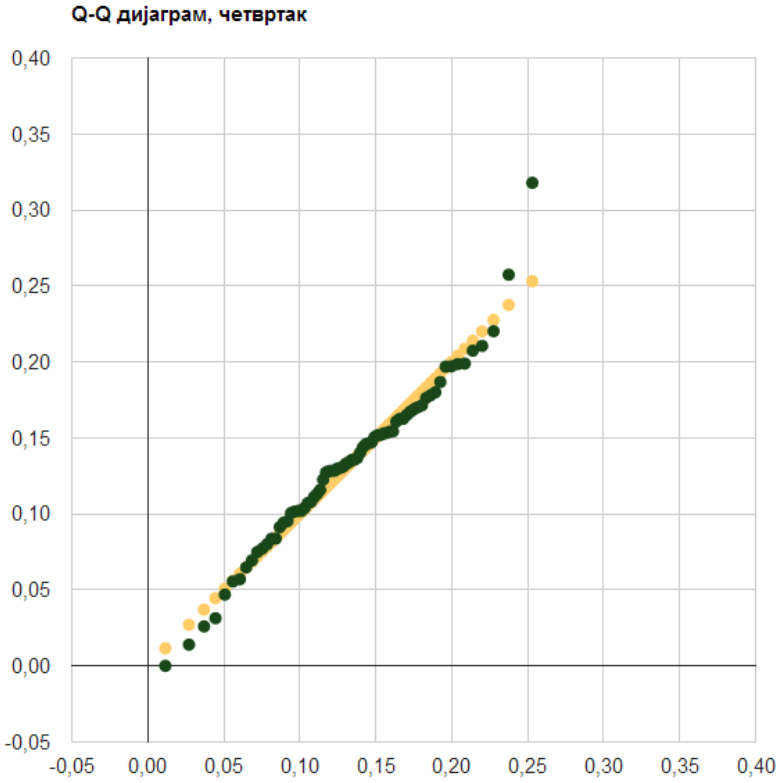
```
SELECT r.rev_timestamp, page_namespace, page_title, r.rev_len, r1.rev_len
parent_len, r.rev_user
FROM `user`
JOIN `revision` `r`
       ON ((r.rev_user = user.user_id))
JOIN `page`
       ON ((r.rev_page = page.page_id))
LEFT JOIN `revision` `r1`
       ON ((r1.rev_id = r.rev_parent_id))
LEFT JOIN `user_groups`
       ON ((ug_user = user_id AND ug_group = 'bot'))
WHERE (user_id = r.rev_user)
       AND (ug_group IS NULL)
       AND (r.rev_len IS NOT NULL)
       AND (r.rev_user IN (SELECT user_id
                                       FROM user
                                       WHERE user_id > 2000
                                              AND user_id < 4000))
ORDER BY rev_timestamp DESC
```

Подршка за статистичке функције у програмском језику *PHP* не долази са инсталацијом већ је неопходно инсталисати библиотеку *PECL*.[14] Функција за рачунање инверзне кумулативне вероватноће је доступна у овој библиотеци, али није документована[15]. Због овога је *PHP* коришћен само за прикупљање података, а њихова анализа и генерисање дијаграма је спроведено у програмском језику Јава. QQ дијаграми нису део проширења за МедијаВики већ су служили само за анализу приликом израде овог рада, тако да код за њихово генерисање не улази у код проширења, и стога постоји слобода у избору технологије. Јава је одабрана зато што је за њу доступна широко коришћена и добро документована класа за рад са нормалном расподелом, и стога је у овом програмском језику било могуће најбрже написати поуздан програм за исцртавање дијаграма.

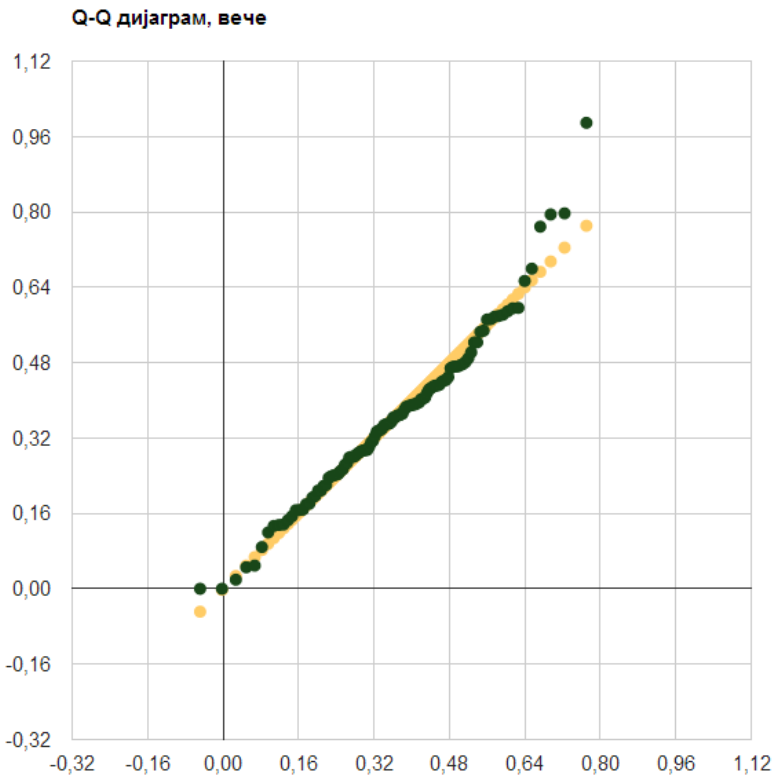
Написани Јава програм учитава податке које је у датотеци записао *PHP* скрипт, и на основу њих генерише *HTML* странице са QQ дијаграмима. За исцртавање дијаграма је коришћен *Scatter Chart* из Гугловог *JavaScript* АПИја за визуализацију.[16]

5.1.2 Резултати

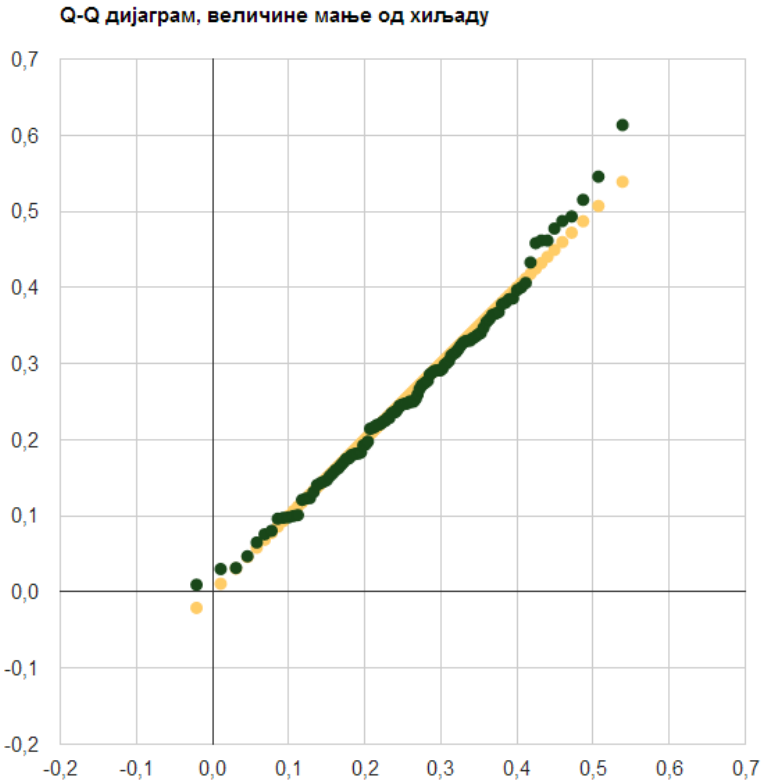
Добијени дијаграми показују да подаци начелно одговарају нормалној расподели. На сликама 4 до 7 су приказани примери таквих дијаграма: проценат измена четвртком, проценат измена у вечерњим сатима и величине измена од 100 до 1000 бајтова.



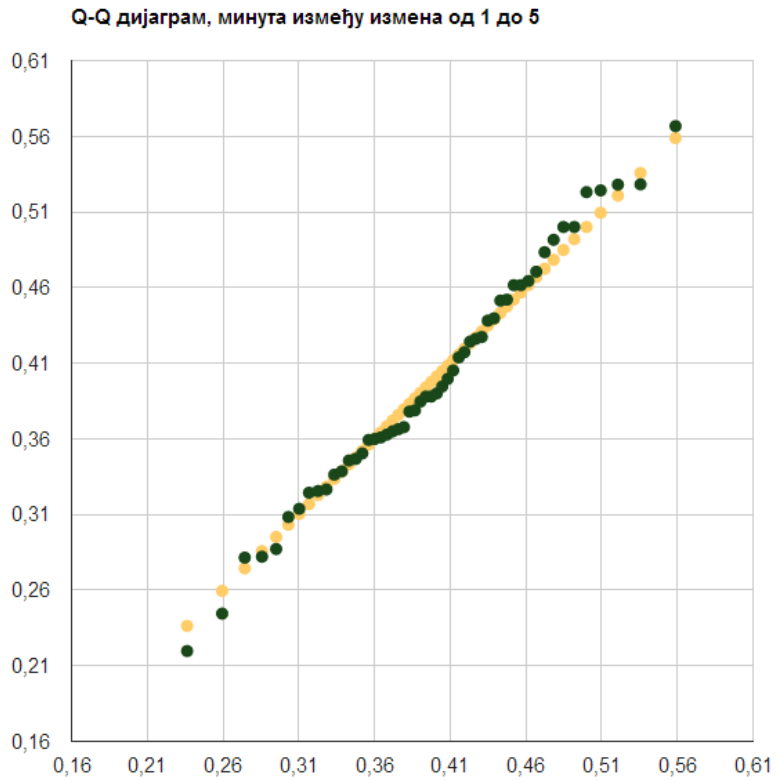
Слика 4 — QQ дијаграм процента измена четвртком, најмање 100 измена, из узорка од 2000 корисника (укупно 73 корисника)



Слика 5 — QQ дијаграм процента измена у вечерњим сатима, најмање 50 измена, из узорка од 2000 корисника (укупно 111 корисника)



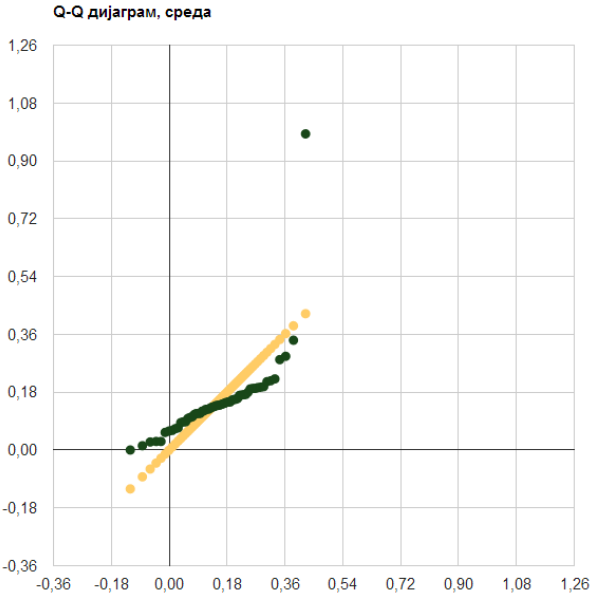
Слика 6 — QQ дијаграм процента измена величине од 100 до 1000 бајтова, најмање 50 измена, из узорка од 2000 корисника (укупно 111 корисника)



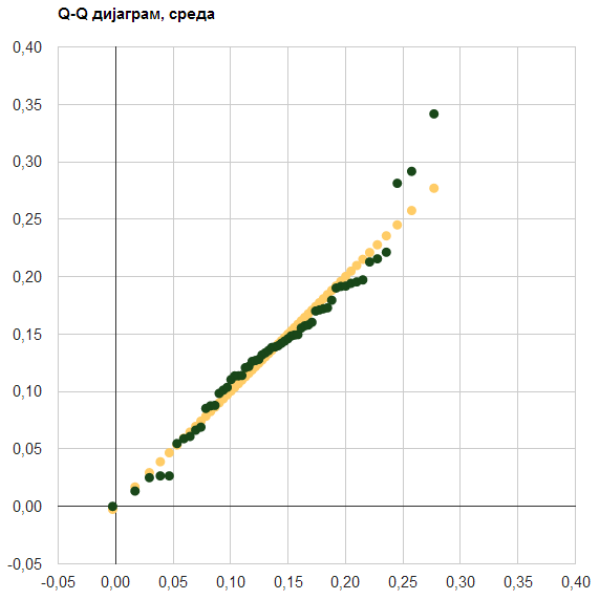
Слика 7 — QQ дијаграм удела измена начињених 1 до 5 минута након претходне, најмање 50 измена, из узорка од 1000 корисника (укупно 53 корисника)

Међутим, на неким дијаграмима су уочени различити типови „аномалија“ (одступања од нормалне расподеле).

Први тип аномалије се односи на одступање нагиба линије од очекиваног за податке који одговарају нормалној расподели. На пример, дијаграм процента измена које су извршене средом за кориснике који имају преко 50 измена (из узорка од 1000 корисника) показује линију која има блажи нагиб од нормалне расподеле (слика 8). На дијаграму се може уочити један „уљез“ (енгл. *outlier*). Управо је овај податак довео до искривљења линије. Када се уљез уклони и график исцрта са новим подацима (слика 9), види се да они врло добро (у односу на величину узорка) одговарају нормалној расподели.

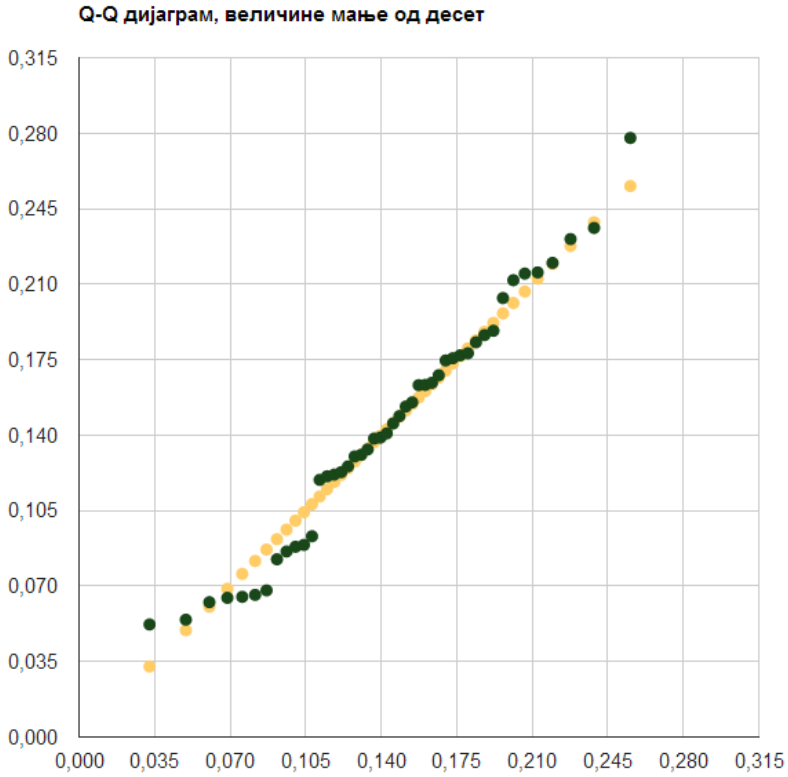


Слика 8 — QQ дијаграм процента измена начињених средом, са уљезом



Слика 9 — QQ дијаграм процента измена начињених средом, без уљеза

Други уочени тип аномалије је неочекивано степенасто груписање података које се јавља у неким графицима. На слици 10 се види дијаграм процента измена које имају од 0 до 10 бајтова, за кориснике који имају најмање 100 измена, извучене из узорка од 1000 корисника (укупно 46 корисника). Општи правац линије одговара нормалној расподели, али се може видети више мањих група које стварају хоризонталније линије, између којих се налазе прореди.



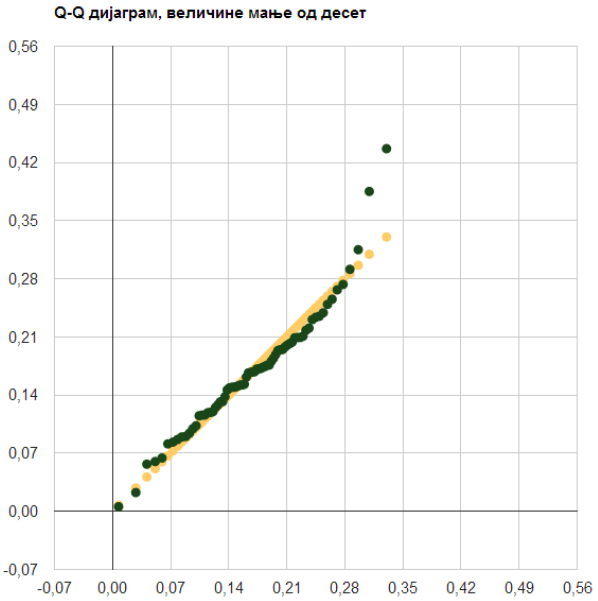
Слика 10 — Пример „хоризонталног“ груписања

Прво је проверено да ли постоји грешка у извлачењу или израчунавању података. Како је за извлачење података коришћен *PHP* (упит је писан преко МедијаВики АПИја), а за обраду података је коришћена Јава, поступак је поновљен тако што су подаци извучени помоћу *MySQL* упита, а за обраду података је искоришћен *Excel*.

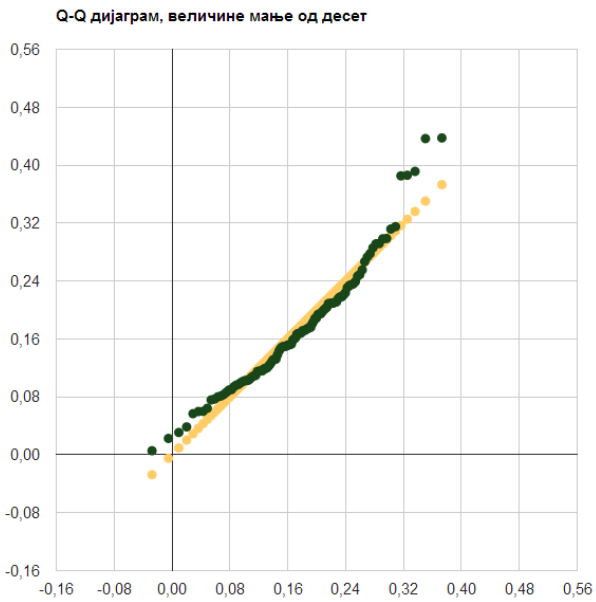
Приликом поновног израчунавања су добијени исти резултати и тиме је установљено да не постоји грешка у подацима, њиховој обради или приказу. Након тога је утврђено да се ова појава јавља услед природе QQ дијаграма, и мале величине скупа.

Узорак је сортиран по величини и извучени су одговарајући квантили, а како је узорак релативно мали, очекивано је да на неким сегментима буде гушћи него на другим. Због начина исцртавања, та повећана густина и околне „празнине“ доводе до формирања „хоризонталнијих“ линија са размацима између (уколико су подаци згуснути више него квантили, нагиб линије је испод 45 степени, а уколико су разређенији од квантила, нагиб линије је већи од 45 степени).

Међутим, нормална расподела не би требало да има сегменте који су гушћи и сегменте који су разређенији уколико је узорак довољно велик. Заиста, на већим узорцима (слике 11 и 12) се може уочити да су ове групе мање истакнуте.

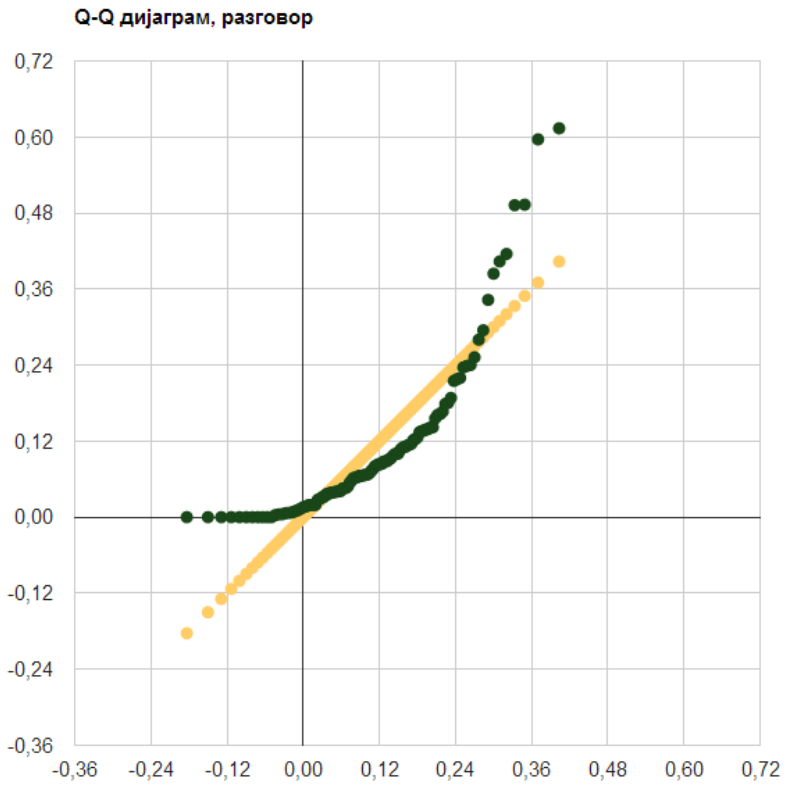


Слика 11 — 73 корисника, минимум 100 измена

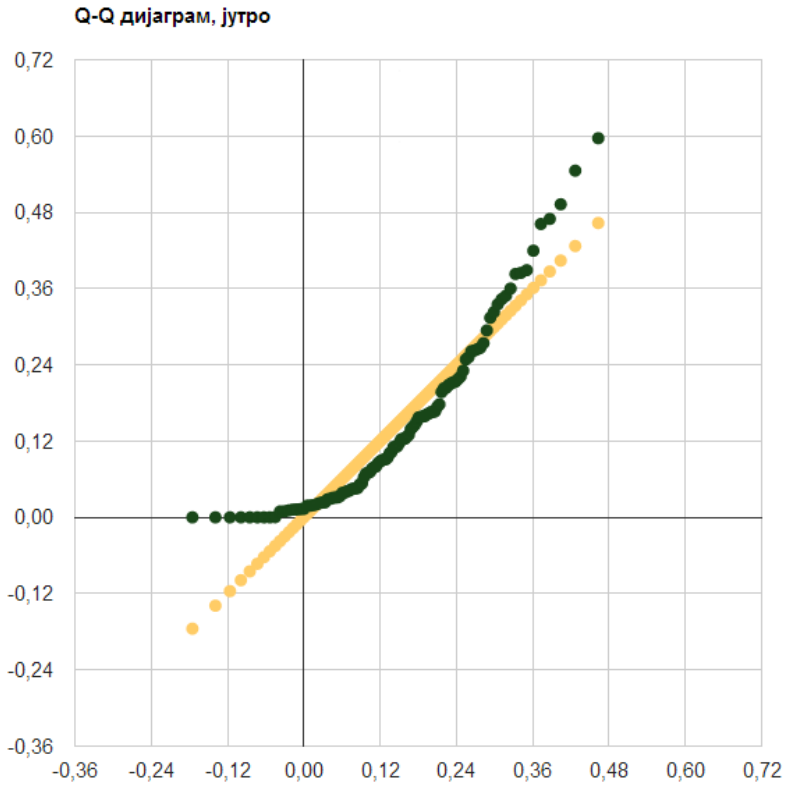


Слика 12 — 111 корисника, минимум 50 измена

Објашњења за ове две аномалије показују да се не ради о одступањима од нормалне расподеле. Ипак, на неким дијаграмима (најчешће онима који се тичу доба дана, именских простора и процента исправки) су уочена значајна закривљења у линијама које представљају податке (слике 13 и 14).



Слика 13 — QQ дијаграм процента измена на странама за разговор, корисници са најмање 50 измена из скупа величине 2000 (укупно 111 корисника).



Слика 14 — QQ дијаграм процента измена у јутарњим часовима, корисници са најмање 50 измена из скупа величине 2000 (укупно 111 корисника).

Овакво закривљење је присутно код података чија су аритметичка средина, стандардна девијација и величина узорка (број квантила) такве да би у случају да се ради о нормалној расподели неке од вредности биле негативне, што по природи података није могуће, јер се ради о процентима.

За разлику од претходне две, ова аномалија очигледно представља значајно одступање од нормалне расподеле, што значи да треба бити обазрив приликом избора расподеле која описује понашање посматране случајне променљиве.

6 Просеци

Основна идеја новог приступа упоређивању корисника је да сличност профила измена два корисничка налога представља индикацију да иза оба налога стоји иста особа. Међутим, уколико ти профили нису особени већ одговарају „просеку“, онда уочена сличност има мању тежину него уколико су профили специфични а међусобно слични. Како би се утврдила међусобна блискост профила измена у односу на просек неопходно је израчунати просечан профил измена.

Просечан профил измена се израчунава тако што се издвоји одређени број најскоријих измена за одређени број случајно одабраних корисника, и израчунају се вредности посматраних индикатора (на пример проценат измена по данима).

Када су у питању индикатори који се рачунају над појединачним изменама, све измене се могу посматрати као један скуп. Међутим, када су у питању индикатори који се рачунају над сесијама, неопходно је издвојити сесије за сваког корисника засебно а онда ове индикаторе рачунати над скупом свих сесија. Изузетак је број сесија по дану, где се не може посматрати скуп свих сесија, већ је неопходно рачунати вредности за сваког корисника засебно.

На први поглед има смисла ове просечне вредности израчунати једном и сачувати вредности. Међутим, овај приступ има одређене недостатке. Просечне вредности се вероватно разликују по језичким пројектима (на пример због тога што је неки језик распрострањен у више а неки у мање часовних зона). Затим, поставља се питање да ли податке чувати у бази или у датотеци. Чување у бази доноси компликованију инсталацију проширења, и отежава њено прихватање од стране програмерске заједнице а чување у датотеци отежава одржавање. У оба случаја, неопходно је периодично израчунавати просеке изнова јер је могуће да се временом мењају, а периодична извршавања није једноставно имплементирати у веб-апликацијама.

Из ових разлога, одлучено је да се просечне измене не чувају већ да се израчунавају изнова приликом сваког упоређивања два корисника. Наравно, и овај приступ носи са собом разне недостатке. Могуће је да просеци варирају у некој мери између два израчунавања. Осим тога, извршавање траје знатно дуже јер израчунавање просека

није јевтино. Због тога је потребно одредити оптималну величину узорка тако да израчунавање буде што јевтиније а да израчунати просеци буду што стабилнији.

Бирање случајног узорка се састоји из три корака. Први корак је проналажење скупа корисника који имају довољан број измена да би могли да фигуришу у узорку. Други корак је налажење случајног узорка одговарајуће величине из овог скупа. Трећи корак је узимање потребног броја најскоријих измена за сваког од случајно одабраних корисника.

Прва два корака су једноставна, али трећи представља изазов. Упит који за сваког корисника из задатог скупа враћа одређени број најскоријих измена се може написати на више начина:

- коришћењем клаузе `LIMIT`, на пример:

```
SELECT …
FROM revision r1
...
WHERE rev_id IN (SELECT rev_id
                     FROM revision r2
                     WHERE r2.rev_user = r1.rev_user
                     LIMIT n)
```

- коришћењем аналитичке функције `OVER()`, на пример:

```
SELECT …, ROW NUMBER OVER(PARTITION BY rev_user) rbr
FROM revision r
…
WHERE rev_user IN (...)
       AND rbr < n
```

На жалост, *MySQL* не подржава ни функцију `OVER()`, нити коришћење клаузе `LIMIT` унутар подупита (који нису у `JOIN` клаузи), тако да ова два приступа нису могућа.

Још један приступ је[17]:

```
select
    (count(*) from revision r1 where r2.rev_timestamp > r1.rev_timestamp and
r1.rev_id <> r2.rev_id) as rbr,
    ...
from revision p2
...
```

али није прихватљив јер је исувише скуп (налажење само две најскорије измене за два корисника може да траје 2 минута уколико корисници имају велики број измена).

Једина два решења која преостају су:

- налажење измена сваког корисника у засебном упиту.
- коришћење променљивих сесије:

```
select rev_user, rev_id, convert(rev_timestamp using utf8)
from (select rev_user, rev_id, rev_timestamp, case when @prevuser !=
rev_user then @rownum:=0 else @rownum:=@rownum+1 end as rownum,
@prevuser:=rev_user as prevuser
       from (select *
              from revision
              where rev_user in (...)
              order by rev_user, rev_timestamp desc) u,
(select @rownum:=0) r, (select @prevuser:=-1) x) y
where rownum <= n
```

Прво решење је очигледно неефикасно а друго је неелегантно (коришћење променљивих у функционалном програмском језику), а такође се приликом тестирања показало као скупо - налажење по 1.000 измена за 7 корисника који имају пуно измена (укупно ~140.000) је на тест-серверу трајало око 40 секунди.

Преостало је да се тестирањем утврди које је од ових решења временски ефикасније и да се изабере оно које је мање лоше. Како би ова два приступа била упоређена, издвојено је по 1.000 измена за случајно одабраних истих 50 корисника на оба начина и измерено је време извршавања. Поступак је поновљен још једном са других 50 случајно одабраних корисника.

- први приступ - први узорак 75 секунди, други узорак 68 секунди;
- други приступ - први узорак 155 секунди, други узорак 225 секунди.

Очигледно је да је први приступ (налажење података за сваког корисника у засебном упиту) ефикаснији и да мање зависи од узорка (од броја измена корисника у узорку). Такође, овај приступ омогућава и потенцијалну хеуристику у виду израчунавања међупросека и прекидања израчунавања у тренутку кад се међупросеци стабилизују.

6.1 Величина узорка

Величина узорка за израчунавање просека има три димензије: број корисника, минимални број измена за корисника (кориснике са малим бројем измена не узимамо у обзир), и максималан број измена по кориснику (неки корисници имају и преко 100.000 измена, и ако би се све оне узеле у обзир израчунавање би било скупо, а такви корисници би због великог броја измена у искривљивали просечне вредности).

Потребно је да вредности ових димензија буду такве да узорак буде стабилан, то јест што сличинији у узастопним израчунавањима, а због брзине израчунавања је битно да број корисника имају што мању вредност. Број измена по кориснику и минималан број измена за корисника не утичу значајно на време израчунавања, тако да је ове параметре могуће подешавати тако да се нађе што већа стабилност за што мањи број корисника.

Стабилност узорка се мери величином промене просека када се у њега дода још један корисник. Што више корисника се налази у узорку, новододати корисник има мањи утицај на свеукупну вредност просека. Метрика која се користи за оцену стабилности је:

$$d_i = \frac{\max_{1 \le j \le k} |v_{ij} - v_{i-1j}|}{\sum_{j=1}^{k} v_{ij}} * 100$$

где је:

- d_i стабилност просека када се дода i-ти корисник
- k број категорија
- v_{ij} вредност за j-ту категорију

Вредности се независно израчунавају за сваки од индикатора.

Ова метрика одговара највећој апсолутној разлици вредности индикатора за неку категорију након додавања новог корисника у узорак у односу на збир вредности свих категорија.

Како би се испитала стабилност узорака различитих структура и величина, упоређивани су узорци са различитим вредностима за минималан број измена које корисник мора да има и максималан број измена које се узимају у обзир за сваког корисника. Укупно су испитивана четири типа узорака:

- мали корисници, мали узорак
- мали корисници, велики узорак
- велики корисници, мали узорак
- велики корисници, велики узорак

„Мали корисници“ представљају кориснике који имају најмање 200 измена, „велики корисници“ имају најмање 1.000 измена, „мали узорак“ подразумева највише 300 измена по кориснику док „велики узорак“ подразумева највише 1.000 измена по кориснику.

За сваки тип израчунавање је поновљено три пута, а у табелама 1 до 4 су дате репрезентативне вредности за сваки индикатор за типове узорака са просечним временом израчунавања.

	10	20	30	40	50	60	70	80	90	100
1.	2,08	1,28	0,92	0,54	0,31	0,47	0,41	0,5	0,27	0,42
2.	9,17	2,86	3,37	2,33	1,45	1,4	1,28	1,09	1,05	0,61
3.	5,47	3,08	1,71	1,36	0,7	0,72	0,48	0,67	0,5	0,54
4.	3,62	1,63	0,97	0,76	0,63	0,51	0,42	0,62	0,35	0,5
5.	9,43	2,83	0,54	0,46	2,95	0,66	0,37	0,07	0,29	0,13
6.	13,36	4,32	0,98	0,82	3,74	1,05	0,59	0,09	0,44	0,16
7.	19,5	5,62	1,26	1,4	4,74	0,68	0,44	0,19	0,35	0,07
8.	4,79	2,17	2,33	1,77	0,24	0,89	0,76	0,78	0,64	0,62

Табела 1 — Мали корисници, мали узорак (просечно време извршавања: 65,5 секунди)

	10	20	30	40	50	60	70	80	90	100
1.	3,65	0,72	0,83	0,21	0,35	0,11	0,38	0,19	0,27	0,21
2.	9,13	2,00	2,81	0,43	0,66	0,32	1,70	0,62	0,96	0,21
3.	7,41	0,93	2,10	0,32	0,84	0,41	0,79	0,30	0,84	0,19
4.	4,29	0,87	1,46	0,36	0,48	0,29	0,83	0,32	0,46	0,12
5.	1,89	0,59	1,09	0,13	0,52	0,29	0,66	0,26	0,99	0,19
6.	2,63	0,91	1,76	0,16	0,74	0,34	1,12	0,47	1,51	0,31
7.	1,26	0,96	1,54	0,39	1,30	0,48	1,17	0,68	0,73	0,32
8.	7,27	1,21	1,64	0,40	0,43	0,25	0,82	0,37	0,54	0,14

Табела 2 — Мали корисници, велики узорак (просечно време изврчавања: 135 секунди)

	10	20	30	40	50	60	70	80	90	100
1.	2,48	1,23	0,79	0,56	0,37	0,45	0,38	0,28	0,21	0,25
2.	9,93	2,81	2,08	2,02	1,82	1,13	0,99	0,89	0,64	0,67
3.	5,48	3,39	1,45	1,09	0,9	1,03	0,6	0,54	0,55	0,45
4.	4,15	1,82	0,99	0,74	0,56	0,56	0,43	0,6	0,43	0,47
5.	2,13	3,45	1,5	2,95	1,2	1,54	0,25	1,65	0,16	0,2
6.	4,6	6,18	2,03	4,56	1,62	2,16	0,35	1,78	0,19	0,32
7.	7,14	8,89	0,88	4,16	0,81	1,7	0,35	1,52	0,22	0,14
8.	6,38	1,86	1,38	1	0,72	0,58	0,84	0,4	0,52	0,5

Табела 3 — Велики корисници, мали узорак (просечно време извршавања: 64 секунде)

	10	20	30	40	50	60	70	80	90	100
1.	2,17	1,44	0,77	0,56	0,48	0,34	0,30	0,28	0,22	0,34
2.	7,98	4,48	2,77	1,33	1,73	1,33	1,21	1,07	1,02	0,66
3.	4,16	2,31	1,34	1,07	0,83	0,61	0,68	0,68	0,51	0,56
4.	4,08	3,06	1,25	0,76	0,61	0,64	0,47	0,34	0,28	0,34
5.	9,19	0,90	1,11	1,24	0,73	0,61	0,41	0,73	0,56	0,11
6.	12,44	1,32	2,02	2,03	1,58	1,07	0,60	1,30	0,94	0,10
7.	12,95	1,93	3,32	1,25	2,40	0,52	0,34	1,04	1,21	0,04
8.	3,94	2,94	1,82	0,85	0,86	0,82	0,53	0,63	0,49	0,52

Табела 4 — Велики корисници, велики узорак (просечно време извршавања: 177 секунди)

Као што се може видети, величина корисника незнатно утиче на време извршавања, док величина узорка значајно утиче на време извршавања.

Тражена одлика довољно стабилног просека је да се приликом више узастопних израчунавања испитивани корисник налази „са исте стране просека“, то јест да се ни за једну категорију не дешава да у једном израчунавању корисник буде изнад просека а у следећем буде испод просека.

На жалост, ни код једног типа узорка стабилност просека није задовољавајућа, чак ни за 100 корисника. Нежељене ситуације, да се корисник у једном израчунавању за неку категорију налази са једне, а у следећем са друге стране просека се јављају у скоро сваком израчунавању. Ово је донекле могуће објаснити чињеницом да постоји 8 индикатора са укупно чак 37 категорија[5]. Ипак, такву ситуацију би по сваку цену ваљало избећи, јер би у супротном корисник који врши проверу могао да понавља израчунавање све док не добије исход који он жели, чак и ако тај исход не одговара у потпуности стварном стању.

Израчунавања су поновљена и са знатно већим узорцима (200 и 300 корисника), али ни са том величином узорка није постигнута прихватљива стабилност.

Даље повећање узорка не би било прихватљиво због предугог времена извршавања, тако да је било неопходно применити неку другу стратегију за решавање овог проблема. Стога је уведен интервал поверења. Експерименталним путем је утврђено да појас око просечне вредности величине по 1% са сваке стране повећава стабилност узорка у довољној мери да се ретко дешава да се у узастопним израчунавањима за исте кориснике добијају различити резултати. Уколико се вредност корисникове статистике нађе у овом појасу, не зна се са сигурношћу да ли је изнад или испод просека. Како би се избегла могућност да ово доведе до лажних позитивних резултата, узима се да је вредност корисникове статистике у овим случајевима са оне стране просека која повећава разлику између два корисника.

Величина узорка за просечне вредности износи 100 корисника. Разматрају се само корисници који имају више од 100 измена, а за сваког корисника се узима у обзир највише 1.000 измена.

7 Величина корисниковог узорка

Потребно је одредити доњу и горњу границу броја корисникових измена које се узимају у обзир.

Када је у питању горња граница, основни разлог за њено ограничавање је да се убрза израчунавање. Међутим, како се израчунавање просечних вредности састоји из рачунања тестова за стотињак корисника од којих сваки има пар стотина до хиљаду измена, израчунавање већег броја измена за још два корисника неће превише

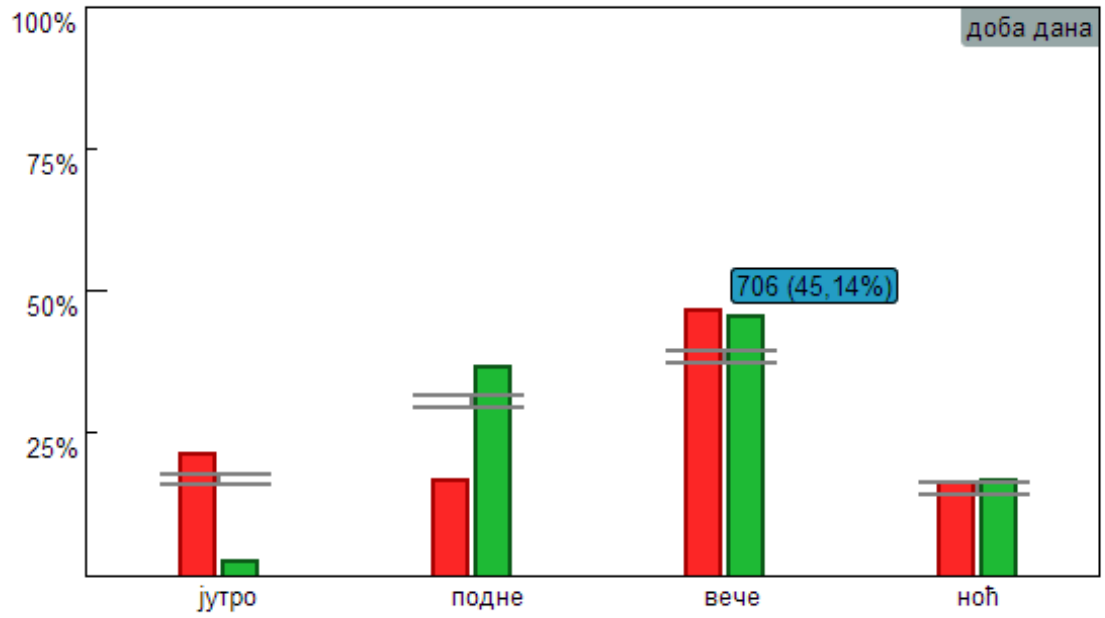
[5] За опис индикатора и категорија види поглавље 9.

допринети укупном времену израчунавања, тако да се горња граница може поставити на разумно велику вредност од 5.000 измена.

Када је у питању доња граница, битно је обезбедити да број корисникових сесија не буде премали. Број измена је већи или једнак броју сесија, тако да ако се обезбеди значајан број сесија, и број измена ће нужно прелазити тај праг. Као доњи праг за број сесија сесија је узето 50 сесија. Да би се одредила граница која задовољава такав услов, извучен је узорак од 100 корисника, за свакога је извучено 500 најскоријих измена, и израчунато је колико измена улази у 50 последњих сесија. Од 100 одабраних корисника испоставило се да 4 корисника имају толики број измена по сесији да 500 измена не покрива 50 сесија. Они су стога одбачени као екстремни случајеви. Када су у питању осталих 96 корисника, аритметичка средина броја измена у 50 сесија износи 348,69 док медијана износи 293,5 док за 75% узорка 50 сесија има мање од 448 измена. Стога је минималан број измена које корисник мора да има да би се могао упоређивати са другим корисником постављен на 500. Ово није услов који ограничава примену алата, јер је он првенствено и намењен за поређење активних корисника, који имају велики број измена.

8 Графици

Ради графичког приказа индикатора су развијени графици који приказују сличност између два налога, на основу конкретног индикатора. Сваки график упоредно приказује вредности индикатора за оба посматрана корисника, по категоријама.

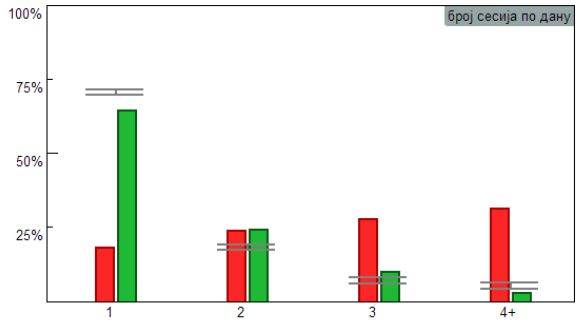


Слика 15 — График расподеле броја измена по временима дана

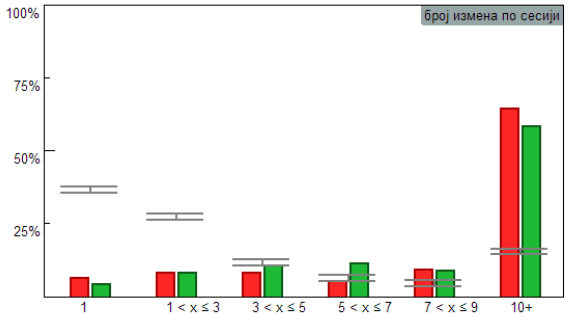
Црвени стубићи приказују вредности за првог корисника а зелени стубићи за другог корисника. Сиве линије представљају појас у коме се налазе просечне вредности. Висина стубића одговара проценту измена или сесија корисника које потпадају под одговарајућу категорију (збир процената за све категорије износи 100%). Као што се може видети на слици 15, ако се курсор миша постави изнад неког од стубића или цртице која представља просечне вредности, појављује се натпис који приказује колико измена или сесија спада у ту категорију, као и колики је њихов удео у укупном броју посматраних измена или сесија.

Вредности се приказују у процентима како би графици за све индикаторе били униформни, и у смислу програмерске имплементације, и у смислу коришћења. На тај начин се омогућава једноставно додавање нових тестова, а корисницима се омогућава да на исти начин интерпретирају податке за све индикаторе.

На слици 16 је приказан график броја сесија по дану за два корисника. Из графика је очигледно да се кориснички профили по овом критеријуму значајно разликују. На слици 17 је приказан график броја измена по сесији за два корисника чије понашање је слично. Може се уочити да су подаци за два корисничка налога међусобно слични, а да значајно одступају од просечних вредности, што је наводи да би иза оба налога могла стајати иста особа.



Слика 16 — Поређење броја сесија по дану, пример веома различитих профила



Слика 17 — Поређење броја измена по сесији, пример веома сличних профила

9 Индикатори

Развијен је скуп индикатора за упоређивање понашања сумњивих корисничких налога. Овај скуп се дели у две категорије: (а) индикатори који се тичу појединачних измена; и (б) индикатори који се тичу уређивачких сесија.

- расподела измена по данима у недељи
- распоред по именским просторима
- доба дана
- преклопљена времена дана
- величине измена

- број измена по сесији
- дужина сесије
- број сесија по дану
- време између измена
- проценат исправки

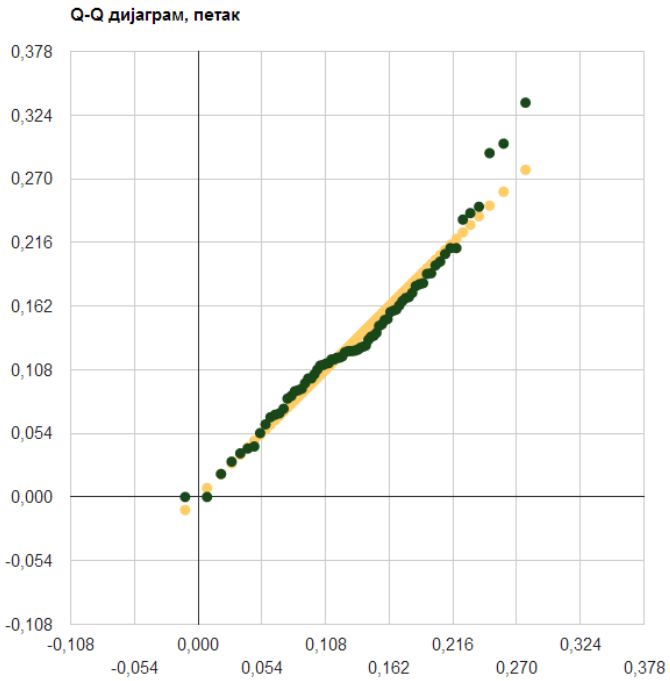
Индикатори су имплементирани на униформан начин, како би их било лакше анализирати (машински и од стране корисника), и како би било једноставније додавање нових индикатора у будућности.

Сваки индикатор разврстава корисникове измене/сесије у одговарајуће категорије, и исцртава график који приказује колики проценат измена/сесија припада којој категорији.

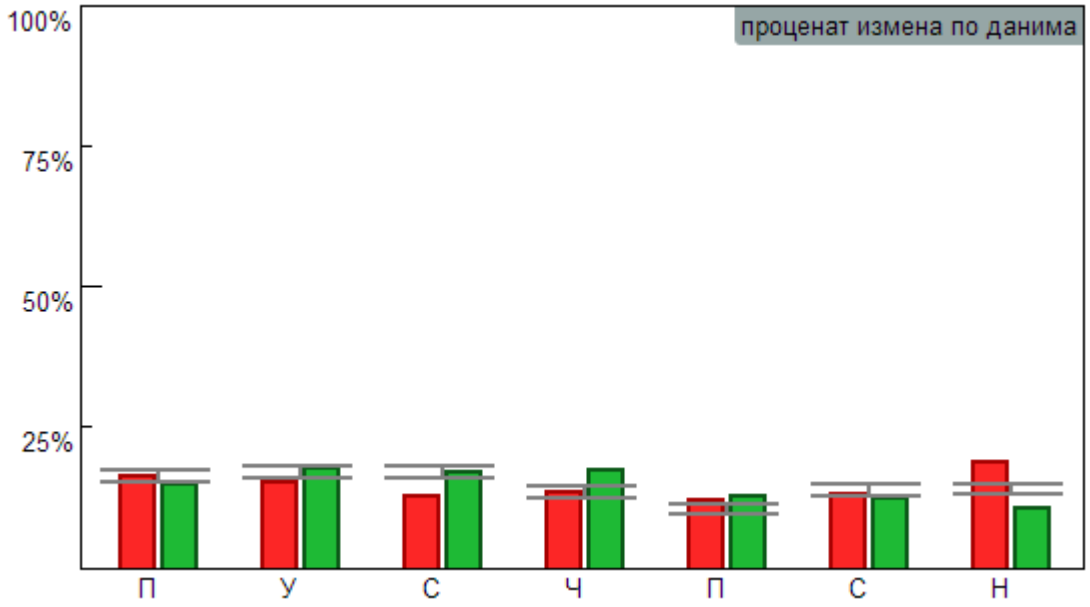
Следи опис индикатора праћен примером QQ дијаграма за неку од категорија индикатора и примером дијаграма који приказује резултате за тај индикатор. Дијаграми који приказују резултате за индикаторе су добијени упоређивањем два случајно изабрана корисничка налога која припадају различитим особама.

Расподела измена по данима у недељи даје груби временски профил корисничких измена. На пример, запослена особа са уобичајеним радним временом највише слободног времена има током викенда, док код пензионера не постоји значајна разлика у количини слободног времена по данима. Уколико две особе имају различит распоред слободног времена, вероватно ће се и временски профил њихових измена разликовати. Такође, уколико појединац који уређује са више налога има особен распоред слободног времена, вероватно ће сви ови налози имати неуобичајену расподелу измена по данима. Овај индикатор измене дели у седам категорија које одговарају данима у недељи.

QQ дијаграм за пример категорије овог индикатора дат је на слици 18, а пример дијаграма израчунатих резултата упоређивања два корисника за овај индикатор је дат на слици 19.



Слика 18 — QQ дијаграм за проценат измена петком

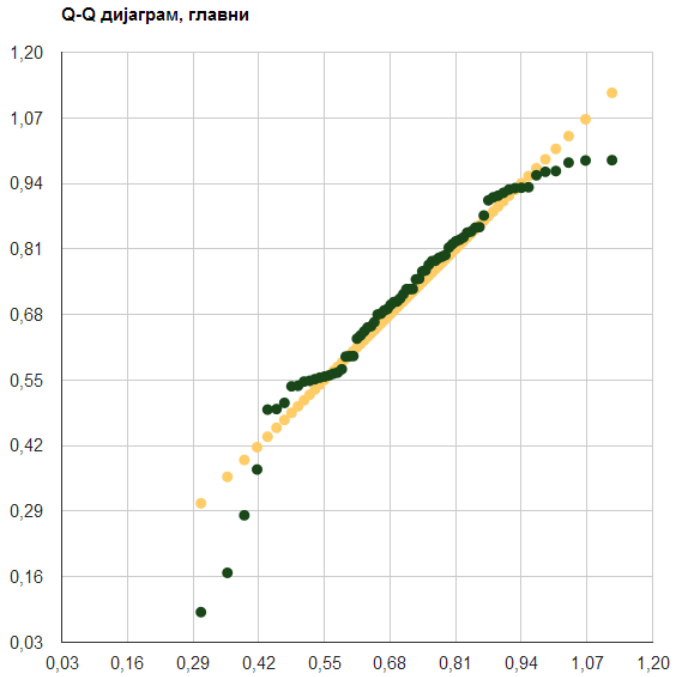


Слика 19 — Проценат измена по данима

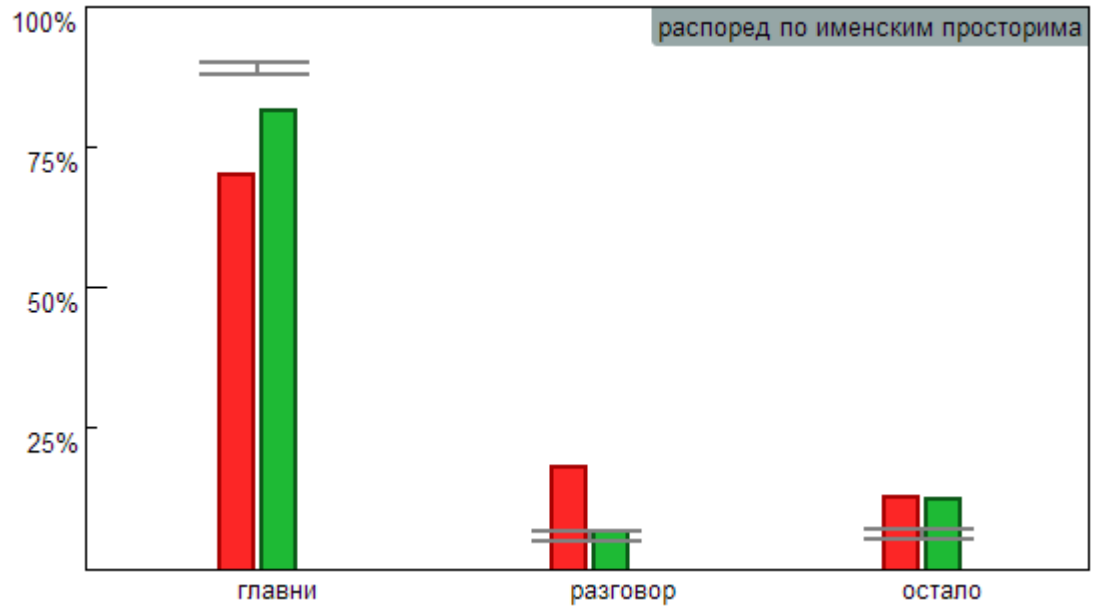
„Најкоришћенији“ **именски простори** су главни именски простор, стране за разговор, и „википедија“ именски простор (на коме се налазе стране за опште расправе, гласања, усвојене политике и слично). Велики удео измена неких корисника отпада на измене на странама за разговор, док се други клоне дуготрајних расправа и концентришу се на рад у главном именском простору. Како удео страна за разговор у укупном броју измена зависи од склоности и карактера особе, има смисла укључити и овај параметар у анализу. Википедија има око двадесет именских простора, али је активност корисника у већини њих врло мала, тако да овај индикатор дели измене у три категорије: главни именски простор, странице за разговор и „остало“.

Како постоје само три категорије од којих је једна доминантна, вероватно је да кад су у питању нумеричке оцене сличности корисника овај индикатор неће бити од превелике користи. Међутим, он ће бити од несумњиве користи особи која врши упоређивање, јер јој омогућава да визуелно на дијаграму утврди специфичности понашања два корисничка налога.

QQ дијаграм за пример категорије овог индикатора дат је на слици 20, а пример дијаграма израчунатих резултата упоређивања два корисника за овај индикатор је дат на слици 21.



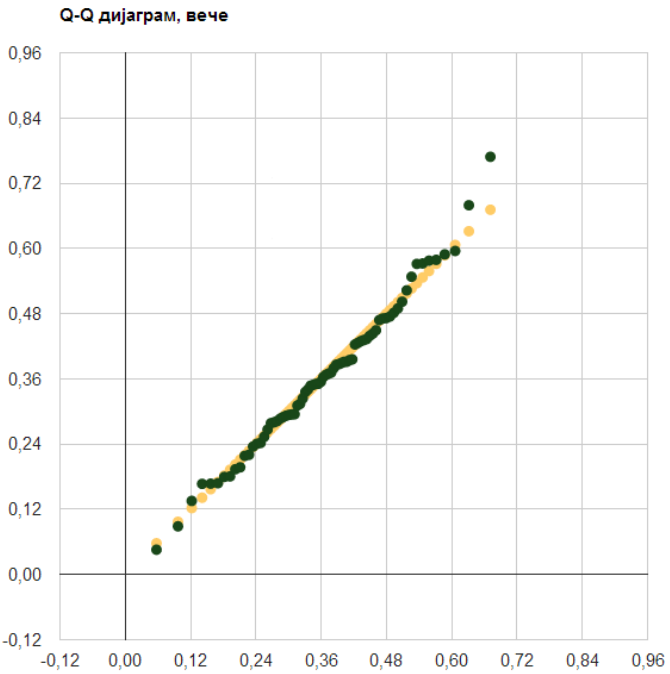
Слика 20 — QQ дијаграм за проценат измена у главном именском простору



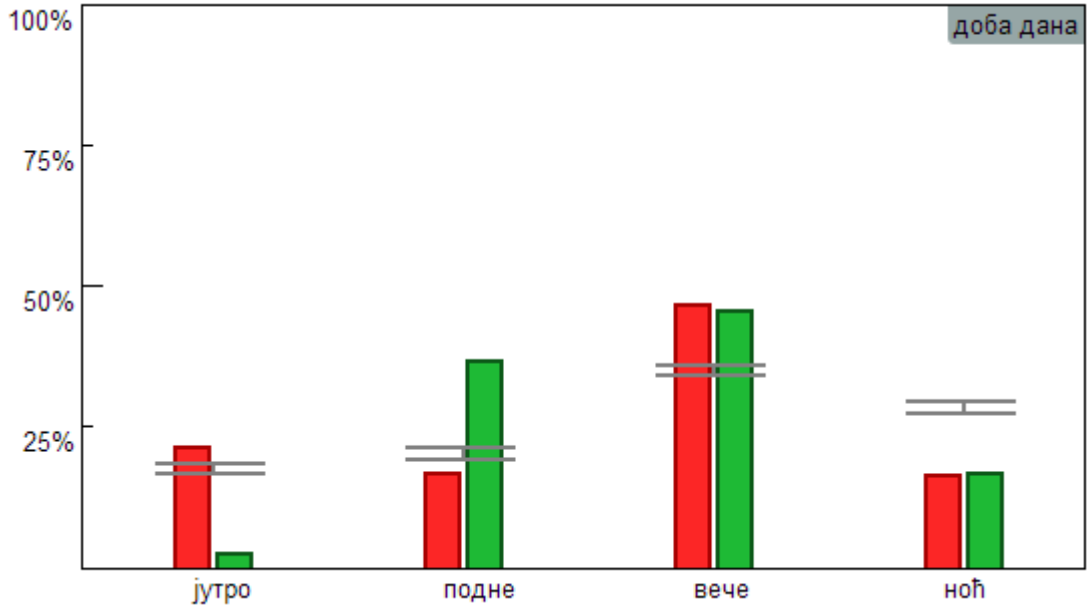
Слика 21 — Распоред по именским просторима

Расподела по добима дана даје финији временски профил корисничких измена. Уколико је особа раноранилац, за очекивати је да ће већи проценат њених измена бити током јутарњих часова, док је код особе која касније устаје и касније леже за очекивати да ће више измена бити током вечерњих и ноћних сати. Ово би могао бити посебно користан индикатор јер директно зависи од особине коју није лако променити. Такође, уколико (што је често случај) су сумњиви налози активни у истом временском периоду (истих дана), није једноставно учинити да се распоред измена по добима дана значајније разликује. Овај индикатор дели измене у четири категорије: јутро (06.00 до 12.00), подне (12.00 до 18.00), вече (18.00 до 00.00) и ноћ (00.00 до 06.00).

QQ дијаграм за пример категорије овог индикатора дат је на слици 22, а пример дијаграма израчунатих резултата упоређивања два корисника за овај индикатор је дат на слици 23.



Слика 22 — QQ дијаграм за проценат измена у вечерњим часовима

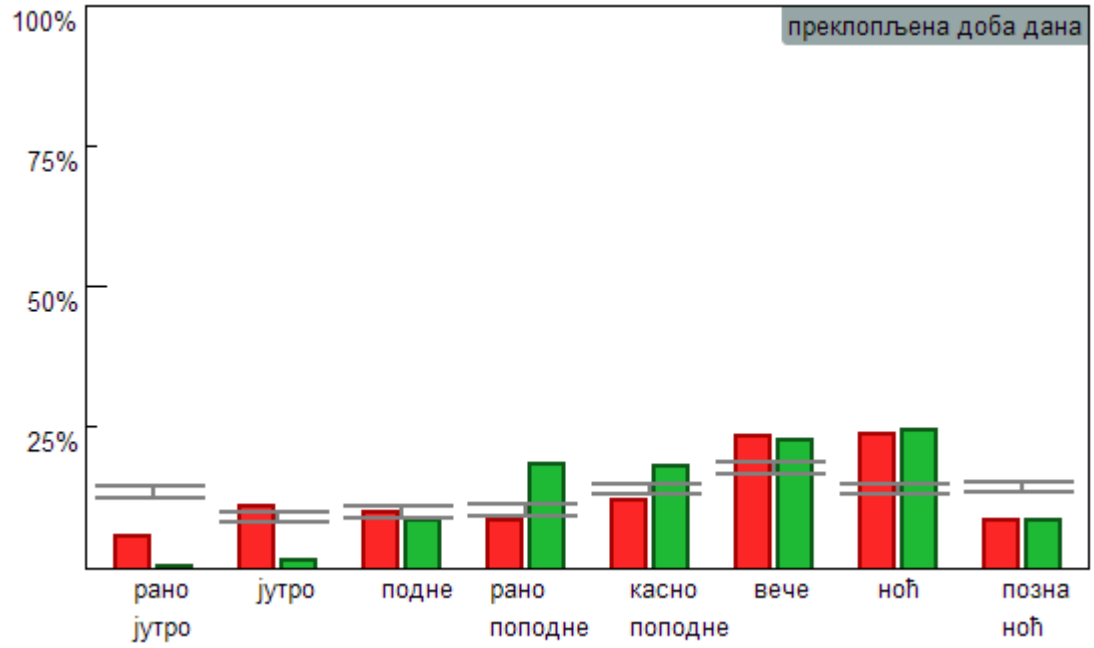


Слика 23 — Доба дана

Расподела по преклопљеним добима дана даје преглед истих података као и расподела по временима дана, али је уместо четири категорије присутно осам, које се међусобно преклапају. Сваки интервал се једном својом половином преклапа са суседом који му претходи, а другом половином се преклапа са суседу који следи иза њега. Категорије су: рано јутро (03.00 до 09.00 часова), јутро (06.00 до 12.00), подне (09.00 до 15.00), рано поподне (12.00 до 18.00), касно поподне (15.00 до 21.00), вече (18.00 до 00.00), ноћ (21.00 до 03.00) и позна ноћ (00.00 до 06.00). Како се интервали равномерно преклапају, сваки податак упада у тачно два интервала, тако да преклапање интервала не доводи до искривљења резултата анализе, а овакав индикатор

би могао да прецизније моделује понашање корисника чије тежиште активности је између два интервала.

Пример дијаграма израчунатих резултата упоређивања два корисника за овај индикатор је дат на слици 24.

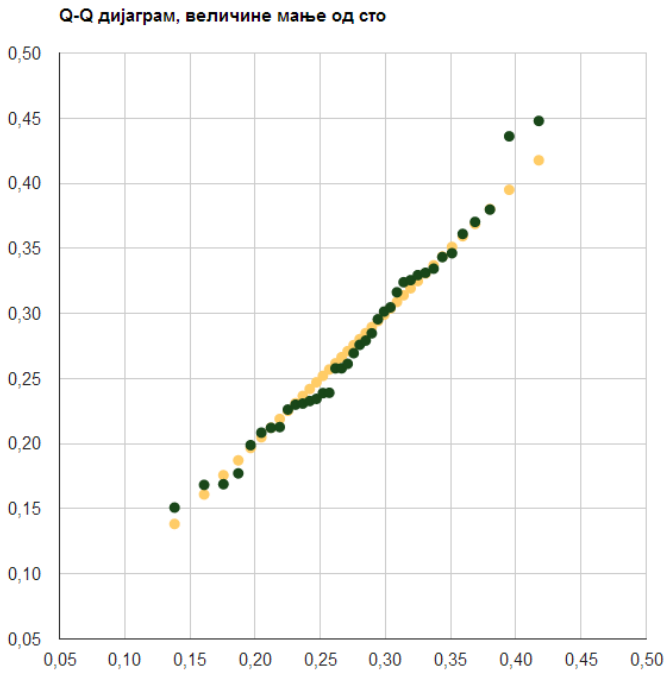


Слика 24 — Преклопљена доба дана

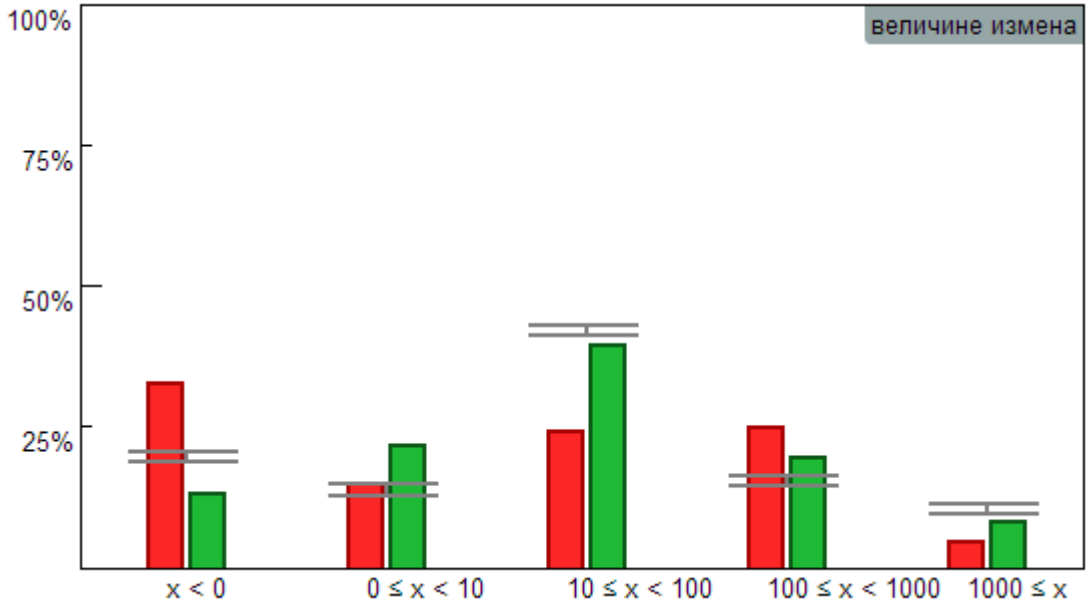
Величина измена се мери у бајтовима (може да буде негативна). Неки корисници начелно праве веће измене, на пример пишући нове чланке, док други који се баве сређивањем чланака (додавањем категорија, вики веза и слично) праве мање измене. Такође, неки корисници, док пишу чланак, праве више мањих измена за разлику од других, који сниме садржај тек кад је у потпуности припремљен. Због тога је овај индикатор повезан са бројем измена по сесији. Овај индикатор дели измене у пет категорија: мање од 0 бајтова[6], од нула до 10 бајтова, од 10 до 100 бајтова, од 100 до 1000 бајтова, и више од 1000 бајтова.

QQ дијаграм за пример категорије овог индикатора дат је на слици 25, а пример дијаграма израчунатих резултата упоређивања два корисника за овај индикатор је дат на слици 26.

[6] У случају да је део текста обрисан а не дописан.



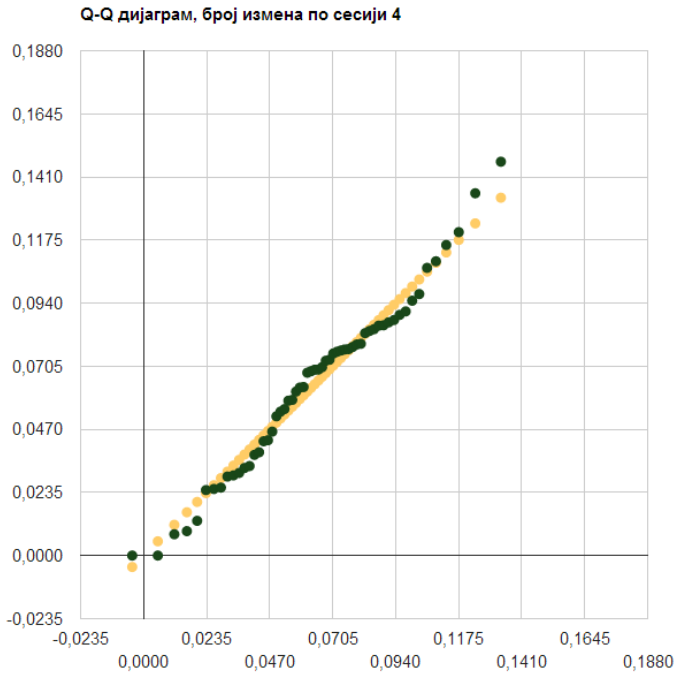
Слика 25 — QQ дијаграм процента измена које су мање од 100 бајтова



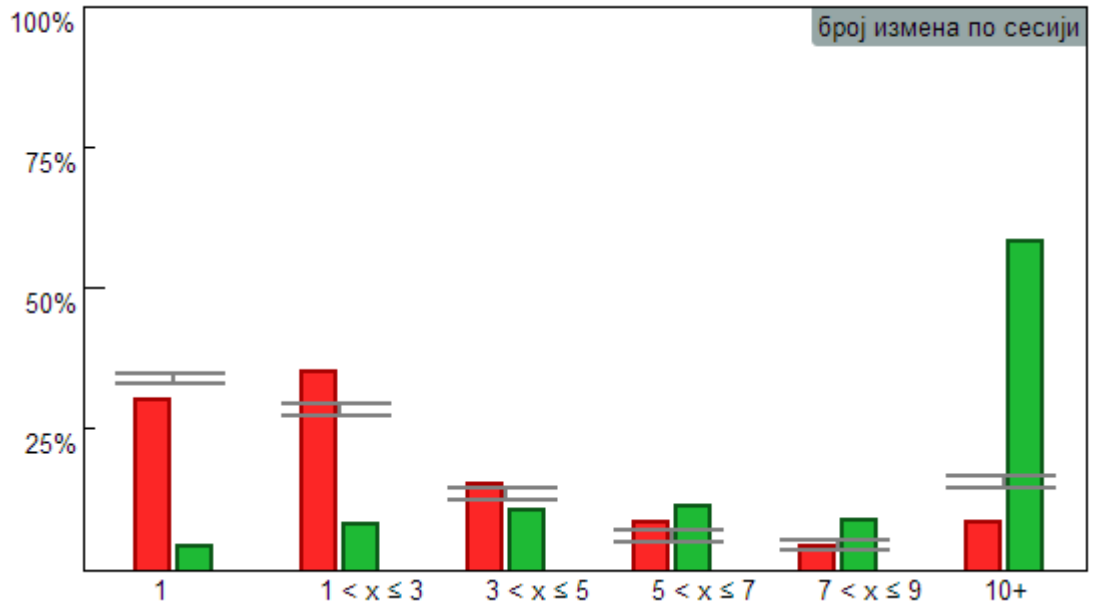
Слика 26 — Величине измена

Број измена по сесији је индикатор који описује ниво активности корисника. Са једне стране су корисници који праве пуно измена по сесији што значи да у просеку не троше превише времена по измени или им сесије трају дуго, а са друге стране су корисници који праве мало измена по сесији, што значи да су њихове измене обимне (временски захтевне), или да једноставно нису превише активни на Википедији (кад посете Википедију начине само по једну или пар измена). Овај индикатор дели сесије у шест категорија: 1 измена по сесији, 2 - 3, 4 - 5, 6 - 7, 8 - 9 и 10 или више измена по сесији.

QQ дијаграм за пример категорије овог индикатора дат је на слици 27, а пример дијаграма израчунатих резултата упоређивања два корисника за овај индикатор је дат на слици 28.



Слика 27 — QQ дијаграм процента сесија које чине 4 измене

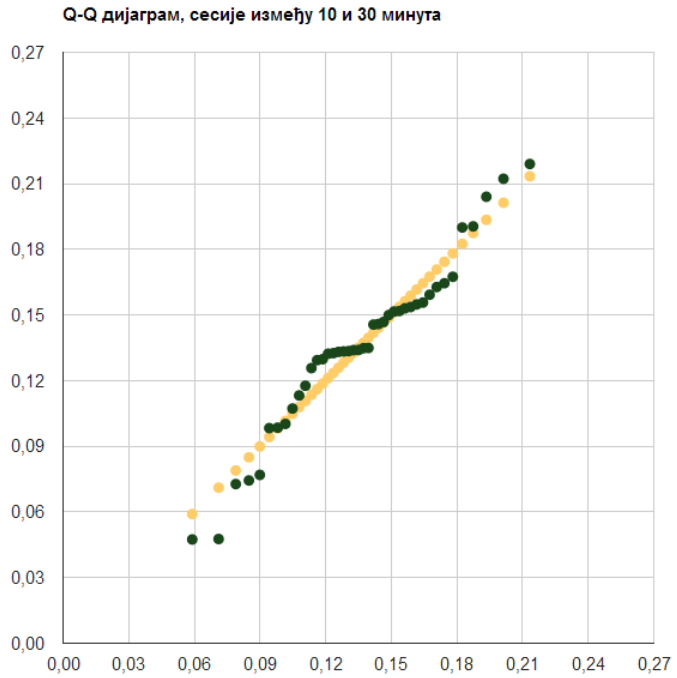


Слика 28 — Број измена по сесији

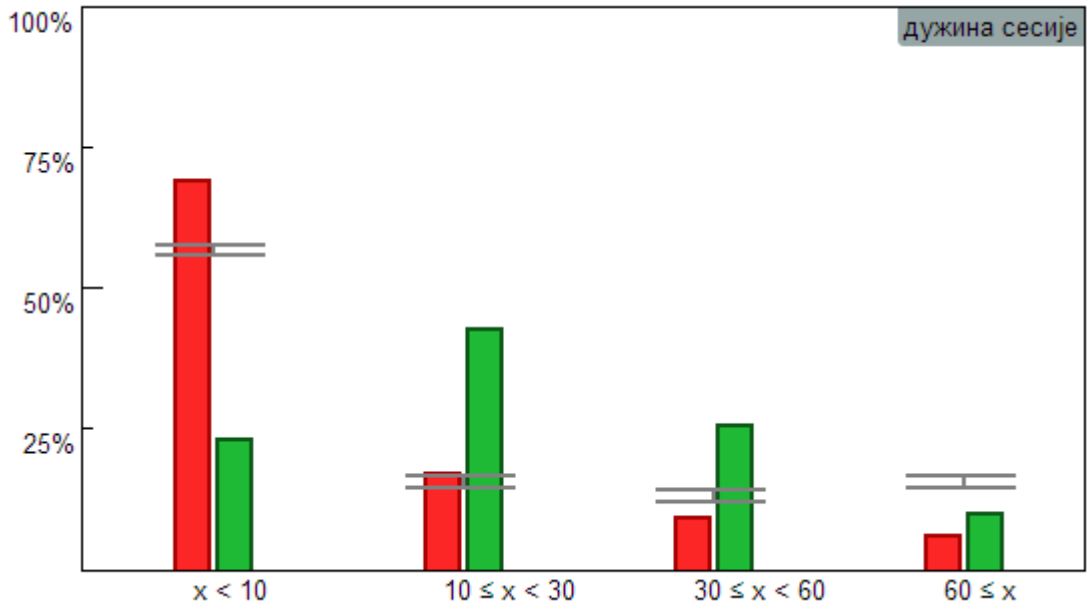
Просечна дужина сесије у минутима одређује колико времена корисник проводи на Википедији. Разумна је претпоставка да се овај индикатор значајно разликује од корисника до корисника, тако да може бити од користи приликом одређивања профила понашања корисника. Овај индикатор дели сесије у четири категорије: сесије краће од десет минута, сесије од 10 до 30 минута, сесије од 30 до 60 минута, и сесије дуже од 60

минута. Наравно, у дужину сесије није могуће урачунати време које је корисник утрошио у припремању прве измене у сесији.

QQ дијаграм за пример категорије овог индикатора дат је на слици 29, а пример дијаграма израчунатих резултата упоређивања два корисника за овај индикатор је дат на слици 30.



Слика 29 — QQ дијаграм процента сесија које трају између 10 и 30 минута

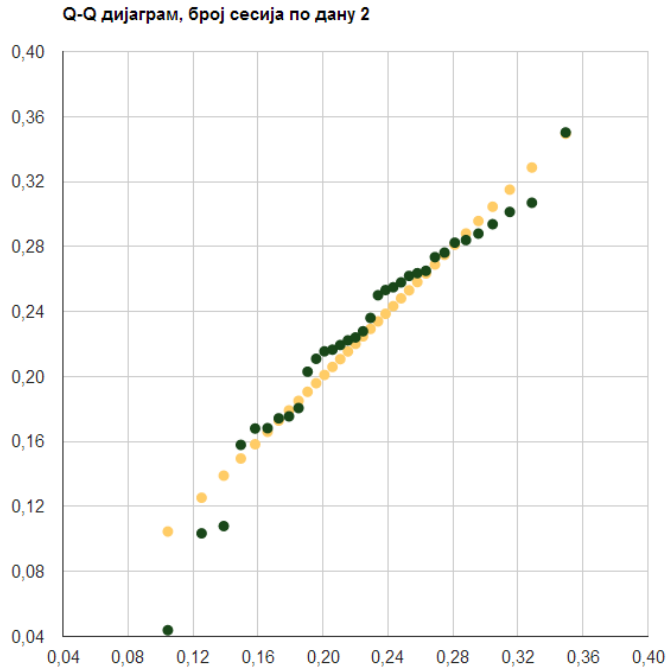


Слика 30 — Дужина сесије

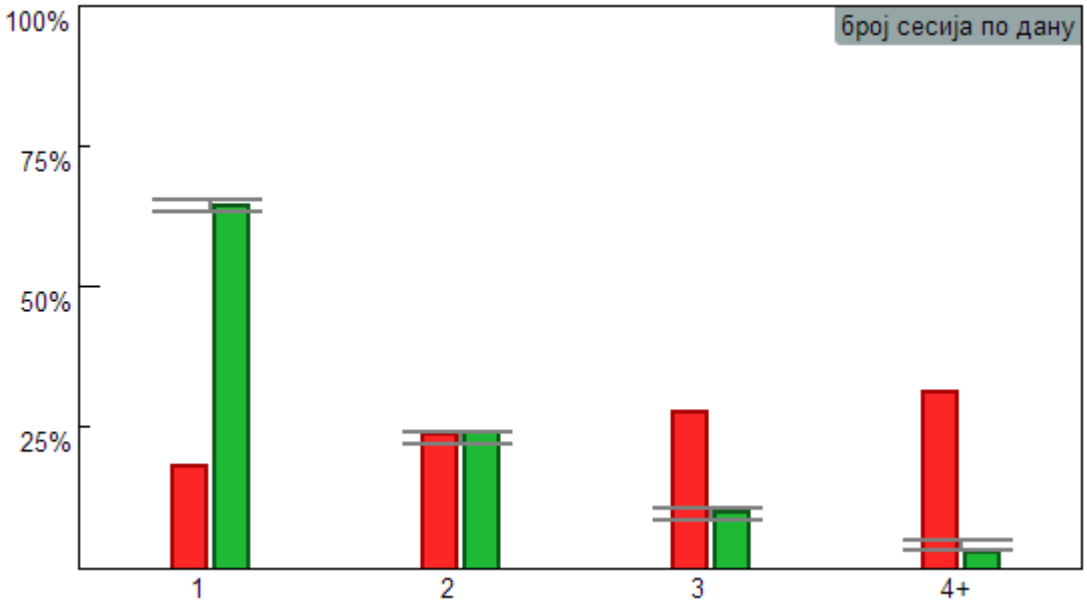
Број сесија по дану дели кориснике у две групе. Једну чине корисници који „успут“ посете Википедију пар пута дневно, и направе по једну или неколико измена, и

корисници који су пасионирани уредници и који Википедију уређују „кад год стигну“, више пута дневно. Другу групу чине корисници који своје измене групишу у једној већој (по броју измена и по дужини) сесији и корисници који мање уређују Википедију, можда не ни сваки дан, а мало вероватно више пута дневно. Овај индикатор дели дане у којима су корисници активни у четири категорије: дани током којих је постојала једна сесија, две, три или четири и више сесија.

QQ дијаграм за пример категорије овог индикатора дат је на слици 31, а пример дијаграма израчунатих резултата упоређивања два корисника за овај индикатор је дат на слици 32.



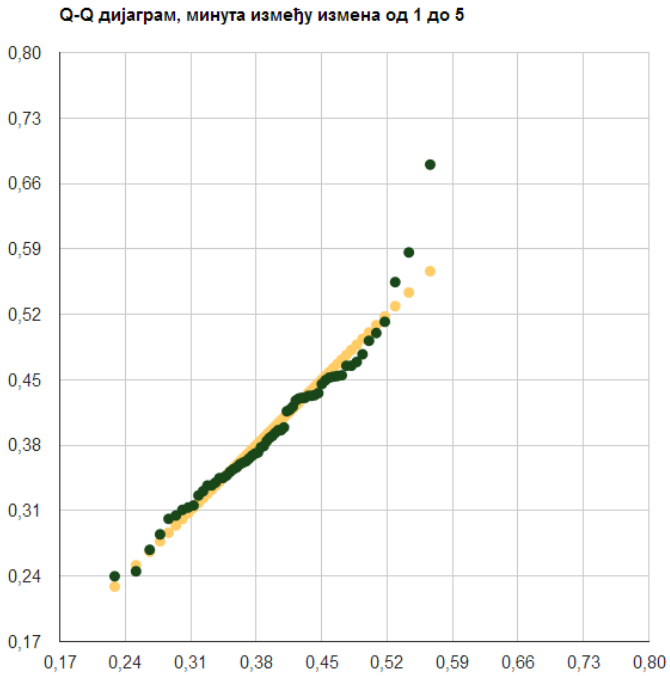
Слика 31 — QQ дијаграм процента дана са по две сесије



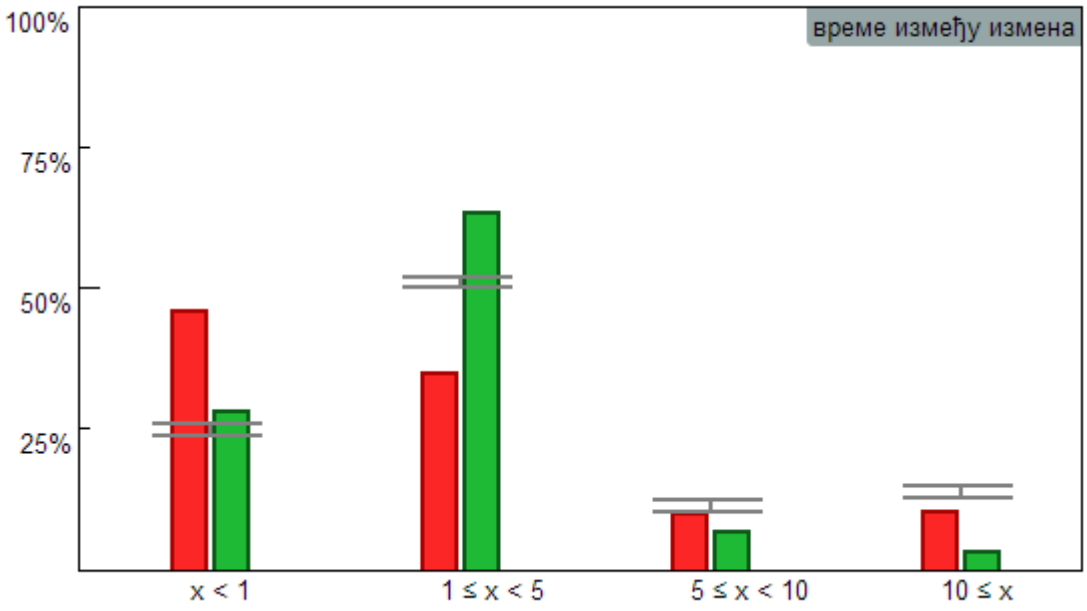
Слика 32 — Број сесија по дану

Време између две измене у оквиру сесије може да буде индикатор две особине корисника. Велики број малих, репетитивних измена (прављење преусмерења, додавање категорија) омогућава да време између њих буде краће. Такође, уколико две особе пишу текст исте дужине, време које ће им бити потребно зависи од њихове брзине писања (а можда чак и од времена које им је потребно да осмисле реченице). Брзина писања вероватно више утиче на измене на странама за разговор него на измене у чланцима, јер је уређивање чланака по својој природи сложеније (тражење извора, форматирање текста) од остављања коментара на странама за разговор. Овај индикатор дели измене у четири категорије: измене које су начињене мање од једног минута након претходне, оне које су начињене један до пет минута након претходне, оне које су начињене пет до десет минута након претходне и оне које су начињене више од десет минута након претходне.

QQ дијаграм за пример категорије овог индикатора дат је на слици 33, а пример дијаграма израчунатих резултата упоређивања два корисника за овај индикатор је дат на слици 34.



Слика 33 — QQ дијаграм процента измена насталих 1 до 5 минута након претходне измене



Слика 34 — Време између измена

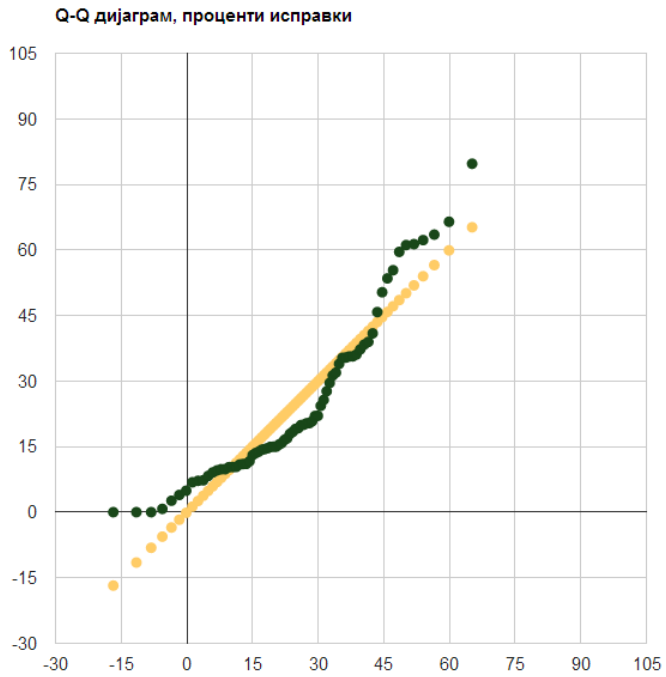
Приликом уређивања Википедије, кориснику је на располагању дугме „прикажи претпреглед“, које му омогућава да види ефекте своје измене пре него што је сачува у бази. На овај начин корисник може да се постара да је његова измена исправна и исправно форматирана пре него што је сачува, и стога не мора да начин више узастопних измена како би исправио недостатке у првобитној измени. Неки корисници употребљавају ово дугме и чак и врло сложене интервенције на чланку спроводе у оквиру једне измене, док други корисници немају обичај да користе претпреглед већ тек након што измену сачувају у бази проверавају њену исправност.

За потребе анализе нису доступни подаци о коришћењу претпрегледа јер се ти подаци не бележе у бази. Навика коришћења дугмета за претпреглед се имплицитно испољава у продуженом времену између две измене и у умањеној учесталости исправки.

Учесталост исправки је параметар по коме се корисници могу врло јасно сврстати у категорије на основу свог понашања. На жалост, не постоји начин да се прецизно утврди да нека измена представља исправку претходне измене. Уколико корисник направи две узастопне измене на страни за разговор, могуће је да се ради о реплици другом корисницу у дискусији. Са друге стране, могуће је да корисник уочи грешку у својој измени и после више сати након првобитне измене. Стога је „исправка“ дефинисана као измена која је начињена на истом чланку као и претходна измена у оквиру исте сесије. Узима се у обзир само главни именски простор (чланци) и не рачунају се измене из две различите сесије (размак између оригиналне измене и исправке не сме бити већи од 60 минута и између њих не сме бити измена на другим страницама).

Ово је једини индикатор који нема више категорија, већ даје само једну нумеричку вредност (проценат свих измена које су оцењене као исправке), и стога се приликом упоређивања корисника по овом критеријуму не исцртава график већ се само израчунавају проценти.

QQ дијаграм за овај индикатор дат је на слици 35.



Слика 35 — QQ дијаграм процента исправки

10 Метрике за поређење профила корисника

Ради добијања нумеричке оцене сличности понашања два корисничка налога је разматрано и имплементирано неколико метрика. Метрике рачунају сличност на нивоу индикатора, сумирајући вредност разлика по категоријама.

- Сума апсолутних разлика
- Сума квадрата разлика
- Растојање Чебишева
- Ранговска метрика

Прве три метрике представљају „наивни“ приступ и служе више за грубу анализу и тестирање исправности кода и смислености извучених података. Ове метрике су засноване на растојању Минковског[18]. Сума апсолутних разлика представља растојање Минковског реда 1, растојање Чебишева представља растојање Минковског бесконачног греда, док је сума квадрата разлика заснована на растојању Минковског реда 2.

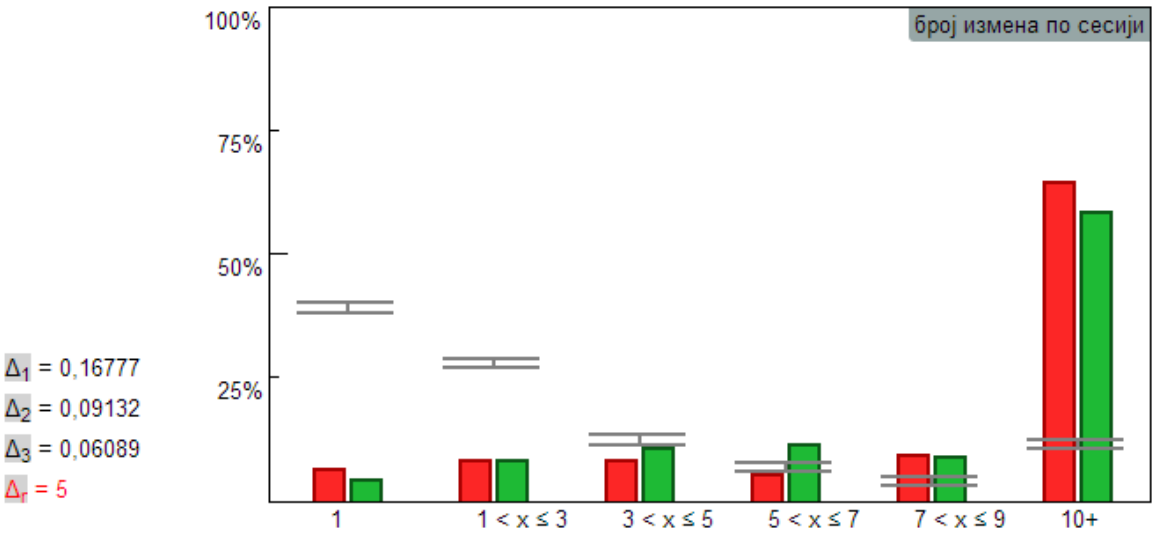
Ранговска метрика у пракси даје много употребљивије резултате. Ова метрика се рачуна као број категорија теста код којих оба корисничка налога одступају од просечне вредности на исту страну. Уколико је вредност неке категорије за оба налога већа од просечне или мања од просечне, онда та категорија добија ранг +, а ако је вредност за један налог већа од просека а за други налог мања од просека онда та категорија добија ранг -. Збир категорија које имају ранг + даје ранговску метрику.

Ради лакше интерпретације података од стране корисника, резултати које даје ранговска метрика су класификовани у четири класе (снажна сличност, слаба сличност, слаба различитост и јака различитост), и приказани су различитим бојама у зависности од интензитета сличности (нијансе зелене уколико подаци указују на различитост а нијансе црвене уколико указују на сличност).

Треба имати у виду да ако индикатор има n категорија, број степени слободе је такође n. На први поглед би се могло учинити да податак са које стране просека се налази првих $n-1$ категорија једнозначно одређује са које стране просека се налази n-та категорија. Да се ради о нумеричким вредностима, то би заиста и био случај, али како се ради о ранговским вредностима, ако k категорија има ранг + и j категорија има ранг - $(k + j = n-1)$, тада n-та категорија може да има вредност и + и - (осим у посебном случају када је $k = 0$ или $j = 0$).

На слици 36 је приказан пример дијаграма броја измена по сесији, и метрика за тај индикатор за иста два корисника. Вредност Δ_1 (0,1678) је сума апсолутних разлика, вредност Δ_2 (0,091) је сума квадрата разлика, вредност Δ_3 (0,061) је максимална апсолутна разлика, а вредност Δ_r (5) је ранговска метрика. Као што се може видети,

вредности индикатора за кориснике одступају од просека на исту страну код свих категорија осим код 4. категорије.



Слика 36 — Дијаграм броја измена по сесији, и пратеће метрике

11 Анализа резултата

Циљ анализе резултата је да се утврди да ли су основне претпоставке које су коришћене у изради овог рада биле оправдане, и да ли је њихова имплементација довела до алата који је заиста употребљив за упоређивање корисника Википедије, и утврђивање да ли два корисничка налога припадају истој особи. Жељени исход је алат који ретко доводи до лажних позитивних и лажних негативних резултата. Лажни позитивни резултати представљају случајеве када два корисничка налога припадају различитим особама, али резултати анализе указују на то да иза њих стоји иста особа. Лажни негативни резултати представљају случајеве када два корисничка налога припадају истој особи, али резултати указују на то да иза њих стоје две различите особе.

Да би се добила оцена квалитета индикатора и метрика, потребно је упоредити резултате узорака из скупа парова корисничких налога који припадају различитим особама и скупа парова корисничких налога који припадају истој особи.

Као парови налога који не припадају истој особи су узети случајно одабрани парови налога.

Када су у питању налози који припадају истој особи, јавља се проблем проналажења довољног броја оваквих парова налога зато што често сумњу да два налога припадају истој особи није могуће са потпуном сигурношћу потврдити или одбацити. Такође, многе ситуације су из разних разлога специфичне. На пример, дешава се да неки корисник има више група корисничких налога за које међусобно није доказано да припадају истој особи, али за налоге унутар групе то јесте потврђено. Из ових разлога

су као парови „налога“ који припадају истој особи узети дисјунктни подскупови скупа измена једног корисничког налога. На пример, подскуп најскоријих 1.500 измена представља првог „корисника“, а подскуп 1.500 измена пре њих представља другог „корисника“. Тачно да два подскупа скупа измена једног корисника нису савршена апроксимација за два скупа измена различитих корисника који припадају истој особи, али могу послужити у циљу утврђивања да ли су одабране карактеристике понашања особе довољно константне да би њихова анлиза послужила сврси.

Оба узорка су узимана из скупа свих корисника који имају преко 3.000 измена. Величина овог скупа у тест-бази износи 135 корисника.

Узорак резултата за корисничке налоге који не припадају истој особи је добијен тако што су из скупа корисника који имају преко 3.000 измена узимани случајни парови корисника. Издвојен је узорак од 30 парова са понављањем (један корисник се може наћи у више парова, али је вођено рачуна да ниједна два неуређена пара не чине исти корисници). За сваки пар су спроведене анализе и забележени рангови за сваки од индикатора.

Узорак резултата за корисничке налоге који припадају истој особи је добијен тако што је из скупа корисника који имају преко 3.000 измена случајним одабиром узето 30 корисника. За сваког од ових корисника је пронађено 3.000 најскоријих измена. Првих 1.500 измена представља првог корисника а других 1.500 измена представља другог корисника. За ова два корисника су спроведене анализе и забележени рангови за сваки од индикатора. Поступак је поновљен 30 пута.

Резултати су дати у табели 5 која садржи следеће колоне:

- Прва колона (аритметичка средина) приказује аритметичку средину вредности индикатора за узорак од 30 парова различитих особа.
- Друга колона (стандардна девијација) приказује стандардну девијацију вредности индикатора за узорак од 30 парова различитих особа.
- Трећа и четврта колона приказују аритметичку средину и стандардну девијацију вредности индикатора за узорак од 30 парова корисничких налога који припадају истим особама.
- Пета колона приказује број категорија индикатора.
- Шеста колона приказује апсолутну разлику аритметичких средина узорака за исте и различите особе, подељену бројем категорија, то јест апсолутну разлику прве и треће колоне подељену бројем категорија.

	различите особе		иста особа		број кат.	апсолутна разлика/ број категорија
	аритметичка средина	стандардна девијација	аритметичка средина	стандардна девијација		
проценат измена по данима	2,63	1,43	2,93	1,46	7	0,043
распоред по именским просторима	0,97	1,00	1,97	1,10	3	0,333
доба дана	1,97	1,00	3,20	0,96	4	0,308
преклопљена доба дана	3,33	1,84	6,20	1,58	8	0,359
величина измена	2,00	1,11	3,57	1,10	5	0,314
број измена по сесији	2,60	1,94	3,23	1,45	6	0,105
дужина сесије	1,80	1,49	2,20	1,44	4	0,100
број сесија по дану	2,10	1,27	2,43	1,36	4	0,083
време између измена	1,47	1,17	2,77	1,10	4	0,325

Табела 5 — Приказ резултата анализе над узорком величине 30 парова корисника

Као што је и очекивано, код свих индикатора, аритметичка средина узорка за различите особе је мања од аритметичке средине узорка за исту особу. Апсолутна разлика је највећа код преклопљених доба дана (2,87), док је најмања код процента измена по данима (0,30), броја сесија по дану (0,33) и дужина сесије (0,40).

Међутим, разлике аритметичких средина је неопходно посматрати у односу на број категорија одговарајућег индикатора, јер на пример ранг 3 има много већи значај ако се ради о индикатору са 4 категорије него ако се ради о индикатору са 8 категорија. У последњој колони табеле је дата апсолутна разлика аритметичких средина узорака истих и различитих особа, подељена бројем категорија одговарајућег индикатора. Вредност за већину индикатора износи око 0,3.

Изузетак је проценат измена по данима, са вредношћу од само 0,043. Ипак, овај индикатор је значајан упркос томе што прави врло малу разлику између два скупа корисничких налога, јер дијаграм који приказује удео корисникових измена по данима у недељи може особи која анализира податке да открије потенцијално врло корисне податке. На пример, уколико свим данима активности оба корисника одговарају просечним, само једним даном значајно одударају на исту страну, иако ранг индикатора може да буде врло мали, ово свеједно наводи на то да би оба налога могла припадати истој особи која одређеног дана у недељи има значајно више или значајно мање слободног времена него осталим данима.

Индикатор број сесија по дану такође има малу вредност апсолутне разлике подељене бројем категорија – само 0,083. Како се ради о индикатору чији визуелни приказ за разлику од удела измена по данима у недељи нема посебан значај у упоређивању корисника, овај индикатор би се могао уклонити.

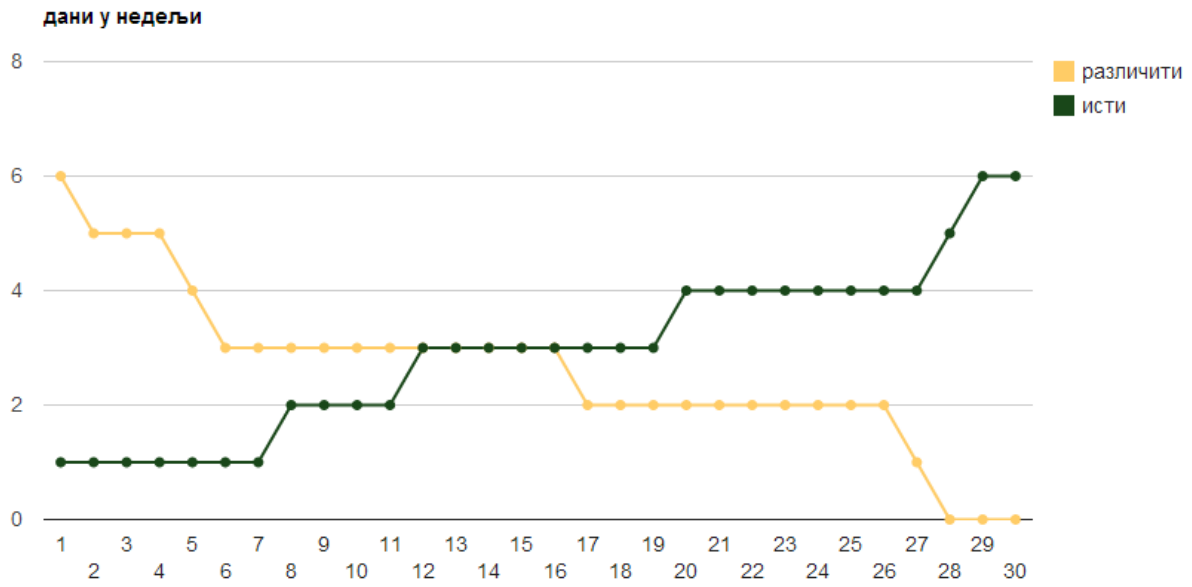
Остали индикатори дају задовољавајуће резултате, тако да могу да се користе у анализи понашања два корисника.

Укупно посматрано, скуп имплементираних индикатора даје смислен и потенцијално врло користан алат који се приликом истрага на Википедији може користити као значајно допунско средство уз постојеће алате.

11.1 Прагови

Подаци у табели 5 показују да је методолошки приступ који је коришћен у овом раду исправан. Ипак, како су кумулативне природе, нису погодни за одређивање прагова поклапања два корисничка налога које би указивало на то да оба налога припадају истој особи. Због тога су за сваки индикатор упоређене вредности из узорака како би се утврдило који праг у односу на број категорија са највећом сигурношћу одваја две популације (корисничке налоге који припадају истој особи и налоге који припадају различитим особама).

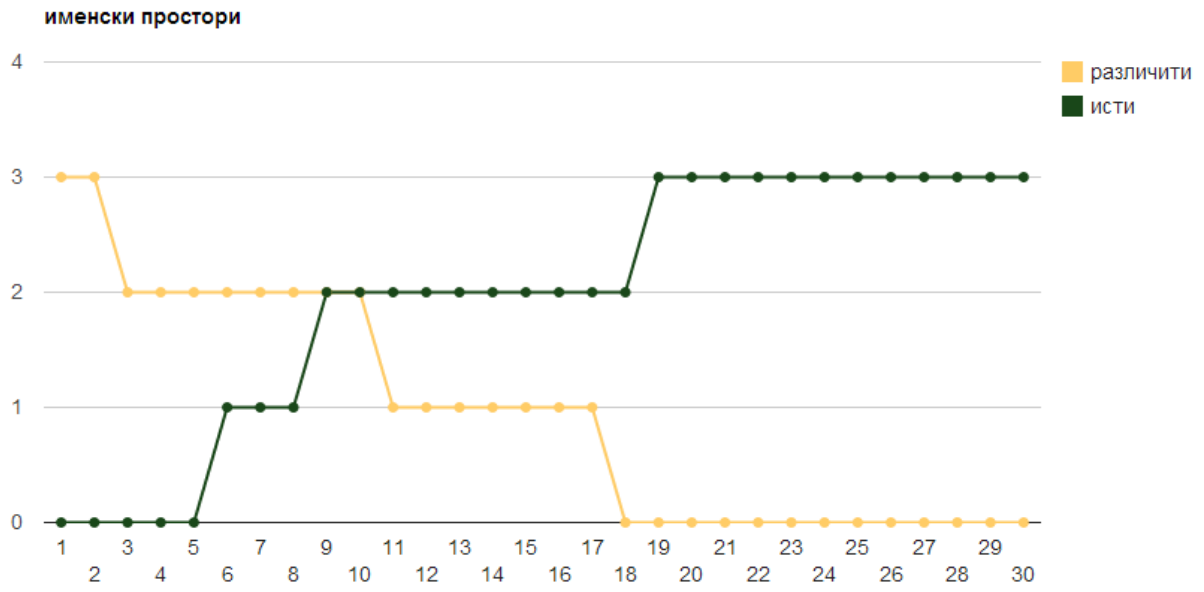
На сликама 37 до 45 су приказани дијаграми који податке из узорка налога који припадају истој особи (**узорак А**) приказују зеленом бојом а податке из узорка који припадају различитим (**узорак Б**) особама приказују жутом бојом. На *y* оси је приказана вредност ранговске метрике за пар упоређиваних корисника, а на *x* оси је приказан редни број пара унутар узорка.



Слика 37 — Распоред по данима у недељи

На слици 37 је приказан дијаграм који упоређује податке из узорака А и Б за распоред по данима у недељи.

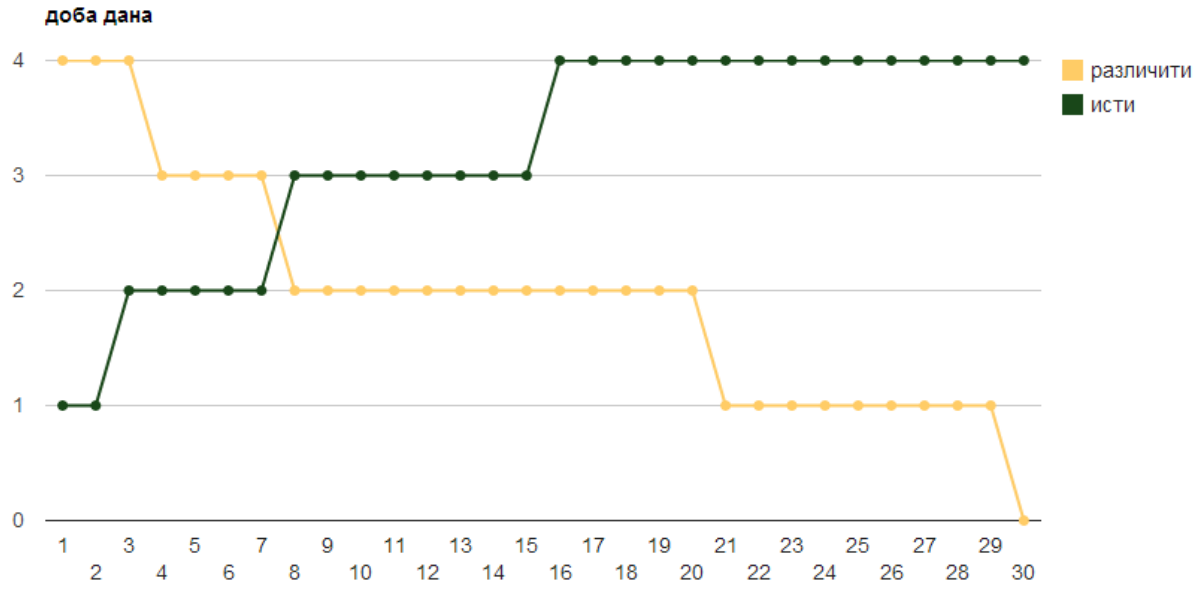
Распоред измена по данима у недељи има 7 категорија. Најбољи резултат даје ранг 4 који задовољава 11 елемената из узорка А и 5 елемената из узорка Б.



Слика 38 — Распоред по именским просторима

На слици 38 је приказан дијаграм који упоређује податке из узорака А и Б за распоред по именским просторима.

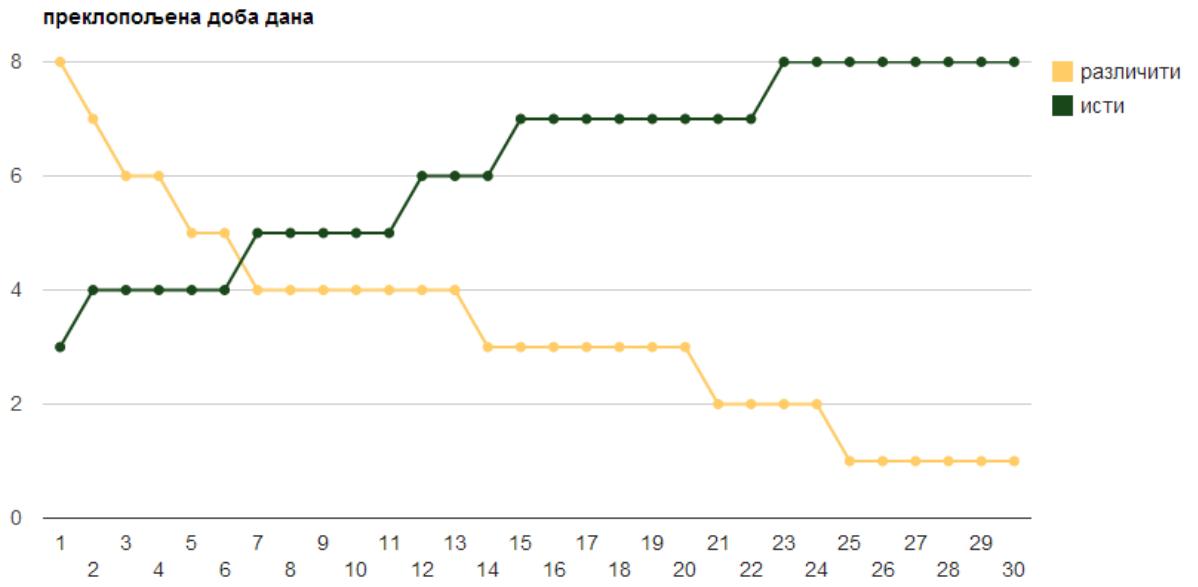
Распоред по именским просторима има 3 категорије. Најбољи резултат даје ранг који 3 који задовољава 12 елемената из узорка А и 2 елемента из узорка Б.



Слика 39 — Распоред по добима дана

На слици 39 је приказан дијаграм који упоређује податке из узорака А и Б за распоред по добима дана.

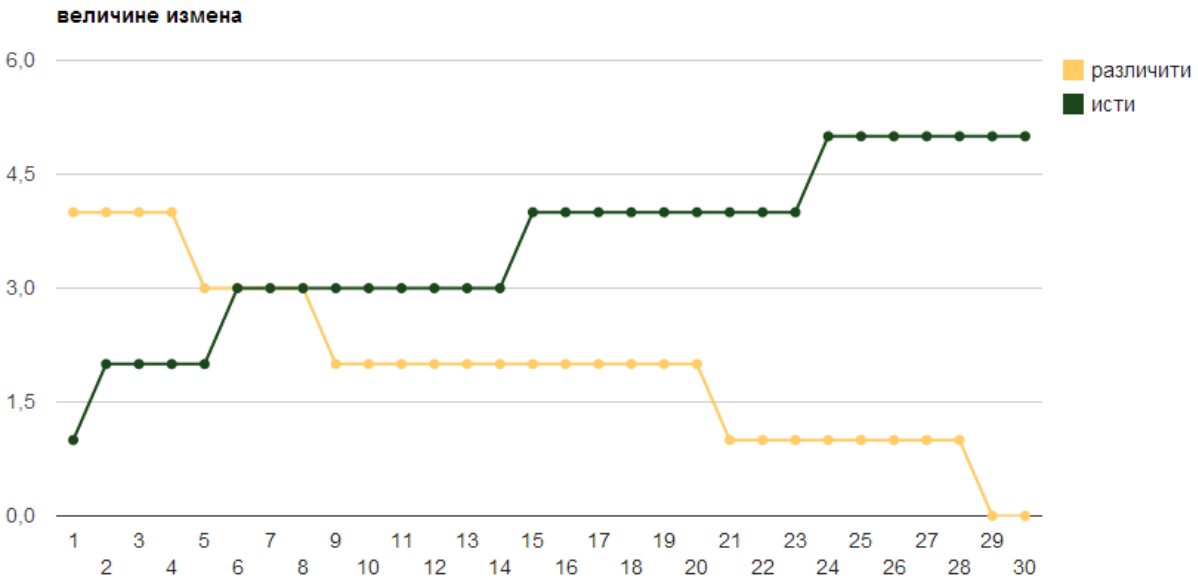
Распоред по добима дана има 4 категорије. Најбољи резултат даје ранг 4 који задовољава 15 елемената из узорка А и 3 елемента из узорка Б.



Слика 40 — Распоред по преклопљеним добима дана

На слици 40 је приказан дијаграм који упоређује податке из узорака А и Б за распоред по преклопљеним добима дана.

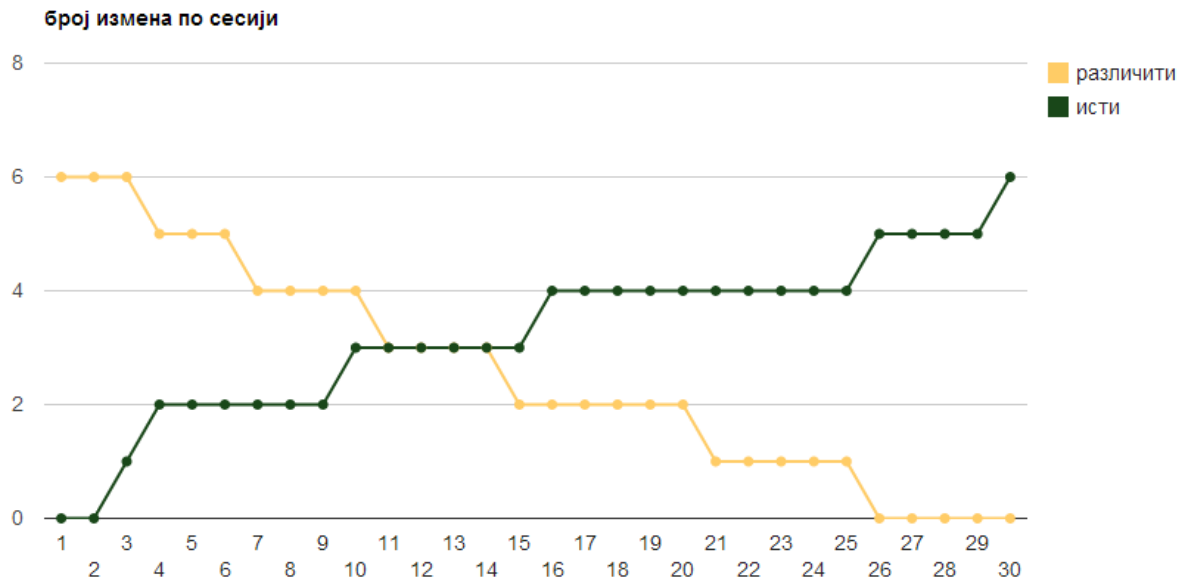
Распоред по преклопљеним добима дана има 8 категорија. Најбољи резултат даје ранг 7 који задовољава 16 елемената из узорка А и 2 елемента из узорка Б.



Слика 41 — Распоред по величинама измена

На слици 41 је приказан дијаграм који упоређује податке из узорака А и Б за распоред по величинама измена.

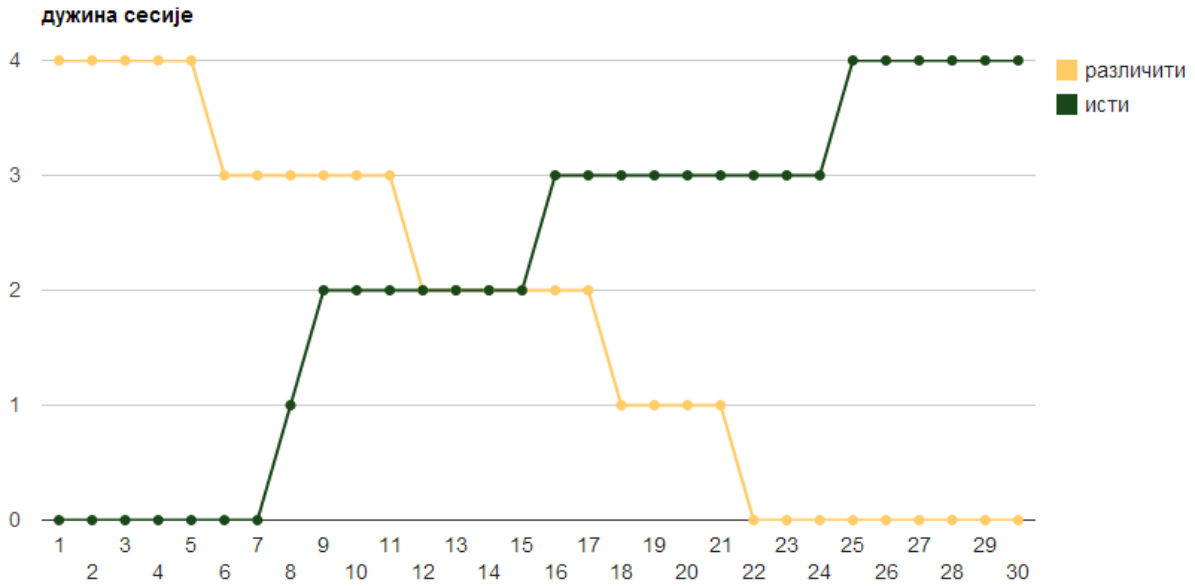
Распоред по величинама измена има 5 категорија. Најбољи резултат даје ранг 5 који задовољава 7 елемената из узорка А и ниједан елемент из узорка Б.



Слика 42 — Распоред по броју измена по сесији

На слици 42 је приказан дијаграм који упоређује податке из узорака А и Б за распоред по броју измена по сесији.

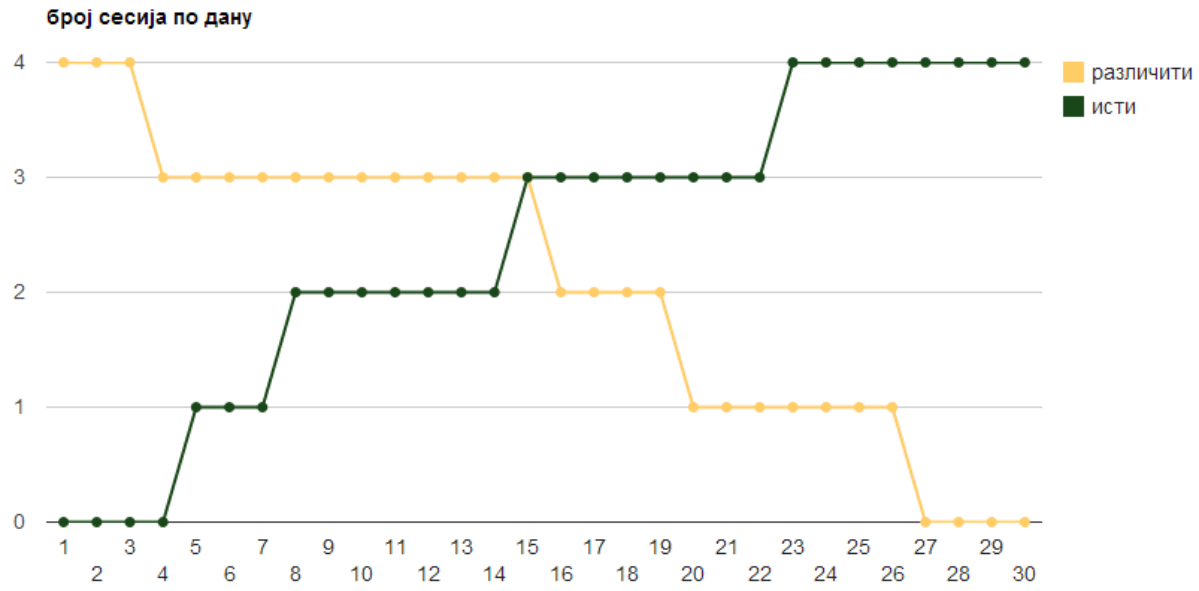
Распоред по броју измена по сесији има 6 категорија. Најбољи резултат даје ранг 4 који задовољава 15 елемената из узорка А и 10 елемената из узорка Б.



Слика 43 — Распоред по дужини сесија

На слици 43 је приказан дијаграм који упоређује податке из узорака А и Б за распоред по дужини сесија.

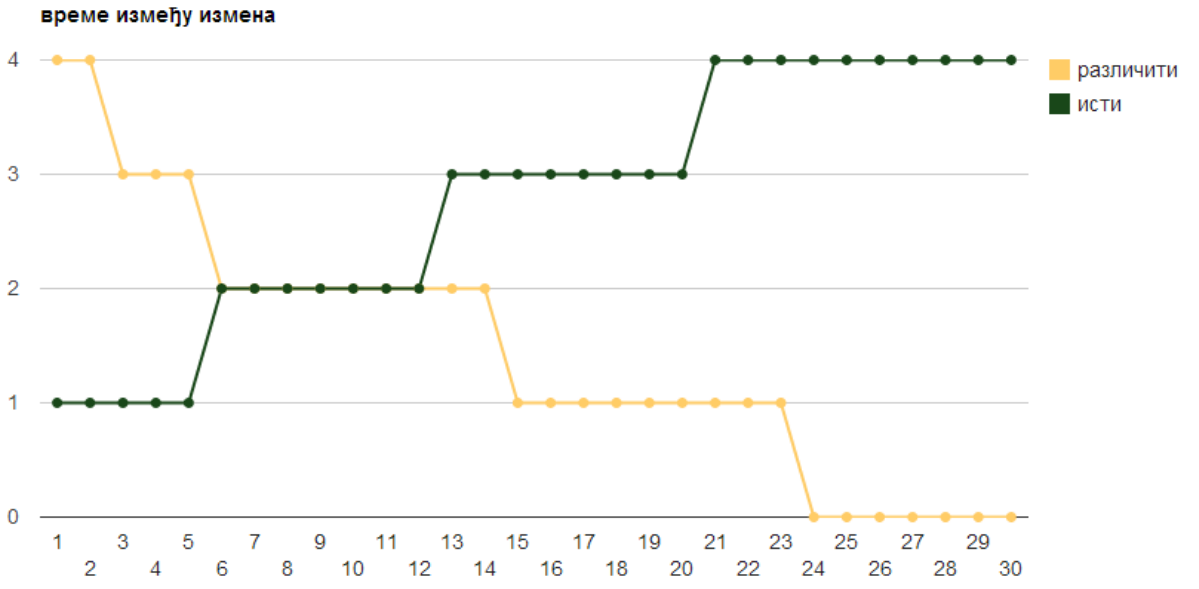
Распоред по дужини сесија има 4 категорије. Најбољи резултат даје ранг 3 који задовољава 15 елемената из узорка А и 11 елемената из узорка Б.



Слика 44 — Распоред по броју сесија по дану

На слици 44 је приказан дијаграм који упоређује податке из узорака А и Б за распоред по броју сесија по дану.

Распоред по броју сесија по дану има 4 категорије. Најбољи резултат даје ранг 4 који задовољава 8 елемената из узорка А и 3 елемента из узорка Б.



Слика 45 — Распоред по времену између измена

На слици 45 је приказан дијаграм који упоређује податке из узорака А и Б за распоред по времену између измена.

Распоред по времену између измена има 4 категорије. Најбољи резултат даје ранг 4 који задовољава 10 елемената из узорка А и 2 елемента из узорка Б.

За све индикаторе осим за распоред по броју измена по сесији и распоред по дужини сесија су пронађени прагови који дају задовољавајуће резултате. Ако се резултати индикатора посматрају заједно, то јест уколико већи број индикатора наводи на закључак да посматрани налози припадају истој особи, вероватноћа да је дошло до лажног позитивног резултата је мала.

11.2 Јединствена оцена сличности

Развијено проширење за МедијаВики је намењено употреби од стране корисника који су информатички писмени и који су способни за самостално доношење закључака на основу скупа презентованих података. Како проширење неће користити општа популација корисника Википедије, није неопходно да се након израчунавања кориснику исписује експлицитан закључак да ли два корисничка налога припадају истој особи или не. Штавише, ово није ни пожељно, јер би оваква „пресуда“ одвраћала истражитеља од дубље анализе приказаних графика која може да открије детаље значајне за његов коначни закључак.

Ипак, како би се испитала заједничка информативност коришћених индикатора, неопходно је одредити јединствену оцену сличности два профила, и на узорку испитати колико је та оцена прецизна у утврђивању да ли корисничким налозима управља иста особа.

Како су за сваки индикатор већ дефинисани прагови, за јединствену оцену сличности је узет број индикатора чија вредност прелази те прагове.

За потребе анализе јединствене оцене су узети нови узорци, на исти начин на који су прављени узорци за потребе анализе резултата и дефинисања прагова. Како у популацији има 135 корисника, узорци су величине 30, и узимани су из исте популације, узорци из претходне и ове фазе се делимично поклапају.

У табели 6 је дат приказ процента парова који имају јединствену оцену већу или једнаку n, за узорак парова који припадају истој особи (узорак А), и за узорак парова који не припадају истој особи (узорак Б). У табели 7 је дат приказ процента парова који имају јединствену оцену мању или једнаку n, за узорак А и за узорак Б.

разлика ≥ n	различити	исти
1	60%	93%
2	37%	80%
3	10%	57%
4	0%	33%

Табела 6 — Проценат парова корисника из узорака који имају јединствену оцену већу или једнаку n.

разлика ≤ n	различити	исти
1	63%	20%
2	90%	43%
3	100%	67%
4	100%	83%

Табела 7 — Проценат парова из узорака који имају јединствену оцену мању или једнаку n.

Како у узорку Б ниједан пар нема јединствену оцену већу од 3, а у узорку А таквих парова има 33%, са високом вероватноћом се може тврдити да парови корисничких налога који имају јединствену оцену већу од 3 припадају истој особи.

У узорку Б само 10% парова има јединствену оцену већу од 2, а у узорку А таквих парова има чак 57%, па за парове корисничких налога који имају јединсвену оцену већу од 2 постоји значајна вероватноћа да припадају истој особи.

У узорку А само 20% парова има јединствену оцену мању од 2, а у узорку Б таквих парова има 63%, тако да се за парове корисничких налога који имају јединствену оцену мању од 2 са високом вероватноћом може рећи да не припадају истој особи.

Како у узорку А 43% парова има јединствену оцену мању од 3, а у узорку Б чак 90% парова има јединствену оцену мању од 3, постоји значајна вероватноћа да парови корисничких налога који имају јединствену оцену мању од 3 не припадају истој особи.

12 Дискусија и закључак

У оквиру овог рада је написан код проширења за МедијаВики и припремљена је база података за тестирање. Подаци су затим анализирани како би се утврдиле њихове карактеристике а након тога је омогућено прављење профила просечног корисника. Затим је развијен је скуп индикатора за упоређивање корисничких налога са просеком, и метрике за интерпретацију резултата индикатора. Коначно су имплементирани дијаграми за приказ резултата како би се омогућило боље разумевање добијених резултата.

Анализом резултата је утврђено да су резултати задовољавајући и да добијени алат може наћи примену у пракси. Индикатори, метрике и дијаграми омогућавају да се понашање корисника посматра из различитих углова и да се утврди присуство евентуалних сумњивих образаца понашања. Ипак, уочено је да решење које је примењенео за израчунавање профила просечног корисника није у потпуности задовољавајуће ни са аспекта брзине израчунавања, нити са аспекта стабилности, те да би стога требало изнаћи нове приступе у решавању овог проблема.

Планови за даљи рад подразумевају првенствено укључивање овог проширења у стандардну инсталацију на свим језичким издањима Википедије. То укључује одређене бирократске али и техничке кораке, као што су превод документације, слање кода у централни репозиторијум и пролазак кроз поступак прегледања и евентуалне дораде кода.

Након што проширење буде пуштено у рад на Википедијама, биће потребно пратити и анализирати у пракси квалитет и резултате развијених индикатора и метрика, као и њихово дорађивање у складу са прикупљеним искуствима. Уколико чланови заједнице изнесу квалитетне предлоге, могуће је укључивање нових индикатора у проширење.

У консултацији са заједницом истражитеља и програмера, може се размотрити да се у испис који проширење приказује ипак укључи и јединствена оцена сличности, која у прво време неће бити део извештаја који се кориснику приказује. Такође, кроз практично искуство и комуникацију са корисницима се може променити начин израчунавања јединствене оцене сличности тако да се неким индикаторима да већи а другим индикаторима мањи значај.

13 Референце

[1] WikimediaFoundation, http://wikimediafoundation.org/wiki/Home 5. јун 2013.
[2] MediaWiki, http://www.mediawiki.org 5. јун 2013.
[3] Manual:Database access, MediaWiki, http://www.mediawiki.org/wiki/Manual:Database_access 5. јун 2013.
[4] Википедија:Кориснички нивои, http://sr.wikipedia.org/wiki/Vikipedija:Korisnički_nivoi 5. јун 2013.
[5] Top Sites, Alexa, http://www.alexa.com/topsites 5. јун 2013.
[6] Википедија:Именски простор, http://sr.wikipedia.org/sr/Vikipedija:Imenski_prostor 5. јун 2013.
[7] C0 Controls and Basic Latin, Unicode, http://www.unicode.org/charts/PDF/U0000.pdf 5. јун 2013.
[8] Wikimedia downloads, http://dumps.wikimedia.org/ 5. јун 2013.
[9] Toolserver, https://wiki.toolserver.org/view/Main_Page 5. јун 2013.
[10] Manual:MWDumper, MediaWiki, http://www.mediawiki.org/wiki/Manual:MWDumper 5. јун 2013.
[11] Performance Tips, Manual:MWDumper, MediaWiki, http://www.mediawiki.org/wiki/Manual:MWDumper#Performance_Tips 7. јун 2013.
[12] [Toolserver-l], Архива мејлинг листе, http://lists.wikimedia.org/pipermail/toolserver-l/2013-April/005873.html 7. јун 2013.

[13] Natallia V. Katenka, Statistics I, QQ Plot for Checking Normality, http://math.bu.edu/people/nkatenka/MA115_FALL2010/QQPlot.pdf 7. јун 2013.
[14] PHP documentation, Statistics, http://www.php.net/manual/en/book.stats.php 7. јун 2013.
[15] PHP documentation, Statistic Functions, comment, http://www.php.net/manual/en/ref.stats.php#84680 7. јун 2013.
[16] Google Developers, Visualization: Scatter Chart https://developers.google.com/chart/interactive/docs/gallery/scatterchart 7. јун 2013.
[17] StackOverflow, MySql Row Number, http://stackoverflow.com/questions/3126972/mysql-row-number/9842345#9842345 7. јун 2013.
[18] Code 10, Minkowski distance, http://www.code10.info/index.php?view=article&id=61 24. септембар 2013.