بسمه تعالی

گزارش نهایی

عنوان طرح:

انطباق آماری رکوردها و بکارگیری آن در بهنگام سازی چارچوب کارگاههای کشور

مجری طرح:

محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

آبان ۱۳۸۷

عنوان طرح:

انطباق آماری رکوردها و بکارگیری آن در بهنگام سازی چارچوب کارگاههای کشور

مجری طرح: محسن محمدزاده

همکاران طرح: افشین فلاح، محسن ملانوری، شروین عسگری، حسن رنجی، شهرزاد پیشکاری

مشاور طرح: علی‌رضا زاهدیان

چکیده

هنگامی که اطلاعات جامع در مورد یک موضوع در چند مجموعه داده یا فایل قرار دارند، استفاده از یک مجموعه داده به معنی از دست دادن اطلاعات موجود در سایر مجموعه داده‌ها است. بنابراین یکپارچه ساختن اطلاعات پراکنده در مجموعه داده‌های مختلف می‌تواند بسیار سودمند باشد. در این راستا لازم است به نحوی رکوردهای یکسان در مجموعه داده‌های متفاوت شناسایی و فایلی حاوی اطلاعات کامل و منحصر به فرد تهیه گردد. علاوه براین گاهی در یک مجموعه داده رکوردهای تکراری وجود دارند و لازم است موارد تکراری شناسایی و حذف شوند. شناسایی واحدهای تکراری درون هر مجموعه داده یا واحدهای یکسان بین مجموعه داده‌های متفاوت را پیوند رکوردها نامند. در این تحقیق مراحل مختلف آماده سازی فایل‌ها از جمله بلوک‌بندی، استاندارسازی و مقایسه رشته فیلدها برای فایل‌های فارسی که با مشکلات خاصی همراه می‌باشند مورد بررسی قرار گرفته و راه حل‌های مناسب ارائه گردیده است. همچنین معیارهای تعیین فیلدهای قابل مقایسه، انتخاب آستانه‌ها، تعیین سطوح خطاهای پذیرفتنی، مبانی نظری پیوند رکوردها، الگوریتم‌های پیوند، نحوه برآورد پارامترها، انواع خطاهای انطباق و تحلیل رگرسیونی رکوردهای پیوندیافته بطور کامل مورد مطالعه قرار گرفته و نهایتاً روشی مناسب برای پیوند رکوردها پیشنهاد شده است. سپس براساس مدل پیشنهادی رکوردهای دو فایل سرشماری کارگاهی سال‌های ۱۳۷۳ و ۱۳۸۱ پیوند داده شده‌اند و از طریق بازبینی دستی نتایج حاصل، میزان دقت پیوند رکوردها ارزیابی شده و با استناد به آنها راهکارهایی برای افزایش دقت الگوریتم پیوند پیشنهاد شده است.

واژه‌های کلیدی: فیلد، انطباق، متغیر شناساگر، پیوند رکوردها، بازبینی دستی.

فهرست مندرجات

لیست اشکال

لیست جداول

فصل ۱

مقدمه

گاهی اطلاعات جامع در مورد یک موضوع در چند مجموعه داده یا فایل قرار دارند. از طرفی جمع‌آوری اطلاعات کامل مستلزم صرف زمان و هزینه زیاد خواهد بود. بنابراین یکپارچه ساختن اطلاعات پراکنده در مجموعه داده‌های مختلف می‌تواند بسیار سودمند باشد. در این راستا لازم است به نحوی رکوردهای یکسان در مجموعه داده‌های متفاوت شناسایی و فایلی حاوی اطلاعات کامل تهیه گردد. شناسایی واحدهای تکراری و یکسان درون و بین جوامع پیوند رکوردها نامیده می‌شود. چنانچه شناساگرهایی یکتا، خالی از خطا و قابل دسترسی برای هر یک از رکوردها وجود داشته باشد، مقایسه آنها برای شناسایی رکوردهای یکسان امکان پذیر خواهد بود. اما بدلیل آنکه فایلهای مختلف داده توسط افراد یا سازمانهای مختلف و با اهداف متفاوت تهیه می‌شوند، معمولاً شناساگرهایی با ویژگیهای یاد شده وجود ندارند و در صورت وجود آغشته به خطا هستند. بنابراین معمولاً از فیلدهای مشترک رکوردها که متغیرهای شناساگر نامیده می‌شوند، برای قضاوت در مورد تشابه رکوردها استفاده می‌شود. برای فایلهای حاوی اطلاعات شخصی، مهمترین متغیرهای شناساگر ممکن است نام، سن، جنس، آدرس و متغیرهایی از این قبیل باشند.

مفهوم پیوند رکوردها اولین بار توسط نیوکمپ (۱۹۵۹) برای ردیابی بیماریهای ارثی مورد استفاده قرار گرفت. در دهۀ ۱۹۶۰ پایه‌های نظری پیوند رکوردها توسط محققانی مانند ناتان (۱۹۶۸)، تپینگ (۱۹۶۸)، دوبویس (۱۹۶۹) و فلگی و سانتر (۱۹۶۹) بنا نهاده شد. این محققان روشهای ریاضی گوناگونی را برای پیوند رکوردها ارائه نمودند. نظریه فلگی و سانتر مهمترین نظریه در این زمینه است و در دهه‌های اخیر بیشترین توجه را به خود جلب نموده است. آرمسترانگ و همکاران (۱۹۹۳)، بلین (۱۹۹۳) و بلین و همکاران (۱۹۹۳) مساله برآورد نرخهای خطا را مورد بررسی قرار دادند. وینکلر(۱۹۹۳، ۱۹۹۴، ۱۹۹۵ و ۱۹۹۸) و جارو (۱۹۸۹ و ۱۹۹۵) مساله برآورد پارامترهای مدل فلگی و سانتر و امکان بهبود آنها را مورد مطالعه قرار دادند. لارسن (۱۹۹۹ و ۲۰۰۱) و لارسن و روبین (۲۰۰۱) استفاده از توزیعهای آمیخته را در پیوند رکوردها مطرح کردند. لاهیری و لارسن (۲۰۰۵) تحلیل رگرسیونی داده‌های پیوند یافته را مورد توجه قرار دادند. فورتینی و همکاران (۲۰۰۰ و ۲۰۰۲) و لارسن (۲۰۰۵) پیوند رکوردها با استفاده از رهیافت بیز و بیز سلسله مراتبی را مطرح نمودند.

بسیاری از کشورهای جهان از مدلهای مختلف پیوند رکوردها برای یکپارچه ساختن مجموعه داده‌های متفاوت در مورد یک موضوع یا شناسایی رکوردهای تکراری در یک مجموعه داده سود جسته‌اند، که در این بخش به چند نمونه از آنها اشاره می‌شود.

۱- اداره آمار آلمان در تلاش است که به جای سرشماری به روش سنتی، از این پس اطلاعات مورد نیاز خود را بصورت ثبتی و با جمع‌آوری اطلاعاتی که در ادارات و سازمانهای مختلف موجودند، بدست آورد. انگیزه اصلی برای اجرای سرشماری‌های ثبتی[1] آن است که در آلمان پس از جنگ جهانی دوم سه بار سرشماری نفوس و مسکن در سالهای ۱۹۵۰، ۱۹۶۱ و ۱۹۷۰ با موفقیت انجام شد. اما سرشماری بعدی که قرار بود در بهار سال ۱۹۸۱ اجرا شود، توسط دادگاه عالی این کشور لغو گردید. دلیل اصلی این تصمیم آن بود که در این کشور یک جریان عمومی بر علیه قانون

[1] Registration-Based Census

سرشماری عمومی نفوس و مسکن شکل گرفت و افراد زیادی از این قانون به دادگاه عالی شکایت کردند. به موجب این قانون انتقال اطلاعات حاصل از سرشماری به شهرداری و سایر ارگانهای حکومتی آزاد است. مخالفان این مساله را افشای اطلاعات فردی و عدول از امانت داری تلقی می‌کردند. چنین برداشتی هنوز هم وجود دارد، با این تفاوت که انتقال اطلاعات آزاد است، اما این اطلاعات نباید بصورت فردی باشد و اسامی اشخاص و هر اطلاعات دیگری که از طریق آن بتوان افراد را شناسایی کرد، باید حذف شوند. بیم سیاستمداران از شکل‌گیری جریانات مخالفی شبیه آنچه در قرن هجدهم شکل گرفت و هزینه‌های بسیار زیاد سرشماری‌های سنتی، آنها را به سمت طراحی روش‌هایی برای جمع‌آوری اطلاعات از مراکز مختلف انداخته است. پیوند رکوردها یکی از روشهایی است که در این زمینه مورد توجه و استفاده فراوان قرار می‌گیرد (زنزنشتاین، ۲۰۰۴).

۲- داده‌هایی که پس از وقوع تصادفات ثبت می‌شوند، معمولاً شامل اطلاعات کاملی در مورد صدمات ناشی از تصادفات نیستند. در سیستم ثبت داده‌های حاصل از تصادفات ایالت مینسوتای آمریکا[۲] فیلدهای اولیه‌ای مربوط به تصادفات براساس گزارشات پلیسی در مقیاس KABCO ثبت می‌شوند و اطلاعات مفید تنها زمانی بدست می‌آیند که این داده‌ها با داده‌های مربوط به جراحات و سوانح که در بیمارستانها و مراکز درمانی ثبت می‌شوند، پیوند داده شوند. فایل داده‌ای که از این پیوند حاصل می‌شود، می‌تواند در جهت شناسایی الگوهای احتمالاتی صدمات مربوط به تصادفات و سوانح خاص و مخارج مربوط به آنها بکار گرفته شود. بناونت و همکاران(۲۰۰۶) در مطالعه‌ای با استفاده از پیوند رکوردهای دو فایل گزارشات پلیس و گزارشات بیمارستانی به ارزیابی صدمات و خسارتهای ناشی از تصادفات پرداختند، که با بکارگیری هیچ یک از این دو گزارش به تنهایی ممکن نبود.

۳- مرکز جمعیت ایالت مینسوتا[۳] نمونه‌هایی از افراد و خانواده‌هایی که در سرشماریهای سالهای

[۲] CDS

[۳] (MPC)Minnesota Population Center

۱۸۶۰، ۱۸۷۰، ۱۹۰۰ و ۱۹۰۰ سرشماری شده‌اند را با نتایج حاصل از سرشماری سال ۱۸۸۰ پیوند داده است. نمونه‌های پیوند یافته امکان مطالعه بسیاری از ویژگی‌های جمعیت‌شناسی از جمله تحرک جغرافیایی، و مهاجرتها را فراهم آورده است(راگلس، ۲۰۰۲).

۴- بخش مراقبتهای همه‌گیرشناسی دانشگاه آکسفورد واقع در بخش سلامت عمومی فایل ملی شامل رکوردهای بیماری و مرگ و میر را پیوند داده است. فایل حاصل از این پیوند، می‌تواند در مطالعات همه‌گیرشناسی بسیار سودمند باشد. البته در این مطالعه برای جلوگیری از افشای اطلاعات، نامها و آدرسها از کلیه فایلهای پیوند یافته حذف شده‌اند.

۵- بخش مطالعات اجتماعی اداره آمار کانادا در سرشماری سال ۲۰۰۶ این کشور، پرسشی مبنی برکسب اجازه از پاسخگو در مورد پیوند رکوردهای اطلاعاتی آنها با سئوال ۱۵ فرم درآمد آنها در سرشماری ۲۰۰۵ پرسیده شد. ۸۵ درصد از پاسخگویان به این سئوال پاسخ مثبت دادند. سئوال مربوط به درآمد تنها از افراد بالای ۱۴ سال پرسیده شد. با توجه به اینکه پاسخهایی که در سرشماریهای گذشته به سئوال مربوط به درآمد داده شده است، تقریبی و غیر دقیق بوده است، اما در فرمهای مربوط به درآمد این گونه سئوالات معمولاً بطور دقیق ثبت می‌شوند، اطلاعات حاصل از این پیوند توانسته است بسیار سودمند باشد(بانکایِر، ۲۰۰۶).

۶- کارگروه ثبت اتوماتیک سرطانها[۴] در چارچوب فعالیت‌های فدراسیون اروپایی انجمنهای سرطان و زیر نظر آژانس اروپایی تحقیقات روی سرطان کار می‌کند. مطالعات این گروه در زمینه طراحی فرآیندی است که از طریق آن بتوان موارد سرطان را بدون دخالت انسان و مصاحبه با فرد مبتلا ثبت کرد. از طرفی با توجه به ملاحظات امنیّتی و اصول محرمانگی در کشورهای اروپایی از جمله آلمان، تنها قانونی کسب و دستیابی به این اطلاعات، اداره هشدارهای ایمنی و همچنین از طریق ارتباط با پزشکان است، که در هر دو مورد طبق قانون فیلدهای حاوی اطلاعات شخصی از رکوردها حذف می‌شوند. کارگروه مزبور تلاش می‌کند از طریق پیوند رکوردها اطلاعاتی را که از

[۴] Automated Cancer Registration

دو منبع بدست می‌آیند برای دستیابی به اطلاعات کاملتر مورد نیاز در تحقیقات علمی ترکیب کند. تاکنون این کارگروه کارگاه‌هایی آموزشی در این زمینه در سالهای ۲۰۰۳ و ۲۰۰۵ برگزار کرده است(کمدلمان و همکاران، ۲۰۰۴).

در این تحقیق مسألهٔ پیوند رکوردها مورد بررسی قرار گرفته است. در فصل دوم نحوه پیش پردازش فایل‌ها و مقایسه فیلدهای متناظر یک زوج رکورد قبل از انجام پیوند، مورد توجه قرار گرفته است. در فصل سوم از میان مدلهای موجود، یک مدل احتمالاتی کارا برای پیوند رکوردهای پیشنهاد و نحوهٔ برآورد پارامترهای این مدل برای حالتی که تعداد فیلدهای مورد مقایسه از ۳ بیشتر نباشد، ارائه می‌شود. در فصل چهارم با استفاده از توزیع‌های آمیخته پارامترهای مدل پیشنهادی براساس الگوریتم EM برآورد می‌شوند. در فصل پنجم بیش از ۳۰۰۰ رکورد از سرشماری‌های کارگاهی ۱۳۷۳ و ۱۳۸۱ براساس مدل پیشنهادی پیوند داده می‌شوند و در فصل ششم کارایی پیوند صورت پذیرفته مورد ارزیابی قرار خواهد گرفت. در پایان یافته‌های این تحقیق مورد بحث قرار گرفته و نتیجه‌گیری و پیشنهاداتی ارائه شده است.

فصل ۲

آماده سازی فایلها

قبل از آغاز فرآیند پیوند لازم است داده‌ها مورد پیش پردازش قرار گیرند، تا فایلهای مورد مقایسه با حذف ناسازگاری‌ها تا حد ممکن یک شکل شوند. پس از آن فیلدهای تشکیل دهندهٔ هر زوج رکورد نظیر به نظیر مقایسه و میزان همخوانی آنها مشخص می‌شود. چون فیلدهای تشکیل دهندهٔ یک رکورد دارای ماهیت‌های متفاوتی می‌باشند و ممکن است از کاراکترهای حرفی یا عددی تشکیل شده باشند، برای مقایسه آنها از مقایسه‌گرهای رشته‌ای استفاده می‌شود. در این فصل نحوه پیش پردازش داده‌ها و چگونگی استفاده از مقایسه‌گرهای رشته‌ای مورد بررسی قرار می‌گیرند.

۱.۲ پیش پردازش

پیوند رکوردها در عمل با چالشهای زیادی روبرو است. حجم زیاد اطلاعات، داده‌های آلوده، ناپایداری‌ها، موجود نبودن شناساگرهای یکتا و قابل اعتماد، مقادیر گمشده، خطاهای تایپی و سایر تغییرپذیری‌های غیرقابل اجتناب، طرحها و چهارچوبهای کدگذاری متفاوت، داده‌های تاریخ مصرف

گذشته و ضرورت حفظ محرمانگی از جمله این چالشها هستند. بنابراین قبل از مقایسه رکوردها لازم است داده‌ها مورد پیش پردازش قرار گرفته و برای مقایسه آماده شوند. بلوک‌بندی، تعیین فیلدهای مناسب برای مقایسه رکوردها و استانداردسازی از جمله مراحل مهم پیش پردازش داده‌ها هستند، که در این بخش به آنها پرداخته خواهد شد.

۱.۱.۲ بلوک‌بندی

یکی از مشکلات اساسی پیوند رکوردها، حجم زیاد اطلاعات یا بزرگی فایلهای تحت مطالعه است. هنگام مقایسه دو فایل A و B که به ترتیب دارای n_A و n_B رکورد هستند، تعداد زوج رکوردهای ممکن برابر $n_A \times n_B$ است. این تعداد مقایسه حتی با استفاده از رایانه‌های توانمند نیز کاری بسیار دشوار و وقت‌گیر است. معمولاً برای رفع این مشکل فایلهای تحت مطالعه براساس یک یا چند متغیر بلوکی، به فایلهای کوچکتر افراز یا بلوک‌بندی می‌شوند. متغیرهای بلوکی، متغیرهای مهم و تعیین کننده‌ای هستند که در صورت صحیح بودن اطلاعات آنها می‌توانند مبنای مقایسه رکوردها قرار گیرند. مثلاً برای فایلهای حاوی اطلاعات فردی متغیر بلوکی می‌تواند کد پستی، منطقه جغرافیایی یا چند حرف اول نام خانوادگی باشد. بلوک‌بندی، عملی ظریف است که اگر بدرستی انجام شود، منجر به کاهش قابل ملاحظه محاسبات و افزایش دقت پیوند رکوردها می‌شود. در صورتی که فایلها بلوک‌بندی شوند، فقط زوج رکوردهای درون هر بلوک با هم مقایسه می‌شوند، زیرا فقط برای رکوردهای درون هر بلوک امکان انطباق وجود دارد. نسبت انطباق‌ها و عدم انطباق‌های هر بلوک با سایر بلوکها متفاوت است. به عنوان مثال، چون برخی حاشیه‌نشینان شهرها پایدار و ساکن هستند و برخی جابجایی زیادی دارند، نرخ انطباق برای حاشیه نشینان شهرهای مختلف متفاوت می‌باشد. ویژگیهای هر بلوک در قالب متغیرهای انطباق بین بلوکها تغییر می‌کند. به عنوان مثال برخی اسامی در یک منطقه خاص عمومی‌تر و آگاهی بخش‌تر هستند، مانند نام‌های

عربی در مناطق جنوب غربی کشور و نام' سیروان' در مناطق کردنشین که رایجتر و آگاهی‌بخش‌تر از سایر نام‌ها می‌باشند.

۲.۱.۲ انتخاب فیلدهای مورد مقایسه

انطباق یا عدم انطباق یک زوج رکورد به میزان همخوانی فیلدهای تشکیل دهندهٔ آنها بستگی دارد. هر چه تعداد فیلدهای همخوان بیشتر و میزان ناهمخوانی در فیلدهای ناهمخوان کمتر باشد، دو رکورد انطباق بیشتری خواهند داشت. در این میان فیلدهای متفاوت دارای اهمیّت یکسانی نیستند و تأثیر هر فیلد به میزان آگاهی بخش بودن آن فیلد بستگی دارد. گاهی در نظر گرفتن یک فیلد ناآگاهی بخش، کارائی الگوریتم پیوند را کاهش می‌دهد. در بسیاری از موارد به این دلیل که اطلاعات یک یا هر دو فیلد مورد مقایسه ثبت نشده‌اند، مقایسه فیلدها عملاً امکان‌پذیر نمی‌باشد. این مساله موجب می‌شود تعداد فیلدهای سالمی که بتوان آنها را مورد مقایسه قرار داد، کاهش یابد. هرچه فیلد مهمتر و در پیوند رکوردها تاثیرگذارتر باشد، گم شدن اطلاعات مربوط به آن لطمه بیشتری به فرآیند پیوند رکوردها وارد می‌سازد. از طرفی با توجه به تعداد بسیار زیاد زوج رکوردهایی که در پیوند رکوردها مورد مقایسه قرار می‌گیرند، با افزایش تعداد فیلدهایی که برای مقایسه در نظر گرفته می‌شوند، حجم محاسبات به سرعت افزایش می‌یابد. انتخاب فیلدهای مناسبی که که بتوان آنها را در انطباق رکوردها مبنا قرار داد، یکی از مسائل مهم در مرحله پیش پردازش داده‌ها است ، که در ادبیات پیوند رکوردها انتخاب خصیصه[1] نامیده می‌شود.

[1] Feature Selectiuon

۳.۱.۲ استانداردسازی کدگزاری داده‌ها

در مرحله پیش پردازش لازم است کدگذاری[۲] داده‌ها به شکل‌های استاندارد تبدیل و کاراکترهای دارای چند شکل متفاوت، همشکل شوند. به عنوان مثال در فونت‌های فارسی، حروفی مانند «ک» و «ی» دارای اشکال متنوعی هستند. در بسیاری از موارد نیز به دلیل عدم رعایت قواعد ساده نگارش یا بی‌دقتی، فضاهای خالی و کاراکترهای غیرحرفی اضافی در یک رشته فیلد وجود دارند. بعلاوه برخی کلمات و اسامی دارای چندین نوشتار رایج هستند. بخصوص اگر فایل‌های مورد مقایسه مربوط به زمانهای مختلف باشند، این مشکل بیش از پیش به چشم می‌آید. به عنوان مثال «کاووس» و «داوود» به صورت «کاوس» و«داود» هم نوشته می‌شوند. در همه این موارد یکسان‌سازی شکل‌های متفاوت یک کلمه، حذف فواصل و کاراکترهای اضافی و اعمالی از این دست مرحله پیش پردازش داده‌ها را تشکیل می‌دهند. مشکلات مورد اشاره، برای سایر فیلدها از جمله فیلدهایی شامل تاریخ‌ها و شماره‌های خاص مانند شماره تلفن نیز صادق هستند.

با توجه به اینکه اهمیت پیش پردازش فیلدهای مختلف به یک اندازه نیست، معمولاً مرحله پیش پردازش برای فیلدی مانند آدرس علاوه بر دشواری از اهمیت بیشتری نیز برخوردار است. زیرا علاوه بر آنکه آدرس‌ها معمولاً به صورتهای مختلف نوشته می‌شوند و حتی یک نفر نیز ممکن است یک آدرس را در دو جای مختلف یه دو شکل متفاوت بنویسد، متاسفانه نحوه نوشتن آدرس به زبان فارسی در ایران از یک استاندارد مشخص پیروی نمی‌کنند. معمولاً برای نامگذاری یا شماره‌گذاری خیابانها، میادین و کوچه‌ها و پلاک‌گذاری واحدهای مسکونی، تجاری، فرهنگی و غیره در شهرهای مختلف الگوی یکسانی وجود ندارد. سیستم پستی اغلب کشورهای در حال توسعه از جمله ایران، آمیخته‌ای از روشهای مدرن و سنتی است. این آمیختگی با در نظر گرفتن میزان توسعه‌یافتگی متفاوت مناطق مختلف، بیشتر آشکار می‌شود. در بسیاری مناطق هنوز استفاده

[۲] Encoding

از کد پستی، کدملی و پلاک ثابت و سایر مولفه‌های دقیق پستی فراگیر نشده است.

۲.۲ الگوریتم‌های مقایسه‌گر رشته‌ای

برای مقایسه رشته فیلدها در رکوردهای دو فایل مختلف لازم است از یک الگوریتم مقایسه رشته‌ای استفاده شود. به دلیل عمومیت و گوناگونی خطاهای تایپی و اشتباهاتی که در ثبت اطلاعات وجود دارد، فرآیند پیوند رکوردها به الگوریتم‌های مؤثری برای مقایسه رشته کاراکترها نیازمند است. چگونگی برخورد با اینگونه تغییر پذیریها با توجه به ماهیت تصادفی آنها از اهمیت زیادی برخوردار است. چنانچه مقایسه رشته فیلدها بصورت حرف به حرف و دستی صورت پذیرد، بسیاری از انطباقهای واقعی ممکن است نادیده انگاشته شوند. لذا معمولاً از مقایسه‌گرهای رشته‌ای[۳] برای مقایسه فیلدها استفاده می‌شود. یعنی هر فیلد یک رشته از کاراکترها تلقی می‌شود و میزان شباهت کاراکترهای بکار رفته در دو رشته فیلد مبنای قضاوت در مورد همخوانی فیلدهای متناظر در رکوردها قرار می‌گیرد.

الگوریتم‌های مقایسه‌گر رشته‌ای زیادی با عناوین مختلف وجود دارد، که از آن جمله می‌توان به فاصله ویرایش[۴]، n–گرام‌ها[۵]، الگوریتم اسمایت–واترمن[۶] و الگوریتم جارو– وینکلر[۷] اشاره کرد. از میان این الگوریتمها، دو مورد فاصله ویرایش و الگوریتم جارو–وینکلر که از دقت زیادی برخوردار هستند(کوچین‌والا، ۲۰۰۱)، در اینجا مطرح می‌شوند. ورودی همهٔ این الگوریتم‌ها دو رشته کاراکتر و خروجی آنها معمولاً عددی بین صفر و یک است، که میزان همخوانی بین دو رشته

[۳] String Comparator
[۴] Edit Distance
[۵] n-grams
[۶] Smith-Waterman Algorithm
[۷] Jaro-Winkler Algorithm

کاراکتر را نشان می‌دهد.

۱.۲.۲ فاصله ویرایش

فاصله ویرایش بین دو رشته کاراکتر کمترین تعداد عملیات حذف یا درج کاراکتر بر روی یک رشته است، به نحوی که آنرا به رشته دیگر تبدیل کند. فاصله ویرایش دارای خاصیت تقارن است و فاصله هر رشته از خودش برابر صفر می‌باشد. اگر فاصله ویرایش دو رشته s_1 و s_2 با $ed(s_1, s_2)$ و طول رشته s با $\ell en(s)$ نمایش داده شود، کمیت $D = \frac{ed(s_1,s_2)}{\max(\ell en(s_1),\ell en(s_2))}$ فاصله ویرایش نسبی نامیده می‌شود، که به دو رشته s_1 و s_2، عددی بین صفر و یک اختصاص می‌دهد. این مقدار را می‌توان بصورت مستقیم یا پس از مقایسه با یک مقدار آستانه برای تصمیم‌گیری در مورد همخوانی یا عدم همخوانی دو رشته کاراکتر بکار برد. به عنوان مثال فاصله ویرایش دو رشته 'کوچه' و 'کوی' برابر ۳ است، زیرا برای تبدیل رشته اول به رشته دوم، بایستی دو حرف 'چ' و 'ه' حذف و حرف 'ی' درج شود. مقدار D برای دو رشته «خوار و بار فروشی محمد آقا و دوستان» و «خوار و بار فروشی محمد آقا و شرکاء » تا چهار رقم اعشار برابر ۰/۱۱۷۶ است. که حاکی از فاصله کم و شباهت زیاد بین دو رشته است.

در یک فرآیند پیوند رکوردها کلیه مراحل باید با یکدیگر سازگار باشند. الگوریتم مقایسهٔ رشته کاراکترها نیز باید با روشی که برای پیوند در نظر گرفته می‌شود، سازگار باشد. برخی از این مقایسه‌گرهای رشته‌ای برای پیوند رکوردها به روشهای غیراحتمالاتی مفیدند.

۲.۲.۲ الگوریتم جارو-وینکلر

از آنجا که روش فلگی و سانتر(۱۹۶۹) مبنای احتمالاتی دارد، می‌توان از الگوریتم جارو وینکلر که کوهن و همکاران(۲۰۰۳) نشان دادند در رده کاراترین الگوریتمها برای مقایسه رشته‌ها قرار دارد، استفاده نمود. این الگوریتم سه عمل حذف، درج و جابجایی کاراکترها را برای مقایسه دو رشتهٔ کاراکتری مد نظر قرار می‌دهد. مراحل این الگوریتم عبارتند از:

۱) طول رشته کاراکترها را مشخص سازید.

۲) تعداد کاراکترهای مشترک[۸] دو رشته یعنی کاراکترهای مشابهی که در نصف طول رشته کوتاهتر وجود دارد را مشخص کنید(Com).

۳) تعداد جابجایی‌ها[۹] را مشخص کنید($Trans$). یک جابجایی به کاراکتری اشاره دارد که نسبت به کاراکترهای مشترک متناظرش در رشته دیگر از ترتیب خارج است.

۴) فاصله دو رشته s_1 و s_2 را بصورت زیر محاسبه کنید.

$$Jaro(s_1, s_2) = \frac{1}{3}\left(\frac{Com}{\ell en(s_1)} + \frac{Com}{\ell en(s_2)} + 0/5 \times \frac{Trans}{Com}\right)$$

مقدار فاصله جارو وینکلر نیز عددی بین صفر و یک است و می‌تواند بصورت مستقیم یا پس از مقایسه با یک آستانه برای تصمیم‌گیری بکار رود. جدول۱.۲ فاصله ویرایش نسبی و جارو وینکلر را برای چند رشته فیلد نمونه نشان می‌دهد. در این تحقیق فاصله جارو-وینکلر با استفاده از نرم‌افزار برگرفته از کتابخانه Sim Metrics بدست آمده است، که از آدرس http://sourceforge.net/projects/simmetrics قابل دریافت است.

---

[۸] Common

[۹] Transposition

جدول ۱.۲: فاصله ویرایش نسبی و فاصله جارو–وینکلر برای چند رشته فیلد نمونه

رشته فیلد اول	رشته فیلد دوم	فاصله ویرایش نسبی	فاصله جارو–وینکلر
حسین	حسن	۰/۷۵	۰/۹۱۶
خوارو بار فروشی	خوار بار فروشی	۰/۹۲۴	۰/۸۶۰
خواروبار فروشی	نمایشگاه ماشین	۰/۲۳۱	۰/۷۶۹

۳.۲.۲ انتخاب آستانه برای مقایسه‌گر رشته‌ای

ورودی الگوریتم‌های مقایسه‌گر رشته‌ای، دو رشته کاراکتر و خروجی آنها عددی بین صفر و یک است که میزان شباهت آن دو رشته کاراکتر را نشان می‌دهد و معمولاً پس از مقایسه با یک آستانه، مبنای تصمیم‌گیری در مورد همخوانی یا عدم همخوانی دو رشتنه فیلد قرار می‌گیرد. مساله انتخاب آستانه برای مقایسه‌گرهای رشته‌ای یکی از حساسترین مسائل پیوند رکوردها است. این مقدار که توسط تحلیلگر و بصورت تجربی تعیین می‌شود، به میزان زیادی بر نتیجه پیوند رکوردها تأثیر گذار است. زیرا با کم یا زیاد کردن آن، تعداد فیلدهای همخوان در رکوردها، تعداد زوج رکوردهای منطبق و نامنطبق و در نتیجه نرخهای خطا تغییر می‌کند. مقدار آستانه برای مقایسه‌گر رشته‌ای ممکن است از فیلدی به فیلد دیگر با توجه به نوع فیلد و کیفیت ثبت اطلاعات در آن فیلد، متفاوت باشد.

فصل ۳

پیوند احتمالاتی رکوردها

روشهای مورد استفاده در پیوند رکوردها را می‌توان به دو دسته کلی روشهای معین[۱] و روشهای احتمالاتی[۲] تقسیم کرد. استفاده از روشهای معین، منوط به در اختیار داشتن شناساگرهای یکتا و خالی از خطا برای هر یک از رکوردها است. چون معمولاً چنین شناساگرهایی وجود ندارند، مدلهای احتمالاتی پیوند رکوردها که در آنها از فیلدهای مشترک فایلها برای قضاوت در مورد تشابه رکوردها استفاده می‌شود، مورد استفاده قرار می‌گیرند. روش فلگی و سانتر(۱۹۶۹) که در این بخش به آن پرداخته می‌شود، یکی از مهمترین و کاراترین روشهای احتمالاتی پیوند رکوردها است(گوماتام و همکاران، ۲۰۰۲). در این روش برای هر زوج رکورد برداری شامل وزنهای انطباق محاسبه می‌شود، سپس مجموع این وزنها با استفاده از دو آستانه برای دسته‌بندی زوج رکوردها مورد استفاده قرار می‌گیرد.

---

[۱] Deterministic

[۲] Probabilitistic

۱.۳ نظریه فلگی و سانتر

فرض کنید در دو فایل A و B که به ترتیب دارای n_A و n_B رکورد هستند، برخی رکوردها دارای اطلاعاتی مشترک از یک واحد آماری باشند. رکوردهای این دو فایل بترتیب با a و b و مجموعهٔ همه زوج رکوردهای ممکن آنها با $A \times B = \{(a,b); a \in A, b \in B\}$ نشان داده می‌شود. فرض کنید $M = \{(a,b) \in A \times B, a = b\}$، مجموعه تمام زوج رکوردهای منطبق و $U = \{(a,b) \in A \times B, a \neq b\}$، مجموعه همه زوج رکوردهای نامنطبق باشند. بدیهی است M و U دو مجموعهٔ ناسازگار و $A \times B = M \cup U$ است. با فرض آنکه هر رکورد شامل $k(\geq ۱)$ متغیر شناساگر باشد، مشاهدات آنها در زوج رکورد (a,b) به صورت دو بردار

$$x_a = (x_{a۱}, x_{a۲}, \ldots, x_{ak})$$

$$x_b = (x_{b۱}, x_{b۲}, \ldots, x_{bk})$$

نمایش داده می‌شوند. در حالت کلی مقایسه این متغیرهای شناساگر براساس تابعی مانند $\gamma_{ab} = f(x_a, x_b)$ صورت می‌پذیرد. معمولاً تابع مقایسه $f(x_a, x_b)$ برداری بصورت $\gamma_{ab} = (\gamma^۱_{ab}, \ldots, \gamma^k_{ab})$ در نظر گرفته می‌شود، که در آن

$$\gamma^h_{ab} = \begin{cases} ۱ & if \quad x_{ah} = x_{bh} \\ ۰ & o.w. \end{cases} \quad , \quad h = ۱, \ldots, k$$

فرض کنید Γ مجموعه تمام بردارهای مقایسه γ_{ab} باشد، که فضای مقایسه نامیده می‌شود. فلگی و سانتر(۱۹۶۹) برای تصمیم‌گیری در مورد انطباق یا عدم انطباق زوج رکورد (a,b)، ارزیابی توزیع بردارهای مقایسهٔ در دو مجموعه M و U را پیشنهاد نمودند. بدین منظور برای متغیر شناسایی γ^i_{ab} با توزیع احتمال

$$\begin{aligned} m(\gamma^i_{ab}) &= p(\gamma^i_{ab}|M) \\ &= P(\gamma^i_{ab} = \gamma | (a,b) \in M) \, , \quad \gamma = ۰, ۱ \end{aligned}$$