OLIY TA’LIM MUASSASALARIDA XORIJIY TALABALARGA OID MA’LUMOTLAR SIFATINI BAHOLASH O‘LCHAMLARI VA INTEGRAL KO‘RSATKICHI

Abdullayev Javoxir Jaxongirovich
Buxoro davlat texnika universiteti o‘qituvchisi, i.f.f.d. (PhD)
https://doi.org/10.5281/zenodo.22934631

Annotatsiya. Maqolada xorijiy talabalarga oid ma’lumotlar sifatini baholash uchun beshta o‘lcham — to‘liqlik, yaroqlilik, izchillik, dolzarblik va yagonalik — tanlangan va har biri uchun hisoblash formulasi taklif etilgan. O‘lchamlarni muassasa toifasiga qarab vaznlangan integral ko‘rsatkichga birlashtirish usuli ishlab chiqilgan. Usul 1015 ta yozuvdan iborat sintetik ma’lumotlar to‘plamida sinovdan o‘tkazilgan va eng kuchsiz o‘lcham izchillik ekani, uning asosiy sababi esa muddati o‘tgan vizaga ega talabalarning faol holatda qolishi ekani ko‘rsatilgan.
Kalit so‘zlar: ma’lumotlar sifati, sifat o‘lchamlari, integral ko‘rsatkich, axborot tizimi, xorijiy talabalar, izchillik, to‘liqlik.

Аннотация. В статье для оценки качества данных об иностранных студентах выбраны пять измерений — полнота, валидность, согласованность, актуальность и уникальность — и для каждого предложена формула расчёта. Разработан способ объединения измерений в интегральный показатель, взвешенный с учётом типа учреждения. Метод апробирован на синтетическом наборе из 1015 записей; показано, что наиболее слабым измерением является согласованность, основной причиной чего является сохранение активного статуса у студентов с истёкшей визой.
Ключевые слова: качество данных, измерения качества, интегральный показатель, информационная система, иностранные студенты, согласованность, полнота.

Abstract. Five dimensions for assessing the quality of international student data are selected — completeness, validity, consistency, timeliness and uniqueness — and a computation formula is proposed for each. A method is developed for combining the dimensions into an integral index weighted by institution type. The method is tested on a synthetic dataset of 1,015 records; consistency is shown to be the weakest dimension, mainly because students with expired visas remain in active status.
Keywords: data quality, quality dimensions, integral index, information system, international students, consistency, completeness.

Kirish

Oliy ta’lim muassasalarida xorijiy talabalar haqidagi ma’lumotlar kontingent hisobi, migratsiya ro‘yxati, hisobotlar va boshqaruv qarorlari uchun asos bo‘ladi. Agar bu ma’lumotlarda xato bo‘lsa, undan hosil qilingan barcha ko‘rsatkichlar ham noto‘g‘ri chiqadi. Shu bilan birga, amaliyotda ma’lumotlar sifati odatda umumiy baholanadi yoki umuman o‘lchanmaydi: "ma’lumotlar yaxshi" yoki "ma’lumotlarda kamchilik bor" degan xulosadan nariga o‘tilmaydi.

Sifatni boshqarish uchun avvalo uni o‘lchash kerak. Maqolaning maqsadi xorijiy talabalar ma’lumotlari uchun sifat o‘lchamlarini tanlash, har biri uchun hisoblash formulasini berish va ularni muassasa xususiyatini hisobga oluvchi integral ko‘rsatkichga birlashtirishdan iborat.

Adabiyotlar tahlili

Ma’lumotlar sifati ko‘p o‘lchamli tushuncha ekanligi R. Wang va D. Strong ishida asoslangan [1]. L. Pipino, Y. Lee va R. Wang sifatni obyektiv baholash uchun oddiy nisbat, min/max operatsiyalari va vaznli o‘rtacha kabi funksional shakllarni taklif qilgan [2]. C. Batini va hammualliflar sifatni baholash metodologiyalarini tizimlashtirib, to‘liqlik, aniqlik, izchillik va dolzarblikni eng ko‘p qo‘llaniladigan o‘lchamlar sifatida ajratgan [3]. ISO 8000-8 standarti sifat tushunchalari va ularni o‘lchash tamoyillarini

belgilaydi [4], ISO/IEC 25012 esa ma’lumotlar sifatining o‘n besh xususiyatdan iborat modelini taklif qiladi [5].

Amaliy vositalar orasida Amazon kompaniyasida ishlab chiqilgan Deequ tizimi alohida o‘rin tutadi: unda sifat talablari "ma’lumot uchun unit-testlar" ko‘rinishida deklarativ yoziladi va katta hajmdagi ma’lumotlarda avtomatik tekshiriladi [6]. Sifatni o‘lchash va monitoring qilish vositalarining so‘nggi sharhi [7] ko‘rsatishicha, vositalarning aksariyati o‘lchamlarni alohida hisoblaydi, ularni soha xususiyatiga qarab vaznlash esa foydalanuvchi zimmasida qoladi.

Ta’lim sohasida, xususan xorijiy talabalar ma’lumotlari uchun maxsus moslashtirilgan o‘lchamlar tizimi adabiyotda deyarli uchramaydi.

Tadqiqot metodologiyasi

Xorijiy talaba yozuvi uchun beshta o‘lcham tanlandi. Tanlovda ikki mezonga tayanildi: o‘lcham avtomatik hisoblanishi kerak va u boshqaruv uchun ahamiyatli xatoni aks ettirishi kerak. Formulalar B. Heinrich va hammualliflar sifat metrikalariga qo‘ygan talablarga, xususan qiymatlarning [0; 1] oralig‘ida normallashgani va oson izohlanishiga muvofiq tuzildi [8]. N — yozuvlar soni bo‘lsin.

To‘liqlik — majburiy maydon to‘ldirilgan yozuvlar ulushi:

$$Q(\text{to‘l}) = N(\text{to‘ldirilgan}) / N \quad (1)$$

Yaroqlilik — belgilangan format yoki qiymatlar ro‘yxatiga mos yozuvlar ulushi (masalan, pasport raqami shabloni):

$$Q(\text{yar}) = N(\text{formatga mos}) / N(\text{to‘ldirilgan}) \quad (2)$$

Izchillik — yozuv ichidagi mantiqiy qoidalar bajarilgan ulush. Biz ikki qoidani oldik: fuqarolik pasport bergan davlatga mos bo‘lishi (R1) va viza muddati tugagan talaba "faol" holatda qolmasligi (R2):

$$Q(\text{izch}) = P(R1) \cdot P(R2) \quad (3)$$

Dolzarblik — oxirgi yangilanishdan beri belgilangan muddat (T = 180 kun) o‘tmagan yozuvlar ulushi:

$$Q(\text{dol}) = N(t0 - t(\text{yangilanish}) \le T) / N \quad (4)$$

Yagonalik — takroriy yozuvlarga ega bo‘lmagan ulush:

$$Q(\text{yag}) = 1 - N(\text{takroriy}) / N \quad (5)$$

Integral ko‘rsatkich vaznli yig‘indi ko‘rinishida hisoblanadi:

$$I = \Sigma\, w(j) \cdot Q(j), \quad \Sigma\, w(j) = 1 \quad (6)$$

Vaznlar muassasa toifasiga qarab tanlanadi. Masalan, tibbiyot OTMlarida talabaning fuqaroligi va viza holati klinik amaliyotga qo‘yilish, xorijiy tashkilotlar bilan hisob-kitob kabi jarayonlarga bevosita ta’sir qiladi, shuning uchun izchillikka kattaroq vazn beriladi (2-jadval). Ushbu vaznlar dastlabki ekspert baholashga asoslangan bo‘lib, keyingi bosqichda ekspert so‘rovnomasi orqali aniqlashtiriladi.

Usulni sinash uchun 1015 ta yozuvdan iborat sintetik to‘plam yaratildi. Unda real vaziyatga yaqin xatolar ataylab kiritildi: 4 % yozuvda pasport raqami yo‘q, 2 % da formati noto‘g‘ri, 3 % da fuqarolik pasportga mos emas, 5 % yozuv 180 kundan ortiq yangilanmagan, 15 ta yozuv takrorlangan. Viza muddati tasodifiy taqsimlangan, barcha talabalar esa "faol" holatda qoldirilgan. Hisoblashlar Python dasturlash tilida pandas kutubxonasi [9] yordamida bajarildi.

Tahlil va natijalar

2-jadval. Sintetik to‘plam bo‘yicha sifat o‘lchamlari va integral ko‘rsatkich

O‘lcham	Qiymat	Vazn (tibbiyot)	Vazn (texnika)
To‘liqlik	0,961	0,20	0,25
Yaroqlilik	0,980	0,15	0,20
Izchillik	0,895	0,35	0,25
Dolzarblik	0,951	0,15	0,15
Yagonalik	0,985	0,15	0,15

O‘lcham	Qiymat	Vazn (tibbiyot)	Vazn (texnika)
Integral ko‘rsatkich I	–	0,943	0,950

Natijalar shuni ko‘rsatadiki, eng past qiymat izchillikka to‘g‘ri keladi (0,895). Uning tarkibi tahlil qilinganda, asosiy sabab fuqarolik xatolari emas, balki viza muddati tugagan 79 nafar talabaning (7,8 %) tizimda "faol" holatda qolishi ekani aniqlandi. Boshqacha aytganda, xato alohida maydonda emas, maydonlar orasidagi bog‘liqlikda yashiringan. Oddiy to‘liqlik yoki format tekshiruvi bunday xatoni ko‘rmaydi.

Bir xil ma’lumotlar to‘plami uchun integral ko‘rsatkich tibbiyot toifasida pastroq (0,943), texnika toifasida esa yuqoriroq (0,950) chiqdi. Bu farq kichik ko‘rinsa ham, muhim mazmunga ega: xatoning ahamiyati muassasa turiga bog‘liq va barcha OTMlar uchun yagona sifat me’yorini qo‘llash ularning haqiqiy holatini to‘g‘ri aks ettirmaydi.

Muhokama

Taklif etilgan yondashuvning afzalligi — har bir o‘lcham avtomatik hisoblanadi va integral ko‘rsatkich nima sababdan pasayganini ko‘rsatish mumkin. Bu ma’lumotlar bilan ishlovchi xodimga qaysi yozuvlarni birinchi navbatda tekshirish kerakligini aniq ko‘rsatadi.

Cheklovlar ham bor. Aniqlik o‘lchami, ya’ni qiymatning real hayotdagi holatga mosligi, bitta manba ichida to‘liq baholanmaydi; buning uchun boshqa manbalar bilan solishtirish kerak. Shuningdek, natijalar sintetik ma’lumotlarda olingan, real ma’lumotlarda xatolar taqsimoti boshqacha bo‘lishi mumkin. Bu masalalar keyingi ishlarda ko‘rib chiqiladi.

Xulosa

Xorijiy talabalar ma’lumotlari uchun beshta sifat o‘lchami va ularning hisoblash formulalari taklif etildi, muassasa toifasiga bog‘liq vaznli integral ko‘rsatkich ishlab chiqildi. Sintetik ma’lumotlardagi tajriba eng xavfli xatolar maydonlar orasidagi mantiqiy nomuvofiqlikda ekanini ko‘rsatdi. Olingan natijalar ma’lumotlar sifatini monitoring qiluvchi axborot tizimi modulini loyihalashda foydalaniladi.

Foydalanilgan adabiyotlar

1. Wang R.Y., Strong D.M. Beyond accuracy: What data quality means to data consumers // Journal of Management Information Systems. – 1996. – Vol. 12, No. 4. – P. 5–33. – DOI: 10.1080/07421222.1996.11518099.
2. Pipino L.L., Lee Y.W., Wang R.Y. Data quality assessment // Communications of the ACM. – 2002. – Vol. 45, No. 4. – P. 211–218. – DOI: 10.1145/505248.506010.
3. Batini C., Cappiello C., Francalanci C., Maurino A. Methodologies for data quality assessment and improvement // ACM Computing Surveys. – 2009. – Vol. 41, No. 3. – Article 16. – DOI: 10.1145/1541880.1541883.
4. ISO 8000-8:2015. Data quality — Part 8: Information and data quality: Concepts and measuring. – Geneva: ISO, 2015.
5. ISO/IEC 25012:2008. Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. – Geneva: ISO, 2008.
6. Schelter S., Lange D., Schmidt P., Celikel M., Biessmann F., Grafberger A. Automating large-scale data quality verification // Proceedings of the VLDB Endowment. – 2018. – Vol. 11, No. 12. – P. 1781–1794. – DOI: 10.14778/3229863.3229867.
7. Ehrlinger L., Wöß W. A survey of data quality measurement and monitoring tools // Frontiers in Big Data. – 2022. – Vol. 5. – Article 850611. – DOI: 10.3389/fdata.2022.850611.
8. Heinrich B., Hristova D., Klier M., Schiller A., Szubartowicz M. Requirements for data quality metrics // ACM Journal of Data and Information Quality. – 2018. – Vol. 9, No. 2. – Article 12. – DOI: 10.1145/3148238.