Ilmiy tadqiqotlar va ularning yechimlari jurnali
Aoka guvohnoma № 323759 ISSN: 3060-4818

TURLI MANBALARDAGI XORIJIY TALABALAR YOZUVLARINI IDENTIFIKATSIYALASH VA MOSLASHTIRISH ALGORITMI

Abdullayev Javoxir Jaxongirovich
Buxoro davlat texnika universiteti o‘qituvchisi, i.f.f.d. (PhD)
https://doi.org/10.5281/zenodo.22929552

Annotatsiya. Xorijiy talabalarning ism-familiyalari turli manbalarda lotin va kirill yozuvlarida, turli transliteratsiya an’analariga ko‘ra yoziladi. Bu bir shaxsga tegishli yozuvlarni avtomatik moslashtirishni qiyinlashtiradi. Maqolada ism-familiyalarni yagona shaklga keltiruvchi normallashtirish qoidalari va Jaro–Winkler o‘xshashlik o‘lchoviga asoslangan moslashtirish algoritmi taklif etilgan. Sakkiz juftlikdan iborat nazorat to‘plamida normallashtirishdan keyin bir shaxsga tegishli juftliklar o‘xshashligi 0,964–1,000 oralig‘iga ko‘tarilgani, turli shaxslar esa 0,85 dan oshmagani ko‘rsatilgan.
Kalit so‘zlar: yozuvlarni moslashtirish, transliteratsiya, Jaro–Winkler, normallashtirish, dublikatlarni aniqlash, axborot tizimi, xorijiy talabalar.

Аннотация. Имена иностранных студентов в разных источниках записываются латиницей и кириллицей по различным традициям транслитерации, что затрудняет автоматическое сопоставление записей одного лица. В статье предложены правила нормализации имён и алгоритм сопоставления на основе меры сходства Джаро–Винклера. На контрольном наборе из восьми пар показано, что после нормализации сходство пар, относящихся к одному лицу, возросло до 0,964–1,000, тогда как для разных лиц оно не превысило 0,85.
Ключевые слова: сопоставление записей, транслитерация, Джаро–Винклер, нормализация, выявление дубликатов, информационная система, иностранные студенты.

Abstract. The names of international students are written in Latin and Cyrillic scripts following different transliteration conventions across sources, which makes automatic matching of records belonging to one person difficult. The paper proposes name normalization rules and a matching algorithm based on the Jaro–Winkler similarity measure. On a control set of eight pairs, normalization raised the similarity of same-person pairs to 0.964–1.000, while pairs of different persons did not exceed 0.85.
Keywords: record linkage, transliteration, Jaro–Winkler, normalization, duplicate detection, information system, international students.

Kirish

Xorijiy talaba haqidagi ma’lumotlar bir nechta manbada shakllanadi: ta’lim agentligi arizasi, qabul komissiyasi hujjatlari, migratsiya ro‘yxati, dekanat bazasi. Har bir manbada talabaning ismi o‘z qoidasiga ko‘ra yoziladi. Masalan, pasportdagi "Mohammed Ali Khan" qabul hujjatlarida "Muhammad Ali Xon", migratsiya ro‘yxatida esa "Мухаммад Али Хан" bo‘lishi mumkin. O‘zbekistonda lotin yozuviga asoslangan alifboga o‘tilgan bo‘lsa-da [1], amaliyotda kirill yozuvidagi hujjatlar va bazalar hamon parallel yuritiladi. Inson bu yozuvlar bir shaxsga tegishli ekanini darhol tushunadi, lekin kompyuter uchun ular butunlay boshqa satrlar.

Yozuvlarni to‘g‘ri moslashtirmasdan turib ma’lumotlarni ko‘p manbali tekshirish ham, takroriy yozuvlarni aniqlash ham mumkin emas. Maqolaning maqsadi transliteratsiya farqlarini hisobga oluvchi normallashtirish qoidalari va ularga asoslangan moslashtirish algoritmini ishlab chiqishdan iborat.

Adabiyotlar tahlili

Yozuvlarni moslashtirishning nazariy asosi I. Fellegi va A. Sunter tomonidan yaratilgan bo‘lib, unda juftliklar "mos", "mos emas" va "noaniq" sinflarga ikki chegara yordamida ajratiladi [2]. W. Winkler ushbu modelni ism-familiyalar uchun moslashtirilgan satr o‘xshashlik o‘lchovi bilan boyitgan [3]. A. Elmagarmid va hammualliflar takroriy yozuvlarni aniqlash usullarini tizimlashtirgan [4], P. Christen esa moslashtirish jarayonining to‘liq bosqichlarini — oldindan ishlov berish, bloklash, solishtirish va tasniflashni batafsil yoritgan [5]. Shaxs ismlarini moslashtirish usullarining qiyosiy tahlili [6] va satr masofasi o‘lchovlarini solishtirish natijalari [7] Jaro–Winkler o‘lchovi aynan ism-familiyalar uchun yuqori samaradorlik berishini ko‘rsatgan.

Biroq bu ishlarning aksariyati bitta yozuv tizimi doirasida ishlaydi. Lotin, kirill va arab yozuvidan kelib chiqqan ismlar aralash uchraydigan vaziyat, ya’ni O‘zbekiston OTMlaridagi xorijiy talabalar holati uchun tayyor yechimlar kam. O‘zbek tilidagi matnlarni avtomatik qayta ishlash bo‘yicha mahalliy tadqiqotlar [8] mavjud bo‘lsa-da, ular xorijiy shaxs ismlarini moslashtirish masalasiga qaratilmagan.

Tadqiqot metodologiyasi

Algoritm to‘rt bosqichdan iborat.

1-bosqich. Normallashtirish. Satr kichik harflarga o‘tkaziladi, diakritik belgilar olib tashlanadi, kirill harflari lotinga o‘giriladi. So‘ngra transliteratsiyadagi tipik farqlarni bir xil shaklga keltiruvchi qoidalar qo‘llanadi (1-jadval) va ketma-ket takrorlangan harflar bittaga qisqartiriladi.

1-jadval. Normallashtirish qoidalaridan namunalar

Farq turi	Misollar	Yagona shakl
x / kh / h	Xon, Khan, Хан	h
q / k	Qodir, Kodir	k
j / dj / dzh	Oguljan, Огулджан	dj
y / i (unlilar orasida)	Atayev, Ataiev	i
oo / u, ee / i	Noori, Nuri	u, i
ism variantlari	Mohammed, Muhammed, Mukhammad	muhammad

2-bosqich. Bloklash. Barcha juftliklarni solishtirish ko‘p vaqt talab qiladi (N ta yozuv uchun $N(N-1)/2$ juftlik). Shuning uchun faqat tug‘ilgan yili va fuqaroligi bir xil yozuvlar solishtiriladi. Bloklash solishtirishlar sonini keskin kamaytiradi, biroq bloklash kaliti noto‘g‘ri tanlansa, mos juftliklar turli bloklarga tushib qolishi mumkin [9].

3-bosqich. O‘xshashlikni hisoblash. Normallashtirilgan satrlar Jaro–Winkler o‘lchovi bilan solishtiriladi:

$$\mathrm{sim}(JW) = \mathrm{sim}(J) + l \cdot p \cdot (1 - \mathrm{sim}(J)) \qquad (1)$$

bu yerda $\mathrm{sim}(J)$ — M. Jaro taklif etgan o‘xshashlik o‘lchovi [10], l — satrlar boshidagi umumiy qism uzunligi (4 tagacha), $p = 0{,}1$. Bu o‘lchov ismlarning boshlanishi odatda to‘g‘ri yozilishini hisobga oladi.

4-bosqich. Qaror qabul qilish. Fellegi–Sunter yondashuviga ko‘ra ikki chegara belgilanadi: $\mathrm{sim} \geq \theta(\text{yuqori})$ bo‘lsa juftlik "mos", $\mathrm{sim} < \theta(\text{quyi})$ bo‘lsa "mos emas", oraliqda esa "noaniq" deb topilib, xodimga ko‘rib chiqish uchun yuboriladi. Dastlabki qiymatlar $\theta(\text{yuqori}) = 0{,}92$ va $\theta(\text{quyi}) = 0{,}85$.

Algoritm Python tilida amalga oshirildi va sakkiz juftlikdan iborat nazorat to‘plamida sinaldi: beshtasi bir shaxsga tegishli, uchtasi turli shaxslar.

Tahlil va natijalar

2-jadval. Nazorat to‘plamida moslashtirish natijalari

1-yozuv	2-yozuv	Bir shaxs	O‘xshashlik (xom)	O‘xshashlik (normallashtirilgan)
Mohammed Ali Khan	Muhammad Ali Xon	ha	0,792	0,973
Mohammed Ali Khan	Мухаммад Али Хан	ha	0,414	1,000
Gurbanguly Atayev	Gurbanguli Otayev	ha	0,933	0,964
Oguljan Berdiyeva	Огулджан Бердыева	ha	0,373	1,000
Rahul Sharma	Raxul Sharma	ha	0,895	1,000
Ahmad Shah Noori	Akhmad Shoh Nuri	ha	0,882	0,973
Rahul Sharma	Rohit Sharma	yo‘q	0,850	0,850
Aziz Karimov	Azamat Karimov	yo‘q	0,799	0,799

Xom satrlarda bir shaxsga tegishli juftliklar o‘xshashligi 0,373 dan 0,933 gacha tarqoq bo‘lib, ularning bir qismi turli shaxslar juftligidan ham past chiqdi. Masalan, lotin va kirill yozuvidagi bir xil ism 0,373 o‘xshashlik bergan, "Rahul Sharma" va "Rohit Sharma" esa 0,850. Bunday holatda hech qanday chegara to‘g‘ri ajratish bermaydi.

Normallashtirishdan keyin bir shaxsga tegishli barcha juftliklar 0,964–1,000 oralig‘iga ko‘tarildi, turli shaxslar juftligi esa 0,85 dan oshmadi. Natijada θ(yuqori) = 0,92 chegara nazorat to‘plamidagi barcha juftliklarni to‘g‘ri ajratdi. "Rahul Sharma" – "Rohit Sharma" juftligi aynan chegara (0,85) ustida turgani uchun "noaniq" deb belgilanib, xodimga yuboriladi. Bu to‘g‘ri xatti-harakat: familiyasi bir xil, ismi o‘xshash ikki shaxsni avtomatik birlashtirish xavfli.

Muhokama

Tajriba normallashtirish bosqichi moslashtirish sifatini belgilovchi asosiy omil ekanini ko‘rsatdi. O‘xshashlik o‘lchovining o‘zi bir xil qolgan holda, faqat satrlarni oldindan yagona shaklga keltirish natijani tubdan o‘zgartirdi.

Shu bilan birga natijalar kichik nazorat to‘plamida olingan va umumlashtirish uchun yetarli emas. Keyingi bosqichda qoidalar real manbalardan yig‘ilgan kengroq to‘plamda sinaladi, chegaralar esa aniqlik va to‘liqlik ko‘rsatkichlari asosida optimallashtiriladi. Faqat ism-familiyaga tayanmay, tug‘ilgan sana va pasport raqamini ham hisobga oluvchi ko‘p atributli moslashtirishga o‘tish rejalashtirilmoqda. Bunda pasport ma’lumotlari shaxsga doir ma’lumot hisoblangani uchun ularni qayta ishlash qonunchilik talablariga muvofiq tashkil etiladi [11].

Xulosa

Xorijiy talabalar ism-familiyalarining turli yozuv tizimlari va transliteratsiya an’analaridagi farqlarini bartaraf etuvchi normallashtirish qoidalari hamda Jaro–Winkler o‘lchoviga asoslangan moslashtirish algoritmi taklif etildi. Nazorat to‘plamida normallashtirish bir shaxsga tegishli yozuvlarni

turli shaxslar yozuvlaridan aniq ajratish imkonini berdi. Algoritm xorijiy talabalar ma’lumotlarini ko‘p manbali tekshirish tizimining yozuvlarni birlashtirish moduli sifatida qo‘llaniladi.

Foydalanilgan adabiyotlar

1. O‘zbekiston Respublikasining 1993-yil 2-sentyabrdagi "Lotin yozuviga asoslangan o‘zbek alifbosini joriy etish to‘g‘risida"gi Qonuni (1995-yil 6-maydagi o‘zgartirishlar bilan).
2. Fellegi I.P., Sunter A.B. A theory for record linkage // Journal of the American Statistical Association. – 1969. – Vol. 64, No. 328. – P. 1183–1210. – DOI: 10.1080/01621459.1969.10501049.
3. Winkler W.E. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage // Proceedings of the Section on Survey Research Methods. – American Statistical Association, 1990. – P. 354–359.
4. Elmagarmid A.K., Ipeirotis P.G., Verykios V.S. Duplicate record detection: A survey // IEEE Transactions on Knowledge and Data Engineering. – 2007. – Vol. 19, No. 1. – P. 1–16. – DOI: 10.1109/TKDE.2007.250581.
5. Christen P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. – Berlin; Heidelberg: Springer, 2012. – 270 p. – DOI: 10.1007/978-3-642-31164-2.
6. Christen P. A comparison of personal name matching: Techniques and practical issues // Proceedings of the 6th IEEE International Conference on Data Mining Workshops (ICDMW). – 2006. – P. 290–294.
7. Cohen W.W., Ravikumar P., Fienberg S.E. A comparison of string distance metrics for name-matching tasks // Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb). – 2003. – P. 73–78.
8. Elov B.B. NLP texnologiyasiga asoslangan o‘zbek tilidagi matnlarni avtomatik tahlil qilish va ishlov berishning lingvistik usullari, modellari va axborot tizimi: texnika fanlari doktori (DSc) dissertatsiyasi avtoreferati. – Toshkent: TATU, 2026.
9. Papadakis G., Skoutas D., Thanos E., Palpanas T. Blocking and filtering techniques for entity resolution: A survey // ACM Computing Surveys. – 2020. – Vol. 53, No. 2. – Article 31. – DOI: 10.1145/3377455.
10. Jaro M.A. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida // Journal of the American Statistical Association. – 1989. – Vol. 84, No. 406. – P. 414–420. – DOI: 10.1080/01621459.1989.10478785.
11. O‘zbekiston Respublikasining 2019-yil 2-iyuldagi O‘RQ-547-son "Shaxsga doir ma’lumotlar to‘g‘risida"gi Qonuni.