Գլուխ 1․ Ընդհանուր փաստեր վեբ scraping-ի մասին

Գրական հետազոտությունների ժամանակ հայտնաբերվել է վեբ scraping-ի մի քանի սահմանումներ: Ստորև ներկայացված բոլոր երեք հասկացությունները հիշատակվում են տվյալների ստացման տարբեր աղբյուրներում: Նրանք տարբերվում են ստացված տեղեկատվության համար նախնական աղբյուրների տեսքով:

Երբեմն անհրաժեշտ է տեղեկատվություն հավաքել վեբ կայքերից, որոնք նախատեսված են ոչ թե ծրագրային գործակալների, այլ մարդկային ընթերցողների համար:

Այս գործընթացը հայտնի է որպես «վեբ scraping»:

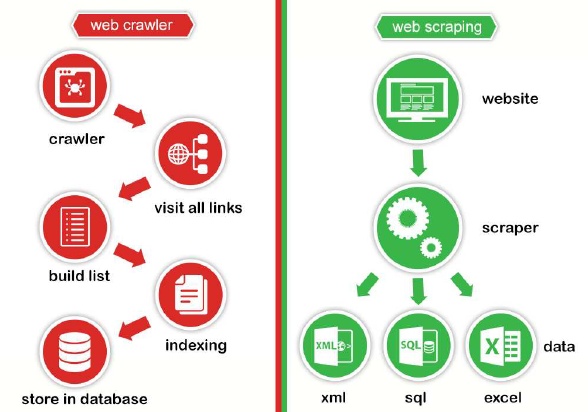
(Ապրիլ, 2009)

Առաջին սահմանումը նշում է տվյալների աղբյուրները, որոնք սկզբնապես նախատեսված են ընթերցողների համար: Նման որոշումը արդեն երկար ժամանակ է համարում են արդեն հնացած: Ավտոմատացված տեխնիկայի զարգացումը հնարավորություն է ընձեռում նաև ծրագրային ապահովման ընթեռնելի աղբյուրներից արդյունահանման միջոցով: Այնուամենայնիվ, պետք է հաշվի առնել, որ հրապարակման ամսաթիվը 2009 թվականն էր: Դիմումների ծրագրավորման ինտերֆեյսը (API) Աղբյուրները այդ ժամանակ սահմանափակ էին: Հանրային API- ի դիրեկտորը, որը հասանելի է Programmable Web-ում (Berlind, 2015) ունի մոտավորապես 750 առկա աղբյուրները , համեմատած 2017 թվականի 17175-ի:

Վեբ scraping-ը, որը նաև հայտնի է որպես վեբ extraction կամ harvesting, այնպիսի մեթոդ է, որ համացանցից (WWW) տվյալները հանում և պահպանում է այլ ֆայլային համակարգում կամ տվյալների բազայում, հետագայում որոնման կամ վերլուծության համար: Ընդհանրապես, վեբ-ի տվյալները ստացվում են (HTTP) պրոտոկոլի կամ վեբ բրաուզերի միջոցով: Դա կատարվում է ձեռքով՝ օգտագործողի կողմից կամ ավտոմատ կերպով բոտի կամ վեբ զննարկչի կողմից: Շնորհիվ այն բանի, որ WWW- ում անընդհատ գեներացվում է հսկայական քանակի տվյալներ, վեբ scraping-ը լայնորեն ճանաչվում է որպես մեծ տվյալների հավաքման արդյունավետ և հզոր տեխնիկա:

Ներկայիս իրավիճակը ավելի կոնկրետ կերպով նկարագրված է երկրորդ սահմանման մեջ, երբ վեբ scraping-ը հիշատակվում է որպես մեծ տվյալների հավաքման աղբյուրներից մեկը: Սահմանում է նաև մեկ այլ տերմին - վեբ Crawler:

 Վեբ Crawler: Վեբ Crawling- ը կատարվում է ուրիշ տարբերակով և տալիս է ուրիշ արդյունք: Նկար 1-ում նկարագրվում են երկու գործողությունները: Ձախ կողմում տեսանելի գործընթացը ցույց է տալիս, որ Վեբ Crawling- ը չունի յուրահատուկ թիրախ և ընդհանրացված տեղեկատվություն է պարունակում `առանց կոնկրետ տեղեկատվության նպատակաուղղվածության: Համեմատելով աջ կողմի նկարի հետ, որտեղ Վեբ Scraper- ը ստանում է տվյաներ, գործընթացներ է իրականացնում այդ նույն տվյալների հետ և վերլուծում տվյալ աղբյուրներից ստացված տվյալները: Web Crawling-ը չի դիտարկվում այս թեմայի մեջ:



Նկար 4.

Ստորև բերված սահմանման մեջ շատ մանրամասներ չեն նշվում:

Այնուամենայնիվ, դա հստակորեն գրավում է վեբ scraping -ի գործունեությունը:

Վեբ Scraping-ը ներառում է աղբյուրի հարցման, արդյունքների էջը վերցնելու և արդյունքների ձեռք բերման համար էջը վերլուծելու գործընթացը:

1.2 Վեբ Scraping -ի նպատակը

Համաշխարհային ցանցում առկա տեղեկատվության հսկայական մասը դեռեւս Հիպերտեքստի նշարկման լեզու (HTML) էջի ձեւաչափով են: Ավտոմատացված արդյունահանումը դժվար է, քանի որ նախատեսված էր մարդկանց ընթերցման համար: Այս գլխում ներկայացվում է տեղեկատվության արդյունահանման դրդապատճառները և Web Scraping- ի նպատակը :

Համաշխարհային ցանցի արագ զարգացումը զգալիորեն փոխվել է, թե ինչպես ենք մենք կիսվում, հավաքում և հրապարակում մեր տվյալները: Տեղեկատվության մեծ մասը պահվում է առցանց, ինչպես կառուցվածքային այնպես էլ ոչ կառուցվածքային ձևերով: Ինչ վերաբերում է որոշակի հարցերի կամ հետազոտությունների թեմաներին, ապա դա նոր խնդիր է առաջացրել – արդեն իսկ անհանգստացնում է տվյալների անհամապատասխանության և անհասանելիության խնդիրը, այլ ոչ թե ինտերնետի խոշոր զանգվածների հաղթահարումը:

(Բ. Կ. 2016)

Համաշխարհային ցանցից օգտվելը հաճախ հնարավոր է միայն ավտոմատացված վեբ Scraping-ի գոյության պատճառով: Առանց այդ տեխնիկայի, անհնար է հավաքել տվյալների մեծ քանակությունը հաճախակի և ողջամիտ ժամկետներում:

1.2.1 Շուկայական վերլուծություն և հետազոտություն

Ինտերնետային աղբյուրներից տվյալների հավաքումը դարձել է շուկայական հետազոտության մեթոդներից մեկը: Այն առաջարկում է ավելի արագ արձագանք, համեմատած դասական հետազոտության հետ:

Վայլ Նովիթը լավագույնն է համարում ավանդական հետազոտությունների օգտագործումը, իսկ վեբ Scraping -ը համարվում է որպես այդպիսի գործիքների համար ծախսարդյունավետ օժանդակություն: Համապարփակ պատկերացում ստանալու և շուկաներում գործիքներ ձեռք բերելու համար պետք է օգտագործել բազմաթիվ աղբյուրներ:

Սպառողները ակտիվ են առցանց աշխարհում եւ կիսվում են իրենց փորձով, վրդովմունքներով ու մոտիվացիաներով: Ընկերությունները, որոնք ցանկանում են ավելին իմանալ սպառողների մասին կարող են ավելացնել առցանց տեղեկատվության աղբյուրները:

Վեբ scraping-ը նման տվյալների հավաքման եղանակներից մեկն է:

1.2.2 Կորպորատիվ տեխնոլոգիաներ

Անհամատեղելի կորպորատիվ տեխնոլոգիաները բնորոշ են խոշոր նախագծերի համար։ Դեռ պետք է միասնական տվյալների ներկայացում մի քանի համակարգերից: Որոշ հատուկ դեպքերում, լուծումը հիմնված է վեբ scraping –ի վրա: Ստորեւ բերված օրինակը ցույց է տալիս մի քանի անկախ աղբյուրներից չափումների հավաքման եւ գնահատման հնարավոր մոտեցումը:

Սուրաբայի ջրամատակարարման ընկերությունը, որը կոչվում է PDAM Surabaya, ունի մի քանի ջրամբարներ, որոնք վերահսկվում են WTW IQ SensorNet 2020 XT-ի կողմից: Սակայն այս սենսորային սարքը կարող է ապահովել ջրի որակի պարամետրերի որոշ արժեքնեի տեղեկատվությունը, բայց սենսոր պասիվ է, եւ ներքին տվյալները դեռ պահվում է սենսորում: Խնդիրը լուծելու համար առաջարկվում է կիրառել տվյալների գրանցիչ՝ վեբ scraping -ի միջոցով ջրի որակի մոնիտորինգի առցանց համակարգի տվյալների հավաքագրման կառավարման համար:

1.2.3 Հասարակական Կարծիքի Հետազոտություն

Ֆիլմերի պրոդյուսերները տվյալներ են հավաքում իրենց ընթացիկ բլոկբաստերների մասին։ Նման տվյալները ներառում օգտատերերի մեկնաբանությունները, եթե դրանք առկա են:

1.2.4Մարդկային ռեսուրսների Գործակալություններ

Մարդկային ռեսուրսների բաժինները (HR) խոշոր ընկերություններում մշակում են իրենց ընկերությունների համար աշխատանքի բազմաթիվ դիմումներ եւ փորձում են գտնել համապատասխան թեքնածուներ:

Միայն ընդունելության դիմումները բավարար չեն թափուր թեկնածուների հաստիքները լրացնելու համար: HR բաժինները նույնպես համագործակցում են 3-րդ կողմի կազմակերպությունների հետ, որոնք կարող են առաջարկել նրանց սեփական մասնագետների հասցեագրքեր։ Հետադարձ կապ-Մայնինգ կարեւոր գործունեություն է նման գործակալությունների:

1.2.5 Սոցիալական ցանց

Վերջին տասնամյակում սոցիալական լրատվամիջոցները (բլոգերը, սոցիալական ցանցերը, միկրո բլոգները) դարձել են խոշոր տվյալների աղբյուր հետազոտությունների համար: Կիրառելով պարզ ծրագրավորման գործիքներ, հետազոտողները կարող են դուրս հանել համապատասխան հաղորդագրությունները սոցիալական լրատվամիջոցների հարթակներից տարբեր հետազոտական նպատակներով:

Սոցիալական ցանցերի տվյալները կարող են ստացվել հանրային աղբյուրներից տարբեր միջոցներով, ինչպիսիք են Scraping -ը, կայքի կողմից տրամադրվող ծրագրերի օգտագործումը եւ շրջանցումը:

1.2.6 Կառավարության ծառայություններ

Հանցագործությունների մոնիտորինգը սոցիալական կայքերում ,կարևորագույն աղբյուր է պետական մարմինների եւ իրավապահ մարմինների համար:

1.2.7 Կորպորատիվ լրտեսություն

Կորպորատիվ կոնտեքստում, վեբ գրառումները թույլ են տալիս ընկերությանը վերանայել ինչպես սեփական, այնպես էլ մրցակիցների հայտնվելը լրատվամիջոցներում: Ընկերությունը կարող է նաեւ հավաքել մանրամասներ մրցակիցների եւ նույնիսկ սեփական աշխատողների մասին:

Ձեռնարկությունների մակարդակում Web Data Extraction- ի մեթոդները առաջանում են որպես բիզնեսի եւ մրցակցային հետախուզության համակարգերում տվյալների վերլուծության, ինչպես նաեւ բիզնեսի վերամշակման համար կարեւոր գործիք:

* 1. Վեբ scraping -ի մեթոդներ

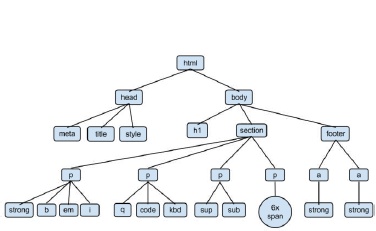
1.3.1 Ջեռքով տվըալների ստացում

Ջեռքով տվյալների ստացումը հնարավոր է, երբ՝

* տվյալների ծավալը շատ փոքր է,
* տվյալները հավաքվում են մեկ անգամ,
* ավտոմատացված համակագրի կարգավորումը ավելի երկար է տևում քան տվյալների հավաքագրումը,
* կայքը պաշտպանված է ավտոմատ համակարգերից կամ հատուկ ձև ունի:

1.3.2HTML վերլուծություն

Կայքերը միշտ չէ, որ տալիս են տվյալները հարմարավետ ձեւաչափերով, ինչպիսիք են .csv կամ .json ֆայլերը: HTML- ի էջերը ստեղծվում են սերվերի կողմից, որպես պատասխան օգտվողի հարցմանը: Այս պահին սերվերի ծրագրային ապահովումը կարևոր չէ, այլ կարեւոր է բրաուզերում արտադրանքը: HTML- ի կառուցվածքի վերլուծությունը (նկար 5-ում ներկայացված պարզ էջի նմուշը) տրամադրված ինտերնետային էջում ցույց կտա կրկնվող տարրերը: Ծրագրային լեզվի սցենարով կամ վեբ գրառման գործիքով, յուրաքանչյուր էջը նմանատիպ օրինակով կարող է օգտագործվել որպես տվյալների աղբյուր:

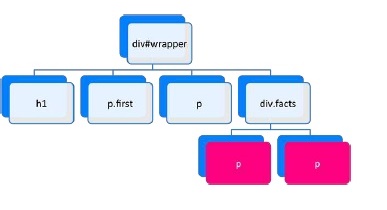


Նկար 5.

1.3.2 DOM վերլուծություն

Փաստաթղթի օբյեկտ մոդելի (DOM) վերլուծությունը HTML- ի վերլուծության էվոլյուցիա է, որը հիմնված է լեզվի եւ բրաուզերների զարգացումների վրա, որոնք հանգեցնում են փաստաթղթի օբյեկտի մոդելի ներդրմանը: DOM- ն մեծապես օգտագործվում է Cascading Stylesheets (CSS) եւ JavaScript- ի համար:

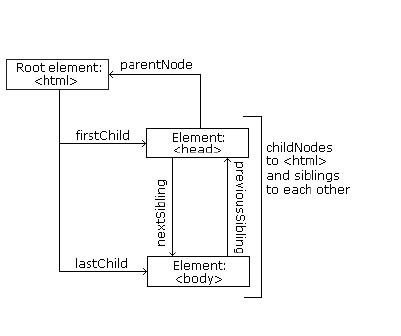
Նկար 6-ը ցույց է տալիս կոնտեյներները սեփական DOM հասցեներով: Դրանք օգտագործվում են վեբ Scraping- ում, վեբ էջի բովանդակության միջոցով ավելի հեշտ տեղաշարժի համար:



Նկար 6.

1.3.3XPath- ը

Հասցեավորման հնարավորությունը նման է այն բանին, որ DOM-ն ապահովում է XPath (XML Path Language): Վերնագիրը առաջարկում XML փաստաթղթերի օգտագործումը։ Սա կիրառելի է նաեւ HTML ձեւաչափով։ XPath պահանջում է ավելի հստակ համակարգված վեբ էջ, քան DOM, եւ ունի նույն հնարավորությունը հասցեագրելով հատվածները վեբ էջի ներսում։ Նկար 7-ը ցույց է տալիս XPath-ի մեկնաբանության փաստաթղթի կառուցվածքը:



Նկար 7.

1.3.4 API

Մինչ նախորդ մեթոդները աշխատում են հավաքել հարմարավետ արդյունքներ, application Programm-ING Interface (API) ակնկալում է դիմումը որպես կապի գործընկեր։ Այսպիսով, API-ները հաճախ կոչվում են մեքենայաընթեռնելի ինտերֆեյսներ (համեմատած մարդակենտրոն): API-ի վերջնական կետ ուղարկված ստանդարտ HTTP հարցումը վերադարձնում է պատասխան սերվերից: Յուրաքանչյուր API ունի իր հստակեցումը եւ պարամետրերը։ Պատասխանի ձեւաչափը կարող է սահմանվել որպես հարցման տարբերակ: API-ի հաղորդակցության ամենատարածված ձեւաչափը JSON-ն է: