مروری بر روش های گمنام‌سازی در پاسخگویی به پرس‌وجوها

مائده مهرآوران[۱]،محمد-رضا زارع‌میرک‌آباد[۲]

[۱] دانشجوی دکتری رشته مهندسی کامپیوتر- گروه مهندسی کامپیوتر-دانشگاه یزد- یزد-ایران
[۲] استادیار – گروه مهندسی کامپیوتر-دانشگاه یزد- یزد-ایران

چکیده

امروزه با توجه به عصر اطلاعات و رشد سریع آن، حجم زیاد داده‌ها همراه با جزئیات کافی در دسترس افراد مختلف است که می‌تواند حریم خصوصی افراد را به مخاطره اندازد و باعث فاش‌شدن اطلاعات خصوصی افراد شود. برای مفید بودن اطلاعات، همراه با تضمین این محافظت، با یک وضعیت متناقضی مواجه هستیم. از یک طرف نیاز به انتشار جزئیات داده‌ها است تا داده‌ها مفید و مورد استفاده واقع گردد. از طرف دیگر نیاز به جلوگیری از انتشار اطلاعات و صفات حساس وجود دارد تا منجر به فاش‌شدن اطلاعات خصوصی و محرمانه افراد نگردد. بنابراین نیاز به روش‌هایی وجود دارد که امکان دسترسی به اطلاعات فرد خاص وجود نداشته باشد. گمنام‌سازی یکی از این روش‌ها است. همراه با گمنام‌سازی داده‌ها، بحث میزان اطلاعاتی که در اختیار افراد قرار می‌دهند وجود دارد. یکی از جنبه‌های مهم در مورد داده‌های گمنام‌سازی شده قابلیت آنها در جواب دادن به پرس‌وجوهای کاربران می‌باشد. در این مقاله، مروری بر کارهای انجام شده در زمینه میزان نزدیک بودن پاسخ پرس‌وجوهای کاربران به واقعیت، در روش‌های مختلف، خواهیم داشت و در نهایت مقایسه‌ای بین روش‌ها انجام می‌شود.

کلمات کلیدی

محافظت از محرمانگی[۱]، گمنام‌سازی[۲]، گمشدگی اطلاعات[۳]، پاسخگویی به پرس‌وجو[۴]

۱. مقدمه

امروزه با توجه به عصر اطلاعات و رشد سریع وسایل جمع‌آوری داده‌ها، همچنان نیاز به اطلاعات فردی افراد وجود دارد. این اطلاعات فقط در زمینه‌های داده‌کاوی و تشخیص تقلب مورد نیاز نیست، بلکه در زمینه‌هایی دیگر چون پزشکی، بررسی خطرات و... نیز مورد استفاده قرار می‌گیرد. حجم زیاد داده‌ها با جزئیات سطح بالا از منابع مختلف که به صورت عام قابل دسترس افراد مختلف است، حریم خصوصی افراد را به مخاطره می‌اندازد.

از طرفی برای مفید بودن اطلاعات، و در عین حال تضمین این محافظت، با یک وضعیت متناقضی مواجه هستیم. از یک طرف نیاز به انتشار جزئیات داده‌ها است تا داده‌ها مفید و مورد استفاده واقع گردد. از طرف دیگر نیاز به جلوگیری از انتشار اطلاعات و صفات حساس وجود دارد تا منجر به فاش شدن اطلاعات خصوصی و محرمانه افراد نگردد.

با توجه به مقدمه فوق، بحث محافظت از محرمانگی[۵] در انتشار داده‌ها جزء مباحث روز می‌باشد. مسأله اصلی در اینجا این است که محافظت از محرمانگی داده‌ها به چه صورت انجام گیرد و چه جزئیاتی از داده‌ها منتشر شود که تا حد امکان به ارزش دادها دست پیدا کنیم و درضمن گمشدگی اطلاعات[۶] کمی داشته باشیم.

در این راستا دو دیدگاه کاملا مخالف مطرح شده است. یکی این که بعضی بدون توجه به تهدیدات ناشی از انتشار داده‌ها، با حذف فقط مولفه‌های کلیدی، داده‌ها را در اختیار دیگران می‌گذارند که به طور قطع می‌تواند مورد سوء استفاده قرار گیرد. از طرف دیگر بعضی به خاطر ترس از افشای اطلاعات محرمانه از انتشار داده‌های خود امتناع می‌کنند که این امر علاوه بر اینکه دست بسیاری از پژوهشگران را در انجام پروژه‌هایی مثل داده‌کاوی می‌بندد، دسترسی آنها به اطلاعات مدیریتی را نیز به تعویق می‌اندازد.

[۱] نویسنده مسئول: دانشجوی دکتری دانشگاه یزد-m.mehravaran@stu.yazd.ac.ir

Archive of SID

۲. پیشینه تحقیق

برای محافظت از محرمانگی داده‌ها و جلوگیری از تهدیدات افشا شدن داده‌ها، راه‌حل‌های مختلفی ارائه شده است که موضوع گمنام‌سازی داده‌ها (k-anonymization) یکی از موضوعاتی است که بسیار کار شده است و الگوریتم‌های مختلفی روی آن ارائه شده است. هدف این الگوریتم‌ها تغییر داده‌ها (مثلا عمومی‌سازی مقادیر صفات خاصه) به سطح بالاتری از امنیت، در انتشار داده‌ها است.

با وجود جلوگیری از افشای اطلاعات با روش‌های گمنام‌سازی، گاهی توزیع صفات حساس به گونه‌ای است که بدون شناسایی افراد می‌توان صفت محرمانه را کشف کرد. این امر سبب مطرح شدن مدلهای دیگری در محافظت از محرمانگی اطلاعات گردید. به این دسته از مدلها تنوع بخشی (l-diversity) گویند.

برای توضیح بیشتر روش‌ها و مفاهیم فرض می‌کنیم داده‌های شخصی افراد در جدولی ذخیره شده است به طوری که هر سطر مربوط به یک فرد است، صفات جدول را به چهار دسته می‌توان تقسیم کرد[1]:

۱- صفت شناسه[۷]: صفاتی همچون کد مشتری در بانک، شماره گواهینامه رانندگی، شماره دانشجویی، کد ملی که دقیقا یک فرد را مشخص می‌کند، که به این صفات شناسه یا کلید گفته می‌شود.

۲- صفت حساس[۸]: صفاتی چون درآمد مشتریان بانک یا بیماری افراد که برایشان با اهمیت است و افراد دوست ندارند کسی از این اطلاعات با خبر شود. به این صفات، صفات حساس می‌گویند.

۳- صفت شبه شناسه[۹]: مجموعه صفاتی مثل سن، جنسیت و کد محل زندگی که به تنهایی منحصر بفرد نیستند ولی به صورت ترکیبی می‌توانند برای مشخص کردن افراد استفاده شوند. به این صفات شبه شناسه گفته می‌شود.

۴- صفت نرمال[۱۰]: بقیه صفات که در دسته‌های بیان‌شده قرار نمی‌گیرند و داده‌هایشان بدون در نظر گرفتن ملاحظات امنیتی خاص قابل انتشار است، به عنوان صفات نرمال در نظر گرفته می‌شوند.

اولین مرحله محافظت از محرمانگی داده‌ها حذف صفات کلیدی است. به این عمل de-identification گویند. این عمل مسلما کافی نخواهد بود و با ترکیب داده‌های مختلف می‌توان دوباره شخص را شناسایی کرد. در واقع صفات شبه شناسه هستند که ترکیب آن‌ها باعث فاش شدن اطلاعات حساس افراد می‌شود. شکل۱-الف اطلاعات بیماران یک بیمارستان را نشان می‌دهد. با فرآیند de-identification روی شکل ۱-الف به شکل ۱-ب می‌رسیم.[2]

Name	Age	Gender	Zip	Disease
Alice	21	Female	17651	Cancer
Jack	22	Male	17652	Flu
Jan	23	Male	17661	HIV
Bob	24	Male	17662	HIV

شکل ۱-الف-جدول اصلی [2]

Age	Gender	Zip	Disease
21	Female	17651	Cancer
22	Male	17652	Flu
23	Male	17661	HIV
24	Male	17662	HIV

شکل۱-ب-جدول اصلی با حذف شناسه[2]

با توجه به اینکه با الحاق جدول شکل ۱- ب با جداول دیگر یا داشتن اطلاعاتی راجع به افراد مثل اینکه Bob ۲۴ سال دارد، قابلیت فاش‌سازی داده‌های حساس وجود دارد. بنابراین مفهوم k-anonymity و l-diversity را توضیح می‌دهیم.

k-anonymity: این مفهوم تضمین می‌کند که هر رکورد در داده‌های منتشر شونده از حداقل k-1 رکورد دیگر بر اساس شبه شناسه‌ها قابل تشخیص نباشد. بنابراین یک الحاق با جدول گمنام‌سازی شده حداقل k مورد را شامل می‌شود. بنابراین شخص بین k نفر مبهم می‌ماند. برای این عمل لازم است یکسری از ویژگی‌ها را عمومی‌سازی[۱۱] کنیم. در واقع مقادیر واقعی را با مقادیری که کمتر خاص هستند جایگزین کنیم. شکل ۲ نسخه 2-anonymous از رکوردهای شکل ۱ بعد از عمومی‌سازی است. در این حالت نمی‌توان تشخیص داد که آلیس کدام یک از دو فرد اول گروه است وبنابراین نمی‌توان گفت کدام بیماری را دارد.

Age	Gender	Zip	Disease
[21-22]	*	1765*	Cancer
[21-22]	*	1765*	Flu
[23-24]	Male	1766*	HIV
[23-24]	Male	1766*	HIV

شکل ۲-جدول 2-anonymous[2]

با وجود اینکه گمنام‌سازی جلو شناسایی موجودیت‌های منتشر شده را می‌گیرد ولی گاهی جهت حفاظت از اطلاعات حساس اشخاص درست کار نمی‌کند. در مثال قبلی اگر Bob یکی از دو رکورد آخر شکل۲ باشد، حتی اگر نتوان دقیقا مشخص کرد که کدام یک است ولی چون هر دو رکورد دارای بیماری HIV هستند با احتمال ۱۰۰٪ می‌توان گفت که Bob بیماری HIV دارد.

l-diversity : این مفهوم جلو استنتاج‌های اطلاعات ناخواسته را می‌گیرد. در واقع ضمانت می‌کند که اطلاعات حساس اشخاص با احتمال مشخصی قابل تشخیص نیست. فرآیند تغییر نتایج در شکل۳ نشان دهنده نسخه 2-diversity از شکل ۱ است. در این جدول هر گروه از رکوردها دو مقدار حساس مختلف دارند که احتمال تشخیص بیماری را به ۵۰٪ می‌رساند.

Archive of SID
www.SID.ir

Age	Gender	Zip	Disease
[21-23]	*	176**	Cancer
[21-23]	*	176**	HIV
[22-24]	Male	176**	Flu
[22-24]	Male	176**	HIV

شکل۳- 2-diversity [2]

در بسیاری مقالات مثل[10][11][9][1] روی بحث k-anonymity و l-diversity و روش‌های دیگر گمنام‌سازی تکمیلی بحث شده است. اما بحث بسیار مهم، راجع به این است که بعد از گمنام‌سازی، تا چه حد پاسخ به پرس‌وجو به واقعیت نزدیک است. در واقع برای مقایسه دو روش گمنام‌سازی مختلف از دو معیار میزان محرمانگی و بهروری داده‌ها استفاده می‌کنند که البته دو معیار متناقض است. معیار اول برای محاسبه محرمانگی در مقاله سالاری و همکاران به صورت حاصل جمع فاصله هر رکورد تا مرکز هر دسته محاسبه می‌شود. مسلما این مقدار هر چه بیشتر باشد داده‌های از دست‌رفته بیشتری خواهیم داشت[12]. معیار دوم بهره‌وری از داده‌ها است که در مقاله سالاری به صورت زیر تعریف شده است. فاصله اقلیدوسی هر رکورد تا رکورد تغییر یافته متناظرش نسبت به فاصله همه رکوردها به آن رکورد تغییر یافته کمتر باشد. در واقع رکورد تغییر یافته بسیار به رکورد اصلی متناظرش نزدیک باشد تا از بتوان از داده‌ها بهتر استفاده کرد[12]. معیار بهره‌وری را می‌توان بوسیله نزدیک بودن پاسخ پرس‌وجو به واقعیت بررسی کرد. در ادامه مروری خواهیم داشت به کارهایی که در زمینه پاسخگویی به پرس و جوهای کاربران در پایگاه‌داده گمنام‌سازی شده انجام شده است.

۳. پاسخگویی به پرس‌وجو در مجموعه داده گمنام‌سازی شده

Xiao و همکاران کار روی پایگاه داده‌های گمنام‌سازی شده برای جواب به پرس‌وجوهای تجمعی را بررسی می‌کنند [3]. در واقع به جای آنکه جواب پرس‌وجو یک رکورد باشد، مجموعه رکوردها را بوسیله یک تابعی از توابع sum,average,max,min مورد ارزیابی قرار می‌دهد. در این پرس‌وجوها احتمال اینکه فردی با زدن چند پرس‌وجو پشت سرهم بتواند صفت حساسی را فاش کند، وجود دارد.

Xiao و همکاران سه روش برای پاسخگویی به پرس‌وجوها بیان کرده‌اند[3]. روش اول سیستم باید تاریخچه پاسخگویی فرد را داشته باشد تا بتواند نسبت به پاسخ دادن به پرس‌وجو آن فرد تصمیم بگیرد. در واقع اگر قبلا پرس‌وجو داده باشد که با پرس‌وجو جدید آن امکان فاش شدن اطلاعات حساس وجود داشته باشد، پرس‌وجو رد و در غیر این صورت پاسخ داده می‌شود. روش دوم پرس‌وجو را سیستم می‌گیرد و جواب درست را پیدا می‌کند و با اضافه کردن نویز به پاسخ نتایج را تغییر می‌دهد و اعلام می‌کند. روش سوم بحث تغییر داده‌های اصلی و روش‌های گمنام‌سازی(k-anonymity) است. در واقع بعد از گمنام‌سازی، پاسخ به پرس‌وجو داده می‌شود. در مقاله[3] روی روش سوم بحث شده است چون روش اول به خاطر نگه‌داشتن سابقه پرس‌وجوها نیاز به فضای زیادی دارد و عملا امکان پذیر نیست و روش دوم، باید نویز مناسب[۱۲] برای اضافه شدن به پاسخ پرس‌وجو پیدا کند.

با توجه به استفاده از روش گمنام‌سازی، بحث قابل بررسی چگونگی گروه‌بندی داده‌ها به گونه‌ای است که جهت پاسخگویی به پرس‌وجوهای چند بعدی پاسخ بهتری حاصل شود. برای مثال شکل ۴-الف جدول اصلی را نشان می‌دهد و شکل۴-ب و ۴-ج گمنام‌سازی مختلف از جدول۴-الف را نشان می‌دهد. در هر دو 2-anonymous و l-diversity رعایت شده اما کدام بهتر است؟

Name	Age	Zip.	Disease
Alice	20	12000	flu
Bob	23	58000	gastritis
David	38	41000	flu
Helen	42	23000	gastritis
Jack	46	25000	flu
Ken	48	13000	gastritis
Linda	49	51000	flu
Mary	52	52000	insomnia
Paul	53	49000	gastritis
Ray	59	61000	flu
Tom	61	39000	gastritis

شکل۴-الف-جدول اصلی [3]

Tuple ID	Age	Zip.	Disease
1 (Alice)	[20, 23]	[12k, 58k]	flu
2 (Bob)	[20, 23]	[12k, 58k]	gastritis
3 (David)	[38, 42]	[23k, 41k]	flu
4 (Helen)	[38, 42]	[23k, 41k]	gastritis
5 (Jack)	[46, 48]	[13k, 25k]	flu
6 (Ken)	[46, 48]	[13k, 25k]	gastritis
7 (Linda)	[49, 53]	[49k, 52k]	flu
8 (Mary)	[49, 53]	[49k, 52k]	insomnia
9 (Paul)	[49, 53]	[49k, 52k]	gastritis
10 (Ray)	[59, 61]	[39k, 61k]	flu
11 (Tom)	[59, 61]	[39k, 61k]	gastritis

شکل۴-ب-جدول گمنام‌سازی۱[3]

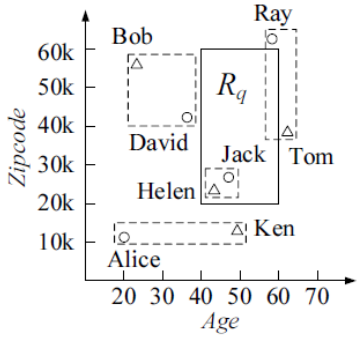
Tuple ID	Age	Zip.	Disease
1 (Alice)	[20, 48]	[12k, 13k]	flu
2 (Ken)	[20, 48]	[12k, 13k]	gastritis
3 (Jack)	[42, 46]	[23k, 25k]	flu
4 (Helen)	[42, 46]	[23k, 25k]	gastritis
5 (David)	[38, 61]	[39k, 41k]	flu
6 (Tom)	[38, 61]	[39k, 41k]	gastritis
7 (Linda)	[49, 53]	[49k, 52k]	flu
8 (Mary)	[49, 53]	[49k, 52k]	insomnia
9 (Paul)	[49, 53]	[49k, 52k]	gastritis
10 (Ray)	[23, 59]	[58k, 61k]	flu
11 (Bob)	[23, 59]	[58k, 61k]	gastritis

شکل۴-ج-جدول گمنام‌سازی ۲[3]

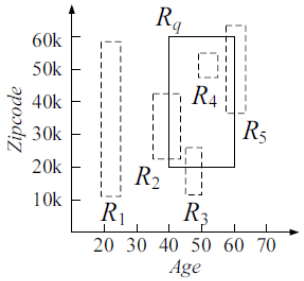
معیار مهم، جهت خوب بودن نحوه ابهام‌سازی، نزدیک بودن جواب پرس‌وجوها به جواب واقعی است. بهتر است تقسیم بندی فضا طوری انجام شود که هم‌پوشانی کمتری با پرس‌وجو داشته باشد. گمنام‌سازی در داده‌ها باعث گروه‌بندی داده‌ها به یکسری بازه‌های خاص می‌شود. زمان پاسخگویی به پرس‌وجو، گمنام‌سازیی خوب است که یا بازه‌ها، همپوشانی کامل با پرس‌وجو داشته باشد یا کاملا مجزا باشد. در شکل۵ مستطیل پر رنگ نشان دهنده محدوده پرس‌وجو است و مستطیل‌های خط‌چین نشان دهنده بازه‌های گمنام‌سازی شده هستند. در شکل ۵-الف می‌بینیم که گمنام‌سازی طوری انجام شده است که پرس‌وجو، پاسخ نزدیک به واقعیت می‌دهد و با بازه‌های

Archive of SID
www.SID.ir

کمتری وجه مشترک دارد. در حالیکه در شکل۵–ب هم پوشانی بین بازه‌ها و پرس‌وجو دیده می‌شود.



شکل۵–الف[3]



شکل۵–ب [3]

درواقع برای هر پرس‌وجو یک مدل گمنام‌سازی استفاده می‌کند و طبق گفته مقاله نیازی به نگه‌داری نسخه‌های قبلی نیست[3].

Kumari وهمکاران مسئله گمشدگی‌داده‌ها[۱۳] را بررسی کرده‌اند[4]. در واقع روشی بیان شده که از مجموعه‌های فازی برای افزایش اطلاعات منتشر شده استفاده کرده است. این روش هم برای صفات عددی و هم برای صفات دسته‌ای مناسب است. برای هر صفت که مقدار عمومی‌تر آن در نظر گرفته می‌شود، یک درجه عضویت(میزان تعلق) تعریف می‌کند. برای مثال شکل ۶ میزان حقوق افراد را به سه مقدار {low,medium,high} تبدیل می‌کند و میزان عضویت هر مقدار را در دسته مورد نظر بیان می‌کند. در واقع با بازه‌بندی هر مقدار، میزان عضویت هر مقدار به بازه مشخص می‌شود.

Income	10000	23000	58000	85000	94000
Weight	1.0	0.71	0.73	0.66	0.86
changed to	low	low	Medium	high	high

شکل۶– جدول داده‌های گمنام‌سازی با فازی[4]

برای مثال ۱۰۰۰۰ و ۲۳۰۰۰ هر دو در دسته low قرار گرفته است ولی چون ۱۰۰۰۰ کمتر است پس درجه عضویت بیشتری به low دارد. در مورد دو مقدار ۸۵۰۰۰ و ۹۴۰۰۰ چون ۹۴۰۰۰ بیشتر است پس درجه عضویت بیشتری به high دارد.

در این حالت میزان اطلاعاتی که در اختیار کاربر قرار می‌گیرد بیشتر است. درحالیکه به مقدار دقیق صفت حساس فرد دست نمی‌یابد.

Mukkamala و همکاران، با توجه به اینکه بازه‌های تعریف شده برای مجموعه‌های فازی ممکن است هم‌پوشانی داشته باشد، بحث تعیین میزان درجه عضویت هر صفت را مطرح می‌کنند[5]. هر مقدار، می‌تواند به چند مقدار درجه عضویت داشته باشد. برای مثال بازه سن ۰–۱۰۰ را به سه بازه (۰–۵۰) مجموعه۱ و بازه(۲۵–۷۵) مجموعه۲ و بازه(۵۰–۱۰۰) مجموعه۳ افراز می‌کنند. سن۴۵ سال هم به مجموعه ۱ هم به مجموعه۲ تعلق دارد. میزان عضویت ۴۵ در مجموعه۱، ۰.۹و میزان تعلق ۴۵ به مجموعه۲، ۰.۱ است. یک روش این است که بزرگترین را در نظر بگیریم و ۴۵ را متعلق به مجموعه۱ بدانیم. اما روش‌های مطرح شده در مقالهMukkamalaو همکاران از ترکیب مقادیر استفاده می‌کند تا یک مقدار واحد بدست آورد[5]. برای آنکه بتواند چند مقدار را به یک مقدار تبدیل کند، از روش‌های میانگین‌گیری مقادیر، مینیمم ، ماکزیمم و .. برای ترکیب مقادیر استفاده می‌کند.

Liu وهمکاران از روش گمنام‌سازی مجموعه‌ای استفاده می‌کند تا بتواند به پرس‌وجوها بهتر پاسخ دهد[6]. در واقع به جای آنکه در گمنام‌سازی از مقدار عمومی هر قسمت استفاده کند، مقدار صفات را به صورت بازه‌هایی در می‌آورد و مقدار صفت را با مقادیر مجموعه‌ای موجود مدل می‌کند. نمونه‌ای از گمنام‌سازی مجموعه‌ای را در شکل ۷ می‌بینیم. در شکل۷ صفت‌ها به جای اینکه به صورت بازه در بیاید یا با مقادیر کلی‌تر جایگزین شوند، به صورت مجموعه‌ای بیان شده است. در حالت مجموعه‌ای اطلاعات بیشتری در اختیار استفاده کننده قرار می‌گیرد و اطلاعات از دست رفته کمتری داریم.

	Age	Education	Gender	Disease
	{21,25,32}	{High-School,Bachelors, Prof-School}	{F}	Asthma
G_1	{21,25,32}	{High-School,Bachelors, Prof-School}	{F}	Asthma
	{21,25,32}	{High-School,Bachelors, Prof-School}	{F}	Asthma
	{57,60,62}	{11th,11th,Assoc-acdm}	{F,M,M}	Asthma
G_2	{57,60,62}	{11th,11th,Assoc-acdm}	{F,M,M}	Obesity
	{57,60,62}	{11th,11th,Assoc-acdm}	{F,M,M}	Flu
	{68,70,71,73}	{7-8th,10th,12th,12th}	{F}	Gastritis
G_3	{68,70,71,73}	{7-8th,10th,12th,12th}	{F}	Cancer
	{68,70,71,73}	{7-8th,10th,12th,12th}	{F}	Obesity
	{68,70,71,73}	{7-8th,10th,12th,12th}	{F}	Cancer

شکل۷–جدول بازه‌ای شد[6]

برای مثال در پاسخگویی به پرس‌وجو زیر که تعداد افراد زیر ۵۵ سال و سطح تحصیلات بالای ۱۲ سال را می‌خواهد، در گمنام‌سازی مجموعه‌ای فقط گروه ۱ در شکل بالا شرایط پرس‌وجو را دارد. در حالیکه در گمنام‌سازی بازه‌ای گروه ۲ هم علاوه بر گروه ۱ باید گزارش شود. بنابراین می بینیم که در حالت بازه‌ای پاسخ به واقعیت نزدیک‌تر است.

```
Select COUNT(*) From T*
Where Age<55 And Education>12
```

Srisungsittisunti وهمکاران در رابطه با مدل (k-e) anonymity بحث می‌کند[7]. در این مدل علاوه بر دسته‌بندی داده‌ها به k دسته، صفات حساس داخل هر گروه به صورت تصادفی جابجا می‌شوند و در هر دسته حداقل تفاضل صفت حساس باید e باشد. این تفاضل به این دلیل است که صفات حساس عددی موجود در هر گروه ممکن است به علت نزدیک بودن داده‌ها، قابلیت فاش‌سازی داشته باشند. این روش برای داده‌های عددی مناسب است.

Srisungsittisunti و همکاران مدل را برای پرس‌وجوهای تجمعی طراحی کرده‌اند[7]. بطوریکه زمان ارسال نسخه‌های مختلف پایگاه‌داده، بوسیله پرس‌وجوهای اشتراکی و تفاضلی نتوان صفات حساس فرد خاصی را فاش کرد. به این مشکل نشت محرمانگی افزایشی[۱۴] گویند. در شکل زیر این فاش شدن را بوسیله عملیات اشتراک و تفاضل می‌توان دید. در شکل ۸

Archive of SID
www.SID.ir

قسمتa داده‌های منتشر شده در زمان$t1$ را به دسته‌های ۳ تایی دسته بندی می‌کند و صفت حساس حقوق را در هر گروه جابجا می‌کند. در شکل۸ قسمتb داده‌های منتشر شده زمان$t2$ را مشاهده می‌کنیم که به دسته‌های چهارتایی تقسیم شده و صفت حساس در هر گروه جایگشت داشته است. در شکل۸ قسمت c نتیجه تفاضل پارتیشن P2 در زمان $t1$و در زمان$t2$ را نشان می‌دهد که یک فرد اضافه شده و اگر حمله کننده بداند که Meg در زمان $t2$ اضافه شده است با توجه به لیست حقوق‌ها می‌تواند حقوق فرد را کشف کند. درشکل۸ قسمتd اشتراک نسخه‌های منتشر شده زمان $t1$و $t2$ را می‌بینیم که به راحتی می‌توان حقوق را یکی از دو عدد ۸۴۰۰۰ یا ۸۵۰۰۰ دانست[7].

ID	Quasi-identifiers		Sensitive
Name	Age	Gender	Salary
Meg	35	female	87,000

c) Result of difference between p_2 time t_1 and p_2 time t_2

ID	Quasi-identifiers		Sensitive
Name	Age	Gender	Salary
Tom	21	male	84,000
Mike	23	male	85,000

d) Result of intersection between p_1 time t_1 and p_1 time t_2

شکل۸–فاش شدن داده‌های فرد [7]

	ID	Quasi-identifiers		Sensitive
	Name	Age	Gender	Salary
p_1	Tom	21	male	85,000
	Mike	23	male	84,000
	Bob	52	male	87,000
p_2	Alice	27	female	89,000
	Hank	53	male	88,000
	Kate	49	female	91000

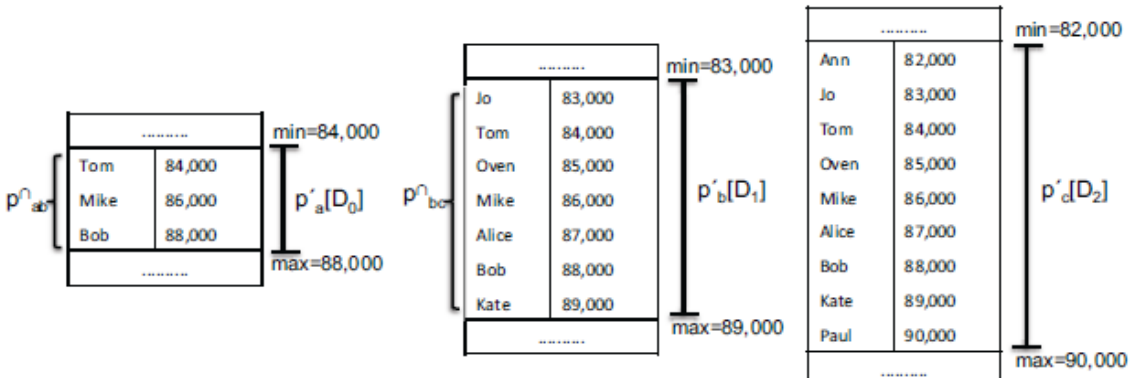
a) Shuffled data at time t_1

	ID	Quasi-identifiers		Sensitive
	Name	Age	Gender	Salary
p_1	Tom	21	male	84,000
	Mike	23	male	83,000
	Robin	29	male	85,000
	Oven	31	male	86,000
p_2	Alice	27	female	87,000
	Meg	35	female	89,000
	Hank	53	male	91,000
	Kate	49	female	88,000
p_3	Bob	52	male	87,000
	.....			

b) Shuffled data at time t_2

برای جلوگیری از نشت محرمانگی شرط زیر مجموعه بودن نسخه‌های مختلف را مطرح می‌کند. شکل۹ نسخهD_0 (نسخه منتشر شده در زمان$t1$) و D_1 (نسخه منتشر شده در زمان $t2$) را نشان می‌دهد. نسخه‌ها طوری باید منتشر شود که در هر قسمت، بازه صفات حساس، زیر مجموعه بازه صفات حساس نسخه بعدی باشد. در شکل ۹ می‌بینیم که در نسخهD_0 بازه صفت حساس در یک گروه بین ۸۴۰۰۰ تا ۸۸۰۰۰ است. در نسخهD_1 بازه صفت حساس بین۸۳۰۰۰ تا ۸۹۰۰۰ است. بازه اولی زیرمجموعه بازه دومی است.در این صورت است که اشتراک و تفاضل قابلیت فاش‌سازی صفات حساس را ندارد. نکته قابل توجه بیان شده در مقاله[7] این است که الگوریتم نیازی به همه نسخه‌های منتشر شده ندارد و فقط به نسخه آخر نیاز دارد.

Zhang و همکاران مقایسه روش جایگشت[۱۰] و روش گمنام‌سازی (k-anonymity) در پاسخگویی به پرس‌وجو را مورد بحث قرار می‌دهند[8]. در روش گمنام‌سازی علاوه بر دسته‌بندی داده‌ها، یکسان‌سازی صفات شبه شناسه هم صورت می‌گیرد، در حالیکه در روش جایگشت بعد از دسته‌بندی داده‌ها صفت حساس در هر دسته جابه‌جا می‌شود و یکسان‌سازی در هر دسته صورت نمی‌گیرد. برای مقایسه دو روش از دو پرس‌وجو استفاده کرده و به این نتیجه رسیده است که روش جایگشت بسیار بهتر از روش گمنام‌سازی جواب می‌دهد و پاسخ‌ها به واقعیت نزدیک‌تر است. در شکل۱۰ داده‌های اصلی دیده می‌شود و شکل۱۱–الف گمنام‌سازی معمولی را نشان می‌دهد و شکل۱۱–ب بر اساس جایگشت داده‌ها را آماده کرده است. اگر پرس‌وجویی مطرح شود که



شکل۹–روش مناسب برای نسخه‌های مختلف[7]

Archive of SID
www.SID.ir

مینییمم حقوق کارمندان خانم چقدر است، در روش گمنام‌سازی[16] چون ستون جنسیت کلی شده[17]، نمی‌توان تشخیص داد کدام رکوردها خانم هستند. در حالیکه در روش جایگشت در هر گروه تعداد خانم‌ها مشخص است و جواب پرس‌وجو به واقعیت نزدیک‌تر خواهد بود.

	ID	Quasi-identifiers			Sensitive
tuple ID	name	age	zipcode	gender	salary
1	Alex	35	27101	M	$54,000
2	Bob	38	27120	M	$55,000
3	Carl	40	27130	M	$56,000
4	Debra	41	27229	F	$65,000
5	Elain	43	27269	F	$75,000
6	Frank	47	27243	M	$70,000
7	Gary	52	27656	M	$80,000
8	Helen	53	27686	F	$75,000
9	Jason	58	27635	M	$85,000

شکل۱۰-داده‌های اصلی[8]

		Quasi-identifiers			Sensitive
group ID	tuple ID	age	zipcode	gender	salary
1	1	[31-40]	271*	*	$56,000
1	2	[31-40]	271*	*	$54,000
1	3	[31-40]	271*	*	$55,000
2	4	[41-50]	272*	*	$65,000
2	5	[41-50]	272*	*	$75,000
2	6	[41-50]	272*	*	$70,000
3	7	[51-60]	276*	*	$80,000
3	8	[51-60]	276*	*	$75,000
3	9	[51-60]	276*	*	$85,000

شکل۱۱-الف روش (k-e)-anonymity[8]

		Quasi-identifiers			Sensitive
group ID	tuple ID	age	zipcode	gender	salary
1	1	40	27130	M	$54,000
1	2	38	27120	M	$55,000
1	3	35	27101	M	$56,000
2	4	41	27229	F	$65,000
2	5	43	27269	F	$70,000
2	6	47	27243	M	$75,000
3	7	52	27656	M	$75,000
3	8	53	27686	F	$80,000
3	9	58	27635	M	$85,000

شکل۱۱-ب روش جایگشت[8]

همچنین در مقاله[8] جدول کمکی[18] مطرح شده است که می‌تواند با توجه به اینکه در هر گروه چند رکورد پاسخ پرس‌وجو است، بازه مینیمم و ماکزیمم را برای صفت حساس مشخص کند. در شکل۱۲ جدول کمکی برای دو تابع تجمعی(min,sum) روی جداول شکل۱۱ نشان داده شده است. برای مثال سطر اول مشخص می‌کند اگر جواب پرس‌وجو یکی از رکوردهای گروه۱ باشد و تابع تجمعی min برای پرس‌وجو انتخاب شود، حد بالای صفت حساس حقوق $56K و حد پایین$54K خواهد بود.

group ID	hits	sum-l-b	sum-u-b	min-l-b	min-u-b
1	1	$54K	$56K	$54K	$56K
1	2	$109K	$111K	$54K	$55K
1	3	$165K	$165K	$54K	$54K
2	1	$65K	$75k	$65K	$75K
2	2	$135K	$145K	$65K	$70K
2	3	$210K	$210K	$65K	$65K
3	1	$75K	$85K	$75K	$85K
3	2	$155K	$165K	$75K	$80K
3	3	$240K	$240K	$75K	$75K

شکل۱۲-جدول کمکی[8]

٤. مقایسه روش‌ها

همانطور که بیان شد، در مقالات مختلف الگوریتم‌های گمنام‌سازی گوناگون بیان شده و بوسیله میزان واقعی بودن پاسخ پرس‌وجو در روش‌ها، روش‌ها مورد بررسی قرار گرفتند. در روش‌های بیان شده، روش گمنام‌سازی مجموعه‌ای[6] از روش گمنام‌سازی بازه‌ای[5] میزان اطلاعات بیشتری در اختیار کاربران قرار می‌دهد. چون به جای اینکه در هر دسته مقادیر رکورد با بازه جایگزین کند، آن را به صورت مجموعه‌ای بیان می‌کند. در روش مجموعه‌ای[6] به جای یکسان‌سازی رکوردها، از قرار دادن مجموعه مقادیر برای صفات استفاده می‌کند که در پاسخ به پرس‌وجو، جواب‌ها نزدیک‌تر به واقعیت خواهد بود.

پاسخ به پرس‌وجو در روش جایگشت[8] نسبت به روش گمنام‌سازی معمولی به واقعیت نزدیک‌تر است. در واقع به جای تغییر داده‌ها به مقادیر کلی، از جابه‌جایی مقادیر حساس رکورد استفاده می‌کند که در پاسخگویی به پرس‌وجو، جواب‌های نزدیک‌تر به واقعیت می‌دهد.

اگر چه روش‌های فازی بیان شده در مقالات[5][4] میزان اطلاعات بیشتری به کاربر می‌دهد و از دست رفتگی اطلاعات کمتر است ولی در مقابل احتمال فاش‌سازی را افزایش می‌دهد.

٥. نتیجه‌گیری و پیشنهادات

جهت جلوگیری از افشای اطلاعات، باید داده‌ها طوری تغییر یابد که علاوه بر نگه داشتن اطلاعات مفید، بحث حفظ حریم خصوصی افراد رعایت شود. البته هرچه داده‌های بیشتری را پنهان کنیم مسلما حریم خصوصی افراد بیشتر حفظ خواهد شد ولی در مقابل داده‌های مفید را از دست خواهیم داد. یک تعادل بین میزان اطلاعاتی که در اختیار کاربران قرار می‌گیرد و احتمال فاش‌شدن داده‌های حساس باید برقرار شود. در این زمینه یکی از مهمترین بحث‌ها این است که چطور بتوانیم پاسخ به یک پرس‌وجو را از روی داده‌های گمنام‌سازی شده با بهترین تقریب به جواب درست، در اختیار کاربر قرار دهیم. همچنان که دیدیم در این زمینه کارهای زیادی انجام شده است و هنوز هم این مسأله جزء مسائل باز در این حوزه می‌باشد.

منابع

[1] L. Yongcheng, Le. Jiajin and W.Jian, *Survey of Anonymity Techniques for Privacy Preserving*, International Symposium on Computing, Communication, and Control, Singapore, IACSIT,2011.

[2] M.Zare Mirakabad ,*A Framework for Privacy diagnosis and preservation in data publishing*, thesis for Doctor of philosophy, April 2010.

[3]X.Xiao, Y.Tao, *Dynamic Anonymization: Accurate Statistical Analysis with Privacy Preservation*, SIGMOD'08, June 9–12, Vancouver, Canada, 2008.

[4] V.Kumari, S.Rao, K.Raju, K. Ramana and B. Avadhani, *Fuzzy based approach for privacy preserving publication of data*, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.1, January 2008

Archive of SID
www.SID.ir

[5] R. Mukkamala ,V.G. Ashok, *Fuzzy-based Methods for Privacy-Preserving Data Mining*, Eighth International Conference on Information Technology,2011.

[6] Y.Liu, D.Lv, Y.Ye, J.Feng and Q. Hong, *Set-Expression based Method for Effective Privacy Preservation*, The Ninth International Conference on Web-Age Information Management,IEEE2008

[7] B.Srisungsittisunti and J.Natwichai, *An Efficient Algorithm for Incremental Privacy Breach on (k, e)-Anonymous Model*, 16th International Conference on Network-Based Information Systems,2013.

[8] Q.Zhang, N.Koudas,, D.Srivastava ,T. Yu, *Aggregate Query Answering on Anonymized Tables*, 23rd IEEE International Conference on Data Engineering, ICDE 2007

[9] S.Gokila, P.Venkateswari, *A Survey on Privacy Preservation Data Publishing*, International Journal on Cybernetics & Informatics (IJCI) Vol. 3, No. 1, February 2014.

[10] B.Funk, K.Wang, R.Chen and P. Yu, *Privacy-Preserving Data Publishing: A Survey on Recent Developments*, ACM Computing Surveys, 2011.

[11]P. R Bhaladhare, D. Jinwala, *A Sensitive Attribute based Clustering Method for k-anonymization*, Springer-Verlag Berlin Heidelberg 2012.

[12] M.Salari, S.Jalili and R.Mortazavi ,*A Utility Preserving Data-oriented Anonymization Method Based on Data Ordering*, 7th International Symposium on Telecommunications (IST'2014).

زیر نویس

[1] preservation privacy
[2] k-anonymity
[3] information loss
[4] query answerability
[5] privacy preservation
[6] information loss
[7] Identifier
[8] Sensitive attributes
[9] Quasi-identifier
[10] Normal attributes
[11] generalize
[12] نویز مناسب می‌تواند اضافه کردن مقداری به صفت حساس یا اضافه کردن چند تاپل باشد.
[13] Information loss
[14] Incremental privacy breach
[15] permutation
[16] k-anonymity
[17] Generalization
[18] help table

Archive of SID

www.SID.ir