پالایش داده رکورد های هویتی افراد بکمک تطبیق تقریبی آنها

رامین رهنمون
سازمان تامین اجتمائی- دفتر راهبری سیستم ها

چکیده

داده ها در دنیای واقعی پر از خطا، ناقص و ناسازگار هستند. هدف از پالایش داده ها[۱] رفع این نواقص اطلاعاتی است. سیستم هویتی در یک سازمان وظیفه شناسائی افراد شرکت کننده در آن سازمان را بر عهده دارد. بدلیل وجود خطا در هنگام تولید اطلاعات مشخصات هویتی، در عمل رکورد های هویتی متفاوتی برای یک فرد واحد تولید می گردد.

این مقاله بدنبال یافتن الگوریتمی جهت شناسائی رکوردهای متعددی است که به یک فرد در سیستم تعلق دارد تا امکان پالایش داده ها میسر گردد. برای این منظور از ایده فاصله لونشتین[۲] جهت میزان مشابهت بین دو رکورد استفاده شده است. فاصله لونشتین برای تعیین مشابهت بین هر دو رشته ای قابل استفاده است. اما در اینجا این الگوریتم در کاربرد خاص ذکر شده تخصصی شده است. اشتباهات رایج تایپی در زبان فارسی، اهمیت مکان کاراکتر ها در رشته ها، حذف اطلاعات رج زده شده از جمله مواردی است که به ایده اولیه فاصله لونشتین اضافه شده است.

از آنجا که در این مقاله مسئله تطبیق تقریبی رکوردها[۳] مطرح است، اهمیت فیلدهای هر رکورد در میزان مشابهت از طریق ضرایبی تعیین می شود. برای بدست آوردن ضرایب بهینه الگوریتم ژنتیکی مورد استفاده قرار گرفته تا از طریق یادگیری از نمونه های موجود بهینه ترین ضرایب محاسبه گردند. نتایج شبیه سازی موید بهبود قابل توجه سیستم در شناسائی رکورد های مشابه پس از اعمال نکات فوق است.

واژه های کلیدی : *فاصله لونشتین، الگوریتم ژنتیکی، ضرایب تطبیق شونده، مشابهت حروف فارسی*

۱ - مقدمه

پالایش داده ها یکی از اولین قدم ها در راه ساخت انباره داده ها[۴] است. داده های تولید شده در دنیای واقعی بدلایل متعدد ناقص، خطا دار و ناسازگار هستند. الگوریتم های پالایش داده برآنند تا با پر کردن مقادیر تهی در جداول، حذف خطاها و تصحیح ناسازگاری ها داده های پالایش شده برای ایجاد انباره داده ها را تولید کنند. [۵][۲]

یکی از منابع اصلی اطلاعاتی در سازمان های بزرگ سیستم اطلاعاتی مشخصات هویتی افراد تحت پوشش آن سازمان است. سازمان تامین اجتمائی ایران بعنوان بزرگترین سازمان بیمه ای کشور، میلیون ها ایرانی را تحت پوشش خود دارد و برای ارائه سرویس های مختلف به این افراد ناچار است تا از طریق سیستم هویتی مشخصات این افراد را برای شناسائی آنها نگه داری کند. بدیهی است هر بیمه شده از طریق شماره بیمه ای که به وی اختصاص داده شده قادر به استفاده از سرویس های این سازمان است. از آنجا که گستردگی واحدهای بیمه ای آن سازمان سراسر کشور را تحت پوشش خود قرار داده و شبکه انتقال داده متمرکزی نیز بین این واحدها وجود ندارد، این امکان بوجود آمده تا یک بیمه شده در سیستم هویتی چندین شماره بیمه داشته باشد.

---

۱ Data Cleaning
۲ Levenshtein
۳ Approximate Record Matching
۴ Data Warehouse

در این مقاله از ایده فاصله لونشتین برای شناسائی رکوردهای هویتی مختلفی که در واقع به یک فرد تعلق دارند، استفاده شده است. فاصله لونشتین تعداد حروف حذف ،اضافه یا جابجا شده بین دو رشته را محاسبه می کند. اما در کاربرد واقعی نکات دیگری نیز اهمیت پیدا میکنند. برای نمونه چه اشتباهات تایپی رایجی در زبان فارسی اتفاق می افتد؟ آیا برخی از اطلاعات توسط کاربران رج زده شده اند یا خیر؟ و سئوالاتی از این قبیل. عدم توجه به این نکات درصد موفقیت در شناسائی تقریبی رکوردهای مشابه را بسیار کم می کند.

در این مقاله ایده اولیه فاصله لونشتین برای شناسائی فیلد های دو یا چند رکورد هویتی که احتمال یکی بودن آنها وجود دارد بکار گرفته شده است. اما تغییراتی در این الگوریتم بر مبنای ماهیت داده ها داده شده است و در ضمن پیش پردازش هائی نیز برروی اطلاعات صورت می گیرد تا احتمال تطبیق افزایش یابد. باید توجه داشت که هدف از این مقاله تطبیق تقریبی رکوردهای هویتی است. پس تطبیق تقریبی دو فیلد مشکل را حل نخواهد کرد. در تطبیق بین دو رکورد مهمترین مسئله تعیین اهمیت هر یک از فیلدها است. کنترل اهمیت هر فیلد از طریق ضرایبی میسر است اما پرسش اصلی اینست که این ضرایب باید چه مقادیری به خود گیرد تا بهترین تطبیق تقریبی حاصل شود؟

برای پاسخ به این پرسش بود که ازیک الگوریتم یادگیری ماشینی با سرپرست استفاده شد. بدین ترتیب که نمونه هائی به سیستم یادگیری (بر مبنای الگوریتم های ژنتیکی) داده شد که یکی بودن آنها محرز بود ولی محتویات رکورد ها دارای تفاوت های جزئی بودند. الگوریتم ژنتیکی بر این پایه اقدام به یادگیری ضرایب مناسب کرد که مقایسه نتایج با تنظیم تجربی ضرایب مشخص کرد که روش ماشینی موفق تر عمل می کند.

در ادامه این بحث ابتدا در بخش ۲ به تعریف صورت مسئله کاربردی پرداخته شده، در بخش ۳ تعریفی ریاضی از فاصله لونشتین مطرح شده و تغییرات لازم نیز عنوان می شود. در بخش ۴ پس زمینه مطالعاتی در این حوزه عنوان شده و در بخش ۵ الگوریتم های ژنتیکی بحث شده است. در بخش ۶ به شرح الگوریتم پیشنهادی پرداخته و در نهایت پس از بیان شبیه سازی های انجام شده نتیجه گیری مطرح شده است.

۲-پایگاه اطلاعات هویتی بیمه شدگان سازمان تامین اجتمائی، مشکلات پیش رو

یکی از وظایف اصلی هر سازمان بیمه ای شناسائی بیمه شدگان خود است. سازمان تامین اجتمائی بعنوان بزرگترین سازمان بیمه ای کشور نیازمند شناسائی دقیق هویت بیمه شدگان خود می باشد. بدیهی است در صورت عدم شناسائی صحیح بیمه شدگان سرویس های ارائه شده به آنها از قبیل کمک های کوتاه مدت، بازنشستگی و مستمری، سابقه و غیره دچار خدشه های جدی خواهد شد. بدلیل گستردگی جغرافیائی واحد های بیمه ای در سراسر کشور و عدم ارتباط سیستم های کامپیوتری موجود در این مراکز، بیمه شدگان ممکن است از طریق مراکز متفاوت مکررا شناسائی شوند. به همین علت یک فرد ممکن است شماره های بیمه متعددی دریافت کند که این امر به دلایل ذکر شده سیستم های پایه را دچار خدشه جدی می سازد.

ازسوی دیگر برخی از بیمه شدگان نیز بدلیل نفع شخصی و یا عدم آگاهی در زمان تغییر کارگاه، خود را بعنوان بیمه شده جدید معرفی میکنند. باید توجه داشت که روند ثبت اطلاعات بعلت خطا در ورود اطلاعات همواره بدرستی صورت نمی گیرد. در عمل برای فرد معینی شماره های بیمه متعددی (که کلید دسترسی به فرد است) تولید شده و مشخصات هویتی فرد (همانند نام، نام خانوادگی، نام پدر، شماره شناسنامه، تاریخ تولد) در هر کدام از رکوردهای هویتی منتسب به آن شماره دارای اختلافاتی از نظر محتوی اطلاعات با دیگر رکوردها است. در جدول شماره یک آمار مربوط به خطاهای بارز ورود اطلاعات در یکی از شعبات بزرگ سازمان تامین اجتماعی ذکر شده است.

جدول یک – آمار مربوط به خطاهای بارز در فایل هویتی یکی از شعب سازمان تامین اجتماعی. تعداد کل شماره های بیمه ثبت شده در این شعبه ۶۷۳۹۱۴ بوده است. منظور از خطاهای بارز فیلد های خالی، پر شده با حروف بی معنی(برای مثال در فیلد نام وجود کاراکتر هائی مثل -، ؟، _ و غیره)، فیلدهائی با اطلاعات غیر قابل قبول(برای مثال نام پدر با محتویات پدر یا نامشخص) و فیلدهای با طول بسیار کم(برای مثال نام خانوادگی یک یا دو حرفی) است

مشخصه فیلد هویتی	درصد خطا های بارز در فیلد (بر حسب درصد)
نام خانوادگی	۰.۴۳
نام	۰.۵۵
نام پدر	۸.۹۳۴
شماره شناسنامه	۴.۸۰۰
تاریخ تولد	۲.۶۱۵
محل تولد	۹.۱۳۶

مطالعه آمار موجود در جدول یک نشان دهنده میزان گسترده خطا های آشکار در اطلاعات است که عمده دلیل آن عدم اطلاعات درست در زمان ورود داده به سیستم بوده است. بدلیل وجود این میزان خطای گسترده شناسائی مشخصات هویتی یک فرد که چندین بار به سیستم داده شده(و مسلما در هربار با خطاهائی صورت گرفته) نیازمند الگوریتمی هوشمند است که قادربه شناسائی رکوردهای تقریبا مشابه گردد.

۳ - فاصله لونشتین، ایده اولیه و چگونگی توسعه آن برای تطبیق رکوردها

فاصله لونشتین، روشی برای تشخیص مشابهت مابین دو رشته است. اگر یک رشته را با s و رشته دیگر را با t نمایش دهیم، آنگاه تعداد درج، حذف و تغییرات کاراکتر هائی که s را به t تبدیل میکند، فاصله لونشتین بین این دو رشته می نامند. برای مثال فاصله بین دو رشته " محمد "و" احمد" یک است. نوع توسعه یافته این روش به الگوریتم اسمیت-واترمن شهرت دارد که بویژه در زمینه آنالیزDNA مورد استفاده قرار می گیرد. در این الگوریتم میزان مشابهت بین عناصر DNA از طریق ماتریس مشابهت عناصر قابل تعیین است. وجود این ماتریس باعث می گردد تا رشته هائی که دارای کاراکترهای نزدیک به هم باشند نیز بعنوان رشته های یکسان در نظر گرفته شوند.[۱۰][۷]

از این ایده برای تطبیق دو رکورد استفاده می شود. در فرایند تطبیق رکورد هدف شناسائی رکوردهای متعددی است که به یک موجودیت تعلق دارند. این مسئله از سال ها پیش مورد توجه قرار گرفته است. تطبیق رکورد در دو فاز صورت می گیرد، فاز اول جستجو و فاز دوم تطبیق نام دارد. در فاز جستجو، ابتدا باید رکورد هائی را یافت که بلقوه احتمال یکی بودن آنها وجود داشته باشد و در فاز تطبیق این احتمال یکی بودن از طریق الگوریتمی(همانند الگوریتم اسمیت-واترمن) مورد تحقیق قرار می گیرد. این مقاله صرفا به فاز تطبیق پرداخته و چندان توجهی به فاز جستجو ندارد. اگر چه فاز جستجو خود اهمیت جدا گانه ای دارد اما نیازمند بررسی مستقلی است.[۹]

بدیهی است در تطبیق رکورد ها، هر رکورد شامل چندین فیلد بوده و درزمان تطبیق ابتدا محاسبه فاصله بین تک تک فیلدها صورت می گیرد. اما در زمان تعیین فاصله بین دو رکورد باید اهمیت هر یک از فیلدها را با کمک ضرایبی مشخص کرد. برای مثال اگرفیلد نام خانوادگی دو رکورد مشابهت زیادی دارند ولی نام پدر آنها تفاوت بیشتری دارد آیا این دو رکورد یکی هستند یا خیر؟

از سوی دیگر، فاصله لونشتین برای مقایسه بین دو رشته بدون توجه به ماهیت اطلاعاتی آن تعریف شده است. اما در کاربرد واقعی نکاتی وجود دارد که عدم توجه به آن منجر به عدم دقت در عملکرد شناسائی خواهد شد. برای مثال آیا در ورود اطلاعات نام خانوادگی احتمال اینکه پسوند یک فامیل در حین تایپ فراموش شده و یا بدرستی تایپ نشود بیشتر از اشتباه در تایپ حروف اول نام خانوادگی نیست؟ یا اینکه چه اشتباهات تایپی رایجی در حین ورود اطلاعات

وجود دارد. آیا حروف ر،ز و ژ اشتباها به جای یکدیگر بیشتر از دیگر حروف، تایپ می شوند یا خیر؟ توجه به این نکات کاربردی می تواند میزان دقت مشابهت تقریبی بین رکوردها را افزایش دهد.

۴ - پس زمینه علمی و مطالعات مربوط

همانطور که قبلا نیز اشاره شد، فرایند تطبیق تقریبی رکوردها معمولا در دو فاز جستجو و تطبیق انجام می شود. ایده های متعددی در فاز جستجو مطرح شده که از آن جمله می توان به رهیافت همسایگی مرتب شده، کد soundex ، ادغام و حذف تدریجی [۳]یا موازی و الگوریتم صف اولویت [۷]اشاره نمود. از آنجا که این مقاله بحث مهمی در این راستا ندارد، لذا تنها به ذکر نام روش ها بسنده می گردد.[۱۰]

در فاز تطبیق ایده های مختلفی مطرح شده که در این مقاله روش پایه فاصله لونشتین یا فاصله تصحیح[۵] مورد استفاده قرار گرفته است. علاوه بر این روش (که در بخش قبل توضیح داده شد) روش های دیگری نیز مطرح هستند[۶]:

- N-gramsشامل ترای گرم، بی گرم و یونی گرم بطور گسترده ای در زمینه تشخیص متن و روش های تصحیح تلفظ مورد استفاده قرار می گیرد N-gram .نمایش برداری است که شامل تمامی ترکیباتn حرفی در یک رشته است. بردار n-gram بردار مولفه ای برای تمامی ترکیبات ممکن از n حرف است. الگوریتم تطبیق با استفاده از تشکیل بردار n-gram برای دو رشته ورودی، تفاضل این دو بردار را محاسبه می کند. اگر میزان تفاضل از آستانه معینی کوچکتر بود دو رشته یکسان شناخته می شوند.
- الگوریتم : Jaro این الگوریتم بر پایه محاسبه طول رشته و مقایسه بین دو رشته از نظر تعداد کاراکتر های مشترک و تعداد جابجائی ها بنا نهاده شده است. حال بر اساس فرمول مشروحه در این الگوریتم میزان فاصله بین دو رشته محاسبه می گردد.
- کد گذاری : Soundex هدف عمده این کد گذاری ها دسته بندی کلمات با تلفظ مشابه در یک گروه یا کد واحد است بگونه ای که واژه های باتلفظ یکسان دارای کد یکسانی نیز باشند. بدیهی است که این کدگذاری برای هر زبان باید بصورت جداگانه تعریف گردد و کد گذاری پیشنهادی برای زبان انگلیسی قابل استفاده برای زبان فارسی نخواهد بود. این روش متکی به اشتباهات رایج در تلفظ واژه ها است.
- الگوریتم تطبیق بازگشتی فیلدها: در این الگوریتم توجه خاصی به ماهیت بازگشتی فیلد ها شده است. برای یک زوج رشته، میزان تطابق ۱ است اگر رشته های اتمیک یکسان باشند یا یکی مخفف دیگری باشد، در غیر اینصورت میزان تطبیق صفر خواهد بود. حال هر زیر فیلد از یک رشته با زیر فیلد رشته دیگر مقایسه شده تا بهترین میزان تطبیق بدست آید. [۸]

۵ - الگوریتم های ژنتیکی

اساس اندیشه الگوریتم ژنتیکی، استفاده از مفهوم وراثت بین نسل های جمعیت و بکارگیری آن بعنوان یک الگوریتم است. بنیان وراثت بر کروموزوم ها گذاشته شده است که صفات ارثی بین نسل ها را انتقال می دهند. اما این صفات بر اثر اشتباهاتی در عمل کپی گرفتن با تفاوت هائی به نسل بعد منتقل می شود که همین اشتباهات باعث تنوع در جمعیت می گردد. بر اساس تئوری انتخاب طبیعی داروین از بین یک جمعیت موجوداتی باقی می مانند که بیشترین تطبیق با شرایط محیطی خود را داشته باشند. اما وجود عوامل تصادفی که باعث تولید اشتباهات در کد های ژنتیکی می شوند، تنوع ژنتیکی در جمعیت را همواره ایجاد می کنند.

حال همین ایده در الگوریتم های ژنتیکی بکار رفته است. ابتدا جمعیتی از کروموزوم ها تعریف می گردد که هر کروموزوم بسته به مسئله کاربردی معنی خاص خود را دارد که معمولا بصورت رشته بیتی کد شده است. حال چرخه

---

[۵] Edit Distance

اصلی آغاز شده و در تولید هر نسل جدید با کمک اعمال عملگر های ژنتیکی کـه بـاعث تولیـد تغیـیرات تصـادفی مـی شوند، گروه جدیدی از کروموزوم ها تولید می گردند.سپس مکانیزم تطبیق با شرایط محیطی از طریق تابع تطبیق فعال شده و میزان انطباق هر کروموزوم با شرایط محیطی را تعیین می کند. در این چرخه کروموزوم هائی باقی خواهند ماند که تطابق بیشتری با شرایط محیطی خود را داشته باشند.

عوامل اصلی بکارگیری الگوریتم ژنتیکی در یک کاربرد خاص، تعریف مناسب کروموزوم، تعریف تـابع تطبیـق و از همـه مهمتر تعیین عملگر های ژنتیکی مناسبی است که قادر به تولید اعضا جدید مفیـدی بـرای جمعیـت جـاری باشـند. از جمله عملگر های مشهور می توان به جهش[٦]، کراس اوور، درج ، حذف و ... اشاره نمود.[١]

۶ – الگوریتم پیشنهادی برای تطبیق تقریبی رکورد های هویتی

بنیان الگوریتم پیشنهادی بر مبنای فاصله لونشتین قرار داده شده است. این الگوریتم را می توان بـه دو بخـش تقسـیم نمود. بخش اول میزان مشابهت بین دو فیلد معین از رکورد هویتی و بخش دوم مشابهت بین دو رکورد است.

در مشابهت بین دو فیلد در این کاربرد خاص چند نکته حائز اهمیت است و عدم توجه به آن منجـر بـه کـارائی پـایین الگوریتم اولیه خواهد شد.

- در هر فیلد ممکن است داده های غیر قابل قبول بصورت عمدی و یـا سـهوی وارد سیسـتم شـده باشـند. در موارد سهوی اشتباه تایپی و یا نا آگاهی پانچیست عامل اصلی ورود اطلاعات بوده و از آنجا که برخـی کنـترل های منطقی در سیستم اجرائی موجود در نظر گرفته نشده در طول زمان اینگونه خطاهـا وارد سیسـتم شـده اند. اما موارد عمدی در دو حالت روی می دهند. حالت اول زمانی است که بدلیل ناقص بـودن اطلاعـات داده شده به پانچیست و الزام پر شدن فیلد در نرم افزار اجرائی، پانچیست به سـلیقه شـخصی خـود داده هـائی را وارد سیستم کرده باشد. اما در حالت دوم داده ها بعلت اهمال پانچیست رج زده شده اند. در جدول دو نمونـه های واقعی از این موارد ذکر شده است. شناسائی و حذف این داده ها از سیستم تا حد زیادی از مشابهت های بیمورد یا عدم تشابه غلط جلوگیری خواهد کرد.
- در الگوریتم توسعه یافته اسمیت-واترمن میزان تشابه بین کاراکترها در ماتریسـی ذخـیره میـگردد. حـال در زبان فارسی باید معین نمود چه اشتباهات رایجی در حین تـایپ صـورت مـی گیـرد و بـر ایـن مبنـا مـیزان مشابهت بین کاراکترها را تعیین کرد.

بطور تجربی میتوان حدس زد که برای مثال در نام خانوادگی بیشتر اشتباهات تایپی در پسوند فامیلی صـورت میگیـرد (به جدول سه نگاه کنید). تعیین دقیق این مکان می تواند دقت الگوریتم تطبیق تقریبی را بالا تـر بـرد. امـا در زمینـه تطبیق دو رکورد، بدیهی است هر رکورد شامل چندین فیلد است. میزان مشابهت بین دو فیلد مشابه از دو رکورد را می توان تعیین نمود. اما در تطبیق دو رکورد مسئله مهم تعیین اهمیت هر یک از فیلد ها در میزان مشابهت است. این امـر از طریق تعیین ضرایب برای هر فیلد امکان پذیر است. اما نکته مهم معین نمودن ضریب هر فیلد است. در بررسی اولیه این ضرایب بصورت تجربی تعیین شد. اما در عمل تعیین تجربی ضرایب منجر به عدم دقت در میزان مشـابهت بیـن دو رکورد می گردد. به همین دلیل ضرورت تعیین این ضرایب از طریق یک الگوریتم یادگیری ماشینی دیده شد. برای این منظور از الگوریتم های ژنتیکی استفاده شده که در بخش های بعدی به شرح آن می پردازیم.

---

[٦] Mutation

جدول دو – نمونه ای از چند رکورد هویتی با مشخصات مشابه

شماره شناسنامه	تاریخ تولد	نام پدر	نام	نام خانوادگی
۳۸۳۰	۱۳۵۸/۰۷/۱۲	وجیهه اله	حمید	چراغی
۲۸۳۰	۱۳۳۷/۰۱/۰۱		جمید	چراغی
۱۵۳	۱۳۴۹/۰۹/۰۲		محمدرضا	چراغعلی
۱۵۳	۱۳۴۷/۰۹/۰۲	۰	محمدرضا	چراغعلی خانی
۱۲	۱۳۵۶/۰۵/۰۶	میراسد	سیدمهدی	چاوش باشی
۱۱	۱۳۵۶/۰۵/۰۶	میراسد	سیدمهدی	چاوشباس
۲۳۸۰	۱۳۰۰/۰۰/۰۰		۰جلیل	چرمی فر
	۱۳۰۰/۰۰/۰۰		جلیل	چرمی

جدول سه – نمونه هائی از اشتباهات تایپی فامیل افراد

درخوش بایع کلائی	حـــق شـــناس انزابی	حق شناس امیر هنده
درخوش بایع کلانی	حـــق شـــناس انزالی	حـــق شـــناس امـــیر هندهی
درخـــوش بـــایع کلایی		

۶ - ۱- مشابهت بین کاراکتر های فارسی

یکی از نکات مهم در الگوریتم اسمیت-واترمن تعیین میزان مشابهت بین کاراکترهای رشته است. چـون ایـن مقالـه در مورد رشته های فارسی به بحث پرداخته، می بایست میزان این مشابهت در این نوع کاراکتر ها معین گردد. بـرای ایـن منظور ابتدا توده ای از نمونه های تایپ شده رشته های فارسی تهیه شده و سپس اشتباهات تـایپی (مربـوط بـه تـایپ اشتباهی یک حرف به جای حرف دیگر)آن مشخص شد. بر پایه این آزمون عملی حروفی که بیشترین اشتباه تایپی بین آنان صورت گرفته بود بعنوان مشابه ترین حروف شناخته شدند. نمونه ای از مشابهت های یافت شـده بیـن حـروف در جدول شماره چهار ذکر شده است.

همانطور که از محتویات رایج ترین اشتباهات می توان دریافت(جدول چهار) عمده اشتباهات تایپی در نمونه گیری فوق مربوط به نزدیکی مکان کاراکتر ها برروی صفحه کلید است. البته این نمونه گیری مربوط به تایپ اطلاعات در سیسـتم هویتی افراد است و اطلاعات بر مبنای سند های دستی که در اختیار کاربران قرار داده شده وارد سیسـتم شـده اسـت. بدیهی است در مواردی که ورود اطلاعات بصورت شنیداری است ممکن است اشتباهات متفاوت باشند. اما ایـن آزمـون نشان می دهد در مواردیکه ورود اطلاعات بر پایه سند های دستی انجام می گیرد عمـده اشـتباهات تـایپی مربـوط بـه نزدیکی حروف تایپ شده برروی صفحه کلید بوده است.

۶- ۲ - تطبیق تقریبی رکوردها وتعیین ضرایب اهمیت هر یک از فیلد ها

پس از تکمیل الگوریتم تطبیق تقریبی دو فیلد متناظر از یک رکورد باید به تطبیق تقریبی بین رکورد ها پرداخـت. اگر S_{ijk}معرف میزان مشابهت فیلد k از دو رکورد i و j باشد، آنگاه $\sum_{k\in R} S_{ijk}$: مـیزان تطبیـق تقریبـی کلـی دو رکـورد خواهد بود که در رابطه R معرف دامنه فیلد های یک رکورد است. اما جمع ساده این مقادیر نمی تواند نتیجـه منطقـی به همراه داشته باشد. برای مثال در رکورد های هویتی که در این مقاله مورد مطالعه قرار گرفته معمولا در اکـثر مـوارد تاریخ تولد به اشتباه وارد سیستم شده، چرا که سیستم ورود اطلاعات کاربر را وادار به پر کـردن ایـن فیلـد کـرده و در اسناد دستی موجود اکثرا این فیلد فاقد اطلاعات است. بدیهی است اپراتور در چنین مواردی مجبـور بـه رج زدن شـده

است. پس باید از ضرایبی برای این منظور استفاده کرد. رابطه اصلی تطبیق تقریبی رکوردها که در این مقاله پیشنهاد شده بصورت زیر است:

$$Sim\ (i,j) = \frac{\sum_{k \in R} \left[1 - \frac{d_{ijk}}{Max\ \{Len_{ik}, Len_{jk}\}} \right] B_k}{\sum_{k \in R} B_k}$$

در این رابطه $Sim(i,j)$ میزان مشابهت بین دو رکورد i و j، d_{ijk} فاصله لونشتین توسعه یافته بین دو رکورد i و j در فیلد k، Len_{ik} معرف طول فیلد k از رکورد i بوده و B_k ضریب تقویت یا تضعیف فیلد k است(که در بازه صفر تا صد تغییر میکند.)

برای یافتن مقادیر ضرایب برای هر فیلد رکورد (B_k) از بین 100^5 ترکیب مختلف از ۵ ضریب متفاوت برای هر فیلد یک رکورد، می بایست ۵ مقدار معین انتخاب گردند. از آنجا که امکان آزمون تمامی این حالات از نظر زمان محاسبه امکان پذیر نیست، لذا نیازمند استفاده از الگوریتمی یادگیرنده هستیم تا قادر به یافتن مقادیر بهینه بر مبنای نمونه آموزشی داده شده به سیستم گردد که در این بررسی از الگوریتم ژنتیکی ساده استفاده شده است.

در الگوریتم ژنتیکی مورد استفاده، هر ژن یک مجموعه پنج تائی است که هر یک از اعضا مقدار یکی از ضرایب پنج گانه تطبیق فیلدهای رکورد است. از آنجا که در این کاربرد ژن از اعداد تشکیل شده عملگر های کراس اوور و جهش بسادگی و بر طبق الگوریتم های کلاسیک مطرح شده برای این دو عملگر پیاده سازی شده است.

تابع تطبیق بر پایه ضرایب انتخابی در هر ژن و بر طبق رابطه مطرح شده برای تطبیق تقریبی رکوردها تعریف شده است. برای این منظور نمونه ای آموزشی از مواردی که رکوردها یکسان هستند به سیستم داده شده و سپس بر مبنای میزان موفقیت الگوریتم تطبیق تقریبی مقدار تابع تطبیق تعیین می گردد.

جدول چهار – نمونه های مهم از مشابهت بین حروف به همراه در صد مشابهت آنها. کل نمونه حروف اشتباه تایپ شده ۹۰۱۱ حرف بوده است. در این جدول تنها ۱۰ مورد از اشتباهات رایج ذکر شده است.

میزان مشابهت به درصد	حرف دوم	حرف اول
۰.۹	ت	ا
۱.۶	ن	ت
۱.۴	ح	ج
۲.۱	خ	چ
۱.۴	خ	ح
۵.۷	ز	ر
۳.۷	ش	س
۰.۹	ض	ص
۲.۴	غ	ع
۷.۵	ن	م

۷ - شبیه سازی

برای انجام شبیه سازی و تعیین ضرایب برروی سیستم ابتدا می بایست نمونه آموزشی را تولید نمود که در آن رکوردهای مشابه معین شده باشند و از طریق آن سیستم قادر به یادگیری ضرایب گردد. برای این منظور از فایل تغییرات مشخصات هویتی استفاده شده است. توضیح آنکه در سیستم کنونی ثبت مشخصات هویتی افراد در سازمان تامین اجتمائی امکان تغییر مشخصات افراد وجود دارد اما این تغییرات انجام شده هر بار بصورت یک رکورد جدید ثبت می شود تا امکان کنترل تغییرات فراهم گردد. از آنجا که کلید شماره بیمه در این رکوردها یکسان است بسادگی می توان رکورد های تغییر یافته متعلق به یک فرد را شناسائی نمود.

در شبیه سازی انجام شده ابتدا کلیه رکوردهای تغییرات هویتی مربوط به یکی از شعب تامین جمع آوری شده است. نمونه تولیدی حاوی ۸۵۰ رکورد تغییرات است که این رکوردها متعلق به مشخصات هویتی ۴۱۲ فرد متفاوت می باشند(توجه کنید تغییرات برای یک فرد ممکن است بیش از دو بار صورت گرفته باشد). جدول پنج نتایج حاصل از اعمال الگوریتم ژنتیکی برای تعیین ضرایب فیلدهای یک رکورد را نشان میدهد.

جدول پنج – مقادیر ضرایب در پنج مورد شبیه سازی انجام شده بکمک الگوریتم های ژنتیکی.

اندازه جمعیت	مقدار تطبیق نهائی	ضریب نام	ضریب نام خانوادگی	ضریب نام پدر	ضریب شماره شناسنامه	ضریب تاریخ تولد
۳۰	۰.۰۳۰۲۷	۴۰%	۲۷%	۱%	۳۰%	۲%
۳۰	۰.۰۲۷۲۷	۱۸%	۲۷%	۲%	۴۸%	۵%
۳۰	۰.۰۲۵۰۰	۲۲%	۲۸%	۲%	۴۴%	۴%
۵۰	۰.۰۲۲۷۰	۲۲%	۲۲%	۱%	۵۴%	۱%
۵۰	۰.۰۲۵۰۰	۲۹%	۳۲%	۱%	۳۷%	۱%

با توجه به ضرایب حاصل از الگوریتم ژنتیکی الگوریتم تطبیق تقریبی پیشنهادی با ضرایب بهینه برروی یک نمونه ۱۶۵ رکوردی مورد آزمون قرار گرفت که دراین نمونه ۷۳ مورد مشابهت وجود داشت. این نمونه فایل از یکی از شعب تامین اجتمائی تهیه شده که اقدام به تصحیح اطلاعات نموده و این نمونه شامل موارد قبل و بعد از تصحیح است که با مقایسه رکوردها توسط انسان و مطابقت با پرونده بیمه شدگان استخراج شده است. نتایج حاصل نشان می دهد که تنها در دو مورد مثبت کاذب و سه مورد منفی کاذب اتفاق افتاده و بقیه موارد شناسائی با موفقیت صورت گرفته است. در جدول شش نمونه این موارد مشخص شده است. منظور از مثبت کاذب مواردی است که الگوریتم به اشتباه دو رکورد را یکسان فرض نموده و منفی کاذب مواردی مشابه بوده که الگوریتم قادر به شناسائی مشابهت بین آنها نشده است.

تحلیل نمونه های جدول شش نشان میدهد که در موارد اشتباه اطلاعات بشدت ناقص است. در مورد مثبت کاذب تنها نام و نام خانوادگی موجود است و در مورد منفی کاذب مشخصات تاریخ تولد و شماره شناسنامه اختلاف قابل تاملی دارند.

جدول شش – دو نمونه از مثبت و منفی کاذب یافت شده توسط الگوریتم تطبیق تقریبی

شماره شناسنامه	تاریخ تولد	نام پدر	نام خانوادگی	نام	
-	-	-	ضامنی	سعید	مثبت کاذب
-	-	-	ضامینی	سعیده	
۳۶۰۵	۸/۳/۱۳۰۲	-	چاوشی	یداله	منفی کاذب
۳	۸/۸/۱۳۱۹	-	چاوشی کیوی	یداله	

۸ – نتیجه گیری

انتساب رکوردهای سیستم هویتی به افراد معین، یکی از مهم ترین وظایف یک سیستم اطلاعاتی است. وجود خطا در حین ورود اطلاعات بویژه برای سازمان هائی که با حجم بزرگی از اطلاعات سر و کار دارند امری بدیهی و غیر قابل اجتناب است. در فرایند پالایش داده ها آماده سازی اطلاعات برای ایجاد انباره داده ها یکی از اولین گام هاست.

تجربه انجام شده دراین بررسی نشان دادکه تصحیح اطلاعات سیستم هویتی بیمه شدگان سازمان تامین اجتمائی بدون توجه به ماهیت اطلاعات و کشف خطا های رایج در تولید داده ها در این نمونه کاربردی چندان موفق نخواهد بود. کشف الگوریتمی مواردی که کاربران اقدام به رج زدن در ورود اطلاعات آنهم در سطح وسیع می زنند خود می تواند موضوع تحقیق جداگانه ای باشد. اما در این مقاله صرفا تلاش شده تا مشابهت تقریبی بین دو یا چند رکورد از مشخصات هویتی افراد مورد توجه قرار گیرد.

برای این منظور از ایده فاصله لونشتین استفاده شده اما توسعه این ایده الزامی است. عدم توجه به نکات خاص کاربرد باعث کاهش دقت مشابهت خواهد شد. برای مثال عدم توجه به این نکته که در فیلد تاریخ تولد بسیاری از کاربران تاریخ تولد ۱۳۰۰/۰۱/۰۱ را مورد استفاده قرار می دهند منجر به تطبیق تقریبی دو رکورد خواهد شد در حالیکه این دو رکورد ممکن است واقعا مشابهتی نداشته اند. در کنار این موارد که با کمک الگوریتم های پیش پردازشی قابل رفع هستند باید به تغییراتی در کارکرد محاسبه فاصله لونشتین مثل اهمیت محل قرارگیری حرف در رشته پرداخت.

تطبیق تقریبی بین دو فیلد مقوله ای متفاوت از تطبیق تقریبی بین دو رکورد است. این عمل تنها از طریق جمع میزان مشابهت بین فیلد های دو رکورد صورت نمی گیرد بلکه عامل مهم تعیین ضرایب تقویت یا تضعیف هر فیلد است. تجربیات قبلی انجام شده نشان داده بود تنظیم دقیق این ضرایب نقش کلیدی در دقت تشخیص الگوریتم به همراه دارد. به همین علت از یک الگوریتم ژنتیکی برای یافتن ضرایب مناسب استفاده شد. نتایج شبیه سازی در این زمینه نشان داد که تنها سه فیلد نام، نام خانوادگی و شماره شناسنامه در تطبیق بین رکوردها اهمیت دارند و عملا با داده های موجود نقش دو فیلد تاریخ تولد و نام پدر تقریبا به صفر نزدیک خواهد بود.

مراجع

[۱] D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison_Wesley, ۱۹۸۹

[۲] J. Han, M. Kamber, *Data Mining : Concepts and Techniques, Morgan Kaufmann*, ۲۰۰۱.

[۳] M. A. Hernadez, S.J.Stolfo, "*Real-world Data is Dirty: Data Cleansing and the Merge/Purge Problem*", Journal of Data Mining and Knowledge Discovery, ۱(۲), ۱۹۹۸.

[۴] J. A. Hylton, *Identifying and Merging Related Bibliographical Records*, Master's Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, ۱۹۹۶.

[۵] M. Kantardzic, *Data Mining : Concepts, Methods, and Algorithms*, IEEE Press, ۲۰۰۳.

[۶] K. Kukich, "*Techniques for Automatically Correcting Words in Text*", ACM Computing Survey vol. ۲۴, No. ۴, ۱۹۹۲.

[۷] A. E. Monge, *Adaptive Detection of Approximately Duplicate Database Records and Database Integration Approach to Information Discovery*, PHD Thesis, University of California, San Diego, ۱۹۹۷.

[۸] A. E. Monge, C. P. Elkan, "*The Field Matching Problem: Algorithms and Applications*" Second International Conference of Knowledge Discovery and Data Mining, AAAI Press, ۱۹۹۶.

[۹] V. S. Verykios, A.K.Elmagarmid, E.H.Houstis, "*Automating the Approximate Record Matching Process*", Information Science , vol ۱۲۶, No ۱-۴, ۲۰۰۰

[۱۰] V. S. Verykios, G.V.Moustakides, "*A Cost Optimal Decision Model for Record Matching*", Workshop on Data Quality, ۲۰۰۱