5. Дастурлаш тили объектга йўналтирилган.

ИНТЕРНЕТ ТАРМОҒИДА МАТНЛАРНИ МАТН ТАРКИБИДАГИ ГАПЛАР ВА СЎЗЛАР АСОСИДА ЎҲШАШЛИККА 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.
Калит 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.*

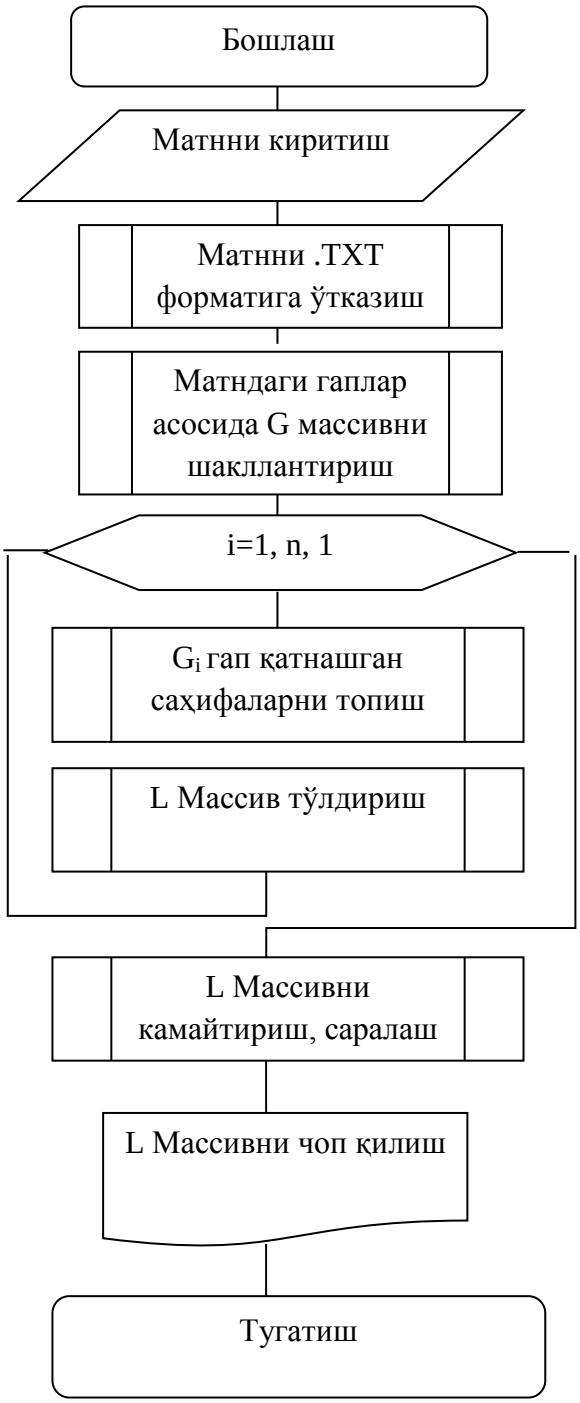
Кириш. Бизга маълумки, интернет тармоғининг маълумотлар базаси ҳар куни, ҳар соат, ҳар дақиқада янада бойиб, ахборотлар тобора ошиб бораётган бир пайтда ўзимизга керак бўлган маълумотларни излаб топиш, таҳлил қилиш мураккаблашиб бормоқда. Шу сабабли ҳам, тез суръатлар билан ўзгариб бораётган ўта шиддатли 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

электрон кутубхоналарнинг маълумотлар базасидаги электрон ахборот-таълим ресурслари билан таққослаш орқали плагиатни аниқлашда экспертларга ёрдам бериш долзарб вазифалардан бири ҳисобланади. Зеро, кўчирмачиликнинг олдини олиш талабаларнинг мустақил ишларини сифатли бажаришларини таъминлайди. Бу эса мамлакатимизнинг маънавий салоҳиятини барқарорлик сари етаклайди.

Асосий қисм

Мазкур муаммони бартараф этишда, яъни интернет тармоғида тўлиқ матнга ўхшаш маълумотларни излашда қуйидаги усулларни талиф этамиз:

Матн таркибидаги гаплар асосида маълумотларни ўҳшашликка текшириш. Мазкур усулда тўлиқ матн гапларга ажратилади ва ўхшаш маълумотлар ҳосил қилинган гаплар асосида амалга оширилади. Қуйида ушбу жараённинг алгоритми келтирилган.



1-расм: Интернет тармоғида тўлиқ матнлаи маълумотларни ўҳшашини аниқлаш алгоритми

a. Матнни гапларга ажратиш. Маълумки, табиий тилда ҳар қандай маълумотлар тўплами гаплардан ташкил топган. Ҳар бир гап “.”, “?”, “!”, “...” каби тиниш белгилар асосида бир биридан ажратилади. Демак, мазкур усулда юқорида келтирилган тиниш белгилар асосида қисмларга ажратиш назарда тутилган;
b. ҳосил бўлган ҳар бир гапга мос маълумотларни глобал тармоқдаги ахборот қидириш тизимлари (Google, Yandex, Yahoo ва бошқ.) асосида ўхшаш матнларни қидириш;
c. топилган саҳифаларни матн таркибидаги гапларнинг қатнашиш сони камайиши тартибида жойлаштириш.

Ҳосил бўлган рўйхатнинг бошида дастлаб, матнга энг юқори ўхшаш бўлган саҳифа манзили ўрин олади, сўнгра матннинг ўхшаш даражасига қараб кетма-кет саҳифа манзиллари жойлашади. Ушбу усулда маълумот излаш тезлиги матн таркибидаги гаплар сонига боғлиқ бўлади. Мазкур жараён алгоритмининг блок схемаси 1-расмда келтирилган. Шу ўринда 1-расмда келтирилган алгоритмнинг айрим қисмларини кенгроқ кўриб чиқайлик:

Ҳар хил форматлардаги матнларни оддий текст матн кўринишига ўтказиш. Маълумки, одатда маълумотлар фойдаланувчига ўқиш қулай бўлиши учун PDF, HTML, MsWORD ёки бошқа форматларда бўлиши мумкин. Интернетдан маълумот излаш тизимлари эса фақат оддий матнлар асосида маълумот излашга мўлжалланган. Мазкур қисмда ҳар хил турдаги тизимлардан фойдаланиш мумкин. Шулардан бири Apache Tika-очиқ кодли тизими бўлиб [2], мазкур тизим pdf, xls, word, ppt, html форматидаги матнларни .TXT форматига ўтказиш учун ишлаб чиқилган.

Матн таркибидаги гаплар қатнашган Web саҳифаларни аниқлаш жараёнида юқорида таъкидлаб ўтилганидек - Google, Yandex, Yahoo ва бошқ. Ахборот тизимларидан фойдаланиш мумкин. Одатда ушбу тизимлар изланаётган маълумотни ўхшашлик даражасига қараб натижаларни саралаб кўрсатади. Яъни изланаётган маълумотга ўхшашлиги юқори бўлган саҳифалар натижасининг юқори пағоналаридан ўрин эгаллайди. Мазкур ҳолат бизга умумий матнга ўхшаш бўлган саҳифалар ўхшашлик даражасини аниқлашда қўл келади. Яъни, матн таркибидаги гаплардан энг кўп ва юқори ўринда жойлашган саҳифалар тўлиқ матн ўхшаш бўлган саҳифалар рўйхатида ҳам юқори ўринда бўлади.

Дейлик, $L_i, (i = \overline{1..m})$ массиви матн таркибидаги гаплар қатнашган Web саҳифалар манзили бўлсин. Матн таркибидаги гапларни эса мос равишда $G_j, (j = \overline{1..n})$ массиви билан белгилаб олсак. Бу ерда, n - матн таркибидаги гаплар сони, m - матн таркибидаги гаплар қатнашган Web саҳифалар манзиллар сони. Интернетдан ахборот қидириш тизимидан олинган

натижаларни эса $R_{ij}, (i=\overline{1..m}, j=\overline{1..n})$ деб белгиласак, бу ерда r_{ij} элемент g_i гапнинг l_j саҳифага ўхшашлик даражасини билдиради. R массивнинг ҳар бир элементи қиймати [0..10] интервалдаги сон бўлиб уни ҳисоблаш қуйидаги алгоритм асосида бўлади.

a. g_i гап интернетда ахборот қидириш тизимларидан бири асосида маълумот изланади.
b. Излаш натижасида l_j гап қатнашган саҳифалар ахборот қидириш тизими қайтарган кетма-кетлигида мос равишда рўйхатнинг биринчи поғонасида келган саҳифасига 9 ва ҳакозо кетма кетлигида қиймат берилади. Дейлик, l_j гап мос равишда a,b,c,d,e,f тартиб саҳифаларда қатнашган бўлсин. У ҳолда R массив элементлари қўйидаги қийматга эга.

$r_{ai}=10$	$r_{ci}=8$	$r_{li}=6$
$r_{bi}=9$	$r_{di}=7$	$r_{fi}=5$

c. Икки ўлчамли массивнинг устун элементлари йиғиндиси асосида бир ўлчамли $R_i^{`}, (i=\overline{1..m})$ массив ҳосил қиламиз ва уни камайиш тартибда тартиблаймиз.

$$r_i^{`} = \sum_{j=1}^{n} r_{ij} \qquad (1)$$

Натижавий $R^{`}$ массив берилган матнга ўхшаш бўлган Web саҳифаларнинг рўйхатини чиқаради. Ушбу алгоритмда хар бир матн таркибидаги гаплар ўҳшашликка текшириш жараёнида тенг кучли ҳисобланади. Мазкур алгоритмда маъно англатмайдиган умумий гаплар таҳлил қилинмайди.

Матн таркибига сўзлар асосида ахборот излашни ташкил қилиш. Мазкур усулда дастлаб тўлиқ матн сўзларга ажратилади ва интернет тармоғидаги ўхшаш саҳифалар эса шу сўзлар асосида аниқланади. Биринчи усулдан фарқли равишда сўзлар орасидан маъно англатмайдиган ва ёрдамчи сўзлар олиб ташланади. Мазкур ҳолат бизга ўхшашликка текшириш жараёнини тезроқ ва сифатлироқ амалга оширилишини таъминлайди. Чунки, ҳар бир матн ёрдамчи сўзлар мавжуд бўлиб, одатда уларнинг гап таркибидаги қатнашиш сони, матн маъносини англатувчи айрим калит сўзлардан кўп бўлиши мумкин. Шу билан бирга мазкур жараённи амалга ошириш учун бизга H ёрдамчи ва маъно англатмайдиган сўзлар тўпламидан иборат луғат керак бўлади. Қуйида мазкур жараённинг алгоритми берилган.

a) Матн таркибидаги сўзлар ва уларни матн таркибида қатнашиш сонини аниқлаш;
b) ҳар бир сўз асосида интернетда ахборот қидириш тизимлари асосида Web саҳифаларни излаш;
c) ҳосил бўлган саҳифалар ва матн таркибида қатнашган сўзлар ҳамда уларнинг қатнашиш сони асосида натижавий массив ҳосил қилинади.

Шу ўринда, юқорида келтириб ўтилган алгоритмга кенгроқ тўхталиб ўтсак. Дастлаб, матн таркибидаги сўзлардан ташкил топган $S_i, (i=\overline{1..n})$ массив тўғрисида сўз юритамиз. Мазкур массив таркибидаги ҳар бир элемент учун

$$s_i \notin H \quad (2)$$

шарт ўринли, яъни матн таркибидаги маъно англатмайдиган ва ёрдамчи сўзлар мазкур массивга кирмайди ҳамда ахборот қидириш жараёнида иштирок этмайди. Бундан ташқари, $C_i, (i=\overline{1..n})$ массиви бўлиб, мазкур массивида ҳар бир сўзнинг матн таркибида такрорланиш сони сақланади, яъни матн учун у ёки бу сўзнинг муҳимлик даражасини кўрсатади. Матн таркибидаги жами сўзлар сони эса l^* бўлиб унинг қиймати C_i массив элементлари йиғиндисига тенг.

$$l^* = \sum_{i=1}^{n} c_i \quad (3)$$

Интернетда ахборот қидириш тизимларидан фойдаланган ҳолда, $S_i, (i=\overline{1..n})$ сўзлар қатнашган Web саҳифалар $W_j, (j=\overline{1..m})$ массивига киритсак. Бу ерда m Web саҳифалар сонини билдиради. Ҳар бир сўз ва у қатнашган Web саҳифалар орасидаги боғланишни эса $P_{ij}, (i=\overline{1..m}, j=\overline{1..n})$ массивига киритамиз. Ҳар бир p_{ij} элемент қиймати учун $p_{ij} \subset [0..10]$ бўлиб, s_i сўзнинг w_j Web саҳифада қатнашиш даражасини англатади. Мазкур ҳолат ҳам гаплардаги каби, ахборот қидириш жараёнида биринчи ўринда учраган саҳифа учун 10, иккинчи ўринда учраган саҳифа учун мос равишда 9 ва ҳакозо каби ҳисобланади. Натижавий $P_j^{`}, (j=\overline{1..m})$ бир ўлчамли массивда матн ўхшаш бўлган Web саҳифалар тўпламидан ташкил топади. Қуйида ушбу массивнинг $j$ элементи $p_j^{`}$ ни ҳисоблаш формуласи келтирилган.

$$p_j^{`} = \left(\sum_{i=1}^{n} \frac{c_i \cdot p_{ij}}{l^* \cdot 10} \right) \cdot 100 \quad (4)$$

Юқорида келтирилган формулада натижалар фоиз ҳисобида кўрсатилиб, уни камайиш тартибда тартиблаш орқали, берилган матнга ўхшаш бўлган сахифалар рўйхатига эга бўламиз. Мазкур алгоритмда биз фақат сўзлар асосида ахборот қидириш жараёнини ташкил қилдик. Ахборот қидириш жараёнида матн таркибидаги сўз бирикмаларидан фойдаланиш янада яхшироқ натижа беради. Чунки, сўз бирикмаси гап таркибида иштирок этаётган сўзларни маъно жихатдан қайси мақсадда қўлланилишини ифодалайди.

Хулоса. Ҳулоса ўрнида шуни айтиш мумкинки, юқорида кўрилган иккита алгоритм ҳам матнга ўхшаш бўлган саҳифаларни рўйхатини олишга ёрдам беради. Мазкур алгоритмларнинг асосий камчилиги Web саҳифалар фақат GET сўрови асосида ҳосил бўлувчи саҳифалардангина маълумот излайди. Бундан ташқари, натижавий массив қиймати тўғридан тўғри интернетда ахборот қидириш тизимлари натижасига боғлиқ. Маълумки, интернетда СЕО муҳандислик ишланмалари мавжуд бўлиб, улар интернет тизимида ахборот қидириш жараёнида Web саҳифаларни олдинги қаторларда

кўринишини таъминлайди. Бу эса матнга ўхшаш бўлган маълумотларни излаш жараёнига салбий таъсир кўрсатади.

Фойдаланилган адабиётлар

1. Daniele Anzelmi, Domenico Carlone, Fabio Rizzello,Robert Thomsen, D. M. Akbar Hussain Plagiarism Detection Based on SCAM Algorithm // Proceedings of the International MultiConference of Engineers and Computer Scientists 2011 Vol I? IMECS 2011, March 16-18, 2011 Hong Kong
2. Erik H., Otis G., Michael McC. Lucene in Action – Covers Apache Lucene v.3.0// Manning Publications.-486p.,2009 y.
3. Седова Я. А., Квятковская И. Ю. Интеллектуальный анализ корпуса документов научной информации // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика.- № 1 / 2011

Мамчич, А.А. Система автоматизированного поиска, индексирования и реферирования научно-технической информации / А.А. Мамчич, Л.В. Степура, Д.А. Черников // Библиотеки в информационном пространстве: проблемы и тенденции развития : материалы II Междунар. науч. конф. молодых ученых и специалистов, Минск, 16 февр. 2010 г.

МИЛЛИЙ КОРПОРАТИВ ЭЛЕКТРОН КУТУБХОНА ШАКЛЛАНТРИШ ИСТИҚБОЛЛАРИ: МУАММОЛАР ВА ЕЧИМЛАР

Каримов У.Ф., Каримов У.У. *(Тошкент ахборот технологиялари университети)*

Мақолада, миллий корпоратив электрон кутубхона шакллантириш бўйича амалга оширилган лойиҳалар таҳлили, таълимга оид корпоратив электрон кутубхоналарнинг ўзига ҳос хусусиятлари ва корпоратив электрон кутубхоналарни ривожланиш истиқболлари ёритилган.

Калит сўзлари: *электрон кутубхона, таълим, стандарт, LOM, MODS, ARMAT, электрон таълим ресурслари*

PROSPECTS FOR THE DEVELOPMENT OF THE CORPORATE NETWORK OF ELECTRONIC LIBRARIES: PROBLEMS AND SOLUTIONS

Karimov U.F., Karimov U.U. *(Tashkent University of Information Technologies)*

The article analyzes the projects on the forming of the national corporate electronic library, the features of educational corporate electronic libraries and the prospects for the development of the corporate network of electronic libraries.