بازبینی

User k-anonymity for privacy preserving data mining of query logs

داریوش حسن‌پور آده

۹۳۰۸۱۶۴

این مقاله به مشکل حفظ حریم شخصی افراد در اطلاعات بدست آمده از موتورهای جستجو پرداخته است. برای نشان دادن اهمیت مساله نمونه مثال از شرکت AOL آورده است که برای منظور کمک به جامعه پژوهشی بازیابی اطلاعات باعث گردید اطلاعات شخصی برخی از کاربرانش افشا شود برای همین مساله برای حفظ اطلاعات شخصی افراد قبل از اینکه گزارشات جستجوها[۱] برای عموم ارائه شود نیاز است که یک عملیات پنهان‌سازی[۲] بروی داده‌ها اعمال شود؛ ولی تعیین اینکه تا چه حد داده‌ها نیاز به پنهان‌سازی[۳] دارند، سخت می باشد. بنابراین سعی دارد که روشی ارائه دهد که توازنی بین حفظ اطلاعات شخصی افراد و عدم از دست رفت اطلاعات واقعا مفید موجود در داده‌های جمع آوری شده از موتورهای جستجو ارائه دهد.

این مقاله با نام بردن چندین کار قبلی مبنی بر اینکه آن‌ها نیز از روش k-anonymity استفاده کرده‌اند ولی بطور کامل تضمین نمی‌کردند که معیار k در k-anonymity رعایت شده است – زیرا با حذف داده‌ها به این منظور می‌رسیدند. نوآوری این مقاله در ارائه‌ی روشی برای تضمین معیار k-anonymity در داده‌های جستجو با استفاده از میکرو–دسته[۴] بدون اینکه هیچ یک از داده‌ها صراحتا از search log حذف شده باشند.

این مقاله سپس به شرح مفهوم micro-aggregates پرداخته است. که شرح داده که شامل دو بخش می شود پارتیشن کردن و بعد متراکم کردن داده‌ها است. و همچنین آورده که برای این منظور یک جنگلی از درختان باید تشکیل شود که که هر درخت متعلق به یکی از کاربران است و شاخه‌های این درختان شامل جستجوهای آن کاربر می‌باشد. و سپس معیار فاصله را برای هر جستجو معرفی کرده است.

بطور خلاصه اگر بخواهیم روش مطرح شده با استفاده از micro-aggregates را بگوییم می‌توان گفت که ابتدا داده‌ها با توجه به معیار فاصله مطرح شده ریز–دسته‌بندی[۵] می‌شوند سپس مرکز دسته‌ها به عنوان نماینده آن دسته می‌شود و یک کوئری[۶] به عنوان نماینده آن کوئری‌ها ارائه می‌شود. که علاوه بر اینکه نمایانگر کوئری های هم دسته خود می‌باشد به علت نوع انتزاعی‌ای این نماینده دارد به خوبی می‌تواند اطلاعات شخصی افراد را مخفی نگه دارد – در اینجا k تعداد کابران موجود در هر دسته می‌باشند.

از مزایای روش ارائه شده در این مقاله علاوه بر سادگی روش می‌توان گفت که تضمین می‌کند معیار k در k-anonymity رعایت شده است و سپس اینکه اطلاعات شخصی افراد را می‌تواند با نسبت خوبی حفظ کند بدین نحو که بعد از پنهان‌سازی داده‌ها به ازای هر کوئری k کاربر مرتبط با آن کوئری می‌توان نسبت داد؛ همچنین مقاله آورده است که روش پیشنهادی از نظر محاسباتی حداکثر کارایی را دارد.

از آنجایی که روش مطرح شده در حالت پایه در واقع دسته‌بندی می‌باشد بنابراین تمامی معایب مطرح شده در دسته‌بندی را به طور ضمنی دارد مانند بررسی داده ها با ابعاد بالا نیاز به زمان بیشتری میخواهد، کارایی روش به فاصله تعریف شده وابسته می‌باشد. و همچنین مقاله در مورد راه‌حل‌هایی در مقابله به کوئری‌های پرت[۷] و تاثیر داده‌های پرت بر میزان پنهان‌سازی داده‌ها حرفی نزده و همان‌طور که می‌دانیم خوشه‌بندی به داده‌های پرت حساس می‌باشد که مقاله از کنار این موضوع بدون بررسی گذشته است. همچنین یکی از معایب دیگری که می توان به روش ارائه شده نسبت داد این است که میزان تعیین k مناسب نیز سخت می‌باشد که در اینجا کل عملیات

---

[۱] Search log
[۲] Anonymize
[۳] Degree of privacy
[۴] micro-aggregates
[۵] Clustering
[۶] Query
[۷] Outlier

بر مبنای مقدار k بنا شده است. از طرف دیگر روش ارائه شده همه‌ی ویژگی‌های موجود در رکوردها با یک دید یکسان نگاه می‌کند در حالی که اهمیت مخفی‌سازی برخی ویژگی‌ها از برخی ارجهیت دارد و این ارجهیت در خوشه‌بندی‌های اعمال شده تاثیر داده نمی‌شود مثلا اهمیت پنهان‌سازی زمان جستجو از لینک کلیک شده بعد از جستجو کمتر است و این ارجهیت می‌تواند در نحوه‌ی شکل یافتن میکرو-دسته‌ها موثر باشد که مقاله به این موضوع نیز نپرداخته است.