

FOYDALANILGAN ADABIYOTLAR

1. Doktorlik dissertatsiyasi mavzusini ro'yxatdan o'tkazish va O'zbekiston Respublikasi OAK Axborotnomasida e'lon qilish tartibi// Vazirlar Mahkamasining 2012 yil 28 dekabrda 365-son qaroriga ilovalar. – 15-b. http://www.fdu.uz/dissertatsiyaga_talablar/Tartib.pdf
2. Higher Education in the Twenty-First Century: Vision and Action// World Conference On Higher Education – UNESCO, 1998. – 27-b.
3. Болотнова Н.С. Текстовая компетенция и пути ее формирования //Материалы науч.-практич. конференции – Томск, 2001. – С.66-76.
4. Вишнякова С.М. *Профессиональное образование Словарь. Ключевые понятия, термины, актуальная лексика.* — М. НМЦ СПО, 1999. — 538 с.

КОРПОРАТИВ ТАРМОҚДА ФАН ВА ТАЪЛИМГА ОИД МАЪЛУМОТЛАРНИ ИНТЕЛЛЕКТУАЛ ИЗЛАШ МОДЕЛИ ВА ВОСИТАСИ

Мўминов Б.Б., Рахматуллаев М.А. (*Тошкент ахборот
технологиялари университети*)

Мақолада фан ва таълимга оид маълумотларни интеллектуал излаш учун математик моделлар, норавшан билимлар базаси, сематик ядро модели ишлаб чиқиш асосида FSV технологиясининг архитектураси келтирилган.
Калит сўзлар: маълумот излаш, излаш элементлари, сўров, мантиқий излаш, векторли алгебра, стохастик назария, норавшан наўария, қоидалар, билимлар базаси.

MODELS AND MEANS OF INTELLECTUAL SEARCH FOR SCIENTIFIC AND EDUCATIONAL INFORMATION IN CORPORATE NETWORK

Muminov B.B., Rakhmatullaev M.A. (*Tashkent University of Information
Technologies*)

In article giving the architecture of FSV technology is based on foundations, search model, fuzzy knowledge base and semantic core for the intellectual searching data of scientific and educational information in corporate network.
Keywords: Search information, searching elements, query, logical searching, vector algebra, stochastic theory, fuzzy theory, rules and knowledge base.

Бугунги кунда корпоратив тармоқларда фан ва таълимга оид ахборот муҳитларининг асосий манбалари электрон шаклда бўлиб, уларни манбалар деб ажратиш оламиз. Бундай тизимларда маълумотларни интеллектуал излаш модуллари самарали ишлаши учун бир нечта базавий босқичларнинг

бажарилиши лозим бўлади. Улардан энг муҳимлари қуйидагилар деб ҳисоблаймиз:

- излаш сўровига ишлов бериш;
- излаш натижаларининг сўровга мослиги;
- топилган манбаларни тўғри ва оқилона даражалаш имконияти.

Google, Yahoo, Bing, Яндекс, Rambler каби излаш модуллари бўлган катта қувватли ва энг машҳур бўлган МИТлари миллиардлаб веб-саҳифаларни қамраб олади. Бундай тизимлар сифатли ва тез излашни таъминлашга имкон берадиган махсус алгоритмлари билан бир-биридан ажралиб туради. Лекин бу алгоритмларнинг барчаси асосий ёндашувлар – излаш моделларининг модификациялари ҳисобланади.

Тадқиқот ишларимизда модификацияланган ва ишлаб чиқилган математик моделлар, усуллар ва алгоритмлар асосида маълумотларни интеллектуал излаш (МИИ)ни инструментал дастурий воситасини комплекс қамраб олган, интеграциялаш учун ягона тизим сифатида инструментал платформани ишлаб чиқиш зарурияти бор. Бу инструментал платформа бугунги куннинг замонавий дастурлаш технологиялари DOM, XML, ORM, MVC ва турли Framework технологиялари каби архитектураси, математик асослари, IDEF моделлари ва кутубхоналари, маълумотлар тузилмаси ва бошқа бошқаришнинг инструментал воситасига эга бўлиши лозим деб ўйлаймиз.

МИИнинг ҳар қандай математик модели ёки IDEF моделлари қуйидаги таркибий қисмлардан иборат бўлиши керак:

1. (F) - Сўровни тақдим этиш усули – тизим фойдаланувчисининг ахборот эҳтиёжларини ифодалашнинг шакллаштириш усули (F - forming).

2. (S) - Манбанинг сўровга мувофиқлик функцияси – сўровнинг ва топилган манбанинг мувофиқлиги даражаси (долзарблиги, мослиги) (S - searching).

3. (V) - Манбаларни тақдим этишнинг усули (V - viewing).

Бу уч таркибий қисмни бирлаштириб, корпоратив тармоқларда фан ва таълимга оид ахборот муҳитлари МИИлар учун FSV технологияси (FSV платформи, FSV Framework)деб номлаймиз.

Фараз қилайлик, T тўпламдан t_i терминининг индекси i ($i = 1, \dots, M$) учун, $d^{(j)}$ – D манбалар тўпламига мансуб бўлган j - манба, ҳамда $w^{(i,j)} \geq 0$, $(t_i, d^{(j)})$ жуфтлик билан боғлиқ бўлган вазн катталиқ ва $d^{(j)}$ манбага қирмайдиган ҳар бир t_i терминин учун, унинг вазни нолга тенг бўлсин, яъни $w^{(i,j)} = 0$ мавжуд бўлсин.

FSV технологиясида мантиқий моделларга асосланиб, маълумотларни излашнинг математик моделларининг модификациялашган варианты. Излашнинг мантиқий моделлари тўпламлар назариясига ва математик мантиққа асосланади. Манбалар,

сўровлар ва терминларни калит сўзларнинг тўплами сифатида қаралади. Бу тўпландаги ҳар бир термин буль ўзгарувчиси орқали ифодаланади, яъни 0 (сўровдаги термин манбада мавжуд эмас) ёки 1 (сўровдаги термин манбада мавжуд). Бунда терминнинг манбадаги вазнли қийматлари фақат икки қийматга эга бўлади:

$$w^{(i,j)} \in \{0,1\} \quad (1)$$

Излашнинг буль моделларида фойдаланувчи сўровни буль ифодаси шаклида ифодаланиши ва бунинг учун ВА (and \wedge), ЁКИ (or \vee), ЙЎҚ (not \neg) операторларидан фойдаланилади. Маълумки, ҳар қандай мантикий ифодани конъюнкцион операцияси (дизъюнктив нормал шакли) билан ўзаро боғланган маълум ифодаларнинг дизъюнкцияси шаклида ифодалаб тақдим этиш мумкин. Шунинг учун:

$$q \equiv d_{dnf} = \bigvee_{i=1 \dots N} q_{cc}^{(i)}$$

бунда q – сўров, $q_{cc}^{(i)}$ – q_{dnf} сўров шаклининг i -чи конъюнктив элементи. Бу ҳолда, мантикий моделидаги манба $d^{(j)}$ ва сўров q нинг яқинлиги ўлчови $q - sim(d^{(j)}, q)$ билан белгиланади ва у (2) ифодада келтирилган:

$$sim(d^{(j)}, q) = \begin{cases} 1, \text{ agar } \exists q_{cc}^{(i)} : (q_{cc}^{(i)} \in q_{dnf}) \wedge (\forall k, g_k(q_{cc}^{(i)}) = g_k(d^{(j)})) \\ 0 \text{ қолган ҳолларда} \end{cases} \quad (2)$$

бу ерда g_k – термин t_k индексига мос келувчи инверсияли функция бўлиб, у қуйидаги тарзда аниқланади:

Агар ушбу конъюнктив элемент ҳар бир терминнинг инверсияли функцияси $d^{(j)}$ манба учун худди шундай инверсияли функцияга тўлиқ мос келадиган ҳолда дизъюнктив нормал шакли q_{dnf} га кирадиган $q_{cc}^{(i)}$ конъюнктив компонентаси мавжуд бўлса, $g_k(d^{(j)}) = w_k^j$, яъни $sim(d^{(j)}, q) = 1$, акс ҳолда $sim(d^{(j)}, q) = 0$ га тенг бўлади. Шундай қилиб, агар $sim(d^{(j)}, q) = 1$ бўлса, буль моделига мувофиқ ҳолда манба $d^{(j)}$ q сўровига мос деб ҳисобланади. Акс ҳолда манба мос ҳисобланмайди.

FSV технологиясида маълумотларни излашнинг векторли алгебрага асосланган математик моделларининг модефикациялашган варианты. Излашнинг векторли модели алгебраик моделлар синфининг анъанавий вакилидир. Ушбу моделнинг доирасида манбалар ва сўровлар терминларнинг кўп ўлчовли фазосидаги векторлар шаклида таърифланади. Манбада ишлатиладиган ҳар бир терминга унинг вазнли қиймати мос келтирилади. Бу қиймат терминнинг кўриб чиқиладиган манбада ва бутун манбалар массивида пайдо бўлишининг сонлари ҳақидаги статистик ахборот асосида аниқланади. Векторли моделда сўровларда мантикий операциялардан фойдаланиш кўзда тутилмаган. Сўров ва манбанинг яқинлигини баҳолаш учун сўров ва манбанинг тегишли векторларининг скаляр кўпайтмаси ишлатилади.

$d^{(j)}$ манбанинг q сўровга яқинлиги

$$\bar{d}^j = (w_1^{(j)}, w_2^{(j)}, \dots, w_n^{(j)}) \text{ ва } q = (w_1^q, w_2^q, \dots, w_n^q)$$

терминларнинг вазни қийматлари билан ифодаланган ахборот векторларининг скаляр кўпайтмаси сифатида қаралади. Бунда айрим терминларнинг вазини турли усуллар билан ҳисоблаш мумкин. Бунда қўллаш мумкин бўлган энг содда ёндашувлардан бири терминнинг вазни $w_i^{(j)}$ сифатида терминнинг ушбу манбада тез-тез учрашининг меъёрлаштирилган $freq_i^{(j)}$ дан фойдаланишга асосланган, бунда ушбу термин тўпламининг бошқа манбаларида топишининг тезлиги ҳам ҳисобга олинади. Ушбу усул терминнинг дискриминацион кучини ҳисобга олиш деб аталади (3):

$$w_i^{(j)} = freq_i^{(j)} \cdot \log\left(\frac{N}{n_i}\right), \quad (3)$$

бу ерда n_i – термин t_i қўлланган манбаларнинг сони, N эса – манбаларнинг массив ичидаги умумий миқдоридир. Масалан, агар маълум бир сўз массивнинг ҳар бир манбада учрайдиган бўлса, сўровда ундан фойдаланиш самарасиз бўлади. Бунга мос ҳолда, бу ҳолатда $n_i = N$, демак,

$$w_i^{(j)} = freq_i^{(j)} \cdot \log\left(\frac{N}{N}\right) = 0$$

Терминлар вазини ўлчашнинг бундай усули – $TF * IDF$ стандарт белгиланишига эга, бу ерда TF (ингл. Term Frequency –терминин учраши) терминнинг МАНБАда учрашиш (такрорланиш) лар сонини кўрсатади, IDF (ингл. Inverse Document Frequency –манбанинг учрашиши) эса – массивда ушбу терминни ўз ичига олган манбаларнинг миқдорига тескари пропорционал учрашишининг катталигини кўрсатади.

Манба ва сўровнинг мавзули яқинлигини аниқлаш учун, ушбу моделда содда скаляр кўпайтма $sim(d^j, q)$ қўлланилади, у \bar{d}^j ва \bar{q} векторлари ўртасидаги бурчакнинг косинусига мос келади. Манба $d^{(j)}$ ва сўров q яқинлигининг ўлчови бўлган катталик (4) бўйича ҳисобланади:

$$sim(d_j, q) = \frac{\bar{d}^{(j)} \cdot \bar{q}}{|\bar{d}^{(j)}| \cdot |\bar{q}|} = \frac{\sum_{i=1}^n w_i^{(j)} w_i^q}{\sqrt{\sum_{i=1}^n (w_i^{(j)})^2} \sqrt{\sum_{i=1}^n (w_i^q)^2}} \quad (4)$$

Векторли модел амалда энг кўп қўлланади, чунки у анчагина содда амалга оширилади, излашнинг ва даражалашнинг самарадор бўлишини таъминлайди. Бундан ташқари, векторли-фазовий модел излаш тизимларига бундай манбаларни излашнинг режимини осон амалга ошириш имкониятини таъминлайди. Ҳар бир манба сўров каби қаралиши мумкин. Лекин, шу билан бирга, векторли-фазовий модел юқори ўлчамли массивларни ҳисоблаш билан боғлиқ бўлиб, анъанавий шаклида катта маълумот массивларига ишлов бериш унча яроқли бўлмайди.

FSV технологиясида маълумотларни излаш стохастик назарияларга асосланган математик ва IDEF моделлари. Эҳтимоллик асосида излаш моделининг пойдевори сифатида эҳтимоллар назарияси ва математик статистика назарияларидан фойдаланилади. Бу моделдаги мослик манба фойдаланувчига қизиқарли бўлиб чиқишининг эҳтимоли сифатида

қаралади. Бунда фойдаланувчи томонидан танланган ёки бирор-бир содалаштирилган автоматик тарзда олинган мос манбаларнинг мавжуд бўлган бирламчи танланмаси (ўқув танланмаси) борлиги кўзда тутилади.

Ҳар бир кейинги манба учун мос бўлиб чиқиш эҳтимоли мос тўпламида ва тўпланининг қолган, «мос бўлмаган» қисмида терминларнинг учраш тезлигининг нисбати асосида ҳисобланади. Ушбу излаш моделида манба сўровга мос бўлишининг эҳтимоли сўровнинг терминлари мос ва мос бўлмаган манбаларнинг ичида турлича тақсимланган, деган фарзга асосланади. Байеснинг теоремасига мувофиқ, маълум бир эҳтимоллик функцияси бўйича якуний шаклини ҳосил қиламиз, у ўқув танланмасидаги ҳар бир манба учун излаш статуси деб аталадиган мослиги эҳтимоллигининг даражасини баҳолайди (5):

$$SV = \sum_{t_i \in q \cap d} SV = \sum_{t_i \in q \cap d} \log \frac{rel_i(nrel - nrel_i)}{nrel_i(rel - rel_i)} \quad (5)$$

бу ерда rel_i – i индекси бўлган терминни ўз ичига олган мос манбаларнинг сони, $nrel_i$ – мос ҳолда, мос бўлмаган МАНБАларнинг сони, d – манбанинг ўқув танланмаси, сўз тўплами сифатида қаралади, q – сўровга кирувчи сўзларнинг тўплами, $q \cap d$ сўровда ва МАНБАда умумий терминларнинг тўпланини билдиради.

Манбаларнинг фақат матнли мазмунини ҳисобга оладиган математик моделларни кўриб чиқиш муҳим омилардан ҳисобланади, чунки замонавий излаш тизимлари гиперматнли ҳаволаларнинг тузилмасини таҳлил қилади. Бу маълумот эълон қилинган ресурсларнинг долзарб эканлиги ҳисобига амалга оширилади. Долзарблилик кўрсаткичи жуда муҳим ўрин тутаяди. Излаш модуллари фойдаланувчи сўровга мос бўлган манбаларни фойдаланганда топилган ресурсларнинг муҳимлигини қўшимча тарзда ҳисобга олишни талаб қилади. Манбанинг муҳимлигини аниқлашда омиларнинг бир нечта турлари ҳисобга олиниши мумкин. Амалиётда энг кўп ишлатиладиган модел PageRank модели бўлиб ҳисобланади. У қуйидаги (6) формула билан ифодаланади:

$$PR_\alpha = \frac{(1 - d)}{N} + d \sum_{i=1}^n \frac{PR_i}{C_i} \quad (6)$$

бу ерда PR_α – қаралаётган манбанинг PageRank рейтинги, d – камайиш коэффициенти, N – манбаларнинг умумий миқдори, PR_i – қаралаётган хужжатга ҳавола қилаётган i - чи манбанинг PageRank рейтинги, C_i – i - чи манбадаги ҳаволаларнинг умумий сонидир.

PageRank дан ташқари, амалда ҳаволалар орқали даражалашнинг бошқа моделлари камроқ ишлатилади. Уларга BackRank (PageRank нинг ўзгартирилган шакли), HITS, HillTop, SALSA киритилиши мумкин. Санаб ўтилган моделлар web графни бутунлай ёки унинг бир қисмини таҳлил қилишни ишга туширади.

МИИларда манбанинг графини яратиш ва уни даражалаш учун куйидагича математик модел таклиф қилинган. Унинг умумлашган шаклини (7) формула каби ёзиш мумкин.

$$D_r = \frac{1}{2} \left(\frac{\sum_{i=1}^{|I(i)_{DocID}|} D_r^i}{|I(i)_{DocID}| \sum_{count(I(i))} I(i)} + \frac{|O(i)_{DocID}|}{\sum_{count(O(i))} O(i)} \right) \quad (7)$$

Бунда DocID – ҳар бир манбанинг маълумотлар базасидаги идентификатор тартиби, $I(i)$ - манбага кирувчи ҳаволалар, $O(i)$ – манбадан чиқувчи ҳаволалар, $|*|$ - тўпланинг элементлар сони. Ушбу математик модел ёрдамида манбаларни даражалаш орқали манбамларида маълумотларни излашни осонлаштириши мумкин.

Замонавий излаш тизимлари амалга оширилган пайтда излашнинг бир нечта моделларини биргаликда қўллайди. Маълумот излашнинг барча моделларини шартли равишда иккита гуруҳга бўламиз. Биринчисига олдинги натижаларни таҳлил қиладиган моделлар, иккинчи гуруҳга эса – ҳаволаларнинг тузилмасини ҳисобга оладиган моделлар киради.

Одатда, ҳаволаларни ҳисобга олиш манбанинг долзарблиги (муҳимлиги)ни баҳолашга, олдинги натижаларни таҳлил қилиш эса сўровга мослигини баҳолашга имкон беради.

Маълумотларни излаш учун сарфланадиган вақтни камайтириш учун мантиқий семантик излашнинг математик модели ишлаб чиқилган

$$f(q) = \{ \{q_i | q_j\} \} = \frac{|R^j \cup R^i|}{|R^j|} \quad (8)$$

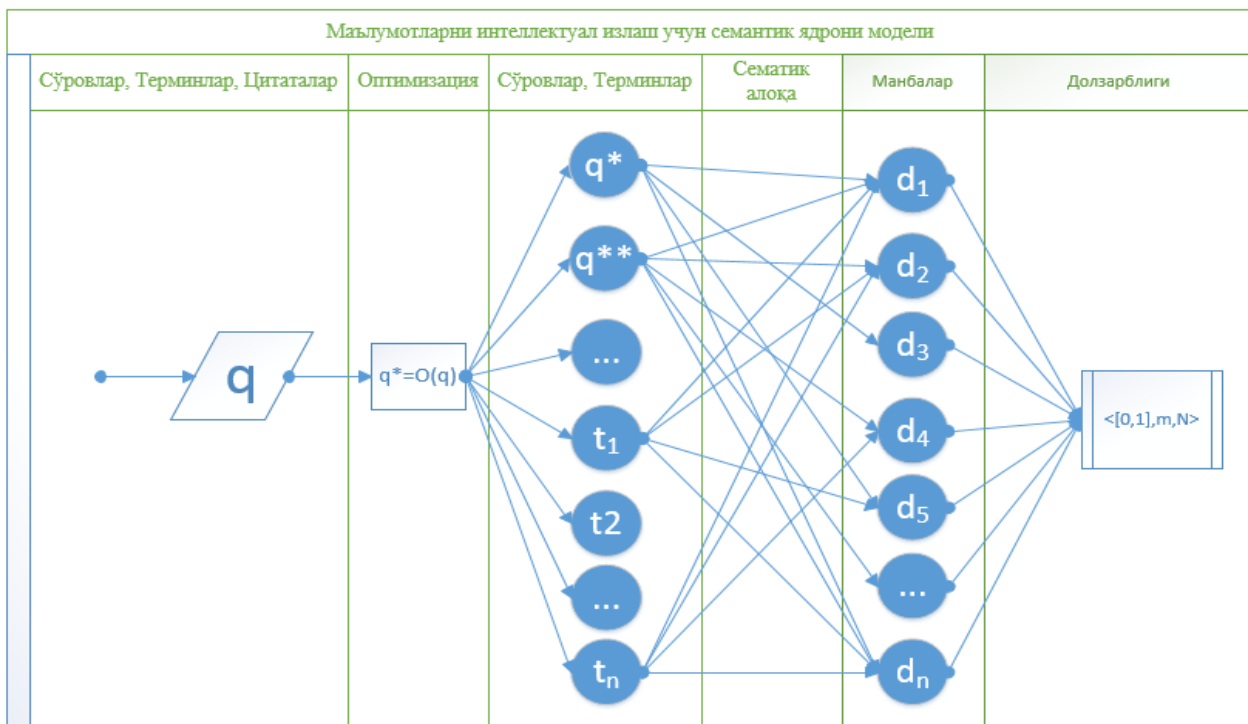
Бунда q сўровлар тўплами, $\{a|b\}$ – | амал бўлиб, a нинг b га семантик боғланганлиги ва ўхшашлигини билдиради ва $\frac{|R^j \cup R^i|}{|R^j|}$ амал билан ҳисобланади.

МИИлар учун инструментал дастурий воситанинг функционал имкониятини моделлаштириш учун 3 та IDEF0 ва реляцион маълумотлар тузилмаси учун IDEF1x модел ишлаб чиқилган.

FSV технологиясида маълумотларни интеллектуал излаш ва қайта ишлашнинг норавшан назарияларга асосланган математик ва IDEF моделлари. НАМда маълумотларни излашни интеллектуал амалга оширишда ихтиёрий лингвистик ўзгарувчи учун (9) кўринишидаги математик модел ва алгоритми ҳамда у билан ифодалаш мумкин бўлган моделлар учун параметрик тегишлилик функцияларини лойиҳалаштириш ва уларнинг параметрлари орасидаги ўзаро муносабатларни ишлаб чиқиш, тегишлилик функцияларнинг асосий синфларига мос норавшан терминларни танлаш механизмларга асосланади.

$$f = F \langle \beta, T(\beta), X, G(\beta), \mu_x \rangle \quad (9)$$

Шунингдек, маълумотларни интеллектуал излашда семантик ядро яратиш бўйича алгоритмлар ишлаб чиқилган, МИИлар учун семантик ядро модели яратилган.



1-расм. МИЙлар учун семантик ядро модели.

Маълумотларни излаш тизимида норавшан билимлар базасининг математик модели (10) ва қоидаларини ишлаб чиқиш механизими, ББси учун ББсининг ядроси ва уни ривожлантириш алгоритми, Билимлар базасининг ядросини IDEF1x модели лойиҳалаштирилган. НАМда маълумотларни излаш тизимларининг 3 та ўзаро бир бири билан боғлиқ IDEF0 ва IDEF1x моделлари ишлаб чиқилган.

$$\bigcup_{i=1}^n \left((q_i^* \subseteq q) \text{ BA } (q_{i1}^* \in d) \text{ BA } (q_{i2}^* \in t^*) \text{ BA } (a_i \in \mu_{i,x_i}(q_{i2}^*)) \right) \rightarrow d \quad (10)$$

FSV технологиясининг ишлаб чиқиш учун юқорида келтирилган математик ва IDEF моделлар асосида сервер иловали мижоз-сервер ахтитектурасига асосланган ҳолда архитектурасини ишлаб чиқамиз (2-расмга қаранг).

Сервер иловали мижоз-сервер ахтитектураси фойдаланувчи ва сервер иловаларини бошқариш ва қайта ишлаш қулай ва асосий имкониятлари қуйдагича:

1. Нозик мижоз.
2. Мижоз ва сервер орасида минимум маълумот берилади параметрлар ва натижалар.
3. Сервер иловалари бир неча нусхада бир нечта компьютерда ишга туширилиши мумкин.



2-расм. FSV технологиясининг архитектураси.

FSV технологияси – бу ахборот муҳитида маълумотларни излаш ва қайта ишлаш моделлари, усуллари ва алгоритмларини интеграция ва модификацияловчи, сервер иловали мижоз-сервер ахтитектураси асосланган инструментал дастурий платформа.

Умуман олганда FSV технологиясини нафақат фан ва таълимга оид корпоратив тармоқларда балки турли ахборот тизимларида МИИ учун фойдаланиш мумкин. Бунинг учун улар қуйидаги талабларга жавоб бериши лозим.

1. Сервер иловали мижоз-сервер ахтитектураси.
2. Уч босқичли тамойил, яъни фойдаланувчи интерфейси, бизнес-логика, маълумотларни бошқаришга асоаланган.
3. MVC ва Framework технологияси асосида ишлаб чиқилган.
4. XML тузилмасини таҳлил қилиш ва МББТга боғлаш мавжуд бўлиши.

5. Дастурлаш тили объектга йўналтирилган.

ИНТЕРНЕТ ТАРМОҒИДА МАТНЛАРНИ МАТН ТАРКИБИДАГИ ГАПЛАР ВА СЎЗЛАР АСОСИДА ЎХШАШЛИККА ТЕКШИРИШ АЛГОРИТМЛАРИ

Атажанов Ж.А. (*“УЗТЕЛЕКОМ” компанияси “Биллинг Телеком” филиали
дастурчилар бўлими бошлиғи, техника фанлари номзоди*)

Мақолада интернет тармоғидаги илмий-таълимий ва илмий-техникавий ахборот ресурслари орасидан ўхшаш матнларни қидириш алгоритми берилган. Шунингдек, изланаётган матнни бошқа тилларда эълон қилинган матнлар орасидан ҳам ўхшашликка текшириш усули баён қилинган.

Калим сўзлар: *Google, Yandex, Yahoo, PDF, HTML, MsWORD, Apache Tika, базавий тил, ахборот излаш, Web саҳифалар*

ALGORITHMS OF CHECKING SIMILARITIES ACCORDING TO WORDS AND PHRASES OF THE TEXTS BY INTERNET

Atadjanov J.A. (*The chairman of the programs' department of “Billing Telecom”
in “UZTELECOM” company*)

In the article there was provided information about the algorithms of looking for similar text among the scientific-educational and scientific-technic informational resources by internet. Also, some methods which are used for checking similar written works the searched text among other languages, were revealed there.

Keywords: *Google, Yandex, Yahoo, PDF, HTML, MsWORD, Apache Tika, Web pages.*

Кириш. Бизга маълумки, интернет тармоғининг маълумотлар базаси ҳар куни, ҳар соат, ҳар дақиқада янада бойиб, ахборотлар тобора ошиб бораётган бир пайтда ўзимизга керак бўлган маълумотларни излаб топиш, таҳлил қилиш мураккаблашиб бормоқда. Шу сабабли ҳам, тез суръатлар билан ўзгариб бораётган ўта шиддатли ҳамда мураккаб бир воқеликни ўзида мужассам этган глобал тармоғида ўхшаш матнларни тезкор қидириб топиш бугунги куннинг энг долзарб масалалар қаторига кирмоқда. Шунингдек, Ўзбекистонда ҳимоя қилинаётган диссертацияларни чет элларда ҳимоя қилинаётган диссертациялар билан таққослаш, талабалар томонидан ёзилган «курс иши», «битирув малакавий иши», «магистрлик диссертацияси»ни профессор-ўқитувчиларнинг илмий нашр ишлари билан таққослаш ҳамда профессор-ўқитувчилар ва тадқиқотчилар томонидан тайёрларган диссертациялар, илмий тўпламлар, илмий - услубий қўлланмалар, илмий журналлардаги мақолалар ва монографияларни интернет тармоғидаги ёки