Modern Methods Of Determining Relevance In Web Search Systems

Eshko'ziyeva Rayhona
Shakhrisabz State Pedagogical Institute, Preschool Education,
1st year master's student

Abstract
This article examines modern methods for relevance determination in web search engines. The analysis covers classical approaches such as TF-IDF and BM25, as well as neural network-based techniques including BERT, dense retrieval, and multi-vector models. Semantic search, Learning to Rank (LTR) technologies, user behavioral signals, and multimodal search systems are discussed. The effectiveness of methods is evaluated based on the practical experience of major systems — Google, Bing, and Elasticsearch.
Keywords: relevance, web search, TF-IDF, BM25, BERT, semantic search, Learning to Rank, dense retrieval, information retrieval, neural network.
Annotatsiya:
Ushbu maqolada veb-qidiruv tizimlarida relevatlikni aniqlashning zamonaviy metodlari ko'rib chiqilgan. TF-IDF va BM25 kabi klassik yondashuvlardan tortib, neyron tarmoqqa asoslangan BERT, dense retrieval va ko'p vektorli modellapraча bo'lgan usullar tahlil qilingan. Semantik qidiruv, Learning to Rank (LTR) texnologiyalari, foydalanuvchi xulq-atvori signallari va multimodal qidiruv tizimlari muhokama etilgan. Zamonaviy amaliy tizimlar — Google, Bing, Elasticsearch — tajribasi asosida usullarning samaradorligi baholangan.
Kalit so'zlar: relevatlik, veb-qidiruv, TF-IDF, BM25, BERT, semantik qidiruv, Learning to Rank, dense retrieval, axborot qidirish, neyron tarmoq.

KIRISH

Zamonaviy axborot jamiyatida internet foydalanuvchilari har kuni milliardlab so'rovlar bilan qidiruv tizimlariga murojaat qiladi. Bunday sharoitda qidiruv natijalari sifati — ya'ni berilgan so'rovga eng mos hujjatlarni topa bilish qobiliyati — muhim texnologik muammo bo'lib qolmoqda. Ushbu muammoning markazida relevatlik tushunchasi yotadi: hujjat foydalanuvchi ehtiyojini qanchalik qondirishini o'lchash.

Relevatlikni aniqlash — axborot qidirish (Information Retrieval, IR) fanining asosiy yo'nalishlaridan biri hisoblanadi. 1950-yillardan boshlab rivojlana boshlagan ushbu soha bugun sun'iy intellekt, katta ma'lumotlar (Big Data) va chuqur o'rganish (Deep Learning) texnologiyalari bilan chambarchas bog'liq holda yangi pog'onaga ko'tarildi [1].

Ushbu maqolaning maqsadi — veb-qidiruv tizimlarida relevatlikni baholashning klassik va zamonaviy metodlarini tizimli tahlil qilish, ularning kuchli va zaif tomonlarini aniqlash hamda kelgusidagi rivojlanish tendensiyalarini ko'rsatishdir.

ADABIYOTLAR SHARHI

Relevatlikni aniqlash muammosi ilmiy adabiyotlarda keng o'rganilgan. Manning va boshqalarning [1] fundamental asarida axborot qidirish nazariyasining asoslari bayon etilgan. Salton tomonidan taklif etilgan vektor fazosi modeli (Vector Space Model) [2] klassik yondashuv sifatida hozirgacha keng qo'llaniladi.

Robertson va Zaragoza [3] tomonidan ishlab chiqilgan BM25 algoritmi hozirgi kunda ham ko'plab tijorat tizimlarining asosini tashkil etadi. Devlin va boshqalar [4] tomonidan taqdim etilgan BERT modeli esa matnni tushunishda paradigmatik o'zgarishni amalga oshirdi: ikki tomonlama transformer arxitekturasi yordamida kontekstga bog'liq so'z vektorlari hosil qilish mumkin bo'ldi.

Karpukhin va boshqalar [5] tomonidan taklif etilgan Dense Passage Retrieval (DPR) usuli zichlik asosidagi qidirishning samaradorligini namoyish etdi. Multimodal qidiruv sohasida Radford va boshqalar [7] tomonidan yaratilgan CLIP modeli matn va tasvirni birgalikda qayta ishlash imkonini berdi.

RELEVATLIKNI BAHOLASHNING KLASSIK METODLARI

TF-IDF modeli

TF-IDF (Term Frequency — Inverse Document Frequency) — axborot qidirishda eng keng tarqalgan klassik usullardan biri. Ushbu metodda so'zning hujjat uchun ahamiyati ikki ko'rsatkich asosida belgilanadi:

TF (Atama Chastotasi) — muayyan so'zning hujjat ichida necha marta uchrashi. So'z qancha ko'p uchrasa, hujjat u bilan shuncha ko'proq bog'liq deb hisoblanadi.

IDF (Teskari Hujjat Chastotasi) — so'zning barcha hujjatlar to'plamida qanchalik umumiy yoki noyob ekanligini ko'rsatadi. Kam hujjatlarda uchraydigan so'zlar yuqori IDF qiymatiga ega bo'ladi.

TF-IDF usulining soddaligi va tezligi uning asosiy afzalligi hisoblanadi. Biroq ushbu yondashuv so'zlarning semantik ma'nosini hisobga olmaydi

— sinonimlar va polisemiya muammolarini hal eta olmaydi [1].

BM25 algoritmi

BM25 (Best Match 25) — TF-IDF ning takomillashtirilgan versiyasi bo'lib, Robertson va Zaragoza tomonidan probabilistik qidirish modeli asosida yaratilgan [3]. BM25 quyidagi muhim omillarni hisobga oladi:

Hujjat uzunligiga normalizatsiya — uzun hujjatlardagi so'z chastotasi qisqaroq hujjatlarga nisbatan tuzatiladi;

Atama chastotasining to'yinish effekti — so'z juda ko'p marta uchrasa ham, ball cheksiz oshmaydi, balki ma'lum chegarada to'xtaydi;

k1 va b parametrlari orqali tizimni sozlash imkoniyati.

BM25 bugun ham Elasticsearch, Apache Solr kabi keng tarqalgan qidiruv platformalarida asosiy rankinг algoritmi sifatida qo'llaniladi. Uning semantik cheklovlariga qaramasdan, amaliyotda yuqori samaradorlik ko'rsatishi uning mashhurligini saqlab kelmoqda.

ZAMONAVIY NEYRON TARMOQ ASOSIDAGI METODLAR

BERT va transformer modellari

2018-yilda Google tomonidan taqdim etilgan BERT (Bidirectional Encoder Representations from Transformers) modeli [4] tabiiy tilni qayta ishlash (NLP) sohasida inqilob yasadi. BERT ikki tomonlama kontekstni hisobga olish orqali so'zning haqiqiy ma'nosini aniq tushunish imkonini beradi.

Relevatlikni baholashda BERT bir necha usulda qo'llaniladi:

Cross-encoder: so'rov va hujjat birgalikda BERT ga kiritiladi, natijada juftlik uchun to'g'ridan-to'g'ri relevatlik balli hosil qilinadi. Eng aniq, lekin sekin usul;

Bi-encoder: so'rov va hujjat alohida kodlanadi, vektorlar o'rtasidagi kosinus o'xshashligi relevatlik o'lchovi sifatida ishlatiladi. Tezroq, lekin biroz kam aniq;

ColBERT (ko'p vektorli): har bir token uchun alohida vektor hosil qilinadi, keyin MaxSim operatori orqali yig'ma ball hisoblanadi [8].

Dense Retrieval (Zich qidirish)

Dense Passage Retrieval (DPR) [5] — so'rovlar va hujjatlarni yuqori o'lchamli vektor fazosida ifodalash orqali relevatlikni aniqlash usuli. DPR ning asosiy afzalliklari:

Semantik o'xshashlikni aniqlash — leksik mos kelmasada, ma'no jihatidan yaqin hujjatlarni topish;

FAISS kabi taqribiy eng yaqin qo'shnini qidirish (ANN) kutubxonalari bilan integratsiya orqali millionlab hujjatlar orasida tez qidirish;

Labeled ma'lumotlar asosida fine-tuning qilish imkoniyati.
Gibrid qidiruv tizimlari
Amaliyotda eng yaxshi natijalar BM25 (sparse) va dense retrieval usullarini birlashtirib ishlatish orqali erishiladi. Bu gibrid yondashuv quyidagi sxema asosida ishlaydi:

Bosqich	Metod	Vazifasi
1-bosqich (Candidate retrieval)	BM25 + Dense	Millionlab hujjatdan 100-1000 kandidat ajratish
2-bosqich (Re-ranking)	Cross-encoder BERT	Kandidatlarni aniqroq saralash
3-bosqich (Post-processing)	Foydalanuvchi signallari	Shaxsiylashtirilgan natijalar

1-jadval. Gibrid qidiruv tizimining bosqichlari

LEARNING TO RANK TEXNOLOGIYASI

Learning to Rank (LTR) — mashina o'rganishi metodlarini qidirish natijalarini saralashga qo'llash yondashuvi. LTR tizimlari bir nechta belgilar (features) asosida hujjatlarning optimal tartibini o'rganadi [6].
LTR da foydalaniladigan asosiy belgilar guruhlari:
So'rovga bog'liq belgilar: BM25 balli, TF-IDF, so'rov-sarlavha o'xshashligi, URL da kalit so'z mavjudligi;
Hujjat sifati belgilari: PageRank, domenning ishonchliligi, sahifa yuklash tezligi, mobil moslashuvchanlik;
Foydalanuvchi xulq-atvor signallari: bosish darajasi (CTR), sahifada o'tkazilgan vaqt, qayta qidirish ko'rsatkichi (pogo-sticking).
LambdaMART, RankNet, ListNet kabi mashhur LTR algoritmlari Google va Bing kabi yirik qidiruv tizimlari tomonidan faol qo'llaniladi. Ma'lumotlarga ko'ra, Google qidiruv algoritmi 200 dan ortiq belgi asosida ishlaydi va LTR ularning muhim qismi hisoblanadi.
SEMANTIK QIDIRUV VA BILIM GRAFLARI
Semantik qidiruv — foydalanuvchi niyatini (intent) va so'rovning kontekstini tushunishga asoslangan yondashuv. Bu yo'nalishda bir necha texnologiya muhim rol o'ynaydi.
Bilim graflari (Knowledge Graphs). Google 2012-yilda Knowledge Graph ni ishga tushirdi — bu real dunyo ob'ektlari (shaxslar, joylar, tushunchalar) va ular orasidagi munosabatlarning strukturalashgan ma'lumotlar bazasi. Knowledge Graph yordamida qidiruv tizimi faqat kalit so'zlarni emas, balki haqiqiy ma'noni tushunib, to'g'ri javobni berishi mumkin.
Word2Vec va GloVe vektorlari. Ushbu modellar har bir so'zni ko'p o'lchamli vektor fazosida ifodalaydi, bunda semantik jihatdan yaqin so'zlar vektor fazosida ham yaqin joylashadi. Masalan, 'qirol — erkak + ayol = malika' kabi algebraik munosabatlar mumkin bo'ladi.
Sentence Transformers. Butun jumlani yoki hujjatni bitta vektorga kodlovchi ushbu modellar semantik o'xshashlikni aniqlashda yuqori samaradorlik ko'rsatadi [9]. all-MiniLM-L6-v2, paraphrase-multilingual-MiniLM kabi modellar ko'p tilli qidiruv uchun maxsus moslashtirilgan.

MULTIMODAL QIDIRUV VA SHAXSIY RELEVATLIK

Multimodal qidiruv
Zamonaviy qidiruv tizimlari faqat matn emas, balki tasvirlar, audio va video kontent bilan ham ishlaydi. OpenAI ning CLIP modeli [7] matn va tasvirni birgalikda o'rganib, ularni umumiy vektor

fazosida ifodalaydi. Bu texnologiya 'it rasmi' kabi so'rovlar bilan tegishli tasvirlarni topish yoki aks holda rasm orqali matnli ma'lumotlarni qidirishga imkon beradi.
Shaxsiy relevatlik va foydalanuvchi niyati
Bir xil so'rovga turli foydalanuvchilar uchun turli natijalar eng relevatli bo'lishi mumkin. Zamonaviy tizimlar quyidagilarni hisobga oladi:
Foydalanuvchining joylashuvi — 'restoran' so'roviga yaqin atrofdagi restoranlar ko'rsatiladi;
Qidiruv tarixi — oldingi qidiruvlar va bosishlar asosida profil tuziladi;
Vaqt konteksti — 'yangiliklar' so'rovi bugungi kunga, 'rekord' esa eng so'nggi tadbirga yo'naltiriladi;
Qurilma turi — mobil qurilmalar uchun boshqa kontentni afzal ko'rish.

METODLARNING QIYOSIY TAHLILI

Metod	Aniqlik	Tezlik	Resurs	Semantika
TF-IDF	O'rta	Yuqori	Kam	Yo'q
BM25	O'rta+	Yuqori	Kam	Yo'q
Word2Vec	O'rta+	Yuqori	O'rta	Qisman
Dense (DPR)	Yuqori	O'rta	Ko'p	Ha
BERT cross-enc.	A'lo	Past	Ko'p	Ha
Gibrid (BM25+DPR)	A'lo	O'rta	Ko'p	Ha
ColBERT	A'lo	O'rta+	Ko'p	Ha

2-jadval. Relevatlikni aniqlash metodlarining qiyosiy tahlili

AMALIY TIZIMLARDAGI QO'LLANILISH TAJRIBASI
Google Search — dunyo bo'yicha eng ko'p ishlatiladigan qidiruv tizimi — relevatlikni aniqlashda yuzlab belgilarni hisobga oladi. 2019-yildan boshlab Google BERT ni ishga tushirdi va bu qidiruv natijalarining sifatini sezilarli darajada yaxshiladi. Neural Matching texnologiyasi so'rovning chuqur ma'nosini tushunish imkonini beradi.
Elasticsearch — korporativ darajadagi ochiq kodli qidiruv platformasi — BM25 ni standart algoritm sifatida ishlatadi va kNN (k-Nearest Neighbors) vektor qidiruvini qo'llab-quvvatlaydi. Gibrid qidiruv rejimlari ham mavjud.
Bing — Microsoft ning qidiruv tizimi — TURING NLR va GPT asosidagi modellarni qo'llab, kontekstni chuqur tushunishga intilyapti. ChatGPT
integratsiyasi esa suhbatga asoslangan qidiruv (conversational search) yangi bosqichini ochdi.
KELGUSIDAGI RIVOJLANISH TENDENSIYALARI
Veb-qidiruv sohasida bir nechta asosiy tendensiya kuzatilmoqda:
Generativ qidiruv (Generative Search) — LLM (Large Language Models) asosida foydalanuvchiga to'g'ridan-to'g'ri javob yaratish, ya'ni havolalar ro'yxati o'rniga sintezlangan ma'lumot berish;
Retrieval-Augmented Generation (RAG) — qidiruv natijalari asosida dinamik ravishda matn generatsiya qilish, ishonchlilik va dolzarblikni oshirish;
Ko'p tilli va o'zaro til qidirish — foydalanuvchi bitta tilda so'rov kiritib, boshqa tillardagi hujjatlardan javob olishi;
Foydalanuvchi maxfiyligi bilan muvozanat — shaxsiylashtirilgan relevatlik va ma'lumotlar himoyasi o'rtasida yangi yechimlar izlash.

XULOSA

Ushbu tadqiqotda veb-qidiruv tizimlarida relevatlikni aniqlashning klassik (TF-IDF, BM25) va zamonaviy (BERT, Dense Retrieval, ColBERT, gibrid) metodlari tahlil qilindi. Olingan xulosalar quyidagicha:

BM25 algoritmi hali ham amaliy tizimlar uchun kuchli asos bo'lib qolmoqda va ko'pincha zamonaviy metodlar bilan birgalikda qo'llaniladi.

BERT asosidagi modellar, ayniqsa cross-encoder arxitekturasi, relevatlikni aniqlashda eng yuqori aniqlikni ta'minlaydi.

Gibrid yondashuv — sparse va dense retrieval kombinatsiyasi — amaliyotda eng muvaffaqiyatli yechim sifatida tan olingan.

Kelajakda generativ qidiruv va RAG texnologiyalari qidiruv sifatini yangi pog'onaga ko'tarishi kutilmoqda.

Kelgusida o'zbek tilidagi qidiruv tizimlarini zamonaviy neyron tarmoq metodlari asosida rivojlantirish muhim vazifa bo'lib qoladi.

FOYDALANILGAN ADABIYOTLAR

Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. — Cambridge University Press, 2008. — 482 b.

Salton G., Wong A., Yang C.S. A vector space model for automatic indexing // Communications of the ACM. — 1975. — Vol. 18, №11. — P. 613–620.

Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond // Foundations and Trends in Information Retrieval. — 2009. — Vol. 3, №4. — P. 333–389.

Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT 2019. — P. 4171–4186.

Karpukhin V. et al. Dense Passage Retrieval for Open-Domain Question Answering // Proceedings of EMNLP 2020. — P. 6769–6781.

Liu T.-Y. Learning to Rank for Information Retrieval. — Springer, 2011. — 123 b.

Radford A. et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP) // Proceedings of ICML 2021. — P. 8748–8763.

Khattab O., Zaharia M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT // Proceedings of SIGIR 2020. — P. 39–48.

Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of EMNLP 2019. — P. 3982–3992.

Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval: The Concepts and Technology behind Search. 2nd ed. — Addison-Wesley, 2011. — 913 b.

Mikolov T. et al. Distributed Representations of Words and Phrases and their Compositionality // Advances in Neural Information Processing Systems (NIPS). — 2013. — P. 3111–3119.

Joachims T. Optimizing Search Engines Using Clickthrough Data // Proceedings of KDD 2002. — P. 133–142.

Yusupov A.A., Mirzayev B.T. O'zbek tilidagi axborot qidirish tizimlarini yaratish muammolari // O'zbekiston Milliy universiteti xabarlari. — 2021. — №3. — B. 45–52.

Karimov O.B. Matnlarni tahlil qilishda sun'iy intellekt usullari. — Toshkent: Fan va texnologiya, 2022. — 180 b.

Elasticsearch Documentation. BM25 Similarity. — https://www.elastic.co/guide/en/elasticsearch/reference (murojaat sanasi: 2024-yil). 77.