UMUMIY LUG`AT ASOSIDA PREDMET SOHALARNING MAZMUNAN HAQQONIYLIGINI BAHOLASH

Xudayberganov Shoxrux Faxriidin o‘g‘li
Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti,Toshkent O`zbekiston;

So`zlarning mazmunan haqqoniyligi: Hozirgi vaqtda so'zlarning semantik yaqinligini aniqlash va analogiyalarni topish muammosini hal qilishning to'g'riligi so'zlarning vektorli tasviri sifatining asosiy mezoni hisoblanadi. So'zlarning vektor ko'rinishlarining shunga o'xshash xususiyatidan matnni tabiiy tilda qayta ishlash sohasidagi boshqa muammolarni hal qilish uchun ham foydalanish mumkin.

So'zlarning vektor ko'rinishlarining modellari bir necha o'n yillar davomida faol o'rganilgan. Bu yo'nalishdagi dastlabki ishlardan biri 1975 yillarda qilina boshlangan. Dastlabki bosqichda bunday modellar "so'z-hujjat" yoki "so'z-kontekst" tipidagi chastota matritsalarini qurish va o'zgartirishga asoslangan edi. Bu davr, xususan, LSA (Latent Semantic Analysis) latent semantik tahlil usulini yaratishni o'z ichiga oladi , uning doirasida so'zlarning vektor ko'rinishini olish uchun chastota matritsasining Singular Value Dekompozitsiyasidan foydalanish birinchi marta asoslandi. Chastota (songa asoslangan) modellar ko'rib chiqish hujjatida batafsil tavsiflangan. Turli ilmiy fanlardagi tematik hujjatlarning o'xshashligi ko'rib chiqiladi. Bunday bo'linish yaxshi yo'lga qo'yilgan va amaliyotda keng tarqalgan deb hisoblanadi. Ushbu taxminga asoslanib, tadqiqot hujjatlarni bir-biriga mos kelmaydigan sinflarga bo'lishdan foydalanadi. O'xshashlik va semantik izchillik ko'rsatkichlari taklif etiladi, ular sinf ob'ektlarining uyg'unligi munosabati asosida ixchamlik o'lchovi qiymatlari hisoblanadi.

Ushbu bosqichda, bir nechta xususiyatlardan foydalanganda, xususiyat vektorini ushbu nomzodning atama ekanligiga ishonchini ko'rsatadigan raqamga aylantirish muammosi paydo bo'ladi.Eng oddiy usul, masalan, TermExtractor usulida qo'llaniladigan chiziqli kombinatsiyadir. lug'atlar, tezauriyalar yoki ontologiyalarni boyitish, ma'lumot qidirish, ma'lumot olish, hujjatlarni tasniflash va klasterlash va hokazo. Bugungi kunga qadar atamalarni avtomatik ravishda ajratib olishning ko'plab usullari ishlab chiqilgan, ammo Ko'pgina hollarda bo'lgani kabi, matnni avtomatik qayta ishlashning boshqa vazifalarida, usullarning aksariyati kirish matnlarining tili va mavzu sohasiga sezilarli darajada bog'liq bo'lib, bu usulning amaliy qo'llanilishini tabiiy ravishda toraytiradi. Bundan tashqari, ko'pgina usullarda ma'lumotlar manbai faqat mavzu sohasidagi matnli hujjatlar to'plamidir. Ba'zi usullar tashqi resurslardan ham foydalanadi, masalan, boshqa fan sohalaridagi matnlar korpusi, qidiruv tizimlari yoki mutaxassislar tomonidan yaratilgan ontologiyalar, ammo bu resurslarning barchasi o'zlarining kamchiliklariga ega. Shunday qilib, tashqi matnli hujjatlar, shu jumladan qidiruv tizimlari tomonidan topilganlar,

tuzilishga ega emas va faqat ko'rib chiqilayotgan mavzu doirasidan tashqarida so'zlar va iboralarning paydo bo'lishi haqidagi statistik ma'lumotlardan foydalanishga imkon beradi; gen ontologiyasi kabi ixtisoslashgan ontologiyalardan foydalanish usulni boshqa fan sohalariga o'tkazish imkoniyatini amalda istisno qiladi; WordNet yoki RuThes kabi universal ontologiyalar kichik (taxminan 100-150 ming atama) va faqat mavzu sohalarining eng umumiy tushunchalarini qamrab oladi.

FOYDALANILGAN ADABIYOTLAR:

1. Ignatyev N. A. Structure Choice for Relations between Objects in Metric Classification Algorithms // Pattern Recognition and Image Analysis. 2018. V. 28. № 4. P. 590–597.

2. Ignatev N. A., Tuliyev U. Y. 2018. Analysis of the similarity and onnectivity degree of thematic documents based on measure of compactness. Problems of Computational and Applied Mathematics. 3(15): 1–10.