استفاده از داده‌های آزاد در تولید آمار رسمی

مطالعه موردی (آمارهای شاخص قیمت)

مجری

رضا هادی‌زاده

همکاران

عباس مرادی

سعید فیاض

سهراب سجادی‌منش

ایوب فرامرزی

آرش فاضلی

خرداد ۱۳۹۹

بسم الله الرحمن الرحیم

کد شناسه:	R-9911
عنوان:	استفاده از داده‌های آزاد در تولید آمار رسمی، مطالعه موردی (آمارهای شاخص قیمت)
مجری:	رضا هادی‌زاده
همکاران:	عباس مرادی، سعید فیاض، سهراب سجادی‌منش، ایوب فرامرزی، آرش فاضلی
گروه پژوهشی	پردازش داده‌ها و اطلاع‌رسانی
تاریخ انتشار:	خرداد ۱۳۹۹
نوبت انتشار:	اول
طراح جلد:	علیرضا رنجبر
صفحه‌آرا:	طاهره امینی، ساناز مهندسی

مرکز آمار ایران پژوهشکده‌ی آمار

تهران، خیابان دکتر فاطمی، خیابان باباطاهر، خیابان سرتیپ فکوری، شماره‌ی ۱۴۵

۰۲۱ ۸۸۶۳۰۴۴۰ -۳

www.srtc.ac.ir

به نام خداوند جان و خرد

پیش‌گفتار

با پیشرفت و تغییرات روزافزون در همه بخش‌های اجتماعی و اقتصادی و فناوری اطلاعات داده‌ها و اطلاعات آماری مختلفی هر لحظه تولید می‌شود. بسیاری از این داده‌ها و اطلاعات آماری به شکل داده‌های با بعد زیاد، مه‌داده‌ها، داده‌های باز و داده‌های ساختارنیافته تولید می‌شوند. از یک سو، این داده‌ها فرصت مناسبی را برای تولید آمارها در اختیار همگان و به‌طور ویژه خبرگان و متخصصین آمار قرار می‌دهد، از سوی دیگر، سازمان‌های آماری با افزایش بار پاسخگویی، بی‌میلی پاسخ گویان و موازی کاری‌ها با مشکلات عدیده‌ایی برای گرداوری داده‌ها مواجه هستند. همچنین محدودیت‌های بودجه، نیروی انسانی متخصص و زمان‌بر بودن اجرای آمارگیری‌ها نیز عواملی هستند که سازمان‌های آماری را به سمت استفاده از داده‌های بهنگام‌تر، در دسترس‌تر و کم‌هزینه‌تر با نیروی انسانی کمتر سوق می‌دهند. کاربران اصلی آمارها نیز به‌دنبال آمارهای بهنگام هستند که در کوتاه‌ترین زمان ممکن در اختیار آنها قرار گیرد. افزایش رو به رشد و تنوع آمارهای مورد نیاز مطرح‌شده توسط کاربران، نیازمند یک سیستم پویا برای گرداوری داده‌های حجیم تولیدشده در فرایندهای اداری و روزمره است. در نظام آماری مدرن روش‌های سنتی از مرحله طراحی، گرداوری، پردازش و انتشار آماری با روش‌های نوین و همگام با تغییرات تکنولوژی بهنگام‌سازی می‌شود. امروزه در روش‌های مدرن طرح‌های آماری، پرسشنامه‌های کاغذی، سرشماری‌ها و ... با منابع داده‌ایی جدید جایگزین خواهند شد. روش‌های نوینی برای گرداوری ریزداده‌ها به منظور طراحی و استقرار پایگاه داده تعریف شده است. با شناسایی و جایگزین کردن منابع جدید داده‌ها و اطلاعات آماری و روش‌های تجزیه و تحلیل نوین می‌توان بسیاری از آمارهای مورد نظر را در راستای طرح‌های مهمی همچون شاخص قیمت، نرخ تورم و آمارهای شاخص قیمت تولیدکننده، آمارهای شاخص قیمت مصرف‌کننده در بخش‌های مهم اقتصادی مانند صنعت، کشاورزی و خدمات تولید کرد. از جمله منابع داده‌ایی جدید می‌توان به منابع زیر اشاره کرد:

۱. داده‌های ثبت‌های اداری [1]
۲. داده‌های آزاد [2]
۳. داده‌های اسکن‌شده [3]
۴. داده‌های تولیدشده در شبکه‌های اجتماعی [4]
۵. داده‌های تولیدشده در سایت‌ها و فروشگاه‌های اینترنتی آنلاین [5]
۶. داده‌های مکانی/ موقعیتی موبایل [6]

---

[1] Administrative Data

[2] Open Data

[3] Scanner Data

[4] Social Network Data

[5] Websites and Online Store Data

[6] Mobile Positioning Data

۷. داده‌های دوربین‌های کنترل ترافیک[7]
۸. و ...

یکی از بزرگترین منابع داده‌ایی که در بالا نیز به آن اشاره شد، داده‌های موجود در سایت‌های تولید و خدماتی که محصولات و خدمات را برای فروش به کاربران ارائه می‌دهند است. داده‌های قابل دسترس یا آزاد در بستر اینترنت فرصت مناسبی برای گرداوری داده‌ها است و بسیاری از مراکز آماری در دنیا از این داده‌ها به‌عنوان منبع جایگزین طرح‌های آمارگیری استفاه می‌کنند. تکنیک واکشی داده‌ها از وب‌گاه‌ها[8] به‌عنوان یکی از روش‌های نوین برای استفاده از داده‌های صفحات اینترنت به منظور ساخت پایگاه داده است که در مطالعه حاضر بررسی و مورد استفاده قرار خواهد گرفت. مراکز آماری در کشورهای مختلف به‌عنوان مرجع اصلی تولید و انتشار آمارهای رسمی در تلاش هستند تا داده‌های موجود در صفحات اینترنت و سایت‌های مختلف را با سرعت، دقت و بهنگام بررسی کنند و پس از طی فرایندهای ادیت و ساخت پایگاه داده در تولید آمارهای رسمی استفاده نمایند. فرایند گرداوری داده‌ها به روش پیشنهادی باید دارای ساختار مناسب و مرتبط با فرایند استاندارد تعریف‌شده از سوی سازمان ملل «**فرایند عمومی کسب وکار آماری**» باشد. در این طرح پژوهشی پس از بررسی ابعاد و تاریخچه‌ی داده‌های باز سایت‌های مرتبط، فرایند گرداوری داده‌ها و ساختار آن بررسی خواهد شد. سپس بر اساس روش‌های پیشنهادی نحوه ساخت و ایجاد پایگاه داده، نحوه بهنگام‌سازی و بروز رسانی آن و نحوه استفاده از این پایگاه داده در تولید آمار رسمی بررسی و تبیین خواهد شد. به منظور درک بهتر از فرایند فوق آمارهای قیمت برای بخش خدمات به‌عنوان مطالعه مورد بررسی قرار گرفت و نتایج حاصل با شیوه‌های سنتی تولید همان نتایج را با نیروی انسانی کمتر و به صورت بازه‌های زمانی کوچکتر در اختیار کاربران و سیاست‌گذاران قرار خواهد داد.

این پژوهش در گروه پردازش داده‌ها و اطلاع‌رسانی پژوهشکده آمار و معاونت اقتصادی مرکز آمار ایران با همکاری آقای رضا هادی‌زاده به عنوان مجری طرح و همکاری آقایان دکتر ایوب فرامرزی، دکتر عباس مرادی، سهراب سجادی‌منش، سعید فیاض و آرش فاضلی به‌عنوان همکاران اصلی طرح پژوهشی محقق شده است که به این وسیله از همه این عزیزان صمیمانه تشکر و قدردانی می‌شود.

معاونت اقتصادی و محاسبات ملی
مرکز آمار ایران

[7] Traffic Controlling Cameras Data

[8] Web Scrapping

فهرست مطالب

فهرست جدول‌ها

فهرست شکل‌ها

۱

کلیات طرح

۱-۱- خلاصه‌ی طرح

داده‌های آماری پیش نیاز لازم برای تولید دانش و نظام آماری بنیادی‌ترین اصل برای نظام برنامه‌ریزی است. برنامه‌ریزی متکی بر آمارهای صحیح و دقیق، یکی از رموز موفقیت و پیشرفت سازمان‌ها است. نقش آمار در عصر حاضر به قدری بدیهی است که وجود نظام آماری مدرن یکی از مهم‌ترین شاخص‌های توسعه‌یافتگی کشورها به شمار می‌رود. از سوی دیگر، با پیشرفت و تغییرات روز افزون در همه بخش‌های اجتماعی و اقتصادی، فناوری اطلاعات داده‌های مختلفی در هر روز، هر ساعت یا هر ثانیه تولید می‌شود. بسیاری از داده‌های فوق به شکل‌های داده‌های بزرگ، مه داده‌ها و داده‌های باز هستند که فرصت مناسبی را برای تولید آمارها در اختیار قرار می‌دهد. سازمان‌های آماری با افزایش بار پاسخگویی، بی‌میلی پاسخ‌گویان و موازی‌کاری‌ها با مشکلات عدیده‌ای برای جمع‌آوری داده‌ها مواجه هستند. محدودیت‌های بودجه، نیروی انسانی متخصص و زمان‌بر بودن اجرای آمارگیری‌ها نیز عواملی هستند که سازمان‌های آماری را به سمت استفاده از داده‌های بهنگام‌تر، در دسترس‌تر و کم هزینه‌تر با نیروی انسانی کمتر سوق می‌دهند. کاربران اصلی آمارها نیز به دنبال آمارهای بهنگام هستند که در کوتاهترین زمان ممکن در اختیار آنها قرار گیرد. افزایش رو به رشد و تنوع آمارهای مورد نیاز مطرح شده توسط کاربران، نیازمند یک سیستم پویا برای جمع‌آوری داده‌های حجیم تولید در فرایندهای اداری و کاری است.

در نظام مدرن آماری روش‌های سنتی از مرحله طراحی، جمع‌آوری، پردازش و انتشار آماری با روش‌های نوین و همگام با تغییرات تکنولوژیکی بهنگام‌سازی می‌شود. در روش‌های مدرن طرح آماری و پرسشنامه‌های کاغذی و حتی سرشماری‌ها با منابع داده‌ای جدید جایگزین می‌شود و داده‌های طرح‌ها با روش‌های نوین کار با داده‌های حجیم و علم پایگاه داده و روش‌های نوین تحلیل داده جایگزین شده است. با جایگزین منابع جدید داده و روش‌های تجزیه و تحلیل نوین می‌توان بسیاری از آمارهای مورد نظر را تولید نمود (از جمله مهم‌ترین آنها نرخ

شاخص قیمت، نرخ تورم و آمارهای شاخص قیمت تولیدکننده/ مصرف‌کننده در بخش‌های مهم اقتصادی مانند صنعت، کشاورزی و خدمات است). از جمله منابع داده‌ای جدید می‌توان به روش‌های زیر اشاره کرد:

- داده‌های ثبت‌های اداری [۹]
- داده‌های آزاد [۱۰]
- داده‌های اسکن شده[۱۱]
- داده‌های تولید شده در شبکه‌های اجتماعی [۱۲]
- داده‌های تولید شده در سایت‌ها و فروشگاه‌های اینترنتی [۱۳] و آنلاین
- داده‌های مکانی/ موقعیتی موبایل [۱۴]
- داده‌های دوربین‌های کنترل ترافیک[۱۵]

و بسیاری از منابع داده‌ای جدید

یکی از بزرگ‌ترین منابع داده‌ای که در بالا نیز به آن اشاره شد، داده‌های موجود در سایت‌های تولید/ خدماتی که محصولات و خدمات را برای فروش به کاربران ارایه می‌دهند. استفاده از داده‌های قابل دسترس یا آزاد در بستر اینترنت فرصت مناسبی برای جمع‌آوری داده‌ها است و بسیاری از مراکز آماری در دنیا از این داده‌ها به‌عنوان منبع جایگزین طرح‌های آمارگیری استفاده می‌کنند. تکنیک واکشی[۱۶] به‌عنوان یکی از روش‌های نوین برای استفاده از داده‌های صفحات اینترنت برای واکشی داده‌ها و ساخت پایگاه داده است که در مطالعه‌ی حاضر بررسی و مورد استفاده قرار خواهد گرفت. مراکز آماری در کشورهای مختلف به‌عنوان مرجع اصلی تولید و انتشار آمارهای رسمی تلاش دارند تا از داده‌های موجود در صفحات اینترنت و سایت‌های مختلف داده‌ها را با سرعت، دقت و بهنگام را بررسی و پس از طی فرایندهای ادیت و ساخت پایگاه داده در تولید آمارهای رسمی استفاده نمایند. فرایند جمع‌آوری داده‌ها به روش پیشنهادی باید دارای ساختار مناسب و مرتبط با فرایند استاندارد تعریف شده از سوی سازمان ملل «فرایند عمومی کسب و کار آماری» باشد. در این مطالعه پس از بررسی ابعاد و تاریخچه از داده‌های باز سایت‌های مرتبط، فرایند جمع‌آوری داده‌ها بررسی و ساختار آن تبیین خواهد شد. سپس بر اساس روش پیشنهادی نحوه ساخت و ایجاد پایگاه داده، نحوه بهنگام‌سازی و به‌روزرسانی آن و نحوه استفاده از این پایگاه داده در تولید آمار رسمی بررسی و تبیین خواهد شد. به‌منظور درک بهتر از فرایند فوق آمارهای قیمت برای بخش خدمات به‌عنوان مطالعه بررسی خواهد شد و نتایج حاصل‌شده با شیوه‌های جاری تولید همان آمارها بررسی و تحلیلی بر روی آن انجام خواهد شد. در پایان نیز، توصیه و پیشنهادی کاربردی برای پیاده‌سازی و استفاده از روش جمع‌آوری داده و ساخت پایگاه داده ارایه خواهد شد.

---

[9] Administrative data

[10] Open data

[11] Scanner data

[12] Social Network data

[13] Websites and Online Store Data

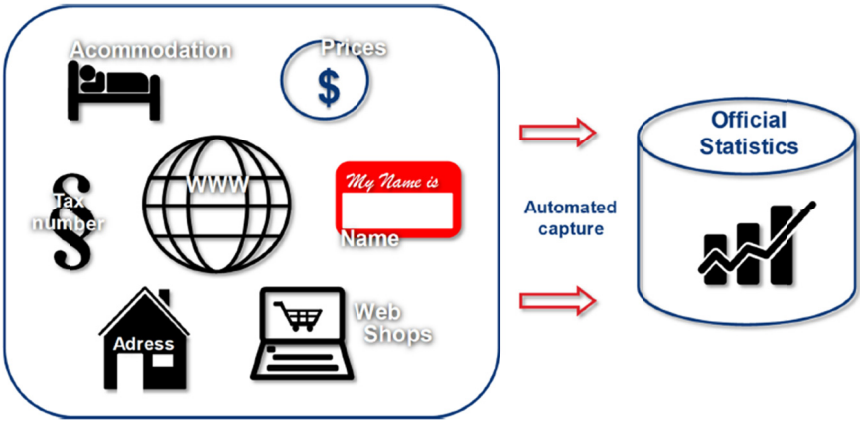
[14] Mobile Positioning Data

[15] Mobile Positioning Data

[16] Web Scrapping

۱-۲-مقدمه

استفاده از منابع داده‌ای جدید، فناوری اطلاعات و روش‌های نوین جمع‌آوری داده یکی از مهم‌ترین رویکردهای مراکز آماری در فرایند مدرن‌سازی نظام آماری است. با توجه به هزینه‌های اجرای طرح‌های آمارگیری و پایین‌آمدن نرخ پاسخگویی، استفاده از منابع داده‌ای جدید نه یک انتخاب بلکه یک ضرورت است. از این رو مراکز آماری در دنیا به سمت استفاده از منابع داده‌ای جدید مانند داده‌های ثبت‌های اداری، داده‌های اسکنر، داده‌های موبایل، مه‌داده‌ها، ترددهای روزانه شهری، داده‌های وب‌سایت‌ها، داده‌های دوربین‌های ثبت‌کننده‌ی تردد، داده‌های آزاد، داده‌های اپلیکیشن‌ها و پلت‌فرم‌های آنلاین، اینترنت اشیاء و بسیاری دیگر از منابع داده که روزانه و لزوماً به خاطر فعالیت‌های اقتصادی و اداری تولید می‌شوند. این منابع داده‌ای باعث می‌شود تا آمارشناسان در مراکز آماری روش‌های نوینی را برای تولید آمارهای رسمی بکار گیرند. ازجمله منابع مهم، داده‌های موجود در سایت‌های اینترنتی است که با تکنیک‌های واکشی در یک پروسه‌ی پیچیده و زمانبر قابل احصا است که نیاز به برنامه‌نویسان حرفه‌ای دارد. اگرچه هر یک از سایت‌های اینترنتی دارای ساختار منحصر به فرد هستند و این موضوع پیچیدگی‌های کار را دوچندان می‌کند. پس از دریافت داده‌ها از سایت‌های آنلاین، این داده‌ها باید در فرایند فنی به پایگاه داده با استانداردهای مورد نیاز برای تولید آمار رسمی منتقل شود. در این مرحله همکاری مشترک و تیمی برنامه‌نویسان و آمارشناسان یک از مهم‌ترین ضرورت‌های فرایند تبدیل داده‌های دریافتی به شاخص‌ها و جداول آماری است.



شکل ۱-۱- واکشی داده‌ها و تولید آمارهای رسمی

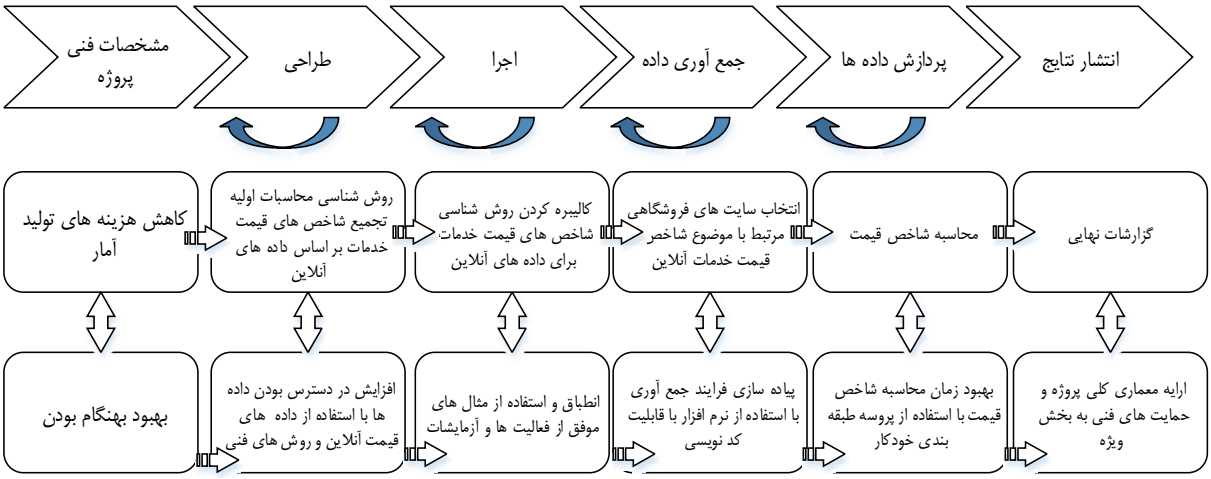
۱-۳- هدف کلّی

در طرح حاضر هدف کلی تشریح فرایند نوین جمع‌آوری داده‌های باز از طریق سایت‌های اینترنتی، چگونگی ساخت پایگاه داده و استفاده از آن در تولید آمار رسمی قیمت است. همچنین مزایای نسبی حاصل از معرفی و پیاده‌سازی روش پیشنهادی با روش‌های جاری برای جمع‌آوری و تولید آمارهای رسمی بخش قیمت مورد بررسی قرار خواهد گرفت.

۱-۴- اهداف تفصیلی طرح

مهم‌ترین اهداف تفضیلی در طرح حاضر عبارت‌اند از:

۱. تبیین فرایند و زیر فرایندها جمع‌آوری داده به روش نوین
۲. تبیین روش‌های نوین جمع‌آوری داده‌ها در مدل عمومی کسب و کار آماری
۳. مزایا و معایب روش‌های نوین جمع‌آوری داده‌ها
۴. روش‌های کاربردی تولید آمارهای رسمی قیمت با استفاده از داده‌های باز آنلاین
۵. برخی از مهم‌ترین کاربردهای داده‌های باز در تولید آمارهای رسمی
۶. یکی از مهم‌ترین اهداف تفضیلی در استفاده روش‌ها نوین جمع‌آوری داده‌های باز آنلاین به‌صورت سیستماتیک و منطبق بر مدل عمومی فرایند کسب و کار آماری GSBPM ۱۷ در مرکز آمار است.



شکل ۱-۲- نمونه‌ای از فرایند پیشنهادی برای جمع‌آوری داده منطبق بر مدل GSBPM

۱-۵- ضرورت

استفاده از داده‌های باز برای مراکز آماری نه تنها یک اولویت انتخاب بلکه یک ضرورت است در ذیل برخی از دلایل این ضرورت ذکر شده است:

۱. نیاز به روشی جهت رفع مشکل کاهش بار پاسخگویی و افزایش عدم همکاری افراد و سازمان‌ها
۲. نیاز به روش جایگزین برای دسترسی سریع و آسان به داده‌ها و اطلاعات جهت تولید آمار رسمی به دلیل هزینه‌بر بودن و زمان‌بر بودن جمع‌آوری داده‌ها در روش‌های سنتی
۳. نیازهای روزافزون کاربران برای آمارهای جزیی‌تر و بهنگام‌تر

---

[17] Generic Statistical Business Process Model (GSBPM)

۴. تقاضای کاربران به انتشار آمارهای در کمترین زمان ممکن
۵. استفاده از سایت‌های و صفحات اینترنتی برای کسب و کار
۶. نیاز به داده‌های با جامعیت و قابلیت اطمینان بالا
۷. تغییر رویکرد سازمان‌های آماری به استفاده بیشتر از داده‌های ثبتی در سیستم نوین آماری
۸. کاهش نیروی انسانی به‌عنوان مامور آمارگیر
۹. کاهش هزینه و زمان مورد نیاز جهت جمع‌آوری داده‌ها
۱۰. جامعیت داده‌ها و امکان دسترسی به محدوده وسیعی از اطلاعات با امکان اعتبار سنجی آنها
۱۱. امکان آنالیز و تحلیل و اعتبارسنجی آنلاین داده‌ها همزمان با جمع‌آوری داده‌ها
۱۲. حذف واسطه‌های نرم‌افزاری و کاغذی و ارسال مستقیم داده‌ها به پایگاه داده
۱۳. امکان جمع‌آوری و تحلیل داده‌ها در بازه‌های زمانی کوتاه ساعتی، روزانه، هفتگی و ماهانه و ...
۱۵. استفاده از تکنولوژی‌های نوین به جای روش‌های سنتی

۱-۶- چالش‌های پیش رو

در فرایند جمع‌آوری داده‌ها در روش نوین پیشنهادی، مشکلات و چالش‌هایی نیز وجود دارد که مهم‌ترین آنها عبارت‌اند از:

۱) محدودیت‌های دسترسی و امنیتی به برخی از صفحات و سایت‌های اینترنتی
۲) مشکل اتصال و واکشی اطلاعات توسط ربات‌ها[18]
۳) تنوع تعداد سایت‌های موجود در یک موضوع خاص (به‌طور مثال در زمینه قیمت محصولات مصرفی: دیجی‌کالا، بامیلو، میله، دیوار و ...)
۴) وجود لایه‌های زیاد در سایت‌ها مورد نظر برای دسترسی به داده‌های مورد نیاز
۵) وجود داده‌های گم‌شده
۶) ساختار متفاوت سایت‌ها و پیچیدگی‌های کدنویسی برای هر ساختار
۷) فرمت‌های مختلف و عمدتاً ناسازگار در داده‌های سایت‌ها
۸) افزایش حجم داده‌های موجود در سایت‌های مختلف با گذشت زمان و مشکلات ذخیره‌سازی و ساخت نسخه‌ی پشتیبان
۹) تنوع و تعداد متغیرهای موجود در سایت‌ها و پیچیدگی‌های زیاد برای انتخاب بهترین متغیرها
۱۰) پیچیدگی و تخصصی بودن هدف واکشی جهت تولید آمارهای مورد نظر
۱۱) محدودیت‌های پایگاه داده
۱۲) کمبود نیروی متخصص و فرایند کاری جدید برای نهادینه‌سازی فرایند پیشنهادی

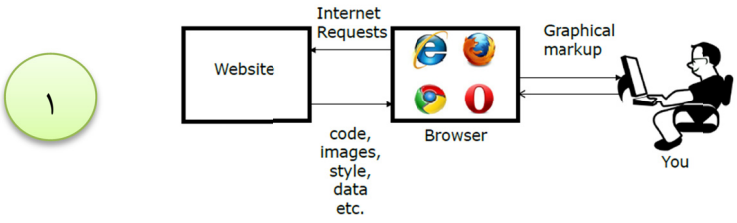
[18] Robot

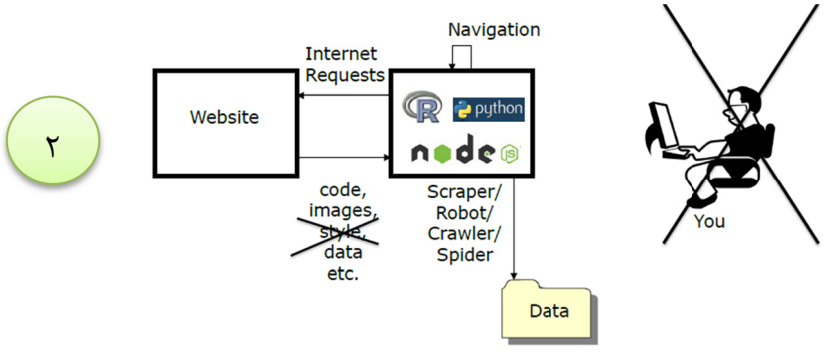
۲

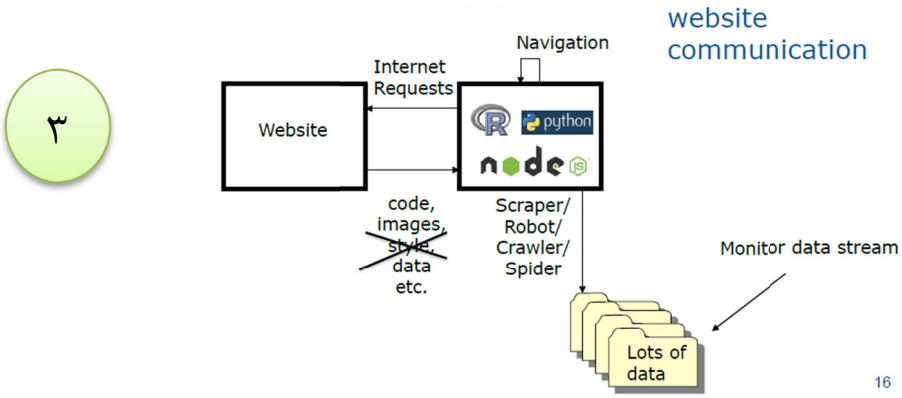
روش واکشی داده‌ها و تاریخچه‌ی آن

۲-۱- مقدمه

استفاده از داده‌های موجود در صفحات وب، منبع مهمی از داده‌ها را فراهم نموده است. استفاده از این منابع داده‌ای که اغلب به‌صورت رایگان و آزاد در اختیار کاربران قرار دارد می‌تواند به‌عنوان منبع مهمی برای تولید محصولات آماری مورد استفاده قرار گیرد. از جمله این داده‌ها موتورهای جستجو در اینترنت هستند که روزانه میلیون‌ها کاربر به آنها مراجعه کرده و به دنبال پیدا کردن صفحه مورد نظر و دریافت اطلاعات مورد نظر است. در روش واکشی ربات‌های برنامه‌نویسی شده به جای کاربر فعالیت جستجوی داده‌ها را دنبال می‌کنند و از داده‌های جمع‌آوری شده اطلاعات مورد نظر خود را استخراج می‌نمایند.







شکل ۲-۱- استفاده از واکشی در موتورهای جستجو در اینترنت

۲-۲- مروری بر تاریخچه

استفاده از داده‌های باز و به ویژه صفحات وب و سایت‌های رسمی و غیر رسمی به‌طور گسترده‌ای در سال‌های اخیر مورد توجه قرار گرفته است. بیساری از مراکز آماری کشورها همواره به دنبال جایگزین منابع داده جدید و مدرن‌سازی فرایندهای جمع‌آوری، تولید و انتشار آمار منطبق با استانداردهای بین المللی بخش آمار سازمان ملل و دیگر نهادهای تخصصی آماری دنیا هستند. در سال ۲۰۱۳ در کنفرانس DGINS کشور هلند و کمیته سیستم آمار اروپا ESSC برای استفاده از مه داده‌ها و استفاده از تولید آمار رسمی و استراتژی‌ها آن توافقاتی را انجام دادند[۱]. در این راستا، دیگر اعضا سند کوتاه مدت و میان مدت و بلند مدت برنامه استفاده از مه داده‌ها و نقشه راه آن، مراحل ضروری و کارهای مورد نیاز برای پیاده‌سازی آن را مورد تصویب قرار دادند [۲] و کمیته مطالعاتی سیستم آمار اروپا ماموریت یافت تا به‌صورت آزمایشی این روش را در برخی از کشورها به‌طور نمونه انجام دهد [۳] تا نیازمندی‌ها و اقدامات لازم جهت پیاده‌سازی این فرایند را بررسی نماید [۴]. در سال ۲۰۱۷، اداره آمار رومانی اقدام به بررسی و پیاده‌سازی روش‌های نوین جمع‌آوری داده‌های موجود در صفحات وب بر اساس توصیه‌ها نمود و نتایج این پروژه را در کنفرانس DGINS سال ۲۰۱۸ در بوداپست ارایه نمود [۵]. بر اساس نتایج حاصله نماینده اداره آمار رومانی [۶] اعلام نمود که منابع داده‌ای جدید، فناوری‌های نوین فناوری اطلاعات و پتانسیل کارشناسی به‌عنوان هسته‌های مرکزی برای تایید آمارهای هوشمند به‌دست آمده از این روش هستند [۷و۸] و اداره آمار اروپا و کمیته سیستم آماری اروپا در سطح اتحادیه اروپا استفاده از منابع داده‌ای جدید و مه داده‌ها برای تولید آمارهای رسمی را به‌طور مشترک پیاده‌سازی نمایند.

در واقع منابع داده‌ای جدید و مه داده‌ها به‌طور کامل جایگزین روش‌های گذشته نخواهد بود (در بسیاری از موارد جایگزینی بسیار دشوار است) و قرار است به‌عنوان بخش‌هایی از فرایندهای گذشته جایگزین گردد [۹و۱۰]. استفاده از مه‌داده‌ها و داده‌های آزاد می‌تواند باعث کاهش بار پاسخگویی موجود در روش‌های گذشته شده و می‌تواند قبل از اینکه حتی طرح آمارگیری با هزینه‌های بالا مورد پیاده‌سازی و ارزیابی قرار گیرد، مورد استفاده قرار گیرد [۱۱].

یکی از مهم‌ترین منابع داده‌ای شبکه گسترده جهانی وب[19] یا اقیانوسی از داده‌ها و اطلاعات است که مراکز آماری نمی‌توانند از این منبع عظیم چشم پوشی نمایند. به‌منظور استفاده از داده‌های منتشر شده در صفحات وب و سایت‌های اینترنتی به‌صورت خودکار و تشکیل پایگاه داده‌های مورد نیاز در تولید آمارهای رسمی باید فرایندهای را تحت عنوان واکشی [20]یا جمع‌آوری اتوماتیک داده‌ها[21] طراحی نمود. اولین بار دانشگاه ام آی تی از داده‌های باز در سایت‌های اینترنتی برای تولید شاخص‌های قیمت محصولات استفاده نمود [۱۳] که در آن از داده‌های جمع‌آوری شده در برخی از کشورهای آفریقای جنوبی برای تولید شاخص CPI استفاده شد. بعد از آن بسیاری از ادارت آمار کشورها به دنبال داده‌های آزاد موجود آنلاین خرده‌فروشی برای تولید آمارها و شاخص‌های قیمت رفتند مانند اداره آمار هلند [۱۴] اداره آمار ایتالیا [۱۵] اداره آمار آلمان [۱۶] از جمله پیشروان این روش در حوزه کشورهای اروپایی بودند هر چند که این ادارات آماری از روش‌های پیشنهادی MIT به‌صورت کامل استفاده نکردند و تنها به دنبال قیمت کالاهایی بودند که در روش‌های سنتی نیز موردنظر فرایند جمع‌آوری داده‌ها بودند.

استفاده از داده‌های باز آنلاین تنها به موضوع داده‌های قیمت محدود نمی‌شود و در برخی از موارد به‌منظور بهبود ثبت‌های اداری [۱۷]، ارزیابی مکان‌ها یا موقعیت‌های شغلی [۱۸] هم مورد استفاده قرار می‌گیرد. صرف نظر از اینکه داده‌های حجیمی از قیمت مورد بررسی قرار گیرد یا تنها برای محصولات خاص و محدودی مورد استفاده قرار گیرد، واکشی یکی از روش‌های بسیار مفید در اختیار آمارشناسان است [۱۹]. در پروژه‌های انجام گرفته هدف پیاده‌سازی روش‌های نوین جمع‌آوری داده‌ها و بررسی اثرات آن بر بار پاسخگویی، هزینه‌های انجام و آزمایش فناوری‌ها و روش‌های جدید بر اساس یک مدل ساختارمند سیستماتیک بوده است [۲۰].

به‌منظور سیتماتیک کردن فرایند نوین جمع‌آوری داده‌های باز آنلاین به روش واکشی الگوریتم‌ها و ساختارهای مناسب و مرتبطی توسعه و تهیه شده است. نمونه‌ای از الگوریتم و ساختار ارایه شده در اداره آمار هلند به شرح زیر است.

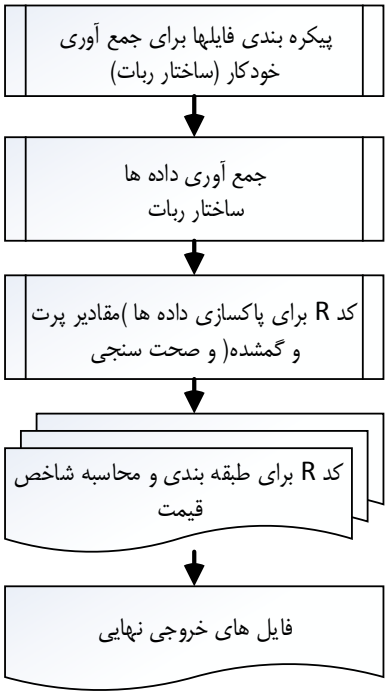
```
Read URL's
Detect cores
Allocate Cluster
for each URL do
    Start node
    procedure RECURSION(URL)
        if URL is dead then
            return
        else
            Visit URL
            text ← get_html[URL]
            URL ← get_url[text]
            Data ← get_content[text]
            return RECURSION(URL)
        end if
        Write Data
    end procedure
    Stop node
end for
```

[19] World Wide Web (WWW)

[20] web scrapping

[21] Automated data collection

همچنین فرایند مذکور دارای گام‌های پیشنهادی است که متناسب با ساختار موجود در اداره آمار هلند طراحی شده است:



شکل ۲-۲- فرایند پیشنهادی جمع‌آورین نوین داده‌های آزاد آنلاین به روش واکشی

در کشور آلمان ۷۲% از شرکت‌هایی که دارای فعالیت‌های اقتصادی هستند دارای وب‌سایت هستند، ۴۶% در رسانه‌های جمعی فعال هستند و ۲۳% از طریق اینترنت کالاها و خدمات خود را به مشتریان عرضه می‌کنند[۲۲]. بسیاری از داده‌های شرکت‌های تولیدی مانند قیمت کالا و خدمات، لیست محصولات/ خدمات و ... از طریق پرسشنامه و مصاحبه تکمیل می‌شوند در حالی که این داده‌ها در وب‌سایت شرکت‌ها برای عموم قابل دسترس است. از آنجا که جمع‌آوری این داده‌ها برای شرکت ضروری است، این داده‌ها می‌تواند از روش واکشی جایگزین شود.

به‌طور معمول داده‌های مربوط به قیمت کالاها و خدمات در بخش فروش و بازرگانی سایت شرکت قابل مشاهده است. در این دسته از شرکت‌ها معمولاً بخش‌هایی با عنوان سیستم سفارش و پرداخت آنلاین، فروشگاه‌های آنلاین، پرتال ثبت سفارشات آنلاین و بخش خدمت‌رسانی وجود دارد. برای بنگاه‌هایی که خدمات خود را به مشتریان عرضه می‌کنند، همچنین قیمت و نوع خدمت قابل ارایه در وب‌سایت برای انتخاب مشتریان ارایه می‌شود (مانند قیمت تعمیرات). در هر نوع از شرکت‌ها/ بنگاه‌ها داده‌های قیمت به‌همراه محصولات/ خدمات در سایت ارایه می‌شود.

داده‌های دیجیتالی می‌تواند دارای ساختارهای متنوع باشد و گستره‌ی وسیعی از موضوعات را شامل شود. این داده‌ها برای اتصال به داده‌های رسمی باید دارای مشخصاتی چون نام، آدرس و شماره مالیاتی، قیمت

[22] See Statistisches Bundesamt, 2017

خرده‌فروشی یا تعداد اتاق‌های موجود در هتل و یا موجود بودن یا نبودن کالاها در فروشگاه‌های آنلاین باشد. با وجود این نوع از داده‌های دیجیتال، باید یک فرایند خودکار تعریف شود تا داده‌ها از طریق این فرایند جمع‌آوری و در اختیار آمارشناسان قرار گیرد. دریافت خودکار داده‌ها به روش‌های مختلفی می‌تواند منجر به بهبود آمارهای رسمی شود:

۱) پشتیبانی از طرح‌های آمارگیری: در روش‌های سنتی، داده‌های توسط مصاحبه رو در رو در طرح‌های آمارگیری و با هزینه قابل توجه جمع‌آوری می‌شود در حالی که با وجود این داده‌ها در اینترنت می‌توان با استفاده از روش‌های واکشی و در یک فرایند خودکار این داده‌ها را جمع‌آوری نمود.
۲) پشتیبانی از داده‌های مرکزی : با توجه به اینکه داده‌های جمع‌آوری شده توسط روش واکشی از لحاظ آماری قابل کمی سازی هستند، این داده‌ها می‌تواند با توجه به اهداف صحت‌سنجی مورد استفاده قرار گیرد.
۳) سرعت بخشی به ارایه داده‌ها و دقت آنها : بر اساس ویژگی‌های منحصر بفرد روش واکشی، دوره‌های ماهانه (هفتگی/ روزانه) می‌تواند به‌عنوان دوره مورد نظر در ارایه داده‌ها باشد. سرعت دریافت اطلاعات در مقایسه با روش سنتی بسیار بالا بوده و می‌توان در بازه زمانی کوتاه‌تر اطلاعات بیشتری را استخراج نمود. دقت این داده‌ها با توجه به شفاف بودن اطلاعات در وب‌سایت‌ها و اعتبار وب‌سایت‌ها بالاتر است.
۴) ارایه محتویات بیشتر: با توجه به اینکه روش واکشی داده‌های آزاد موجود در وب‌سایت شرکت‌ها را دریافت می‌کند، داده‌هایی وجود دارند که در آمارگیری‌های سنتی وجود ندارد. به‌طور مثال موارد مانند استانداردهای امنیتی وبسایت شرکت‌ها، فعالیت‌های تجارت الکترونیکی شرکت‌ها، سرمایه گذاری‌ها در فناوری‌های پایدار را می‌تواند اشاره نمود.
۵) کاهش بار پاسخگویی: استفاده از داده‌هایی که در دسترس عموم قرار دارند، باعث کاهش مراجعات مستقیم به پاسخگویان شده و مواردی مانند موجب کاهش مخاطرات افشای داده‌ها و حفظ محرمانگی آنها می‌شود.

به‌طور خلاصه، دریافت داده‌ها به‌صورت خودکار از وب‌سایت‌های اینترنتی در کمترین زمان، کمترین هزینه، کمترین طرح آمارگیری و تقسیم به زیر بخش‌های متنوع‌تر است. بر اساس تجربیات موجود در اداره ملی آمار آلمان جایگزینی این روش با روش‌های سنتی به سرعت انجام‌پذیر نیست. در این فرایند موتورهای جستجوی هوشمند بسیار مورد توجه واقع شده‌اند. این موتورهای جستجو درخواست‌های مورد نظر را در بین تعداد زیادی از موتورهای جستجو به‌صورت همزمان انجام می‌دهد. اخیراً موتورهای جستجوی از روش‌های واکشی استفاده می‌کنند.

۲-۲-۱- واکشی در سطح اروپا و بین‌المللی

با گسترش ارتباطات و کسب و کارهای آنلاین، از روش واکشی به‌منظور جمع‌آوری داده‌های مورد نیاز برای تولید آمارهای رسمی استفاده می‌شود. مطالعه‌ای که اداره ملی آمار آلمان در خصوص امکان‌سنجی استفاده از