ارایه روش هوشمند تطبیق تقریبی اطلاعات هویتی برای مشتریان بانک بوسیله متن کاوی

جلال‌الدین نصیری[*1]

استادیار، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

امیرمحمود میر[2]

دانشجوی کارشناسی ارشد، دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی، واحد تهران شمال

آرمان ساجدی‌نژاد[3]

استادیار، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

چکیده

شناسایی مشتریان تکراری در بانک‌ها و موسسات مالی، یکی از کلیدی‌ترین مراحل در سامانه‌های ضدپولشویی، رتبه‌بندی اعتباری و ذی‌نفع واحد است. همچنین تکمیل و تصحیح اطلاعات هویتی مشتریان از کاربردهای دیگر استخراج مشتریان مشابه است. کد ملی و کد اقتصادی ثبت شده برای مشتریان حقیقی و حقوقی از اعتبار مناسبی برخوردار نیستند. در نتیجه شناسایی مشتریان یگانه از طریق اطلاعات هویتی مشتریان مانند نام، نام خانوادگی، نام پدر، تاریخ و محل تولد انجام می‌گیرد. اطلاعات هویتی مشتریان به زبان فارسی در پایگاه داده ثبت شده است و همین مسئله چالش‌هایی از قبیل کلمات چند املائی و اشتباهات نگارشی را ایجاد کرده است. در این مقاله، سامانه هوشمند تطبیق تقریبی اطلاعات هویتی برای شناسایی مشتریان تکراری طراحی و پیاده‌سازی شده و الگوریتم لونا شتاین برای تطبیق تقریبی اطلاعات فارسی هویتی مشتریان توسعه داده شده است. سامانه پیشنهادی بر روی مشتریان چندین بانک خصوصی و دولتی طراحی و اجرا شده است. نتایج نشان می‌دهد که سامانه ارائه شده از دقت

[1] j.nasiri@irandoc.ac.ir
[2] mir-am@hotmail.com
[3] sajedinejad@irandoc.ac.ir

و سرعت بالایی برخوردار است. نتایج پیاده‌سازی نشان می‌دهد که با استفاده از هرس هوشمندانه سرعت استخراج مشتریان تکراری تقریبا تا ۲ برابر بهبود یافته است.

واژگان کلیدی: مشتریان تکراری، متن کاوی، برنامه نویسی موازی، تطبیق تقریبی رشته.

۱- مقدمه

شناسایی مشتریان تکراری در بانک‌ها و موسسات مالی یکی از کلیدی‌ترین مراحل در سامانه‌های ضد پولشویی، رتبه‌بندی اعتباری و بطور کلی شناسایی ذی‌نفع واحد است (Sharman 2008). برای پیدا کردن مشتریان تکراری بدون استفاده از کد ملی، الگوریتم‌های هوشمند تطبیق تقریبی متن تنها راه حل موجود می‌باشند.

در چند دهه اخیر، رشد بانکداری الکترونیک باعث رونق در سامانه‌های متعدد شده است (احمدی و سویری ۱۳۹۴). یک مشتری واحد می‌تواند شماره‌های مشتری‌های متفاوتی را دریافت کرده باشد. ورود اطلاعات هویتی پایه به این سامانه ها توسط کاربران انسانی انجام شده است. معمولا این اطلاعات دارای چالش‌های متعددی مانند خالی بودن اطلاعات (مانند اطلاعات پدر)، غلط‌های املائی بسیار زیاد (مخمد ایر به جای محمد امیر) و کلمات چند املایی (داوود و داود) و غیره می باشند. از طرف دیگر کد ملی غیر معتبر و تکراری (مانند ۱۱۱۱۱۱۱۱۱۱) و شماره شناسنامه بدون مقدار باعث شده با شناسه‌های کد ملی و شماره شناسنامه نتوان جستجو و پاک‌سازی انجام داد.

تحقیقات گسترده‌ای در زمینه الگوریتم های تطبیق تقریبی رشته صورت گرفته است که در این میان زبان انگلیسی بیشترین سهم از تحقیقات را به خود اختصاص داده است (Navarro 2001). در خصوص زبان فارسی، به دلیل وجود دشواری های ذاتی زبان از قبیل پیوستگی حروف باعث شده است که این تحقیقات به مرحله کاربردی راه پیدا نکند. الگوریتم های تطبیق رشته دارای مرتبه زمانی بالایی هستند. همچنین در پایگاه داده بانک‌ها نیز تعداد بالای فیلدهای مربوط به اطلاعات هویتی باعث تشدید این مشکل می‌شود. بنابر این برای عملیاتی شدن برنامه باید مشکلات زمان اجرای آن به گونه‌ای حل شود که در زمان منطقی خروجی دهد. همچنین الگوریتم های موجود برای زبان فارسی باید شخصی سازی شود به گونه‌ای که بعضی از حروف (س ص-ج ح و ...) که قرابت ساختاری یا معنایی بیشتری با یکدیگر دارند در الگوریتم تطبیق تقریبی رشته در نظر گرفته شوند.

۲- الگوریتم لون اشتاین

یکی از معروف‌ترین الگوریتم‌های تطبیق تقریبی رشته الگوریتم لون اشتاین[1] است (Navarro 2001). در روش لون اشتاین، فاصله بین دو رشته، بوسیله کمترین تعداد عملیات مورد نیاز برای تبدیل یک رشته به رشته دیگر معین می شود. عملیات مجاز در این روش می‌توانند یکی از عملیات درج، حذف یا جایگزینی باشند. برای پیاده سازی این روش، از روش برنامه‌نویسی پویا بهره گرفته شده است. در این الگوریتم، از ماتریسی با ابعاد $(n + 1) \times (m + 1)$ استفاده شده است، که n و m طول دو رشته ورودی می‌باشند. سپس با استفاده از این ماتریس تعداد عملیات لازم برای رسیدن از رشته اول به رشته دوم محاسبه و ثبت می‌گردد. بدین صورت که هر حرکت افقی در ماتریس d (از سلولی به سلول سمت راست آن) بیانگر یک عمل درج، هر حرکت عمودی (از سلولی به سلول پائین آن) بیانگر یک عمل حذف و هر حرکت مورب (از سلولی به سلول پائین و سمت راست آن) بیانگر یک عمل جایگزینی است. در فرمول ۱ نحوه پرکردن ماتریس و پیدا کردن کمترین هزینه تبدیل رشته دوم به رشته اول نشان داده شده است.

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0, \\ \min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & otherwise. \end{cases} \quad (۱)$$

در شکل ۱ نحوه تطبیق بوسیله الگوریتم لون اشتاین برای دو رشته puzzle و pzzel نشان داده شده است. همچنین مسیر تغییرات در جهت کمترین هزینه مشخص شده است.

		p	u	z	z	l	e
	0	1	2	3	4	5	6
p	1	0	1	2	3	4	5
z	2	1	1	1	2	3	4
z	3	2	2	1	1	2	3
e	4	3	3	2	2	2	2
l	5	4	4	3	3	2	3

شکل ۱. عملکرد الگوریتم لون اشتاین برای تطابق دو رشته

[1] Levenshtein algorithm

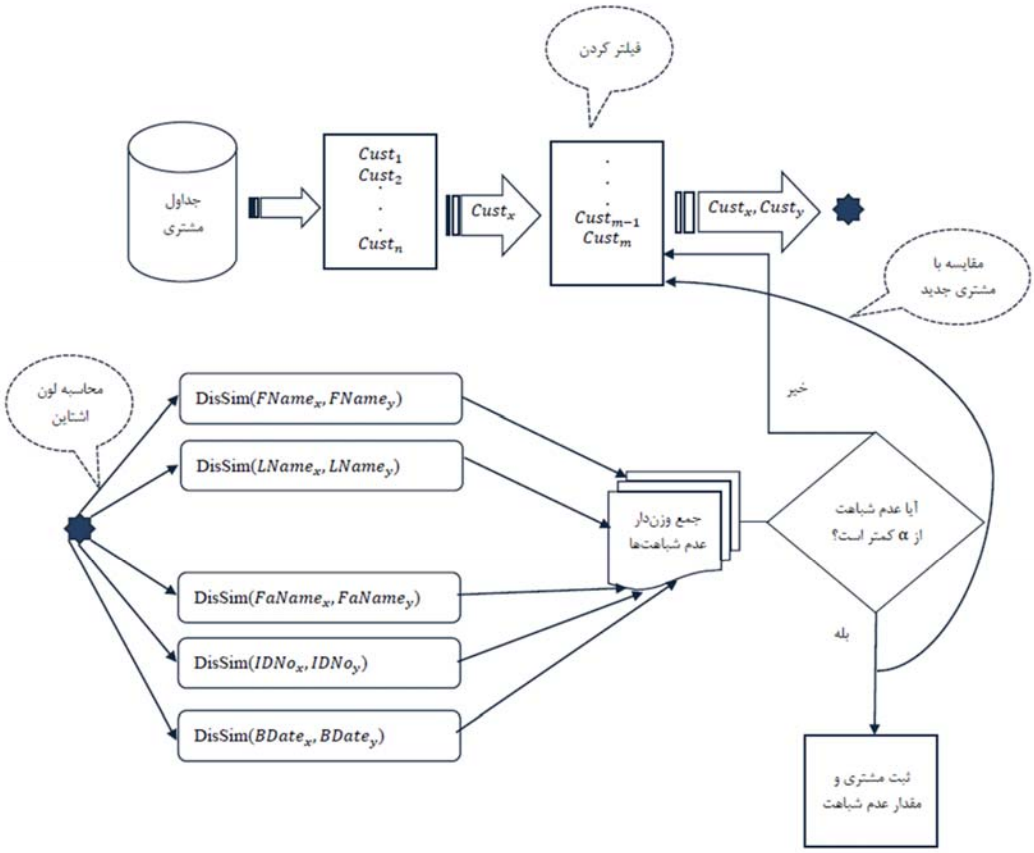
۳- معماری و الگوریتم سامانه هوشمند تطبیق تقریبی مشتری

وجود املاء های متفاوت از یک کلمه، خطاهای انسانی در ورود اطلاعات و همچنین ناقص بودن فیلدهای اطلاعاتی باعث شده است تا یک مشتری مشخص در صورت‌های متفاوتی در جدول مشتریان دیده شود. پس از بررسی فیلدهای متفاوت جدول مشتری، فیلدهای زیر برای منحصر بفرد کردن مشتریان حقیقی انتخاب گردید.

جدول ۱. فیلدهای اطلاعاتی استفاده شده در سامانه

نام فیلد	اعتبار	وزن
نام	خوب	۱
نام خانوادگی	خوب	۰,۹
نام پدر	متوسط	۰,۶
شماره شناسنامه	خوب	۰,۷
تاریخ تولد	متوسط	۰,۵
کد ملی	ضعیف	----

در این الگوریتم که فلوچارت آن در شکل ۲ آورده شده است. ابتدا از جدول مشتریان ، مشتری که میزان شباهت بقیه مشتریان به آن هدف الگوریتم است انتخاب می گردد. این مشتری را $Cust_x$ می‌نامیم. بر اساس ویژگی جنسیت یک فیلتر انجام می‌شود. در واقع فقط مشتریانی که از نظر جنسیت با $Cust_x$ در تضاد نیستند انتخاب می‌شود.



شکل ۲. فلوچارت الگوریتم سامانه هوشمند تطبیق تقریبی

پس از فیلترکردن جدول مشتریان، مرحله پردازش شروع می‌شود. بر اساس الگوریتم فاصله لون اشتاین و ویژگی‌های انتخاب شده، پنج فاصله متفاوت بدست می‌آید. این پنج فاصله از جهت اهمیت یکسان نیستند. بنابر این پنج وزن برای فاصله‌های بدست آمده باید انتخاب گردد. مقادیر این وزن‌ها در جدول ۱ نشان داده شده است.

بو سیله جمع وزن‌دار این فا صله‌ها، یک فا صله بین $Cust_x$ و مشتریان دیگر بد ست می‌آید. اگر این فا صله از حد آستانه α کمتر باشد، در دسته مشتریان مشابه $Cust_x$ ذخیره می‌گردد و در صورتی که از حد آستانه α بیشتر باشد به عنوان متفاوت تلقی خواهد شد. در پیاده‌سازی انجام شده بر اساس مشاهدات فاصله های کمتر از ۱۰۰ معنا دار می‌با شند. ر شته ها با طول‌های متفاوت تاثیر مهمی در خروجی فا صله لون ا شتاین دارند. به عبارت دیگر ر شته های طولانی معمولا فاصله‌های بیشتری را بدست می‌آورند. همچنین ر شته‌های null نیز باید به صورت منا سبی حل شود.

برای رشته‌ای null فاصله لون اشتاین برابر تعداد کاراکترهای رشته غیر null در نظر گرفته شده است. خروجی وزن‌دار فاصله‌ها به وسیله فرمول ۲ نرمال می‌شود که اندازه رشته در خروجی تاثیر منفی نداشته باشد.

$$normalDisSim = \frac{DisSim \times 100}{\max(length(string1), length(string2))} \quad (۲)$$

بعضی از قابلیت‌های جستجوی هوشمند تقریبی به شرح ذیل است:

پیدا نمودن موارد چند املائی (داوود-داود)(اسمعیل-اسماعیل)(اله-الله)

پیدا نمودن موارد اشتباهات نوشتاری (مصطفوی-مصتفوی)

پیدا نمودن مواردی که یک یا چند حرف آن‌ها وارد نشده است (رویا-ریا)(علیرضا-علرضا)

پیدا نمودن موارد چند اسمی ناقص (سیده فاطمه السادات- فاطمه)(جلال الدین-جلال)

سرعت بسیار عالی جستجو

امکان وزن‌دهی به کلماتی که اهمیت بیشتری در جستجو دارند.

٤- نتایج پیاده‌سازی

در این قسمت نتایج پیاده سازی سامانه هو شمند تشخیص تطبیق تقریبی مورد بحث قرار می گیرد. ذکر این نکته مهم به نظر می‌رسد که جهت حفظ محرمانگی، داده‌های نمایش، داده‌های اصلی بانک‌ها نمی‌باشند. به عبارت دیگر، بر روی فیلدهای موجود عملگر درهم‌ریختگی انجام شده است و فیلدهای مهم اطلاعاتی مانند کد ملی، شماره مشتری و شماره شناسنامه یا نمایش داده نشده و یا به صورت مخدوش (تغییر در بعضی از اعداد) نشان داده شده است.

همانطور که اشاره شد یکی از مزایای الگوریتم پیشنهادی پوشش غلط‌های املائی و کلمات چند املایی است. شکل ۳ نمونه‌هایی از مشتریان تکراری پیدا شده توسط سامانه را نشان می‌دهد. قسمت‌هایی که بوسیله مستطیل های سفید محو شده است دارای رشته های کاملا برابر بوده است که برای حفظ حریم خصوصی مشتریان حذف شده است.

غلط املایی (زهر- زهرا)(موسی الرضا – موسی رضا) (عبداله –عبذاله) ، کلمات چند املایی (نصرالله- نصراله)، کلمات با فاصله یا بدون فاصله (آبباریکی- آب باریکی)، اشتباهات (امینی – ایمنی) (فروشان –فروشانی) و پیشوند و پسوندها (کریمی پور قوینالو- کریمی پور) از اشتباهات رایج است که در جدول اطلاعات پایه‌ای مشتریان به وفور یافت می‌شود. همچنین در ورود اطلاعات عددی مانند شماره شناسنامه تایپ اطلاعات اشتباه نیز رایج می‌باشد (۴۱۲۱-۴۱۳۱) (۱۹۸۲-۱۹۸۲۶).

درصد شباهت	نام	نام خانوادگی	نام پدر	شماره شناسنامه	تاریخ تولد	محل صدور شناسنامه	تاریخ صدور شناسنامه
۸۸	علی اصغر	آب باریکی	[redacted]	۴۱۲۱	۶۰۰۶۰۱	سایر موارد	۶۰۰۶۰۱
۸۸	علی اصغر	آبباریکی		۴۱۳۱	۶۰۰۶۰۱	سایر موارد	۶۰۰۶۰۱
۸۸	محمد	[redacted]	عبذاله	۱۹۸۲	۵۵۰۷۰۵	سایر موارد	۵۵۰۷۰۵
۸۸	محمد		عبداله	۱۹۸۲۶	۵۵۰۷۰۵	سایر موارد	۵۵۰۷۰۵
۸۷.۶۶۶۶۶۶۶	[redacted]	امینی فروشانی	نصرالله		۵۹۰۲۱۰	خمینی شهر	۵۹۰۲۱۰
۸۷.۶۶۶۶۶۶۶		ایمنی فروشان	نصراله		۵۹۰۲۱۰	سایر موارد	۵۹۰۲۱۰
۸۷.۶۶۶۶۶۶۶	زهر	سعادت فیروزاباد	[redacted]	[redacted]	۸۱۰۹۱۰	سایر موارد	۸۱۰۹۱۰
۸۷.۶۶۶۶۶۶۶	زهرا	سعادت فیروز آباد			۸۱۰۹۱۰	سایر موارد	۸۱۰۹۱۰
۶۷.۵	[redacted]	شم ابادی	غل امحسین	۳	۵۶۰۴۰۱	سایر موارد	۵۶۰۴۰۱
۶۷.۵		شم ابادی(هیبت جانثار غل امحسین		۳	۵۶۰۴۰۱	سایر موارد	۵۶۰۴۰۱
۶۵	[redacted]	کریمی پور قوینانلو	موسی الرضا	[redacted]	۶۰۰۸۲۶	سایر موارد	۶۰۰۸۲۶
۶۵		کریمی پور	موسی رضا		۶۰۰۸۲۶	سایر موارد	۶۰۰۸۲۶

شکل ۳. نمونه‌ای از نتایج جستجوی دسته‌ای در سامانه هوشمند تطبیق تقریبی

پیدا نمودن مشتریانی که نزدیک‌ترین شباهت را با مشتری مورد جستجو دارند، از مزایای دیگر سامانه هوشمند تطبیق تقریبی است. در صورتی که اطلاعات مشتری به صورت کامل و دقیق در دسترس نباشد می توان از این سامانه در جهت پیدا نمودن مشتریانی که نزدیک‌ترین تشابه را دارند استفاده نمود. در شکل ۴ نمونه جستجو با عنوان فاطمه السادات هاشمی نام پدر محمد در بانک وجود ندارد ولی نزدیکترین مشتریان در بانک نمایش داده می‌شود.

شکل ۴. نمونه ای از نتایج جستجوی موردی در سامانه هوشمند تطبیق تقریبی

پیدا نمودن مشتریانی که کاندید رابطه خواهر/ برادر و رابطه دوقلو بودن هستند، از دیگر مزایای سامانه ارائه شده است. در این سامانه به صورت هوشمند با استفاده از قوانین خبره بر اساس میزان شباهت نام خانوادگی، نام پدر و تاریخ تولد، بعضی از مشتریان تکراری را به عنوان دوقلو یا خواهر برادر بودن معرفی می‌کند.

۵- نتیجه‌گیری

در این مقاله سامانه هوشمند تطبیق تقریبی اطلاعات هویتی که بر روی مشتریان چند بانک خصوصی و دولتی اجرا شده است، تشریح شد . همانطور که گفته شد، متاسفانه کد ملی مشتریان از اعتبار مناسبی برخوردار نمی‌باشد. از این رو برای حل این مشکل استفاده از روش‌های جستجوی تقریبی رشته و متن کاوی تنها راه‌حل به‌نظر می‌رسد. الگوریتم لون اشتاین برای زبان فارسی شخصی سازی شده و با استفاده از بستر برنامه نویسی موازی، از سرعت و دقت بسیار خوبی برخوردار شده است. همچنین در نگارش‌های دیگر این سامانه مشتریان کاندید چندقلو و خواهر/برادری به سامانه اضافه گردیده است. تکمیل و تصحیح اطلاعات مشتریان بوسیله اطلاعات هویتی مشتری مشابه، پیدا نمودن مشتریان تکراری، جستجوی تقریبی یک مشتری و کاربردهای ذینفع واحد از دیگر مزایای این سامانه است.

منابع

احمدی ، سـید محمود و مهدی خندان سـویری . ۱۳۹۴. نظام های پرداخت و بانکداری الکترونیک در ایران . تهران: پژهشکده پولی و بانکی بانک مرکزی جمهوری اسلامی ایران.

اسمعیل پور، ندا، برومندنیا ، علی. ۱۳۹۲. بازشناسی برخط زیر کلمات فارسی براساس کدهای زنجیره ای فازی با استفاده از مدل تطبیق رشته. بیست و یکمین کنفرانس مهندسی برق ایران: ۱-۶.

Sharman, Jason C. 2008. Power and Discourse in Policy Diffusion: Anti-Money Laundering in Developing States. *International Studies Quarterly*, 52(3): 635-656.

Tang, Jun, and Jian Yin. 2005. Developing an intelligent data discriminating system of anti-money laundering based on SVM. *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*. Vol. 6: 3453-3457

Ukkonen, Esko.1985. Algorithms for approximate string matching. *Information and control*, *64*(1-3): 100-118.

Navarro, Gonzalo. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, *33*(1): 31-88.

Diaz, Javier, Camelia Munoz-Caro, and Alfonso Nino. 2012. A survey of parallel programming models and tools in the multi and many-core era. *IEEE Transactions on parallel and distributed systems* 23, no. 8: 1369-1386.

Su, Hung-Cheng, Tsung-Han Wu, and Chun-Jen Tsai. 2014. Temporal multithreading architecture design for a Java processor. In *Circuits and Systems (ISCAS), 2014 IEEE International Symposium on*. 2201-2204.

Intelligent Approximate Personal Identity Information Matching Method for Customer of Bank based on Text Mining

***J. Nasiri**

Assistant Professor, Iranian Research Institute for Information Science and Technology (IRANDOC)

A. Mir

MSc Student, Faculty of Electrical and Computer Engineering, North Tehran Branch, Islamic Azad University

A. Sajedinejad

Assistant Professor, Iranian Research Institute for Information Science and Technology (IRANDOC)

Abstract.

In the banks and financial institutions, Identification of duplicate customers is one of the key steps in anti-money laundering systems, credit rating and the beneficiary identification. In addition, correction and updating of customers' information are one of the application of finding similar customers. National ID and economical ID are not trustworthy for identification of legal and natural persons. Consequently, customers' identity information such as first name, surname, father's name, birthplace and birthdate were used for identification of unique customers. In the bank's database, Customers' identity information was stored in Persian language which creates some challenges like alternate spellings and writing errors. Text mining and approximate string matching algorithms have been used for overcoming those challenges. The Intelligent pruning used for increasing speed of text mining algorithms. In this paper, an intelligent system was designed and implemented for approximate matching of identity information and Levenshtein algorithm was customized for approximate matching in Persian language. The proposed system evaluated on customers of several private and government-owned banks. The result reveals that proposed system is accurate and fast. Moreover, intelligent pruning makes search of duplicate customers roughly two times faster.

Keywords: Duplicate customers, Data mining, Parallel programming, approximate string matching