

TURLI TIPDAGI MA'LUMOTLARNI INTELLEKTUAL TAHLIL QILISH VA DATA MINING TEXNOLOGIYALARINING EVOLYUTSIYASI

Hamroyev Alisher Shodmonqulovich

Alfraganus universiteti dotsenti,

t.f.f.d., alisherkhsh@gmailcom

Annotatsiya.

Mazkur maqolada turli tipdagi ma'lumotlarni intellektual tahlil qilish muammolari hamda Data Mining texnologiyalarining rivojlanish bosqichlari tahlil qilingan. Predmet soha ob'ektlarini tavsiflashda qo'llaniladigan belgilar turlari, ularning turli shkalalarda ifodalanishi va noaniqlik holatlari ko'rib chiqilgan. Ma'lumotlarga dastlabki ishlov berish, belgilar fazosini shakllantirish va ob'ektlarni modellashtirish jarayonlarining intellektual tizimlar yaratishdagi ahamiyati asoslab berilgan. Shuningdek, Data Mining yo'nalishining statistik tahlildan boshlab mashinani o'qitish, sun'iy intellekt va chuqur o'rganish texnologiyalarigacha bo'lgan evolyutsiyasi yoritilgan. Turli tipli ma'lumotlarni tahlil qilishda tasniflash, klasterlash, regressiya va bashoratlash usullarining o'rni ko'rsatib berilgan. Tadqiqot natijalari zamonaviy intellektual tizimlarni ishlab chiqishda ma'lumotlarni intellektual tahlil qilish usullarini takomillashtirish zarurligini ko'rsatadi.

Kalit so'zlar: Data Mining, sun'iy intellekt, mashinani o'qitish, timsollarni aniqlash, belgilar fazosi, Big Data, klasterlash, tasniflash.

Kirish. O'tgan asrning o'rtalaridan boshlab insonning ko'rish, eshitish, his qilish, hid bilish kabi intellektual imkoniyatlarini EHMga tadbiq etishga asoslangan kibernetika fanining vujudga kelishi tabiiy jarayon (ob'ekt, predmet, hodisa, holat, voqea va boshqalar) haqidagi axborotlarni yetarlicha tavsiflay olish xususiyatlariga ega bo'lgan intellektual tizimlarning paydo bo'lishiga olib keldi [1]. Inson xatti-harakati va intellektual xususiyatlarini modellashtirish va ularni intellektual tizimlarga joriy qilish bugungi kun ilm-fanida tadqiqotchilar tomonidan chuqur o'rganishni talab etadigan nihoyasiz masalalar sirasiga kiradi. Zamonaviy kompyuter texnologiyalari, ma'lumotlarni intellektual tahlil qilish, timsollarni aniqlash, matematik va kompyuterli modellashtirish, noravshan to'plamlar nazariyasi kabi yo'nalishlarning rivojlanib borishi bunday muammolarni hal etishga xizmat qiladi. So'nggi yillarda xalq xo'jaligining barcha tarmoqlarida insonning muayyan vaqtda bajarishga qodir bo'lmagan yumushlarini bir zumda hal etishga xizmat qiladigan, intellektual tizimlardan foydalanadigan roboto-texnik qurilmalarni ishlab chiqarish va ulardan keng foydalanish ommalashib bormoqda.

Hozirgi kunda turli manbalardan olinayotgan ma'lumotlar hajmining keskin ortishi hamda ularning turli tipli va noaniq tabiatga ega bo'lishi an'anaviy tahlil usullarining samaradorligini cheklab qo'yimoqda. Shu sababli heterogen ma'lumotlarni yagona belgilar fazosida ifodalash va ularni intellektual tahlil qilish usullarini takomillashtirish dolzarb ilmiy muammolardan biri hisoblanadi. Mazkur maqolaning maqsadi turli tipdagi ma'lumotlarni intellektual tahlil qilish jarayonlarini tizimli o'rganish hamda Data Mining texnologiyalarining rivojlanish bosqichlarini tahlil qilish orqali zamonaviy intellektual tizimlarni yaratishdagi metodologik yondashuvlarni asoslashdan iborat.

Tahliliy qism. Zamonaviy intellektual tizimlarning yanada takomillashtirilishi uchun predmet soha ob'ektlari haqidagi axborotlarni chuqur o'rganishda zamonaviy o'lchov asboblardan foydalanib, belgilar fazosini shakllantirish, dastlabki ishlov berish, tahlil qilish masalalarini yuqori aniqlikda hal etishni taqoza etadi. O'rganilayotgan ob'ekt haqidagi axborotlarni intellektual tahlil qilish maqsadida ularni xarakterlovchi belgilarni to'g'ri aniqlash va o'lchash, ularning qiymatlar sohasini aniqlash va shkalalashtirish muammolarining hal etilishi uning real holatini yoritishda muhim omil hisoblanadi [2, 3].

Bir jinsli $\{S\}$ ob'ektlar to'plamining $\{X\}$ tasnifiy belgilar orqali ifodalanishi – *belgili tavsiflash* deb yuritiladi. Predmet soha ob'ektlari modelini belgili tavsiflashning murakkabligi – bu belgilarning turli, ko'p o'lchovli fizik xususiyatlardan iborat bo'lishiga bog'liq [4]. Masalan, tirik mavjudotlarni belgili tavsiflashda ularning geografik, fizik, morfologik, kimyoviy, genomik kabi turli xususiyatlaridan foydalanish mumkin. Ammo har bir turkum belgilari bir-biriga bog'liq bo'lmagan yo'nalishlar usullari orqali tadqiq qilinadi. Bu uzoq vaqt va ortiqcha mablag'ni talab qiladi. Shu bois, qo'yilgan masalaga ko'ra ob'ektlarning matematik va kompyuterli modellarini shakllantirish maqsadga muvofiqdir.

Bundan tashqari predmet soha ob'ektlarini modellashtirishda quyidagi noaniqlik ko'rinishlari paydo bo'lishi mumkin:

- *ko'p jinslilik*, bir turdagi ob'ektlarning turli o'lchamli belgilar bilan ifodalanishi va ularni bir jinsli ko'rinishga keltirish;
- bir belgining *ko'p sonli (ikki yoki undan ortiq) ost belgilar bilan ifodalanishi*. Bunday holatlarda ob'ektlarni qiyoslash murakkabligi ortib ketadi;
- belgining *ko'p qiymatliligi* (belgining qiymatlari ikki yoki undan ortiq bo'lishi);
- *aprior ma'lumotlardagi noaniqliklar* (ob'ekt belgisidagi axborotning mavjud bo'lmashligi yoki mantiqan xato qiymatlar bilan berilishi);
- *belgilari orasidagi o'zaro bog'liqliklarni* aniqlashning murakkabligi;
- turli shkalalarda o'lchanadigan *ham sifat, ham miqdor* jihatdan bo'ladigan belgilar;

- *belgilarining turli tipliligi* (mantiqiy, sonli, noravshan va.h.k.).

Tadqiqot ishida ob'ektlar haqidagi axborotlarning turli tipliligi va ayniqsa noravshan to'plamlar orqali ifodalanganligiga alohida e'tibor qaratiladi.

Ob'ektlarni identifikatsiyalashda axborotga dastlabki ishlov berish, ularning belgilarini (mantiqiy, nominal, miqdoriy, tartiblangan, lingvistik, intervalli, noravshan ko'rinishda) tiplashtirilishi muhim o'rin tutadi.

Timsollarni aniqlashning klassik usullarida amaliy sohalar ob'ektlarining sifatiy belgilari 1-jadvaldagi singari miqdoriy ko'rinishlarga o'tkazilgan.

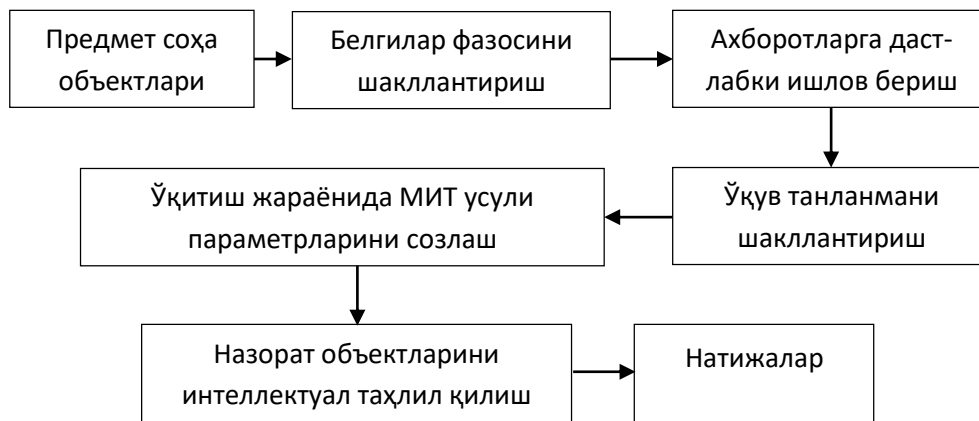
1-jadval

Ob'ekt belgilarining ba'zi tiplari

	Miqdoriy (sonli) belgilar	Sifatiy belgilar
Mantiqiy	{0, 1}	{“mavjud emas”, “mavjud”}
Diskret sonlar	aniq natural yoki haqiqiy sonlar	<i>k</i> tartibli miqdorlar, noravshan sonlar
Uzluksiz miqdorlar	(datchik yoki o'lchov asboblari tomonidan olingan qiymatlar)	oralikli qiymatlar, noravshan sonlar
Nomlanish	tartib raqamga e'tibor berilgan	atamalar, ranglar va h.k. nomi
Lingvistik	ranjirovka qilingan miqdorlar	Noravshan to'plamlar (so'zlar, jumlar)

Klassik matematik usullardan foydalangan ma'lumotlarni tahlil qilish yo'nalishining mantiqiy davomi hisoblangan ma'lumotlarni intellektual tahlil qilish (MIT, Data Mining) yo'nalishida axborotlarning turli tiplilik holatida tahlil qilish muammosiga alohida e'tibor qaratildi.

Inson qabul qiladigan qarorlarga mos bo'lgan ma'lumotlarni intellektual tahlillash (MIT) usullari ishlab chiqish hamda ulardan intellektual tizimlarni rivojlantirishda foydalanish muhim hisoblanadi. MIT – timsollarni aniqlash, sun'iy intellekt, axborot (tasvir, ovoz)larga raqamli ishlov berish, statistika, algoritmlashtirish, vizuallashtirish, ma'lumotlar va bilimlar bazasi kabi fan yo'nalishlarini o'z ichiga qamrab oladi [5]. MITda har bir fan yo'nalishi bo'yicha ishlab chiqilgan usul va algoritmlarning o'zaro integratsiyasidan hosil qilinadigan gibrid usullarni ishlab chiqish ommalashib bormoqda. 1-rasmda intellektual tizimlarning asosiy bosqichlari keltirilgan bo'lib, mazkur ishda MIT usullarini takomillashtirish vazifasiga alohida e'tibor qaratiladi.



1-rasm. Predmet soha ob'ektlarini MIT usullari orqali tahlil qilish

MIT yo'nalishining rivojlanish bosqichlari. Ma'lumotlarni saqlash va qayta ishlash texnologiyasi rivojlanishi ilk bora hisoblash mashinalarida ma'lumotlarni saqlash imkoniyati yaratilishi bilan bog'liq [6].

60-yillarda ilk ma'lumotlar bazasini boshqarish tizimlari (MBBT) iste'molga kiritilib, bir qator ma'lumotlar modellari, jumladan ierarxik, tarmoqli va shu kabilar.

70-yillarda ma'lumotlarning relyatsion modellari va relyatsion MBBTlar iste'molga kiritilgan.

80-yillarda MBBTlarning rivojlantirilgan relyatsion-ob'ektga yo'naltirilgan, kengaytirilgan relyatsion, deduktiv kabi ko'rinishlari, shu qatorda MBBTlarning maxsus yo'nalishlarga "yo'naltirilgan" turlari vujudga keldi.

90-yillarda multimediali ma'lumotlar bazasi (MB), www serverlari, OLAP va nihoyat Data Mining kabi rivojlangan ilg'or texnologiyalar joriy etila boshladi.

MIT texnologiyasi yo'nalishining asoschisi – Grigoriy Piateskiy-Shapiro shunday fikr bildirgan: "MIT– bu inson faoliyatining turli xil sohalarida qaror qabul qilish uchun zaruriy, avvaldan noma'lum bo'lgan, amaliy foydali, ishlov berilmagan ma'lumotlarni oshkor etish jarayonidir". Bu ta'rifda, tushunarsiz va tizimlashmagan axborotlarni ma'lumotlar to'plamidan ajratib olish, ulardan yashirin qonuniyatlarni izlab topish va undan foydalanish uchun yaroqli holda taqdim etish jarayoni ekanligi uqtirilgan [6].

Ba'zi manbalarda Data Mining – ma'lumotlar bazalari, matnlar, rasmlar, video va h.k. ko'rinishdagi ko'p hajmli, har xil ma'lumotlar manbalaridan bilimlar (qonuniyatlar)ni izlab topish jarayonidir. Olingan bilimlar ishonchli, foydali va tadbiq etiladigan bo'lishi kerak, deb keltirilgan [7]. Ushbu sohaning rivojlanishi bir necha ilmiy bosqichlarni o'z ichiga oladi.

1960–1970-yillarda ma'lumotlarni tahlil qilish asosan *statistik analiz* doirasida rivojlandi. Bu davrda ma'lumotlarni eksplorativ tahlil qilish tushunchasi amerikalik statistik olim John Tukey tomonidan ilgari surilib, keyinchalik Data Mining metodologiyasining nazariy asosiga aylandi.

1980-yillarda *ma'lumotlar bazalari texnologiyalarining rivojlanishi* natijasida katta hajmdagi ma'lumotlarni saqlash imkoniyati paydo bo'ldi. Shu davrda bilimlarni ma'lumotlar bazasidan ajratib olish (Knowledge Discovery in Databases — KDD) tushunchasi shakllandi. KDD jarayoni ma'lumotlarni tayyorlash, transformatsiya qilish, modellash va bilim chiqarish bosqichlarini o'z ichiga oldi.

1990-yillarda Data Mining mustaqil ilmiy yo'nalish sifatida shakllandi. Klassifikatsiya, klasterlash, assotsiativ qoidalar va regressiya algoritmlari keng rivojlandi. Ayniqsa, assotsiativ qoidalarni aniqlash uchun Apriori algoritmi sanoat va biznes sohalarida keng qo'llanila boshladi. Bu davrda ma'lumotlar ombori (Data Warehouse) va OLAP texnologiyalari ham rivojlandi.

2000-yillardan boshlab Data Mining mashinaviy o'qitish (Machine Learning) bilan integratsiyalashdi. Tadqiqotchilar, jumladan Jiawei Han tomonidan yaratilgan ilmiy ishlar Data Mining nazariyasi va algoritmlarini tizimlashtirishda muhim rol o'ynadi. Ushbu bosqichda web-ma'lumotlar, matnlar va multimedia ma'lumotlarini tahlil qilish yo'nalishlari paydo bo'ldi.

2010-yillardan hozirgi kungacha Data Mining *Big Data*, *bulutli hisoblash* va *sun'iy intellekt* texnologiyalari bilan chambarchas bog'landi. Chuqur o'rganish (Deep Learning) algoritmlarining paydo bo'lishi tasvirlar, video va sensor ma'lumotlarini avtomatik tahlil qilish imkoniyatini keskin kengaytirdi. Natijada Data Mining sog'liqni saqlash, moliya, qishloq xo'jaligi, kiberxavfsizlik va biometrik identifikatsiya kabi ko'plab sohalarda asosiy analitik vositaga aylandi.

So'nggi o'n yillikda ma'lumotlarning keskin ortib borishi katta hajmli ma'lumotlar ustida ishlash muammosini paydo qildi. Buning natijasida 2008 yilda "Big data" ("Katta ma'lumotlar") termini paydo bo'ldi [8, 9] va bu yo'nalish MIT yo'nalishining deyarli barcha usul va algoritmlarini qamrab oladi. Undan tashqari, ma'lumotlarni yig'ishning avtomatik vositalari, MBBTni keng qamrovli qo'llanilishi, elektron hujjat almashinuvi, www, multimediali arxivlar va boshqalar. Bularning barchasi axborot hajmning o'sishiga va saqlanayotgan axborot tuzilmasini murakkablashishiga olib kelmoqda. Bu kabi muammolarni bartaraf etishda MITning quyidagi masalalari ustida tadqiqot ishlarini olib borish zarurligi asoslangan: tasniflash (classification), klasterlash (clustering), bashoratlash (forecasting), baholash (estimation), assotsiativ qoidalarni izlash (association rule mining), regressiya (regression), vizuallashtirish (visualization, graph mining), web mining, text mining va h.k.

Klassifikatsiya (tasniflash, o'qituvchi yordamida o'qitish, guruhlariga ajratish) masalasidir. Uning yordamida yangi, nazoratdagi ob'ektni avvaldan ma'lum bo'lgan u yoki bu sinfga tegishliligi aniqlanadi. *Klasterizatsiya* (avtomatik klassifikatsiya, taksonomiya, o'qituvchisiz o'qitish) masalasida ob'ektlarning o'zaro yaqinligi bo'yicha avtomatik ravishda klasterlarga ajratiladi va ular ajratilgan klasteriga (sinfiga)

tegishli deb olinadi. Klasterizatsiya usullari bilan aksariyat hollarda katta hajmli turli tipli ma'lumotlarni tadqiq qilish bilan bog'liq masalalarni yechish uchun xizmat qiladi. *Regression tahlil* o'zgaruvchilar o'rtasidagi munosabatlar miqdoriy ifodalangan holda ularning ayrim kombinatsiyalari ko'rinishida ishlatiladi. Unda asosan, standart statistik usullar qo'llaniladi. Bulardan biri chiziqli regressiyadir. Vaqt ketma-ketliklarini *bashorat qilish* boshlang'ich ma'lumotlarni yoki yarim qayta ishlangan ma'lumotlarni tahlil qilish asosida bo'lg'usi bashorat o'zgaruvchilari qiymatini baholash imkonini beradi. *Assotsiatsiya* bir nechta hodisa va faktlarning bog'liqligini topish usuli hisoblanadi. Tartibni keltirib chiqarish tahlil qilinayotgan ma'lumotlarni vaqt tartibi zanjirini aniqlash hamda sochilib yotgan vaqt tartiblarini guruhlashtirishni amalga oshiradi. Ya'ni qo'yilgan maqsad yoki oldindan kutilgan yechimdan xatolik og'ishi aniqlanadi. Bulardan tashqari so'nggi vaqtlarda ma'lumotlarni vizual talqin etish (vizuallashtirish), matnli axborotlarni tahlil etish (web mining, text mining) kabi yo'nalishlar ham rivojlanayotganiga alohida e'tibor qaratilgan.

O'tgan asrning 60 yillaridan boshlab bugungi kungacha MITning asosiy masalalari hisoblanagan tasniflash va klasterizatsiya, asosan, *timsollarni aniqlash* yo'nalishining masalalari sifatida ko'plab manbalarda ularni hal etish yondashuvlari, usullari va algoritmlari batafsil yoritilgan [9]. *Timsollarni aniqlash* – bu muayyan predmet sohaning bir jinsli ob'ektlar to'plamini o'qitish va tanib olishga asoslangan usul va algoritmlarni tadqiq etuvchi fandır. Tirik mavjudot borki, u hayotining har daqiqasida tanib olish masalasini hal etadi. Inson tanib olish masalalarini hal etishi uchun u tashqi muhit haqidagi zaruriy axborotlarga ega bo'lish hamda hisoblash texnikalarini intellektuallashtirish jarayonida ko'p bosqichli o'qitish jarayonining amalga oshirilishi talab etiladi.

Timsollarni aniqlash dastlab signal va tasvirlarni qayta ishlash masalalarida qo'llanilgan bo'lsa, keyinchalik kompyuter ko'rishi, biometrik identifikatsiya va nutqni aniqlash tizimlarining asosiy nazariy bazasiga aylandi.

Sun'iy intellekt yo'nalishi XX asr o'rtalarida matematik mantiq va hisoblash nazariyasining rivojlanishi natijasida shakllana boshladi. Ingliz olimi Alan Turing mashinalarning intellektual imkoniyatlarini baholash g'oyasini ilgari surib, sun'iy intellekt nazariyasining rivojiga muhim hissa qo'shgan. 1956-yilda o'tkazilgan Dartmouth Conference ilmiy seminaridan so'ng sun'iy intellekt mustaqil ilmiy yo'nalish sifatida rivojlana boshladi. Dastlabki tadqiqotlar mantiqiy qoidalar, ekspert tizimlar va bilimlarga asoslangan modellarni yaratishga qaratilgan edi [10].

Mashinani o'qitish — kompyuter tizimlariga ma'lumotlar asosida tajriba orqali o'rganish va mustaqil qaror chiqarish imkonini beruvchi usullar majmuasidir. Ushbu tushuncha 1959-yilda Arthur Samuel tomonidan ilmiy muomalaga kiritilgan. 1990-yillardan boshlab mashinani o'qitish algoritmlari jadal rivojlandi. Jumladan: qaror

daraxtlari, tayanch vektor usullari, sun'iy neyron tarmoqlar, ansambl algoritmlari keng qo'llanila boshlandi [11].

Mashinani o'qitish timsollarni aniqlash muammolarini yechishda asosiy vositaga aylandi, chunki u qoidalarni qo'lda yozish o'rniga ma'lumotlardan avtomatik ravishda model yaratish imkonini berdi.

Chuqur o'rganish (Deep Learning) bosqichi 2010-yillardan boshlab hisoblash texnikasi quvvatining oshishi va katta hajmdagi ma'lumotlarning paydo bo'lishi natijasida ko'p qatlamli neyron tarmoqlarga asoslangan chuqur o'rganish usullari rivojlandi. Ushbu yo'nalishning rivojlanishiga Geoffrey Hinton olib borgan ilmiy tadqiqotlar katta ta'sir ko'rsatdi.

Chuqur o'rganish quyidagi masalalarda yuqori aniqlikka erishdi: tasvirlarni tanish, nutqni aniqlash, matnlarni avtomatik tahlil qilish, biometrik autentifikatsiya. Natijada timsollarni aniqlash jarayonlari klassik statistik usullardan o'rganishga asoslangan neyron modellarga o'ta boshladi.

Timsollarni aniqlash, mashinani o'qitish va sun'iy intellekt dastlab mustaqil yo'nalishlar sifatida rivojlangan bo'lsa-da, hozirgi davrda ular o'zaro integratsiyalashgan yagona ilmiy ekotizimni tashkil etadi. Mashinani o'qitish algoritmlarining rivojlanishi timsollarni aniqlash samaradorligini sezilarli oshirdi, sun'iy intellekt esa ushbu metodlarni real amaliy masalalarni yechuvchi intellektual tizimlarga aylantirdi.

Xulosa. Mazkur tadqiqotda turli tipdagi ma'lumotlarni intellektual tahlil qilish jarayonlari hamda Data Mining texnologiyalarining evolyutsiyasi tahlil qilindi. Belgilar fazosini shakllantirish, ma'lumotlarga dastlabki ishlov berish va modellashtirish bosqichlari intellektual tizimlar samaradorligini ta'minlovchi asosiy omillar ekanligi ko'rsatildi. Data Mining, mashinani o'qitish va sun'iy intellekt texnologiyalarining integratsiyasi murakkab va katta hajmli ma'lumotlarni tahlil qilish imkoniyatlarini kengaytiradi. Kelgusida gibrid intellektual usullarni ishlab chiqish turli predmet sohalarida qaror qabul qilish jarayonlarini yanada takomillashtirishga xizmat qiladi.

Adabiyotlar

1. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. — 4th ed. — Pearson, 2021. — 1136 p.
2. Bishop C. M., Bishop H. Deep Learning: Foundations and Concepts. — Springer, 2024.
3. Goodfellow I., Bengio Y., Courville A. Deep Learning. — MIT Press, 2016.
4. Xu D., Tian Y. "A Comprehensive Survey of Clustering Algorithms." Annals of Data Science, 2020, Vol. 7, pp. 1–42.
5. Aggarwal C. C. Data Mining: The Textbook. — Springer, 2018. — 734 p.

6. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. — 3rd ed. — O'Reilly Media, 2022.
7. Provost F., Fawcett T. Data Science for Business. — 2nd ed. — O'Reilly Media, 2023.
8. Zikopoulos P., deRoos D., Parasuraman K., Deutsch T., Giles J., Corrigan D. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. — McGraw-Hill, 2019.
9. Hashem I.A.T. et al. "The Rise of 'Big Data' on Cloud Computing: Review and Open Research Issues." Information Systems, 2021.
10. J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, Hanover, NH, USA: Dartmouth College, 1955.
11. Murphy K. P. Probabilistic Machine Learning: An Introduction. — MIT Press, 2022.