ИНТЕРНЕТ ТАРМОҒИДА ЎХШАШ МАТНЛАРНИ ИЗЛАШ МОДЕЛИ ВА АЛГОРИТМИ

Атаджанов Ж.

Ушбу мақолада ҳар хил форматдаги тўлиқ матнни гапларга ажратган ҳолда Интернет тармоғидан ўхшашликка текшириш жараёни алгоритми келтириб ўтилган. Шу билан бирга, ҳар хил форматдаги тўлиқ матнни оддий матнга ўтказиш усуллари ҳам кўриб чиқилган. Ахборот қидиришда эса Google, Yandex каби Интернет тармоғида ахборот қидириш тизимларига асосланилади. Мазкур алгоритм асосида сўзларни синонимларини инобатга олувчи матнларни Интернет тармоғида ўхшашликка текширувчи тизим ишлаб чиқиш мумкин.

Калит сўзлар: *матнларни сўзларга ажратиш, тўлиқ матнлардан ахборот қидириш.*

MODELS AND ALGORITHMS SEARCH ANALOG TEXT ON THE INTERNET

Atadjanov J.

This article provides an algorithm for checking the Internet for similarity with the full text of the various types of content. There are also some ways to convert full-text in different formats. The information search process is based on Internet search engines. Based on this algorithm, it is possible to develop a system that checks the synonyms of words in the Internet.

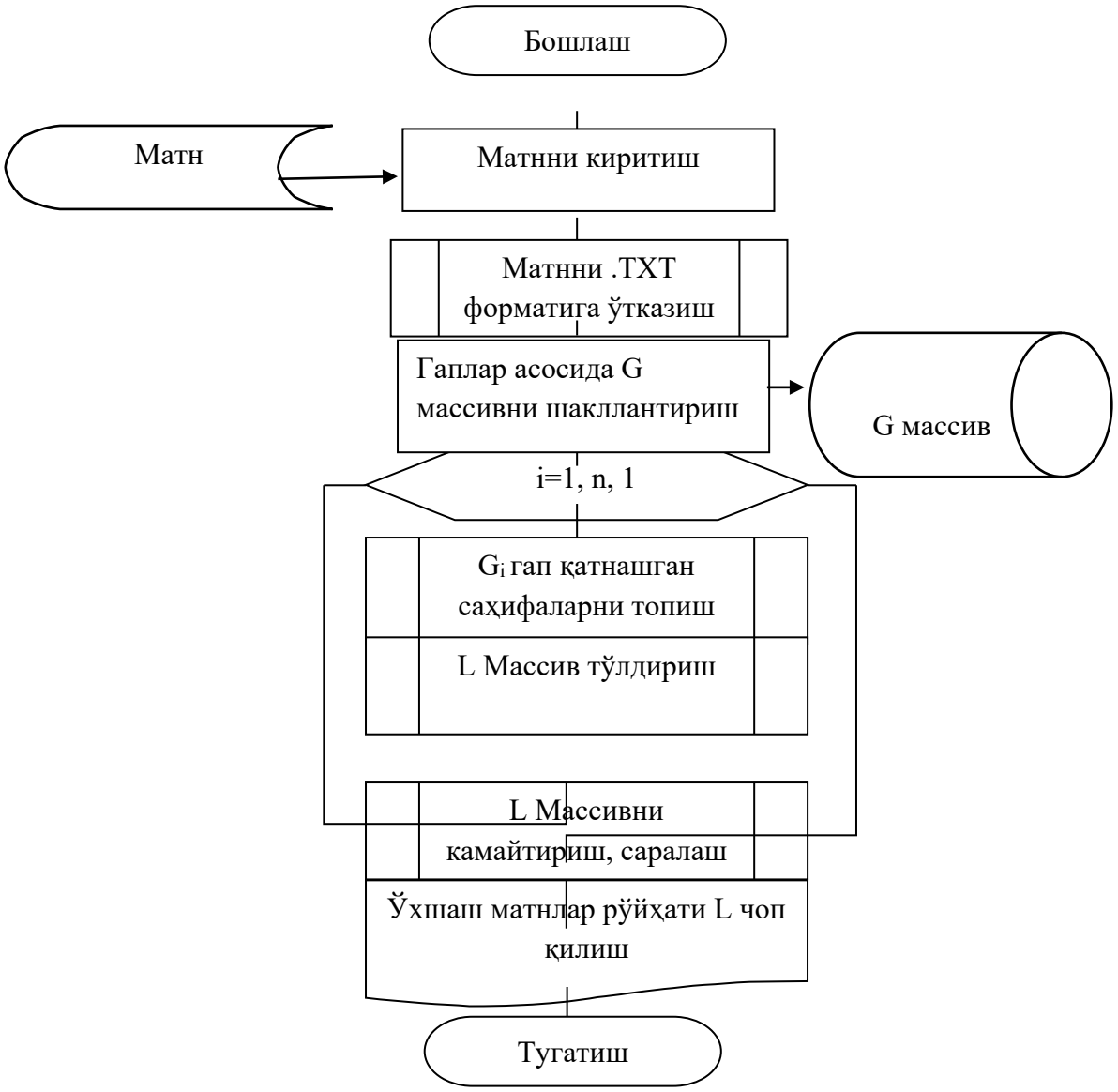
Keywords: *stemming text, search information from text.*

Кириш. Интернет тармоғини ривожланиши, ундаги маълумотлар ҳажми ортишига узвий боғлиқ бўлиб, мазкур маълумотлар ҳозирда дақиқа сайин ортиб бормоқда. Интернет тармоғига жойланган маълумотларда муаллифлик ҳуқуқини ҳимоя қилиш жуда қийин, деярли буни имкони ҳам йўқ [4]. Плагиат – инглизча plagiarism сўзидан олинган бўлиб, бу маълумотни асл манбасини кўрсатмасдан, ўз ишингиз сифатида кўрсатишдир [2, 3]. Компьютер технологиялари ёрдамида плагиатни аниқлаш усуллари – плагиатга текшириш деб юритилиб [1], ҳозирги кунда мазкур муаммони ҳал қилиш учун бир қанча усуллар, алгоритмлар ва дастурий воситалар ишлаб чиқилган. Мисол сифатида «Антиплагиат», Advego Plagiatus, Unplag, miratools.ru, istio.com, Praide Unique Content Analyser II, Plagiatinform тизимларини келтиришимиз мумкин.

Мазкур тизимлар яхлит бўлиб, уларни бошқа тизимларнинг тизим ости сифатида ишлатишни имкони йўқ, бундан ташқари мазкур тизимларда сўзларни синонимлари инобатга олинмаган. Ушбу мақолада биз тўлиқ матнни интернет тармоғида ахборот қидириш тизимлари асосида ўхшашликка текшириш жараёни алгоритмини кўриб чиқамиз.

Матн таркибидаги гаплар асосида ахборот излашни ташкил қилиш. Мазкур усулда тўлиқ матн гапларга ажратилади ва ўхшаш маълумотлар ҳосил қилинган гаплар асосида амалга оширилади. Қуйида ушбу жараённинг алгоритми келтирилган.

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;



1-расм: Интернет тармоғида тўлиқ матнларни ўхшашини аниқлаш алгоритми

b. Ҳосил бўлган ҳар бир гапга мос маълумотларни глобал тармоқдаги ахборот қидириш тизимлари (Google, Yandex, Yahoo ва б.) асосида ўхшаш матнларни қидириш, мазкур жараёнда сўзларни синонимларидан ҳам фойдаланиш мумкин;

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. Шу ўринда 1-расмда келтирилган алгоритмнинг айрим қисмларини кенгроқ кўриб чиқайлик:

Матнни киритиш – ўхшашикка текшириш керак бўлган матн киритилади. Бу ерда матн форматига хеч қандай талаб қўйилмаган.

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, 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 бошқалар. 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 матнга 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 манзиллари 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-кетлигига мос равишда рўйхатнинг биринчи поғонасида келган саҳифасига 9 ва ҳоказо 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.

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

Натижавий $R^{`}$ массив берилган матнга ўхшаш бўлган Web саҳифаларнинг рўйхатини чиқаради. Ушбу алгоритмда ҳар бир матн таркибидаги гаплар ўхшашликка

текшириш жараёнида тенг кучли ҳисобланади. Мазкур алгоритмда маъно англатмайдиган умумий гаплар таҳлил қилинмайди.

Матн таркибидаги 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, матн таркибида 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 белгисини сўзни мазкур матн таркибида қатнашиш сони 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. Ушбу ҳолат ҳам гаплардаги каби, ахборот қидириш жараёнида биринчи ўринда учраган саҳифа учун 10, иккинчи ўринда учраган саҳифа учун мос равишда 9 ва ҳоказо 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.

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

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 жиҳатдан қайси мақсадда қўлланилишини ифодалайди.

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 CEO 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. Аушра А. Научная электронная библиотека как средство борьбы с плагиатом (рус.) // Международный форум Educational Technology & Society 9(3). — 2006. Архивировано 20 сентября 2016 года.
2. Дягилев В. В., Цхай А. А., Бутаков С. В. Архитектура сервиса определения плагиата, исключающая возможность нарушения авторских прав (рус.) // Вестник НГУ. Серия: Информационные технологии. — 2011.
3. Ушакин С. Плагиат? Об этике в науке (рус.) // Общественные науки и современность. — 2001.
4. Седов А. В., Рогов А. А. Анализ неоднородностей в тексте на основе последовательностей частей речи (рус.) // Современные проблемы науки и образования. — 2013. — Вып. 1.
5. Шахрай С. М., Аристер Н. И., Тедеев А. А. О плагиате в произведениях науки (диссертациях на соискание учёной степени): научно-методическое пособие. — М.: МИИ, 2014. — 176 с. — 1000 экз. — ISBN 978-5-00077-056-6.
6. Шарапов Р. В., Шарапова Е. В. Система проверки текстов на заимствования из других источников (рус.) // Всероссийская научная конференция Электронные библиотеки: Перспективные методы и технологии, электронные коллекции. — 2011.
7. Erik H., Otis G., Michael McC. Lucene in Action – Covers Apache Lucene v.3.0// Manning Publications.- 486p., 2009