© Statistical Research and Training Center
مجلّه‌ی پژوهش‌های آماری ایران
دوره‌ی ۴، شماره‌ی ۱، بهار و تابستان ۱۳۸۶، صص ۹۱-۱۰۷
J. Statist. Res. Iran **4** (2007): 91-107

پیوند احتمالاتی رکوردهای فارسی با داده‌های گم‌شده

افشین فلاح و محسن محمدزاده*

دانشگاه تربیت مدرس

چکیده. پیوند رکوردها برای شناسایی واحدهای یکسان در یک یا چند مجموعه‌داده‌ی لاتین در مقالات متعدد مورد بررسی قرار گرفته و روش‌های مناسبی ارائه شده است. اما پیوند رکوردهایی که اطلاعات آن‌ها به‌زبان فارسی ثبت شده است، به‌دلیل ویژگی‌های خاص نوشتارهای فارسی و نبود استاندارد ثبت اطلاعات، با مسائل خاصی مواجه می‌باشد. در این مقاله ضمن معرفی پیوند رکوردها بر اساس یک مدل احتمالاتی، روش‌هایی برای آماده‌سازی فایل‌ها به‌روش استانداردسازی و بلوک‌بندی و انتخاب متغیرهای شناساگر ارائه می‌شوند، که پیوند احتمالاتی رکوردهای فارسی را میسر سازند. برای مقابله با داده‌های گم‌شده که از جمله‌ی مسائل مهم کاربردی در پیوند رکوردها محسوب می‌شوند، روش جدیدی پیشنهاد شده است، که احتمال وجود داده‌های گم‌شده را نیز در مدل پیوند رکوردها لحاظ می‌کند. سپس نحوه‌ی برآورد پارامترهای این مدل با الگوریتم EM ارائه شده است. برای افزایش تعداد فیلدهای قابل مقایسه نیز الگوریتمی مبتنی بر افراز فیلدهای مرکب ارائه گردیده است. سپس نحوه‌ی کاربست روش‌های ارائه‌شده برای پیوند احتمالاتی رکوردهای حاصل از سرشماری‌های کارگاهی در یک منطقه‌ی جغرافیایی ایران، نشان داده شده است.

واژگان کلیدی. رکورد؛ فیلد؛ انطباق؛ پیوند رکوردها؛ نسبت درست‌نمایی؛ الگوریتم EM.

۱ مقدمه

هنگامی که اطلاعات جامع در مورد یک موضوع در چند مجموعه‌داده یا فایل قرار دارند، استفاده از یکی از این مجموعه‌داده‌ها به‌معنی از دست دادن اطلاعات موجود در سایر مجموعه‌داده‌ها است. بنا بر این

* نویسنده‌ی عهده‌دار مکاتبات

یکپارچه ساختن اطلاعات پراکنده در مجموعه‌داده‌های مختلف می‌تواند بسیار سودمند باشد. در این راستا لازم است رکوردهای یکسان در مجموعه‌داده‌های متفاوت یا رکوردهای تکراری در یک مجموعه‌داده، به نحوی شناسایی و فایلی حاوی اطلاعات کامل و غیر تکراری تهیه شود. شناسایی رکوردهای یکسان درون یک مجموعه‌داده یا بین مجموعه‌داده‌های متفاوت، پیوند رکوردها نامیده می‌شود. چنان‌چه شناساگرهایی یکتا، خالی از خطا و قابل دسترسی برای هر یک از رکوردها وجود داشته باشند، با مقایسه‌ی آن‌ها شناسایی رکوردهای یکسان امکان‌پذیر خواهد بود. اما چون فایل‌های داده‌ها 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. بنا بر این معمولاً از فیلدهای مشترک رکوردها که متغیرهای شناساگر[1] نامیده می‌شوند، برای قضاوت در مورد تشابه رکوردها استفاده می‌شود. به‌عنوان مثال، برای فایل‌های حاوی اطلاعات شخصی، مهم‌ترین متغیرهای شناساگر ممکن است نام، سن، جنس، آدرس و متغیرهایی از این قبیل باشند. پیوند رکوردها اولین بار توسط نیوکامب و دیگران (۱۹۵۹) به‌عنوان یک مسئله‌ی آماری و برای ردیابی بیماری‌های ارثی مورد استفاده قرار گرفت. در دهه‌ی ۱۹۶۰ پایه‌های نظری پیوند رکوردها توسط فِلِگی و سانتر (۱۹۶۹) بنا نهاده شد. این محققان یک مدل ریاضی برای پیوند رکوردها ارائه نمودند. آرمسترانگ و مِی‌دا (۱۹۹۲)، بلین (۱۹۹۳) و بلین و روبین (۱۹۹۵) مسئله‌ی برآورد نرخ انواع خطاها را برای ارزیابی پیوند رکوردها مورد بررسی قرار دادند. وینکلر (۱۹۹۳، ۱۹۹۴، ۱۹۹۵ و ۱۹۹۸) و جارو (۱۹۸۹ و ۱۹۹۵) مسئله‌ی برآورد پارامترهای مدل فلگی و سانتر (۱۹۶۹) و امکان بهبود آن‌ها را مورد مطالعه قرار دادند. لارسن و روبین (۲۰۰۱) استفاده از توزیع‌های آمیخته را در پیوند رکوردها مطرح کردند. تمام مطالعاتی که تا کنون انجام شده در خصوص پیوند رکوردها در مجموعه‌ی اطلاعاتی است که به زبان لاتین تهیه شده‌اند؛ اما پیوند رکوردهایی که اطلاعات آن‌ها به زبان فارسی ثبت شده است، به‌دلیل ویژگی‌های خاص نوشتاری زبان فارسی، مانند متصل بودن حروف در کلمات، یکتا نبودن نوشتار حروف و متأثر بودن نوشتار حروف از محل قرار گرفتن آن‌ها در کلمات، دارای دشواری‌های ویژه‌ای است. به‌عنوان مثال، برخی حروف فارسی بسته به این‌که در ابتدا، میانه یا انتهای کلمات قرار گیرند، دارای اشکال متفاوتی هستند. به‌علاوه، برخی کلمات و اسامی دارای چندین نوشتار رایج هستند. به‌عنوان مثال، اسامی «داوود» و «کاووس» به‌صورت «داود» و «کاوس» هم نوشته می‌شوند. نحوه‌ی نوشتن نشانی‌ها نیز در زبان فارسی از استاندارد خاصی پیروی نمی‌کند.

در این مقاله مسائل مبتلابه پیوند رکوردهای فارسی در حضور داده‌های گم‌شده مورد بررسی قرار گرفته و برای هر یک، روشی مناسب ارائه گردیده است، به گونه‌ای که پیوند احتمالاتی رکوردها را میسر سازد. برای این منظور، در بخش ۲ مبانی نظری پیوند رکوردها و مدل احتمالاتی آن ارائه شده است. نحوه‌ی برآورد ماکسیمم درست‌نمایی پارامترهای مدل با استفاده از الگوریتم EM در بخش ۳ ارائه گردیده است. در

بخش ۴ مراحل مختلف آماده‌سازی فایل‌های فارسی شامل استانداردسازی، بلوک‌بندی و تعیین متغیرهای شناساگر، ارائه شده است. در این بخش، روشی برای لحاظ کردن مشاهدات گم‌شده در مدل پیوند رکوردها پیشنهاد شده و برآورد پارامترهای مدل جدید با استفاده از الگوریتم EM به دست آمده است. به‌علاوه برای استانداردسازی فیلد نشانی که از مشکل‌سازترین فیلدها در فرایند پیوند رکوردهای فارسی محسوب می‌شود، الگوریتمی مناسب و کارا پیشنهاد شده است. در بخش ۵ به‌عنوان یک مسئله‌ی کاربردی در مقیاس بزرگ، رکوردهای حاصل از سرشماری‌های کارگاهی کشور در سال‌های ۱۳۷۳ و ۱۳۸۱ برای یافتن رکوردهای مشترک که مورد نیاز مرکز آمار ایران می‌باشد، پیوند داده شده‌اند. سپس کیفیت پیوند صورت‌پذیرفته بر اساس نرخ‌های انطباق درست و نادرست مورد ارزیابی قرارگرفته و پیشنهادهایی برای ارتقای کیفیت پیوند این‌گونه رکوردها در بخش ۶ ارائه گردیده است.

۲ مدل احتمالاتی پیوند رکوردها

فرض کنید در دو فایل A و B که به‌ترتیب دارای n_A و n_B رکورد هستند، برخی رکوردها مربوط به واحدهای آماری یکسانی باشند. مجموعه‌ی همه‌ی جفت رکوردهای ممکن $A \times B = \{(a,b); a \in A, b \in B\}$، به مجموعه‌ی جفت رکوردهای منطبق $M = \{(a,b) \in A \times B, a = b\}$ و مجموعه‌ی جفت رکوردهای نامنطبق $U = \{(a,b) \in A \times B, a \neq b\}$، افراز می‌شود. رکوردهای هر جفت (a,b) توسط فیلدهای مشترک آن‌ها تحت عنوان متغیرهای شناساگر مورد مقایسه قرار می‌گیرند. با فرض آن‌که رکوردهای a و b، شامل $k(\geqslant ۱)$ متغیر شناساگر مشترک با تحقق‌هایی به‌صورت $x_a = (x_{a۱}, x_{a۲}, \ldots, x_{ak})$ و $x_b = (x_{b۱}, x_{b۲}, \ldots, x_{bk})$ باشند، این دو رکورد با استفاده از بردار $\gamma_{ab} = (\gamma_{ab}^۱, \ldots, \gamma_{ab}^k)$ مقایسه می‌شوند، که در آن

$$\gamma_{ab}^h = \begin{cases} ۱, & x_{ah} = x_{bh} \\ ۰, & x_{ah} \neq x_{bh} \end{cases} \quad , \quad h = ۱, \ldots, k$$

و x_{ah} و x_{bh} به‌ترتیب عبارت‌های (strings) مربوط به فیلد hام از رکوردهای a و b را نشان می‌دهند. به‌دلیل عمومیت و گوناگونی خطاهای حروف‌نگاری و اشتباهاتی که در ثبت اطلاعات وجود دارد، فرایند پیوند رکوردها به الگوریتم‌های مؤثری برای مقایسه‌ی عبارتی فیلدها نیازمند است. چگونگی برخورد با این‌گونه تغییرپذیری‌ها با توجه به ماهیت تصادفی آن‌ها از اهمیت زیادی برخوردار است. معمولاً از مقایسه‌گرهای عبارتی[۲] برای مقایسه‌ی عبارت‌ها استفاده می‌شود؛ یعنی هر فیلد، مجموعه‌ای از نویسه‌ها[۳] تلقی می‌شود و میزان شباهت نویسه‌های به کار رفته در دو عبارت، مبنای قضاوت در مورد همخوانی فیلدهای متناظر در رکوردها قرار می‌گیرد. مقایسه‌گرهای عبارتی زیادی با عناوین مختلف وجود دارد، که از آن جمله می‌توان

به فاصله‌ی ویرایش[۴]، n-گرام‌ها[۵]، الگوریتم اسمیت-واترمن[۶] و الگوریتم جارو-وینکلر[۷] اشاره کرد (کوهن و دیگران، ۲۰۰۳). ورودی همه‌ی این الگوریتم‌ها دو عبارت، و خروجی آن‌ها معمولاً عددی بین صفر و یک است که میزان همخوانی دو عبارت را نشان می‌دهد. کوهن و دیگران (۲۰۰۳) نشان دادند الگوریتم جارو-وینکلر در رده‌ی کاراترین الگوریتم‌ها برای مقایسه‌ی عبارت‌ها قرار دارد. این الگوریتم، سه عمل حذف و درج و انتقال نویسه‌ها را برای مقایسه‌ی دو عبارت، مد نظر قرار می‌دهد. مراحل این الگوریتم برای مقایسه‌ی دو رشته‌ی s_1 و s_2 عبارت‌اند از:

۱) مشخص‌سازی طول عبارت‌ها ($\mathrm{len}(s_1)$ و $\mathrm{len}(s_2)$)؛

۲) مشخص ساختن تعداد نویسه‌های مشترک دو عبارت، یعنی نویسه‌های مشابهی که در نصف طول عبارت کوتاه‌تر وجود دارند (com)؛

۳) مشخص ساختن تعداد انتقال‌ها[۸] (trans). یک انتقال، به نویسه‌ای اشاره دارد که نسبت به نویسه‌های مشترک متناظرش در عبارت دیگر، از ترتیب خارج است؛

۴) محاسبه‌ی فاصله‌ی جارو-وینکلر دو رشته به‌صورت زیر:

$$\mathrm{Jaro}(s_1, s_2) = \frac{1}{3}\left(\frac{\mathrm{com}}{\mathrm{len}(s_1)} + \frac{\mathrm{com}}{\mathrm{len}(s_2)} + 0.5 \times \frac{\mathrm{trans}}{\mathrm{com}}\right).$$

مقدار این فاصله، عددی بین صفر و یک است و می‌تواند از طریق مقایسه با یک آستانه‌ی معین، برای تصمیم‌گیری به کار رود.

فرض کنید γ_{ab} بردار مقایسه‌ای باشد که به‌تصادف از فضای مقایسه‌ی $\Gamma = \{\gamma_{ab}; a \in A, b \in B\}$ انتخاب شده است. قاعده‌ی احتمالی برای تعیین تعلق γ_{ab} به یکی از دو مجموعه‌ی M و U احتمال‌های $P(M|\gamma_{ab})$ و $P(U|\gamma_{ab})$ را مورد مقایسه قرار می‌دهد. در این صورت، جفت رکورد (a, b) به مجموعه‌ی M تعلق دارد اگر رابطه‌ی $P(M|\gamma_{ab}) \geqslant P(U|\gamma_{ab})$ یا (بر اساس قاعده‌ی بیز) رابطه‌ی $P(\gamma_{ab}|M) \geqslant P(\gamma_{ab}|U)$ برقرار باشد. در نظریه‌ی فِلِگی و سانتر (۱۹۶۹) برای تعیین انطباق یا عدم انطباق جفت رکورد (a, b)، توزیع شرطی بردار مقایسه‌ی γ_{ab} در دو مجموعه‌ی M و U بر اساس الگوریتم نسبت درست‌نمایی

$$w(\gamma_{ab}) = \log \frac{m(\gamma_{ab})}{u(\gamma_{ab})}, \qquad (۱)$$

تحت عنوان «وزن انطباق کل» مورد مقایسه قرار می‌گیرد، که در آن $m(\gamma_{ab}) = p(\gamma_{ab}|M)$ احتمال مشاهده‌ی γ_{ab} برای جفت رکوردهای منطبق (a, b) است و $u(\gamma_{ab}) = p(\gamma_{ab}|U)$ احتمال مشاهده‌ی

γ_{ab} برای جفت رکوردهای نامنطبق (a,b) می‌باشد. یکی از فرض‌های اساسی در مسئله‌ی انطباق رکوردها که محاسبات را به میزان قابل ملاحظه‌ای ساده می‌سازد، استقلال شرطی متغیرهای شناساگر در یک بردار مقایسه به‌شرط معلوم بودن وضعیت انطباق جفت رکوردهای متناظر با آن بردار است. بر اساس این فرض، توزیع بردار مقایسه‌ی γ_{ab} در مجموعه‌های M و U را می‌توان به‌صورت

$$m(\gamma_{ab}) = \prod_{i=1}^{k} m(\gamma_{ab}^i), \qquad u(\gamma_{ab}) = \prod_{i=1}^{k} u(\gamma_{ab}^i),$$

نوشت، که در آن $m(\gamma_{ab}^i) = p(\gamma_{ab}^i|M)$ و $u(\gamma_{ab}^i) = p(\gamma_{ab}^i|U)$ است و با توجه به (۱)، عبارت

$$w(\gamma_{ab}^i) = \log \frac{m(\gamma_{ab}^i)}{u(\gamma_{ab}^i)}, \qquad i = 1, \ldots, k$$

وزن انطباق فیلد iام نامیده می‌شود. در این صورت، وزن انطباق کل برای یک جفت رکورد، برابر با مجموع وزن‌های انطباق فیلدهای تشکیل‌دهنده‌ی آن رکورد به‌صورت

$$w(\gamma_{ab}) = \sum_{i=1}^{k} \log \frac{m(\gamma_{ab}^i)}{u(\gamma_{ab}^i)} = \sum_{i=1}^{k} w(\gamma_{ab}^i) \qquad (۲)$$

خواهد بود. قاعده‌ی پیوند فلگی و سانتر (۱۹۶۹) بر اساس وزن انطباق کل، هر جفت رکورد را به یکی از سه مجموعه‌ی جفت‌های منطبق، بلاتکلیف و نامنطبق اختصاص می‌دهد. شکل کلی این قاعده‌ی پیوند به‌صورت

$$d(\gamma_{ab}) = \begin{cases} \text{نامنطبق } (a,b), & w(\gamma_{ab}) < C_1 \\ \text{بلاتکلیف } (a,b), & C_1 < w(\gamma_{ab}) < C_2 \\ \text{منطبق } (a,b), & w(\gamma_{ab}) > C_2 \end{cases} \qquad (۳)$$

است، که در آن آستانه‌های بالا و پایین C_1 و C_2 بر مبنای سطوح از پیش انتخاب‌شده‌ی خطا تعیین می‌شوند و فاصله‌ی بین آن‌ها ناحیه‌ی بازبینی دستی نامیده می‌شود. برای این منظور، بردارهای فضای مقایسه‌ی Γ چنان مرتب می‌شوند که دنباله‌ی $\log \frac{m(\gamma_{ab})}{u(\gamma_{ab})}$ به‌طور یکنوا نزولی باشد. بردارهای مرتب‌شده‌ی γ_{ab} با $\gamma_1, \ldots, \gamma_N$ نشان داده می‌شوند. البته برای هر $\gamma_{ab} \in \Gamma$، اگر مقادیر $m(\gamma_{ab})$ و $u(\gamma_{ab})$ برابر با صفر باشند، آن جفت در نظر گرفته نمی‌شود. اگر $m(\gamma_i)$ و $u(\gamma_i)$ برای سادگی، به‌ترتیب با m_i و u_i نشان داده شوند، دو مقدار n و n' $(1 < n < n' - 1 < N)$ چنان انتخاب می‌شوند که نامساوی‌های

$$\sum_{i=1}^{n-1} u_i < \mu \leqslant \sum_{i=1}^{n} u_i, \qquad \sum_{i=n'}^{N} m_i \geqslant \lambda > \sum_{i=n'+1}^{N} m_i,$$

توأماً برقرار باشند. بدیهی است در مقایسه‌ی هر جفت رکورد، دو نوع خطا محتمل است: یکی این‌که دو رکورد را که واقعاً به واحد یکسانی تعلق دارند، نامنطبق بدانیم، و دیگر این‌که دو رکورد متعلق به دو واحد متفاوت را منطبق تلقی کنیم. احتمال‌های این دو نوع خطا به‌ترتیب، نرخ انطباق نادرست[۹] و نرخ عدم انطباق نادرست[۱۰] نامیده و با μ و λ نمایش داده می‌شوند. فلگی و سانتر (۱۹۶۹) نشان دادند که قاعده‌ی پیوند (۳) از این نظر که در سطوح معیّن خطا کوچک‌ترین ناحیه‌ی بازبینی را دارد، بهینه است.

۳ برآورد پارامترهای مدل با استفاده از الگوریتم EM

برای به کارگیری مدل احتمالاتی پیوند رکوردها که در بخش ۲ مورد بحث قرارگرفت، لازم است پارامترهای مدل برآورد شوند. فلگی و سانتر (۱۹۶۹) برای برآورد این پارامترها در حالتی که تعداد فیلدهای مورد مقایسه بیش‌تر از ۳ نباشد، روشی مبتنی بر اطلاعات پیشین در مورد فیلدهای مورد مقایسه ارائه دادند. این روش به‌دلیل دقت ناکافی و محاسبات زیاد از کارایی مطلوبی برخوردار نمی‌باشد. وینکلر (۱۹۹۸) استفاده از الگوریتم EM را برای برآورد پارامترهای مدل پیشنهاد نمود، که برای حالت‌هایی که تعداد فیلدهای مورد مقایسه زیاد باشد، می‌تواند از کارایی مناسبی برخوردار باشد. برای این منظور، توزیع بردار مقایسه‌ی γ_{ab} را درنظر بگیرید، که تحت فرض استقلال شرطی، توزیعی آمیخته به‌صورت

$$\begin{aligned} p(\gamma_{ab}) &= p(\gamma_{ab}|M)P(M) + p(\gamma_{ab}|U)P(U) \\ &= p\prod_{i=1}^{k} m_i^{\gamma_{ab}^i}(1-m_i)^{1-\gamma_{ab}^i} + (1-p)\prod_{i=1}^{k} u_i^{\gamma_{ab}^i}(1-u_i)^{1-\gamma_{ab}^i} \end{aligned}$$

و با پارامترهای $\mathbf{m} = (m_1, \ldots, m_k)$، $\mathbf{u} = (u_1, \ldots, u_k)$ و $p = P(M)$ می‌باشد. فرض کنید r_j جفت رکورد jام، $(j = 1, 2, \ldots, N)$، حاصل از مرتب کردن دنباله‌ی $\log \frac{m(\gamma_{ab})}{u(\gamma_{ab})}$ به‌صورت نانزولی و γ_j بردار مقایسه‌ی متناظر با این جفت رکورد باشد. اگر بردار داده‌ها با $\mathbf{y} = \langle \gamma_j, g_j \rangle$ نشان داده شود، که در آن متغیر نشانگر دوبعدی g_j به‌صورت

$$g_j = \begin{cases} (1, 0), & r_j \in M \\ (0, 1), & r_j \in U \end{cases}$$

تعریف می‌شود، لگاریتم درست‌نمایی به‌صورت

$$\log f(\mathbf{y}|\mathbf{m}, \mathbf{u}, p) = \sum_{j=1}^{N} g_j(\log p(\gamma_j|M), \log p(\gamma_j|U))^{\mathrm{T}} + \sum_{j=1}^{N} g_j(\log p, \log(1-p))^{\mathrm{T}}$$

خواهد بود. بدون آن‌که به کلّیت مسئله خللی وارد شود، فرض کنید $m_i > ۰$ و $u_i < ۱$. در مرحله‌ی Eی الگوریتم EM، عبارت $(P(M|\gamma_j), P(U|\gamma_j))$ به‌جای g جایگزین می‌شود، که در آن

$$P(M|\gamma_j) = \frac{p\prod_{i=۱}^{k} m_i^{\gamma_j^i}(۱-m_i)^{۱-\gamma_j^i}}{p\prod_{i=۱}^{k} m_i^{\gamma_j^i}(۱-m_i)^{۱-\gamma_j^i} + (۱-p)\prod_{i=۱}^{k} u_i^{\gamma_j^i}(۱-u_i)^{۱-\gamma_j^i}}$$

و $P(U|\gamma_j) = ۱ - P(M|\gamma_j)$ است. در مرحله‌ی M، لگاریتم درست‌نمایی بر حسب پارامترهای m_i و u_i و p ماکسیمم می‌شود و برآورد آن‌ها به‌صورت

$$\begin{aligned} \hat{m}_i &= \frac{\sum_{j=۱}^{N} \gamma_j^i \hat{P}(M|\gamma_j)}{\sum_{j=۱}^{N} \hat{P}(M|\gamma_j)}, \qquad i = ۱, \ldots, k \\ \hat{u}_i &= \frac{\sum_{j=۱}^{N} \gamma_j^i \hat{P}(U|\gamma_j)}{\sum_{j=۱}^{N} \hat{P}(U|\gamma_j)}, \qquad i = ۱, \ldots, k \\ \hat{p} &= \frac{\sum_{j=۱}^{N} \hat{P}(M|\gamma_j)}{N} \end{aligned} \tag{۴}$$

به دست می‌آید. در الگوریتم EM این مقادیر تا رسیدن به همگرایی در هر تکرار محاسبه می‌شوند.

۴ پیش‌پردازش فایل‌ها

برای اجرای الگوریتم پیوند رکوردها آماده‌سازی فایل‌ها برای دست‌یابی به نتایج قابل قبول، امری ضروری است. در این بخش، روش‌هایی برای آماده‌سازی فایل‌ها و الگوریتمی برای استانداردسازی نشانی‌های فارسی ارائه می‌شود، که به کمک آن می‌توان علی‌رغم وجود داده‌های گم‌شده، تعداد متغیرهای شناساگر را به نحو قابل ملاحظه‌ای افزایش داد.

آ) **استانداردسازی:** از آن‌جا که داده‌ها توسط افراد مختلف و با نرم‌افزارهای متفاوت ثبت می‌شوند، پیش‌پردازش و تبدیل داده‌های اولیه برای هم‌شکل کردن آن‌ها و از بین بردن تفاوت‌های ظاهری و فریبنده، امری ضروری است. با توجه به نوع داده‌ها و زبان ثبت اطلاعات، لازم است داده‌ها به شکل‌هایی استاندارد تبدیل وکاراکترهای دارای چند شکل متفاوت، هم‌شکل شوند. به‌عنوان مثال، در قلم‌های (fonts) فارسی، حروفی مانند «ک» و «ی» دارای شکل‌های متنوعی هستند. در بسیاری از موارد نیز به‌دلیل عدم رعایت قواعد ساده‌ی نگارش یا بی‌دقتی، فضاهای خالی و نویسه‌های غیر حرفی اضافی در یک عبارت وجود دارند. به‌علاوه برخی کلمه‌ها و اسم‌ها دارای چندین نوشتار رایج هستند. به‌خصوص هنگامی که فایل‌های مورد مقایسه مربوط به زمان‌های مختلف هستند، این مشکل بیش از پیش به چشم می‌آید. در همه‌ی

این موارد، یکسان‌سازی شکل‌های متفاوت یک کلمه، حذف فاصله‌ها و نویسه‌های اضافی، و اعمالی از این دست، مرحله‌ی پیش‌پردازش داده‌ها را تشکیل می‌دهند. مشکلات مورد اشاره، برای سایر فیلدها از جمله فیلدهایی مانند تاریخ‌ها و شماره‌های خاص، مانند شماره‌ی تلفن نیز صادق هستند. اهمیت پیش‌پردازش فیلدهای مختلف به یک اندازه نیست، معمولاً مرحله‌ی پیش‌پردازش برای فیلدی مانند نشانی، علاوه بر دشواری، از اهمیت بیش‌تری نیز برخوردار است؛ زیرا نحوه‌ی نوشتن نشانی به زبان فارسی در ایران از یک استاندارد مشخص پیروی نمی‌کند و ممکن است یک نشانی در دو فایل مختلف به دو شکل متفاوت درج شده باشد. از این رو استانداردسازی نشانی‌ها در پیوند رکوردهای فارسی از اهمیت ویژه‌ای برخوردار است.

ب) بلوک‌بندی: یکی از مشکلات اساسی پیوند رکوردها، حجم زیاد اطلاعات یا بزرگی فایل‌های تحت مطالعه است. هنگام مقایسه‌ی دو فایل A و B که به‌ترتیب دارای n_A و n_B رکورد هستند، تعداد جفت رکوردهای ممکن برابر با $n_A \times n_B$ است. این تعداد مقایسه حتی با استفاده از رایانه‌های توانمند نیز کاری دشوار و هزینه‌بر است. معمولاً برای رفع این مشکل، فایل‌های تحت مطالعه بر اساس یک یا چند متغیر بلوکی، به فایل‌های کوچک‌تر افراز یا بلوک‌بندی می‌شوند. متغیرهای بلوکی، متغیرهای مهم و تعیین‌کننده‌ای هستند که در صورت ثبت دقیق، می‌توانند مبنای مقایسه‌ی رکوردها قرار گیرند. بنا بر این، بلوک‌بندی، عملی ظریف است که اگر به‌درستی انجام شود، منجر به کاهش قابل ملاحظه‌ی محاسبات و افزایش دقت پیوند رکوردها می‌شود؛ زیرا فقط جفت رکوردهای درون هر بلوک با هم مقایسه می‌شوند.

پ) تعیین متغیرهای شناساگر در حضور داده‌های گم‌شده: یکی از مسائل مهم در پیش‌پردازش داده‌ها، تعیین فیلدهای به اندازه‌ی کافی آگاهی‌بخش برای مقایسه‌ی رکوردها است. در بسیاری از موارد، اطلاعات مندرج در فایل‌های مختلف ناقص هستند. وقتی اطلاعات دو فیلد مورد مقایسه، دست‌کم در یک رکورد، ناقص یا گم‌شده باشد، عملاً مقایسه‌ی این فیلدها امکان‌پذیر نمی‌باشد. با توجه به این که فیلدهای مورد مقایسه معمولاً از درجه‌ی اهمیت متفاوتی برخوردارند، هرچه فیلد مهم‌تر و در پیوند رکوردها تأثیرگذارتر باشد، گم‌شدگی اطلاعات مربوط به آن، لطمه‌ی بیش‌تری به فرایند پیوند رکوردها وارد می‌سازد. وجود داده‌های گم‌شده موجب می‌شود تعداد فیلدهای سالم قابل مقایسه و در نتیجه تعداد متغیرهای شناساگر کاهش یابد. از طرفی کارایی الگوریتم EM در برآورد پارامترهای مدل، تحت تأثیر مستقیم تعداد فیلدهای مورد مقایسه است و با کاهش تعداد این فیلدها کارایی این الگوریتم و در نتیجه کارایی فرایند پیوند رکوردها کاهش می‌یابد. در این‌جا برای غلبه بر این مشکل و افزایش تعداد فیلدهای قابل مقایسه، دو راهکار جدید پیشنهاد می‌کنیم.

جدول ۱. افراز فیلد نشانی به فیلدهای جزئی

کلمه‌های کلیدی	اجزای نشانی
محله، کوی	محله
ده، روستا، روستای، آبادی	روستا
میدان، فلکه	میدان
خ، خیابان، بلوار	خیابان
جاده، کمربندی، بزرگراه	جاده
کوچه، ک	کوچه
ده، روستا، پ، پلاک	پلاک
جنب، قبل، بعد، از، روبه‌روی	سایر

راهکار اول. فیلدهای مرکبی را که مقایسه‌ی آن‌ها با مشکلاتی توأم می‌باشد، می‌توان به‌صورتی مناسب به فیلدهای کوچک‌تر افراز نمود. با توجه به مسائلی که در مورد فیلد مرکب نشانی در زبان فارسی عنوان شد، می‌توان این فیلد مرکب را مطابق جدول ۱ به فیلدهای جزئی تجزیه و هر یک از آن‌ها را به‌عنوان متغیر شناساگر منظور نمود. برای این منظور، الگوریتم زیر ارائه می‌گردد:

۱) اجزای فیلد نشانی مطابق جدول ۱ در نظر گرفته می‌شوند؛

۲) کلمه‌های کلیدی مطابق جدول ۱ تعریف و مجموعه‌ی آن‌ها KW نامیده می‌شود؛

۳) اگر کلمه‌ی اول فیلد نشانی در مجموعه‌ی KW نباشد، به‌عنوان اسم شهر در نظر گرفته می‌شود؛

۴) اگر کلمه‌ی iام فیلد نشانی در مجموعه‌ی KW نباشد، تمام کلمات بعدی تا رسیدن به یکی از کلمات کلیدی یا رسیدن به پایان عبارت، به‌عنوان یک فیلد جزئی جدید در نظر گرفته می‌شوند.

راهکار دوم. در مدل فلگی و سانتر (۱۹۶۹) فیلدهای مشترک رکوردها بر اساس برداری دوحالتی مورد مقایسه قرار می‌گیرند، که فقط دو حالت همخوانی و عدم همخوانی را در نظر می‌گیرد؛ در حالی که وقتی در مشاهدات، داده‌های گم‌شده وجود دارند، برخی فیلدهای مشترک به‌دلیل گم شدن اطلاعات دست‌کم یکی از آن‌ها، قابل مقایسه نیستند و از این رو معمولاً فیلدهای حاوی مشاهدات گم‌شده از کلّیه‌ی رکوردها حذف می‌شوند. این عمل، تعداد فیلدهای مورد مقایسه و در نتیجه کارایی فرایند پیوند رکوردها را به‌شدت کاهش می‌دهد. در این‌جا برای جلوگیری از حذف فیلدهای حاوی اطلاعات گم‌شده، به‌جای متغیر نشانگر

دوحالتی، استفاده از متغیر نشانگر سه‌حالتی

$$\gamma_{ab}^{i}=\begin{cases}۱, & x_{ai}=x_{bi}\\ ۰, & x_{ai}\neq x_{bi}\\ -۱, & \text{فیلد } i\text{ام قابل مقایسه نیست}\end{cases}\quad,\quad i=۱,\ldots,k$$

پیشنهاد می‌شود. در این صورت، فیلدهای حاوی اطلاعات گم‌شده را نیز می‌توان به‌عنوان متغیر شناساگر در نظر گرفت و از حذف آن‌ها که کارایی فرایند پیوند رکوردها را کاهش می‌دهد، خودداری کرد. در این صورت، توزیع شرطی γ_{ab}^{i} برای جفت رکورد منطبق (a,b) به‌صورت

γ_{ab}^{i}	-۱	۰	۱
$p(\gamma_{ab}^{i}\|M)$	$۱-p_{mi}-q_{mi}$	q_{mi}	p_{mi}

و برای جفت رکورد نامنطبق (a,b) به‌صورت

γ_{ab}^{i}	-۱	۰	۱
$p(\gamma_{ab}^{i}\|U)$	$۱-p_{ui}-q_{ui}$	q_{ui}	p_{ui}

خواهد بود. با استفاده از نمادهایی که در بخش ۳ معرفی شدند و تحت فرض استقلال شرطی، روابط

$$p(\gamma_j|M)=\prod_{i=۱}^{k}p_{mi}^{I_{\{۱\}}(\gamma_j^i)}q_{mi}^{I_{\{۰\}}(\gamma_j^i)}(۱-p_{mi}-q_{mi})^{I_{\{-۱\}}(\gamma_j^i)}$$
$$p(\gamma_j|U)=\prod_{i=۱}^{k}p_{ui}^{I_{\{۱\}}(\gamma_j^i)}q_{ui}^{I_{\{۰\}}(\gamma_j^i)}(۱-p_{ui}-q_{ui})^{I_{\{-۱\}}(\gamma_j^i)}$$

برقرارند. بنا بر این برای یافتن برآوردهای ماکسیمم درست‌نمایی پارامترهای مدل با استفاده از الگوریتم EM، لگاریتم تابع درست‌نمایی را می‌توان به‌صورت

$$\begin{aligned}\log f(\mathbf{y}|\boldsymbol{\theta})=\sum_{j=۱}^{N}g_j\cdot\Bigg\{\sum_{i=۱}^{k}(&\log p_{mi}^{I_{\{۱\}}(\gamma_j^i)}+\log q_{mi}^{I_{\{۰\}}(\gamma_j^i)}\\&+\log(۱-p_{mi}-q_{mi})^{I_{\{-۱\}}(\gamma_j^i)}),\sum_{i=۱}^{k}(\log p_{ui}^{I_{\{۱\}}(\gamma_j^i)}\log q_{ui}^{I_{\{۰\}}(\gamma_j^i)}\\&+\log(۱-p_{ui}-q_{ui})^{I_{\{-۱\}}(\gamma_j^i)})\Bigg\}^{\mathrm{T}}+\sum_{j=۱}^{N}g_j\cdot(\log p,\log(۱-p))^{\mathrm{T}},\end{aligned}$$

نوشت، که در آن $\boldsymbol{\theta} = (\mathbf{p}_m, \mathbf{q}_m, \mathbf{p}_u, \mathbf{q}_u, p)$ بردار $(۴k + ۱)$-بعدی پارامترها را نشان می‌دهد و از بردارهای $\mathbf{p}_m = (p_{m۱}, \ldots, p_{mk})$، $\mathbf{q}_m = (q_{m۱}, \ldots, q_{mk})$ و $\mathbf{p}_u = (p_{u۱}, \ldots, p_{uk})$ $\mathbf{q}_u = (q_{u۱}, \ldots, q_{uk})$ تشکیل شده است. بدون آن که به کلّیت مسئله خللی وارد شود، فرض کنید برای $i = ۱, \ldots, k$، روابط $p_{mi} + q_{mi} < ۱$ و $p_{ui} + q_{ui} < ۱$ برقرار باشند. مرحله‌ی Eی الگوریتم EM، جایگذاری $E(g_j) = (P(M|\gamma_j), P(U|\gamma_j))$ به‌جای g_j، و مرحله‌ی M این الگوریتم، ماکسیمم‌سازی لگاریتم درست‌نمایی بر حسب پارامترها است که در آن

$$P(M|\gamma_j) = \frac{p \cdot p(\gamma_j|M)}{p \cdot p(\gamma_j|M) + (۱ - p) \cdot (\gamma_j|U)}$$

و $P(U|\gamma_j) = ۱ - P(M|\gamma_j)$. در این صورت می‌توان نشان داد که لگاریتم درست‌نمایی به‌ازای مقادیر $p = \frac{۱}{N} \sum_{j=۱}^{N} P(M|\gamma_j)$ و

$$\begin{aligned} p_{mi} &= \frac{G_{mi}^{-۱}}{\sum_{\alpha \in \mathcal{A}} G_{mi}^{\alpha}}, \qquad q_{mi} = \frac{G_{mi}^{۰}}{\sum_{\alpha \in \mathcal{A}} G_{mi}^{\alpha}}, \\ p_{ui} &= \frac{G_{ui}^{-۱}}{\sum_{\alpha \in \mathcal{A}} G_{ui}^{\alpha}}, \qquad q_{ui} = \frac{G_{ui}^{۰}}{\sum_{\alpha \in \mathcal{A}} G_{ui}^{\alpha}}, \qquad i = ۱, \ldots, k \end{aligned}$$

ماکسیمم می‌شود، که در آن‌ها $\mathcal{A} = \{-۱, ۰, ۱\}$ است و G_{mi}^{α} و G_{ui}^{α} برای $i = ۱, \ldots, k$، به‌صورت

$$\begin{aligned} G_{mi}^{\alpha} &= \sum_{j=۱}^{N} P(M|\gamma_j) I_{\{\alpha\}}(\gamma_j^i), \\ G_{ui}^{\alpha} &= \sum_{j=۱}^{N} P(U|\gamma_j) I_{\{\alpha\}}(\gamma_j^i), \qquad \alpha \in \mathcal{A} \end{aligned}$$

تعریف می‌شوند. در الگوریتم EM، برآورد ماکسیمم درست‌نمایی پارامترها پس از محاسبه‌ی مقادیر p_{mi}، q_{mi}، p_{ui}، q_{ui} و p به‌طور متوالی و تا رسیدن به همگرایی، به دست می‌آیند.

۵ شناسایی رکوردهای مشترک در سرشماری‌های کارگاهی

در این بخش، رکوردهای حاصل از سرشماری‌های کارگاهی در سال‌های ۱۳۷۳ و ۱۳۸۱ پیوند داده می‌شوند. این دو فایل داده‌ها به‌ترتیب حاوی $n_A = ۱۵۸۲۴۹$ و $n_B = ۱۵۱۷۹۰$ رکورد هستند. بر این اساس، پیوند رکوردهای این دو فایل، مستلزم مقایسه‌ی $۲۴,۰۲۰,۶۱۵,۷۱۰$ جفت رکورد است. این تعداد مقایسه حتی با استفاده از رایانه‌های توانمند نیز کاری بسیار

جدول ۲. تعداد رکوردها، مقایسه‌های لازم و نویسه‌های حذف‌شده در بلوک‌های منتخب

بلوک	تعداد رکوردها		تعداد مقایسه‌ها	تعداد نویسه‌های حذف‌شده
	سال ۱۳۷۳	سال ۱۳۸۱		
۱	۱۳۹	۱۵۴	۲۱۴۰۶	۲۶۶۱۲
۲	۳۲۰	۴۵۹	۱۴۶۸۸۰	۱۱۰۸۴۴۲
جمع	۶۱۳	۴۵۹	۱۶۸۲۸۶	۱۱۳۵۰۵۴

دشوار و وقت‌گیر است. هر رکورد در این دو فایل، شامل ۸ فیلد با عناوین نام کارگاه، نشانی، کد نشانی پستی، تلفن، شرح فعالیت، کد فعالیت، وضعیت حقوقی و تعداد شاغلان است. در این مقاله به‌منظور کاهش عملیات محاسباتی و افزایش سرعت محاسبات، فایل‌های دو سرشماری کارگاهی بر اساس سه متغیر کلیدی شهرستان و بخش و شهر، بلوک‌بندی شده‌اند. تعداد کل بلوک‌ها در سرشماری کارگاهی سال ۱۳۷۳ برابر با ۱۸۵ و در سال ۱۳۸۱ برابر با ۱۵۲ است. این به معنی آن است که در سرشماری سال ۱۳۸۱ برخی بلوک‌ها با یکدیگر ادغام شده‌اند و تعداد بلوک‌ها کاهش یافته است. هر بلوک بر اساس یک کد هشت‌رقمی مشخص می‌شود، که دو رقم اول آن بیانگر شهرستان، دو رقم دوم آن بیانگر شهر، و چهار رقم آخر آن بیانگر بخش می‌باشد. نتایجی که ارائه خواهند شد، فقط مربوط به پیوند رکوردهای دو منطقه‌ی جغرافیایی ۱۵-۰۲-۰۰۰۱ و ۰۸-۰۱-۰۰۰۱ به‌عنوان بلوک‌های ۱ و ۲ است. جدول ۲ تعداد رکوردهای این دو بلوک منتخب و تعداد مقایسه‌های لازم برای پیوند رکوردهای آن‌ها را نشان می‌دهد. پیوند رکوردهای این مجموعه‌داده‌ها بدون بلوک‌بندی، مستلزم ۲۸۱۳۶۷ مقایسه می‌باشد، در حالی که جدول ۲ نشانگر آن است که تعداد کل مقایسه‌های لازم برای پیوند رکوردها پس از بلوک‌بندی، برابر با ۱۶۸۲۸۶ است. همان‌طور که ملاحظه می‌شود، اعمال روش بلوک‌بندی منجر به کاهش تقریباً ۶۰ درصدی تعداد مقایسه‌ها شده است.

برای استانداردسازی فایل‌ها از استاندارد یونی‌کد استفاده شده که یک استاندارد بین‌المللی برای حروف مورد استفاده در زبان‌های رایج دنیا است. این استاندارد برای هر یک از زبان‌های رایج دنیا آستانه‌هایی را برای تشخیص نویسه‌های اضافی، ناشناس و نادرست، مانند فاصله‌های اضافی، خطوط فاصله و غیره تعیین می‌نماید. بر این اساس، کلّیه‌ی نویسه‌هایی که کد آن‌ها در جدول یونی‌کد، خارج از محدوده‌ی [۱۵۷۰, ۱۷۲۶] است، از عبارت‌های مورد مقایسه حذف شده‌اند. به این ترتیب، نویسه‌هایی مانند فاصله‌های اضافی و خطوط فاصله در فیلدها حذف شده‌اند. در دو بلوکی که در این تحقیق مورد بررسی قرارگرفته‌اند، در مجموع ۱۱۳۵۰۵۴ نویسه‌ی اضافی شناسایی و حذف شده‌اند. متأسفانه به‌دلیل وجود داده‌های گم‌شده‌ی بسیار زیاد در دو فایل سرشماری کارگاهی، تنها فیلدهایی که می‌توان از آن‌ها در پیوند رکوردها استفاده کرد، چهار فیلدِ نام کارگاه، نشانی، کد فعالیت، و تعداد کارکنان است. بنا بر این، مدل فلگی و سانتر (۱۹۶۹) دارای ۹ پارامتر است که با

جدول ۳. نتایج حاصل از پیوند رکوردهای بلوک ۱

شماره‌ی رکورد		
سال ۱۳۷۳	سال ۱۳۸۱	وضعیت انطباق
۱	۶۵	+
۱۵۱	۱۳۶	+
۵۷	۵۰	+
تعداد جفت رکوردهای منطبق		۳
تعداد جفت رکوردهای بلاتکلیف		۲۵
تعداد جفت رکوردهای نامنطبق		۲۱۳۷۸
نرخ انطباق نادرست		۰

در نظر گرفتن مقادیر اولیه‌ی $m_i = ۰/۱۵$، $u_i = ۰/۹$، $(i = ۱, \ldots, ۴)$، و $p = ۰/۱$ با استفاده از الگوریتم EM برآورد شده‌اند. تأثیر فیلدهای متفاوت بر انطباق یا عدم انطباق یک رکورد، یکسان نمی‌باشد. به‌عنوان مثال، فیلد نشانی غالباً شامل خطاهای بیش‌تری است و کم‌تر پیش می‌آید که دو رکورد منطبق، دارای نشانی‌های دقیقاً یکسانی باشند. در هر حال، یکی از نارسایی‌های روش فلگی و سانتر (۱۹۶۹) این است که به‌دلیل جمعی بودن وزن‌های انطباق، بین دو بردار وزن نوعی $\mathbf{m}_۱ = (۰/۱, ۰/۸, ۰/۶, ۰/۷)$ و $\mathbf{m}_۲ = (۰/۶, ۰/۷, ۰/۸, ۰/۱)$ که دارای مجموع اوزان برابر هستند، تفاوتی قائل نمی‌شود. پس از برآورد پارامترهای مدل، لازم است وزن‌های انطباق هر فیلد و وزن انطباق کل محاسبه شوند.

در مدل فلگی و سانتر (۱۹۶۹)، برای سطوح خطای مدل، μ و λ، نمی‌توان هر مقدار دلخواهی انتخاب نمود؛ بلکه سطوح خطا باید بر اساس تعریف فلگی و سانتر (۱۹۶۹) پذیرفتنی باشند، که به تعداد رکوردها و فیلدها وابسته‌اند و از مسئله‌ای به مسئله‌ی دیگر متفاوت می‌باشند. در ادامه، نتایج حاصل از پیوند رکوردهای درون دو بلوک منتخب ارائه می‌شوند.

آ) بلوک ۱: برآورد پارامترهای مدل با استفاده از الگوریتم EM و با در نظرگرفتن مقادیر اولیه‌ی مذکور به‌صورت $\mathbf{m} = (۰/۰۰۱۲, ۰/۰۰۸۸, ۰/۰۰۵۴, ۰/۹۴۴۰)$، $\mathbf{u} = (۰/۰۰۱۰, ۰/۰۰۸۱, ۰/۰۰۴۶, ۰/۹۴۷۰)$ و $p = ۰/۰۰۱۸$ حاصل شده‌اند. جدول ۳ نتایج حاصل از پیوند رکوردهای بلوک ۱ را برای سطوح $\lambda = ۰/۱ \times ۱۰^{-۳}$ و $\mu = ۰/۱$ و آستانه‌ی $۰/۹$ برای مقایسه‌گر عبارتی نشان می‌دهد. موارد انطباق درست با علامت + نشان داده شده‌اند. همان‌طور که ملاحظه می‌شود، تعداد ۳ جفت رکورد منطبق تشخیص داده شده‌اند، که پس از بررسی مشخص شد هر سه انطباق درست است. بنا بر این، نرخ انطباق نادرست، صفر می‌باشد. به‌علاوه، تعداد ۲۵ جفت رکورد بلاتکلیف تشخیص داده شده‌اند، که در بررسی دستی مشخص شد تمام آن‌ها نامنطبق هستند.

جدول ۴. نتایج حاصل از پیوند رکوردهای بلوک ۲

شماره‌ی رکورد		
سال ۱۳۷۳	سال ۱۳۸۱	وضعیت انطباق
۲۳۲	۵۱	−
۶۱	۸۵	+
۶۳	۱۳	+
۴۴	۲۲	+
۲۱۲	۱۰۶	+
۲۳	۱۵۰	−
۱۶۷	۱۸۵	+
تعداد جفت رکوردهای منطبق		۷
تعداد جفت رکوردهای بلاتکلیف		۶۱۱
تعداد جفت رکوردهای نامنطبق		۱۴۶۲۶۲
نرخ انطباق نادرست		۰٫۲۸۶

ب) بلوک ۲: برآورد پارامترهای مدل با استفاده از الگوریتم EM و با در نظرگرفتن مقادیر اولیه‌ی مذکور به‌صورت $\mathbf{m} = (۰٫۰۰۰۷, ۰٫۰۰۲۳, ۰٫۰۱۹۹, ۰٫۸۹۷۲)$، $\mathbf{u} = (۰٫۰۰۰۳, ۰٫۰۰۱۲, ۰٫۰۱۰۰, ۰٫۸۱۲۵)$ و $p = ۰٫۰۰۱۳$ حاصل شده‌اند. جدول ۴ نتایج حاصل از پیوند رکوردهای بلوک ۲ را برای سطوح $\mu = ۰٫۱$ و $\lambda = ۰٫۱ \times ۱۰^{-۳}$ و آستانه‌ی ۰٫۹ برای مقایسه‌گر عبارتی نشان می‌دهد. موارد انطباق درست و نادرست به‌ترتیب با علامت‌های + و − نشان داده شده‌اند. همان‌طورکه ملاحظه می‌شود، تعداد ۷ جفت رکورد منطبق تشخیص داده شده، که در ۵ مورد از آن‌ها انطباق درست صورت گرفته است؛ بنا بر این، نرخ انطباق نادرست، برابر با ۰٫۲۸۶ است. به‌علاوه، تعداد ۶۱۱ جفت رکورد بلاتکلیف تشخیص داده شده است، که باید مورد بررسی دستی قرار گیرند. تعداد جفت‌های نامنطبق نیز برابر با ۲۱۳۷۸ می‌باشد.

کیفیت داده‌های ثبت‌شده بر کارایی فرایند پیوند رکوردها بسیار مؤثر است و معمولاً عواملی مانند تعداد و نوع فیلدهایی که اطلاعات آن‌ها گم شده است و فاصله‌ی زمانی بین ثبت اطلاعات در دو مجموعه‌داده‌ی مورد مقایسه نیز می‌توانند بر خطاهای انطباق مؤثر باشند. هرچه ثبت داده‌ها دقیق‌تر و بر مبنای الگوهای مناسب‌تر صورت پذیرفته باشد، پیش‌پردازش آن‌ها ساده‌تر و نتایج حاصل دقیق‌تر خواهد بود. این مطلب در مورد فیلدهایی که آگاهی‌بخش‌ترند، مهم‌تر است و ثبت نامناسب اطلاعات یک فیلد از میزان آگاهی‌بخشی آن فیلد می‌کاهد. هرچند انتظار می‌رود «کد فعالیت» فیلدی آگاهی‌بخش باشد، اما تفاوت کدهای ثبت‌شده و وجود خطاهای زیاد از آگاهی‌بخش بودن آن کاسته است. تعداد کارکنان یک کارگاه نیز با محدودتر شدن

یا توسعه‌ی دامنه‌ی فعالیت آن کارگاه تغییر می‌کند. این فیلد فقط در شرایطی می‌تواند از آگاهی‌بخشی لازم برخوردار باشد که در فاصله‌ی زمانی بین دو سرشماری، تغییرات زیادی در آن صورت نپذیرفته باشد. مجموعه‌ی عواملی که به آن‌ها اشاره شد، موجب شده است تا نرخ انطباق نادرست در بلوک ۲ به ۰٫۲۸۶ افزایش یابد. فلاح و محمدزاده (۱۳۸۷) رکوردهای سرشماری عمومی نفوس و مسکن سال ۱۳۸۵ و رکوردهای طرح نمونه‌گیری از هزینه و درامد خانوار سال ۱۳۸۲ را که شامل مشکلات مشابه کم‌تری بودند، با نرخ انطباق نادرست حدود ۰٫۰۶ پیوند دادند. بنا بر این اگر فایل‌های مورد بررسی از کیفیت بهتری برخوردار باشند، انطباق با دقت بیش‌تری صورت می‌گیرد.

۶ نتیجه‌گیری

دقت مدل احتمالاتی پیوند رکوردها تحت تأثیر مستقیم تعداد فیلدهای قابل مقایسه قرار دارد. هنگامی که فیلدهای حاوی داده‌های گم‌شده از مجموعه‌ی رکوردهای مورد مقایسه حذف می‌شوند، تعداد متغیرهای شناساگر و کارایی فرایند پیوند رکوردها به‌شدت کاهش می‌یابد. بنا بر این، افزایش کارایی پیوند رکوردها مستلزم لحاظ نمودن احتمال گم شدن مشاهدات در مدل‌های احتمالاتی است. از طرفی با توجه به مشکلات فیلد مرکب نشانی در زبان فارسی، افراز آن به فیلدهای جزئی منجر به افزایش تعداد فیلدهای قابل مقایسه یا متغیرهای شناساگر و در نتیجه بیش‌تر شدن دقت و کارایی پیوند رکوردها می‌شود. با توجه به افزایش حجم تولید و ثبت اطلاعات، استفاده از پیوند رکوردها در زمینه‌هایی مانند اعتبارسنجی سرشماری‌ها، یکپارچه ساختن اطلاعات پراکنده و دست‌یابی به مجموعه‌داده‌های کامل‌تر، کشف کلاه‌برداری‌ها و اعمال غیر قانونی، یافتن رابطه‌ی بین بیماری‌های مختلف در علوم پزشکی، یافتن عوامل مشترک بین پدیده‌های اجتماعی و غیره توصیه می‌گردد.

سپاس‌گزاری

نویسندگان مقاله از پیشنهادها و نظرهای ارزنده‌ی داوران محترم که باعث اصلاحات سازنده‌ای در محتوا و ارائه‌ی بهتر مقاله شده است، کمال تشکر و قدردانی را دارند. از حمایت قطب علمی داده‌های ترتیبی و فضایی دانشگاه فردوسی مشهد نیز قدردانی می‌گردد.

توضیحات

۱. identifier

۲. string comparator

۳. characters

۴. edit distance

۵. n-gram

۶. Smith-Waterman

۷. Jaro-Winkler

۸. Transpositions

۹. false match rate

۱۰. false non-match rate

مرجع‌ها

فلاح، افشین؛ محمدزاده، محسن (۱۳۸۷). پیوند رکوردهای سرشماری عمومی نفوس و مسکن ۱۳۸۵ و نمونه‌گیری از هزینه و درامد خانوار. در مجموعه مقالات کنفرانس تحلیل یافته‌های سرشماری ۱۳۸۵، تهران.

Armstrong, J.B.; Mayda, J.E. (1992). Estimation of Record Linkage Models Using Dependent Data. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 853-858.

Belin, T. (1993). Evaluation of Sources of Variation in Record Linkage Through a Factorial Experiment. *Survey Methodology* **19**, 13-29.

Belin, T.; Rubin, D.B. (1995). A Method for Calibrating False-Match Rates in Record Linkage. *Journal of the American Statistical Association* **90**, 694-707.

Cohen, W.W.; Ravikumar, P.; Fienberg, S.E. (2003). A Comparison of String Distance Metrics for Name Matching Tasks, In *Proceeding of the IJCAL-2003 Workshop of Integration*, Available at http://secondstring.sourcedforge.net/doc/iiweb.pdf.

Fellegi, I.P.; Sunter, A.B. (1969). A Theory for Record Linkage. *Journal of The American Statistical Association* **64**, 1183-1210.

Jaro, M.A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association* **89**, 414-420.

Jaro, M.A. (1995). Probabilistic Linkage of Large Public Health Data Files. *Statistics in Medicine* **14**, 491-498.

Larsen, M.D.; Rubin, D.B. (2001). Iterative Automated Record Linkage Using Mixture Models. *Journal of the American Statistical Association* **96**, 32-41.

Newcombe, H.B.; Kennedy, J.M.; Axford, S.J.; James, A.P. (1959). Automatic Linkage of Vital Records. *Science* **130**, 954-959.

Winkler, W.E. (1993). Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 274-279.

Winkler, W.E. (1994). Advanced Methods for Record Linkage. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 467-472.

Winkler, W.E. (1995). Matching and Record Linkage. *Business Survey Methods*, Wiley, New York.

Winkler, W.E. (1998). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 667-671.

محسن محمدزاده
گروه آمار، دانشکده‌ی علوم پایه،
دانشگاه تربیت مدرس،
پل نصر، بزرگراه جلال آل احمد،
تهران، ایران.
پیام‌نگار: *mohsen_m@modares.ac.ir*

افشین فلاح
گروه آمار، دانشکده‌ی علوم پایه،
دانشگاه تربیت مدرس،
پل نصر، بزرگراه جلال آل احمد،
تهران، ایران.
پیام‌نگار: *fallahaf@modares.ac.ir*