4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

ASurvey of Threats and Solution to Privacy in Web Search Engines

Hadi Sabouhi[1], Maryam Fathi Ahmad Sharaei[2]

1 Post Doc. and Assistant Professor of Islamic Azad University, Karaj Branch, Faculty of Mechatronics, Computer Department, Alborz, Iran

saboohi@kiau.ac.ir

2 PhD Student of Islamic Azad University, Karaj Branch, Faculty of Mechatronics, Computer Department, Alborz, Iran

fathimaryam2009@gmail.com

Abstract

We use search engines to find our needed information though amount of data. Most of the time people use Internet to get information about every job. Web search engines often confuse people by presenting different results for same queries. Search engines create profiles for users to solve this problem. According to each user's profile, search engines represent results. Often, users' privacy are in endanger by using these tools. In web search engines results, we need to have quality and privacy protecting both together. In this article we present problems that threaten the users' privacy. We also reviewe how to prevent web search engines from violating the users' privacy.

key words: Privacy, personalization, users' profile in search engines.

4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین ۴
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

بررسی تهدیدها و راه حل‌های حفظ حریم خصوصی کاربران در موتورهای جستجوی وب

هادی صبوحی[۱]، مریم فتحی احمدسرائی[۲]

۱ فوق دکتری کامپیوتر، استادیار دانشگاه آزاد اسلامی، واحد کرج، دانشکده مکاترونیک، گروه کامپیوتر، البرز، ایران
saboohi@kiau.ac.ir

۲ دانشجوی دکتری کامپیوتر، دانشگاه آزاد اسلامی، واحد کرج، دانشکده مکاترونیک، گروه کامپیوتر، البرز، ایران
fathimaryam2009@gmail.com

چکیده

موتورهای جستجوی وب برای پیدا کردن اطلاعات خاصی در میان حجم عظیمی از داده‌ها در زمان کم مورد استفاده قرار می‌گیرند. افراد در هر کاری تقریباً به اینترنت متکی هستند و موتورهای جستجوی وب گاهی اوقات با ارائه نتایج مختلف ممکن است مردم را سردرگم کنند. برای حل این مشکل از ساخت پروفایل استفاده می‌شود. نتایج با توجه به پروفایل هر کاربر نمایش داده می‌شود. این ابزارها معمولاً باعث به خطر افتادن حریم خصوصی کاربران می‌شود. محافظت از حریم خصوصی و داشتن کیفیت در نتایج جستجو را باید بتوانیم در کنار هم داشته باشیم. در این مقاله به مشکلاتی که حریم خصوصی کاربران را تهدید می‌کند، اشاره می‌شود. همچنین برای جلوگیری از نقض حریم خصوصی کاربران توسط موتورهای جستجوی وب راه حل‌های ارائه شده اخیر بررسی خواهد شد.

کلمات کلیدی

حریم خصوصی، شخصی سازی، پروفایل کاربران موتورهای جستجو

۱- مقدمه

موتورهای جستجوی وب، کار دشوار استخراج اطلاعات مورد نیاز از حجم انبوهی از داده‌ها را انجام می‌دهند [۱]. بعد از ایمیل، موتورهای جستجوی وب دومین ابزاری هستند که کاربران اینترنت از آن استفاده می‌کنند [۲]. موتورهای جستجوی وب در کم‌ترین زمان، اطلاعات را بازیابی می‌کنند. استفاده از موتورهای جستجوی وب باعث تولید گزارش[۱] از پرس‌وجوهای[۲] کاربران می‌شود [۳]. اگر به گزارش‌ها و تاریخچه جستجوی کاربران توجهی نشود، ممکن است نتایج نادرستی بازگردانده شود [۴و۵]. شخصی‌سازی[۳] جستجوی وب یکی از معمول‌ترین روش‌ها، برای جمع‌آوری و تحلیل اهداف کاربر است. شخصی سازی جستجوی وب دو نوع است [۴]: الف) روش‌های مبتنی بر گزارش کلیک‌های کاربران: که از تاریخچه صفحه‌هایی که کاربر کلیک کرده است، استفاده می کنند. ب) روش‌های مبتنی بر پروفایل کاربران: که با تولید مدل‌های علایق کاربران می‌توان جستجوی بهتری را تجربه کرد. با استفاده از اطلاعات رفتاری کاربر که در پروفایل او وجود دارد می‌توان کیفیت نتایج ارائه شده را بهبود بخشید. این روش‌ها موجب تجاوز به حریم خصوصی کاربران می‌شوند. حمله کننده نیز می‌تواند از این اطلاعات که بین کاربر و موتور جستجوی وب رد و بدل می‌شود، به دلیل فقدان سیاست‌های حریم‌خصوصی سوء استفاده کنند [۱].

در ادامه در بخش ۲ به بررسی تهدیدهای حریم خصوصی می‌پردازیم و در بخش ۳ راه حل‌های حفظ حریم خصوصی را مورد بررسی قرار می‌دهیم. در آخر در بخش ۴، نتیجه گیری و کارهای آینده را ارائه می‌دهیم.

۲ - بررسی تهدیدهای حفظ حریم خصوصی کاربران

در بحث حریم خصوصی، باید بررسی شود که موتورهای جستجو چقدر قابل نفوذ هستند و می‌توانند حریم خصوصی کاربران را تهدید کنند [۲]. زمانی که کاربر برای اولین بار به موتور جستجو متصل می‌شود، ابتدا IP، زمان، تاریخ و

تنظیمات مرورگر کاربر گزارش می‌شود. طرفداران حریم خصوصی ادعا می‌کنند که از جستجوهای یک فرد در یک مدت زمان می‌توان به زندگی شخصی و علایق آن فرد رسید که این مسئله خود، نقض حریم خصوصی است. موتورهای جستجو از کوکی‌ها استفاده می‌کنند [۲]. کوکی‌ها می‌توانند حریم خصوصی کاربران را در معرض خطر قرار دهند. موتورهای جستجوی وب می‌توانند این اطلاعات را رمزنگاری کنند و با این کار می‌توانند هر اطلاعات دیگری را از کاربران ذخیره کنند. کاربران باید توافقنامه‌های موتورهای جستجوی وب را قبول کنند. در این توافقنامه‌ها، حق مدیریت و فروش اطلاعات به شرکت های دیگر ذکر شده است. اینکه چه اطلاعاتی از کاربران آنلاین است و برای شاخص گذاری در اختیار موتورهای جستجو قرار می‌گیرد به میزان مسئولیت‌پذیری مدیر وب سایت بستگی دارد [۲]. یک راه ساده استفاده از پراکسی است. به این ترتیب موتورهای جستجو نمی‌توانند IP کاربران را به‌دست آورند [۷].

۳- بررسی راه حل‌های ارائه شده جهت برقراری حفظ حریم خصوصی کاربران

روش‌های حفظ حریم خصوصی را می‌توان از چند جهت دسته بندی کرد. در این مقاله ما دو روش دسته بندی را بررسی می‌کنیم. یک روش دسته بندی روش‌ها را به دو دسته پردازش دسته‌ای و پردازش زمان واقعی تقسیم بندی می کند. در روش دسته بندی دیگر، روش‌ها به دو دسته رویکردهای سمت کاربر و سمت سرور دسته بندی می‌شوند.

۳-۱- دسته بندی اول روش‌های حفظ حریم خصوصی

در دسته بندی اول، روش‌های حفظ حریم خصوصی عبارتند از [۳]:

۳-۱-۱- پردازش دسته‌ای[۴]

الف) حذف پرس‌وجو[۵]: اطلاعات شناسه‌ای مانند شماره ملی یا شماره بیمه و غیره را حذف کند. نمونه‌هایی از این رویکرد: در [۸]، مجموعه پرس‌وجوی قدیمی را حذف می‌کند، در حالی که فرض می‌کند گزارش پرس‌وجو به اندازه کافی بزرگ نخواهند بود که قادر به افشای هویت شود. در [۹] پرس‌وجوهای نامعمول حذف می‌شوند و فرض می‌کند این پرس‌وجوها به احتمال بیشتر به اطلاعات شناسه‌ای اشاره می‌کنند. انتخاب حدآستانه حذف مناسب می‌تواند کاملا چالش برانگیز باشد. به‌عنوان یک نتیجه، یک روش ممکن است موجب حذف مقدار زیادی از پرس‌وجوهای غیرشناساگر ولی مفید شود. سایر روش‌های مبتنی بر حذف به حذف داده‌های شناسه‌ای مرتبط با پرس‌وجوها و اطلاعات خصوصی مانند شماره‌های تامین اجتماعی، کارت‌های اعتباری، آدرس‌ها و غیره تمرکز دارند. رویکردهای پیچیده‌تر در این زمینه بر حذف آن پرس‌وجوهایی تمرکز دارند که با کلیک کردن کاربر بر روی URLها معمولا خاتمه می‌یابند با این فرض که آن پرس‌وجوها را وابسته در نظر می‌گیرند، یعنی شبه‌شناسه‌ای [۱۰].

ب) کنترل افشای آماری[۶]: پرس‌وجوی کاربران با یکدیگر گروه‌بندی و با یک پرس‌وجوی نماینده جایگزین می‌شود. نمونه‌هایی از این رویکرد در ادامه بحث شده است: به دلیل حجم بالای گزارش‌های پرس‌وجوها، در [۱۱] پیشنهاد داده‌اند تا مجموعه پرس‌وجوهای یک کاربر را به‌وسیله عمومیت بخشیدن پرس‌وجوهای او ناشناس سازی کنند. مشکل این رویکرد این است که یک پرس‌وجو می‌تواند در فرهنگ لغت عمومی بی‌معنی باشد، ولی در یک فرهنگ لغت تخصصی‌تر خطرناک باشد.

۳-۱-۲- پردازش زمان واقعی[۷]:

روش‌های بررسی شده تنها برای داده‌های ایستا[۸] بود. این روش‌ها برای جریان داده‌های[۹] آنلاین مناسب است و دو زیر روش دارد:

الف) ناشناس کردن همه پرس‌وجوها از لحظه صفر تاکنون که با مقدار زیادی داده روبرو هستیم. نمونه‌هایی از این رویکرد: استفاده از تکنیک‌های معروف SDC در این رویکرد به هنگام پارتیشن کردن پایگاه داده، باعث تحمیل هزینه‌های زیادی می‌شود [۱۲]. به‌صورت دقیق‌تر، مولفین ناشناس‌سازی گزارش‌های پرس و جو را اعمال کردند و برداشت آنها این بود که این تکنیک‌ها با مجموعه‌های بزرگ داده‌ها، که نیازمند آپدیت پیوسته و ناشناس‌سازی داده‌های جدید است که همواره افزوده می‌شوند، تناسب خوبی ندارند.

ب) ناشناس کردن پرس‌وجوها در یک زمان محدود که در دوره‌های سه یا شش ماهه انجام می‌شود. این روش؛ هزینه، دقت و حجم مجموعه داده را کاهش می‌دهد. نمونه‌هایی از این رویکرد: این رویکردها، داده‌ها را به‌عنوان یک جریان در نظر می‌گیرند و اطلاعات را به محض دریافت و با ارائه حداقل تاخیر، پردازش می‌کنند [۱۳]. جریان‌های داده می‌توانند نامتناهی باشند و حجم عظیمی از داده‌ها را تولید نمایند. بنابراین، بهره‌برداری از جریان‌های عظیم داده، یک کار چالش برانگیز است. اکثر رویکردهای موجود می‌خواهند تاخیر بین پردازش جریان‌های داده را کاهش دهند. این روش‌ها عمدتا بر مبنای خوشه بندی جریان‌های داده ورودی هستند و همگی نیاز دارند تا جهت ایجاد خوشه‌های ناشناس منتظر داده‌های جدید باشند.

۳-۲- دسته بندی دوم روش‌های حفظ حریم خصوصی

رویکردهای حفظ حریم خصوصی را به دو گروه کلی تقسیم می‌کنیم: الف) رویکردهای سمت کاربر، ب) رویکردهای سمت سرور.

۳-۲-۱- رویکردهای سمت کاربر

رویکردهای سمت کاربر که کاربر به‌طور مستقیم بر روی پروفایل خود کنترل دارد.

روش اول از رویکردهای سمت کاربر

در این سیستم هر کاربر یک حساب کاربری ایجاد می‌کند و اطلاعاتی مانند آدرس، شغل، و غیره را ثبت می‌کند. سپس شروع به جستجو می‌کند و پروفایل کاربر در سرور ذخیره می‌شود. اگر کاربر بر روی خصوصی[۱۰] کلیک کند، لینک‌ها تنها برای خود کاربر قابل دسترس هستند. اگر کاربر بر روی گزینه عمومی[۱۱] کلیک کند، پروفایل او برای همه قابل دسترس است. جریان کاری این روش در شکل ۱ نمایش داده شده است. در این روش از الگوریتم های Spy Naïve Bayes استفاده می‌شود [۴].

روش دوم از رویکردهای سمت کاربر

تاریخچه مرورگر و ایمیل‌های هر فرد می‌تواند منبع اطلاعاتی خوبی برای پروفایل کاربر باشد. فرض ما این است که اگر عبارتی در این اسناد[۱۲] تکرار شود جزء علایق کاربر است. در این رویکرد پروفایل کاربر را به‌صورت سلسله مراتبی بر مبنای تکرار عبارت‌ها ایجاد می‌کنیم. در این سلسله مراتب عبارات با تکرار بیشتر در سطوح بالاتر قرار می‌گیرند. $D(t)$ شامل همه اسنادی است که عبارت t در آنها وجود دارد. برای سازماندهی سلسله مراتبی عبارات از دو قانون هیوریستیک زیر استفاده می‌کنیم: *الف) عبارت مشابه:* دو عبارتی که مجموعه اسناد یکسانی را با همپوشانی[۱۳] زیادی پوشش می‌دهند، معمولاً علایق یکسانی را نشان می‌دهند. *ب) عبارات والد و فرزندی*[۱۴]*:* معمولاً عبارات خاص با عبارات کلی ظاهر می‌شوند ولی عکس این قضیه درست نیست. بنابراین t_B فرزند t_A است اگر احتمال شرطی $p(t_A \mid t_B) > \delta$ اتفاق بیفتد.

الگوریتم بررسی شده را به‌صورت شبه کد در الگوریتم ۱ مشاهده می‌کنیم.

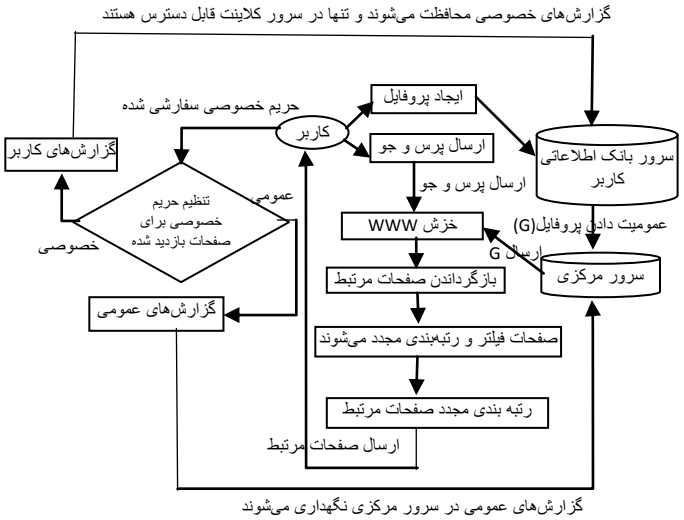
پروفایل کاربر کاملا تحت کنترل کاربر است و می‌تواند عبارت‌هایی را که تمایل دارد در ساختار سلسله مراتبی پنهان کند. کاربر دقیقا می‌داند که کدام بخش از پروفایل او محافظت شده است. برای هر پروفایل یک لیست از زوج مرتب‌های (t,w_t) داریم. به‌طوری‌که t عبارت و w_t وزن عبارت است.

۳-۲-۲- رویکردهای سمت سرور

در این رویکردها کاربر به‌طور مستقیم بر محافظت از حریم خصوصی خود دخالت ندارد بلکه به واسطه یک سرور میانی این کار انجام می‌شود.

روش اول از رویکردهای سمت سرور

پروتکل UUP از دریافت پروفایل معتبر کاربران توسط موتورهای جستجو جلوگیری می‌کند. منظور از U کاربران، C نود مرکزی که الگوریتم UUP را اجرا می‌کند و W همان موتورهای جستجو هستند.



شکل ۱ – نمایش جریان کار سیستمی از رویکردهای سمت کاربر

الگوریتم: Split (n, S(t), minsup, δ)

ورودی: به یک گره n برچسب عبارت t داده می‌شود، اسناد پشتیبانی S(t)، حدآستانه minsup و δ

۱. لیستی از کلمات پرتکرار $\{t_i\}$ تولید می‌شود به‌طوری‌که $D(ti) \geq$ minsup و بر اساس تعداد تکرار به صورت نزولی مرتب شده است.
۲. برای هر عبارت t_i:
۳. اگر $sim(t_i, t_k)>\delta$ و $k<i$
۴. برچسب گره t_i/t_k تنظیم می‌شود و $S(t_i/t_k)=S(t_k) \cup D(t_i)$
۵. در غیراینصورت اگر $P(t_k|t_i)>\delta$ و $k<i$
۶. برچسب گره همان t_k می‌ماند و $S(t_k)=S(t_k) \cup D(t_i)$
۷. در غیراینصورت
۸. یک گره جدید با برچسب t_i ایجاد می‌شود و $S(t_i)=D(t_i)$
۹. $Sup(t_i)$ برای هر گره با برچسب t_i محاسبه می‌شود و کلمات به صورت نزولی مرتب می‌شوند

الگوریتم: BuildUP(n, D, minssup, δ)

ورودی: گره n، استاد پشتیبانی D، حدآستانه minsup و δ

خروجی: پروفایل یک کاربر U

۱. Split (n, D, minsup, δ)
۲. برای هر فرزند c_i با برچسب t_i از گره n:
۳. BuilUP (c_i, $S(t_i)$, minsup, δ)

الگوریتم ۱- الگوریتم Split کردن مجموعه‌ای از داکیومنت‌ها

ایده اصلی این روش به این صورت است: هر کاربر که می‌خواهد پرس‌وجویی را ارسال کند، پرس‌وجوی خود را ارسال نمی‌کند بلکه پرس‌وجوی کاربر دیگر را ارسال می‌کند و پرس‌وجوی او نیز توسط فرد دیگری ارسال می‌شود. کاربران از پرس‌وجوهای یکدیگر اطلاعی ندارند زیرا پرس‌وجوها رمزنگاری می‌شوند. بنابراین موتورهای جستجو نمی‌توانند برای کاربران، پروفایل‌های واقعی ایجاد کنند. نود مرکزی، n کاربر را گروه‌بندی می‌کند. برای هر کاربر U_i به‌طوری‌که $i\in\{1,...,N\}$ یکی از پرس‌وجوهای n-1 کاربر دیگر را ارسال می‌کند. U_i از مبدا پرس‌وجو دریافتی، اطلاعی ندارد. برای رسیدن به این هدف ابتدا همه پرس وجوها بین کاربران به طور تصادفی توزیع می‌شود. این کار با استفاده از عملیات رمزنگاری، بازسازی و جایگشت انجام می‌شود. سپس پرس‌وجوی دریافتی توسط کاربران ارسال می‌شود. موتور جستجو، نتیجه پرس‌وجوها را به کاربران به‌صورت همه پخشی[۱۵] ارسال می‌کند. هر کاربر نتایج مربوط به خود را برمی‌دارد و دیگر نتایج را دور می‌ریزد. در این پروتکل فرض شده است که همه کاربران از پروتکل پیروی می‌کنند و هیچ تبانی بین دو کاربر وجود ندارد. این پیش‌فرض‌ها به‌دلیل رسیدن به این هدف است که موتور جستجو نتوانند برای کاربران پروفایل ایجاد کند [۱]. این پروتکل شامل چهار زیر پروتکل است: ***الف) ایجاد گروه:*** زمانی که کاربر Ui می‌خواهد به موتور جستجوی وب یک پرس‌وجو ارسال کند یک درخواستی را به گره مرکزی C ارسال می‌کند. گره C درخواست‌های n کاربر را دریافت می‌کند و یک گروه n عضوی از کاربران $\{U_1, ..., U_n\}$ ایجاد می کند. به کاربران اعلام می‌کند که عضو یک گروه هستند. کاربران، یک کانال ارتباطی بین خودشان دارند. هر کاربر U_i می‌تواند به بقیه گروه پیام ارسال کند. از این به بعد دیگر نیازی به حضور نود مرکزی C نیست. ***ب) تولید کلید برای گروه:*** کاربران $\{U_1, ..., U_n\}$ بر سر یک کلید p که یک عدد اول است، توافق می‌کنند. به‌طوری‌که $p=2q+1$. q نیز در اینجا یک عدد اول است. کاربران $\{U_1, ..., U_n\}$ با استفاده از الگوریتم ElGamal، کلید عمومی y را تولید می‌کنند. هر کاربر U_i کلید خصوصی $\propto_i$ را برای خود نگه می‌دارد. ***ج) ناشناس کردن پرس‌وجوها:*** برای هر کاربر گروه $\{U_1, ..., U_n\}$ که پرس‌وجوی m_i را ارسال و پرس‌وجوی m_i را از یک عضو دیگر از گروه دریافت می‌کند، مراحل زیر را انجام می‌دهیم: برای هر کاربر U_i به‌طوری‌که $i\in\{1, ..., n\}$است، این کارها را انجام می‌دهیم: هر کاربر U_i، یک مقدار تصادفی ri را تولید می‌کند و پرس‌وجوی m_i خود را با استفاده از کلید عمومی گروه، y رمز نگاری می‌کند (با استفاده از الگوریتم ElGamal). نتیجه به‌صورت رابطه (۱) می‌شود:

$$E_y(m_i, r_i) = (C1i, C2i) = C_i^p \qquad \text{رابطه (۱)}$$

کاربر U_i پرس‌وجوی رمز شده C_i^p را به کاربر دیگر گروه U_j ارسال می کند به‌طوری‌که $i \neq j$. یک ترکیب خاص برای کاربران مثلا از 1 تا n فرض کنید $\{C_1^p, ..., C_n^p\}$. مجموعه پرس‌وجوهای رمز شده پس از ارسال در گروه توسط U_i است. برای هر کاربر U_i از 1 تا n مراحل زیر اجرا می‌شود:

a) کاربر U_i پرس‌وجوهای رمز شده $\{C_1^{i-1}, ..., C_n^{i-1}\}$ از کاربر U_{i-1} را دریافت می‌کند و مجدداً با الگوریتم ElGamal رمز می‌کند و $\{e_1^{i-1}, ..., e_n^{i-1}\}$ را به‌دست می‌آورد. b) سپس ترتیب پرس‌وجوهای رمز شده را عوض می‌کند و مجموعه $\{e_{\delta(1)}^{i-1}, ..., e_{\delta(2)}^{i-1}\}$ را به‌دست می‌آورد. c) U_i مجموعه پرس‌وجوهای رمز شده تا آن مرحله $\{C_1^i, ..., C_n^i\}$ را به U_{i+1} ارسال می‌کنند. در نهایت کاربر U_n مجموعه $\{C_1^n, ..., C_n^n\}$ را به همه اعضای گروه به‌صورت همه پخشی ارسال می‌کند. در این مرحله $\{C_1, ..., C_n\} = \{C_1^n, ..., C_n^n\}$ را در نظر می‌گیریم. U_i پرس‌وجوی $(C1_i)^{\propto_j}$ را از U_j دریافت می‌کند به‌طوری‌که $j=1, ..., n$ و $i\neq j$ است. سپس U_i پرس‌و-جوی $(C1_j)^{\propto_j}$ را با کلید خصوصی خود محاسبه می‌کند. کاربر U_i، پرس‌وجوی رمز شده m_i را به‌صورت رابطه (۲) به‌دست می‌آورد:

$$m_i = \frac{C2_i}{(C1_i)^{\propto_i}\left(\prod_{i\neq j}(C1_i)^{\propto_j}\right)} \qquad \text{رابطه (۲)}$$

د) ارسال پرس‌وجو و بازیابی اطلاعات: هر عضو گروه (U_i)، m_i را به موتور جستجو ارسال می‌کند. زمانی که پاسخ a^i را دریافت کرد، به کل گروه به‌صورت همه پخشی ارسال می کند. هر کاربر با توجه به پرس‌وجوی ارسالی پاسخ a^i خود را برمی‌دارد.

روش دوم از رویکردهای سمت سرور

موتور جستجو، پرس‌وجو را به‌عنوان ورودی می‌گیرد و گزارش پرس‌وجوی ناشناس شده و پروفایل کاربر را به‌عنوان خروجی تولید می‌کند. زیر سیستم‌ها در این بخش توصیف می‌شوند [۳]:

الف) ایجاد پروفایل و ناشناس سازی: این فاز اصلی‌ترین فاز است.

- از توصیه کننده‌ها[۱۶] به منظور تصحیح غلط املایی استفاده می‌شود.
- آنتروپی‌ها[۱۷]: تعداد نتایج این کلمه در مقایسه با بقیه عبارت‌های موجود در پرس‌وجو محاسبه می‌شود تا کلمه اصلی در کل عبارت پرس‌وجو پیدا شود.
- طبقه‌بندی کننده[۱۸]: طبقه یا دسته متناظر با پرس‌وجو پیدا می‌شود.
- ناشناس کننده[۱۹]: گزارش‌های دسته‌بندی شده به‌عنوان ورودی این مرحله است. برای هر دسته دو مجموعه پرس‌وجو و کاربر داریم. هنگامی که یک مجموعه از هر دسته به ماکزیمم مقدار k می‌رسد به‌طور تصادفی یک کاربر و یک پرس و جو انتخاب و به‌یکدیگر متصل می‌شوند. با این کار کلیت حفظ می‌شود ولی ویژگی مخصوص او حذف می‌شود. یک مورد خاص زمانی است که در یک دسته در مجموعه کاربران، همه کاربران مشابه باشند که با توجه به تعداد پرس و جوی دریافتی در ثانیه توسط موتورهای جستجو احتمال خیلی پایینی دارد. در

این شرایط یک ضریب δ در نظر می‌گیریم که در k ضرب می‌کنیم تا سایز مجموعه افزایش یابد. الگوریتم ۲ این فرآیند را توصیف می‌کند.

ب) کاهش ناشناس‌سازی[۲۰]: برعکس فاز ناشناس‌سازی عمل می‌کند. با این کار تست می‌کند که آیا حمله کننده می‌تواند موفق شود یا خیر. در الگوریتم ۳ حمله کننده سعی می‌کند تا هر پرس‌وجو را به کاربر خودش ربط دهد.

۳-۳- مقایسه روش‌های بررسی شده

در جدول ۱ به مقایسه روش‌های بررسی شده می‌پردازیم.

٤- نتیجه گیری و کارهای آینده

در این مقاله به بررسی تهدیدها و راه حل‌های ارائه شده جهت برقراری حفظ حریم خصوصی کاربران پرداختیم. در اکثر الگوریتم‌های ارائه شده، کاربران و سرورهای واسط غیرمهاجم در نظر گرفته شده‌اند. در صورتی‌که در محیط واقعی ممکن است مهاجمی خود را به‌عنوان کاربر عادی معرفی کند. همچنین ممکن است مهاجم به سرورهای واسط نیز حمله کند. در صورتی‌که در روش‌های بررسی شده به نحوه برقراری امنیت فیزیکی و نرم افزاری این سرورها اشاره‌ای نشده است. الگوی جستجوی کاربران ممکن است تغییر کند. در پژوهش‌های آینده باید این شرایط در نظر گرفته شود. تحلیل رفتار کاربران در کنار مشخصات آنها می‌تواند نتایج جستجوی بهتری را ارائه کند. تمام URLهایی که یک کاربر برای یافتن نتایج دنبال می‌کند، دلیل بر مرتبط بودن با موضوع پرس‌وجو را ندارد. باید زمان توقف در یک صفحه، میزان کلمات آن صفحه و مواردی از این قبیل نیز در نظر گرفته شوند.

منابع

[1] J. Roca, A. Viejo and J. Joancomarti, *"Preserving User's Privacy in Web Search Engines"*, Elsevier, Computer Communications, (2009), 1541–1551.
[2] H. Aljifri and D. Navarro, *"Search Engines and Privacy"*, Elsevier, Computers & Security, (2004), 379-388.
[3] D. Estrems, J. Roca and A. Viejo, *"Working at the Web Search Engine Side to Generate Privacy-Preserving User Profiles"*, Elsevier, Expert Systems With Applications, (2016), 523–535.
[4] S. Malthankar and S. Kolte, *"Client Side Privacy Protection Using Personalized Web Search"*, 7th International Conference on Communication, Computing and Virtualization, Mumbai, India, (2016), 1029 – 1035.
[5] J. M. Saji , K. BhongleJ. and et al., *"Advancement in Personalized Web Search Engine With Customized Privacy Protection"*, Springer, Nature Singapore Pte Ltd, (2018), 405-413.

الگوریتم ۲

ورودی: گزارش‌های دسته‌بندی شده، k، δ
خروجی: گزارش‌های ناشناس شده

۱. برای هر گزارش از مجموعه گزارش‌های دسته‌بندی شده این کارها انجام شود:
۲. user, query, category ← log
۳. users [category] ← user;
۴. query [category] ← query;
۵. اگر اندازه users [category]=k بود، سپس
۶. اگر $\forall u \in$ users [category] ،$\exists u \neq$ user سپس
۷. یک کاربر را به صورت تصادفی از مجموعه کاربران انتخاب کن
۸. یک پرس‌وجو را به صورت تصادفی از مجموعه پرس‌وجوها انتخاب کن
۹. ارسال u و q
۱۰. در غیراینصورت k=k* δ
۱۱. پایان

الگوریتم ۲- الگوریتم ناشناس‌سازی

الگوریتم ۳

ورودی: گزارش‌های دسته‌بندی شده، k، δ

خروجی: گزارش پرسوجوها

۱. برای هر گزارش از مجموعه گزارش‌های دسته‌بندی شده این کارها انجام شود:
۲. user, query, category ← log
۳. users [category] ← user;
۴. query [category] ← query;
۵. اگر اندازه users [category]=k بود، سپس
۶. اگر $\forall u \in$ users [category] ،$\exists u \neq$ user سپس
۷. یک کاربر را از مجموعه کاربران انتخاب کن
۸. یک پرس‌وجو را از مجموعه پرسوجوها انتخاب کن
۹. ارسال u و q
۱۰. در غیراینصورت k=k* δ
۱۱. پایان

الگوریتم ۳ – کاهش ناشناس سازی

جدول ۱ – مقایسه راه حل‌های ارائه شده جهت حفظ حریم خصوصی کاربران موتورهای جستجوی وب

نام نویسنده	نام الگوریتم	سمت سرور یا کلاینت	پیش فرض‌ها	مزایا	معایب	عملکرد
[۱] J. Roca, A. Viejo	ندارد	سرور	سرور میانی و کلاینت‌ها تبانی نمی‌کنند	عملکرد بهتر نسبت به روش‌های مشابه که از رمزنگاری استفاده می‌کنند.	تاخیر در پاسخ به پرس وجوها.	هر کاربر پرسوجوی خود را رمز می‌کند و رندم توسط کاربر دیگری ارسال می‌شود. کاربر پاسخ پرسوجوهای دریافتی را رمزگشایی می‌کند تا پاسخ پرس وجو خود را بیابد.
[۳] D. Estrems, J. Roca	UUP	سرور	داده‌های اصلی در یک پایگاه داده امن نگهداری می‌شوند.	احتمال موفقیت حمله‌کننده ۱/۸٪ است، حفظ کلیت علاقه گروهی از کاربران به موضوعی خاص	مصرف زیاد حافظه	جهت ناشناس‌سازی پرسوجوها و کاربران در هر دسته به صورت تصادفی جایگشت داده می‌شوند.
[۴] S. Malthankar و S. Kolte	UPS	کلاینت	سرورها غیرقابل اعتماد هستند و هر کلاینت تنها به خود اعتماد دارد.	هزینه کم، ایجاد پروفایل آنلاین، داشتن حق انتخاب کاربر برای هر پرس‌وجو	کاربر باید ثبت نام کند و اطلاعات هویتی ارائه دهد	کاربر برای هر پرس‌وجو می‌تواند خصوصی و عمومی بودن را برای گزارش آن پرس‌وجو در پروفایل خود تعیین کند.
[۱۴] Y. Xu و B. Zhang	ندارد	کلاینت	اگر عبارتی در داکیومنت کاربر تکرار شود، جز علایق کاربر است.	پنهان کردن هر بخش از پروفایل توسط کاربر	احتمال دسته بندی عبارات به عنوان والد و فرزندی به جای دسته بندی در یک گروه کلی‌تر	ایجاد پروفایل سلسله مراتبی برای کاربران و دسته بندی پرس وجوها در گروه‌های کلی‌تر، کاربر امکان مخفی کردن هر بخش از درخت سلسله مراتبی را دارد.

[6] J. D. Velásquez, *"Web Mining and Privacy Concerns: Some Important Legal Issues to be Consider Before Applying any Data and Information Extraction Technique in Web-Based Environments", Elsevier*, Expert Systems with Applications, (2013), 5228–5239.

[7] J. Wang, and H. Raghav, *"An Exploration of Risk Information Search via a Search Engine: Queries and Clicks in Healthcare and Information Security"*, Elsevier, Decision Support Systems, (2012), 395–405.

[8] A. Cooper, *"A Survey of Query Log Privacy-Enhancing Techniques from a Policy Perspective"*, ACM Transactions on the Web, New York, (2008), 19–27.

[9] S. M. Beitzel, E. Jensen and A.Chowdhury, *"Hourly Analysis of a Very Large Topically Categorized Web Query Log", 27th annual international ACM SIGIR conference on research and devel- opment in information retrieval, Sheffield, United Kngdm*, (2004).

[10] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas, *"Releasing Search Queries and Clicks Privately"*, In Proceedings of the 18th international world wide web conference, Madrid, Spain, (2009).

[11] Y. He and J. F. Naughton, *"Anonymization of Set-Valued Data via Topdown, Local Generalization"*, In *Proceedings of the proceedings of the VLDB endowment*, (2009).

[12] J. Soria and J. Domingo, *"Big Data Privacy: Challenges to Privacy Principles and Models"*, *Data Science and Engineering*, (2015), 1–8.

[13] L. Rutkowski, M. Jaworski, and L. Pietruczuk, *"Decision Trees for Mining Data Streams Based on the Gaussian Approximation"*, IEEE Transactions on Knowledge and Data Engineering, (2014), 108 - 119.

[14] Y. Xu and B. Zhang and K, *"Privacy-Enhancing Personalized Web Search"*, International World Wide Web Conference Committee (IW3C2), Lyon (France), (2007), 8–12.

زیرنویس

[1] Log
[2] Query
[3] Personalize
[4] Batch Processing
[5] Query Removal

[6] Statistical Disclosure Control (SDC)
[7] Real Time Processing
[8] Static
[9] Data Stream
[10] Private
[11] Public
[12] Documents
[13] Overlap
[14] Chide & Parent
[15] Broadcast
[16] Recommender
[17] Entropies
[18] Classifier
[19] Anonymizer
[20] De-anonymizer