ЭЛЕКТРОН КУТУБХОНА ВА АРХИВЛАРДА ТЎЛИҚ МАТНЛИ АХБОРОТ ҚИДИРИШ ЖАРАЁНИНИ ТАШКИЛ ҚИЛИШ

АТАДЖАНОВ Жасур Абдушарибович,
Ўзбектелеком АЖ Ахборот тизимлари филиали
дастурчиси, техника фанлари номзоди

Аннотация: *Мақолада электрон кутубхона ва архивлар маълумотлар базаларидаги тўлиқ матнли ахбротлардан маълумот қидириш, уларни ўзаро солиштириш жараёни кўриб ўтилган. Бу борада биз матнларни кўп тиллик ўхшашликка текшириш учун мўлжалланган jComporator тизимидан фойдаланамиз.*

Калит сўзлар: *Электрон кутубхона, автоматлаштирилган ахброт-кутубхона тизими, библиографик ёзув, MARC формат.*

ATADJANOV Jasur Abdusharibovich,
Uzbektelecom stock company, Infosystems branch, Candidate of Technical Sciences

ORGANIZATION OF FULL TEXT INFORMATION DETAILS IN ELECTRONIC LIBRARIES AND ARCHIVERS

Abstract: *There is a given organization information search process on base full-text files of digital libraries and archives. To implement this functionality used jComporator system which provides cross-language compare full texts.*

Keywords: *Digital library, library system, librarian record, MARC format.*

Биз ахборот алмашиниш ва ахборот қадр-қиммати жуда юқори бўлган бир даврда яшамоқдамиз. Ахборот нафақат илм-фан балки, ишлаб чиқариш, тадбиркорлик ва ижтимой ҳаётда ҳам муҳим омил ҳисобланади. Ахборот ҳажмининг кун сайин ортиши нафақат уни сақлаш, балки уни излаб топиш ва ундан фойдаланиш жараёнида бир қатор муаммолар юзага келтирмоқда. Интернет тармоғининг ривожланиши натижасида маълумотлар алмашиниш жараёни соддалашиш билан бир вақтда унинг суръати кескин ошди.

Одатда локал ёки глобал тармоқда илмий-таълимий маълумотлар элекрон кутубхона ёки махсус ишлаб чиқилган архивларда сақланади. Маълумки, кутубхонашуносликда ҳар бир адабиёт ҳақидаги маълумот библиографик ёзув (БЁ) деб номланиб, уларни тавсифлашда MARC формати оиласига кирувчи стандарт форматлардан фойдаланилади. Ушбу оилага мансуб форматларда (MARC21, USMARC, UZMARC, RUMARC ва бошқалар) ўртача 700–800 та майдон ости мавжуд бўлиб, мазкур майдонлар асосида ихтиёрий турдаги ҳужжатларни каталоглаштириш имкони мавжуд [1]. Лекин мазкур форматларда ишлай олиш учун фойдаланувчидан махсус билим ва кўникмалар талаб қилинади. Бундан ташқари электрон кутубхоналарни шакллантириш жараёнини соддалаштириш мақсадида Dublin Core формати ҳам ишлаб чиқилган бўлиб, мазкур формат 16 та майдондан иборат. Бу эса БЁни электрон каталогга (ЭК) киргизиш, улардан фойдаланиш жараёнини жуда соддалаштиради. Ҳозирги кунда интернет тармоғида айнан мазкур формат асосида минглаб электрон кутубхоналарни кўриш мумкин.

Мазкур форматларни ишлаб чиқишдан асосий мақсад, БЁнинг барча параметрларини тавсифлаш ҳамда улар асосида маълумот қидиришни ташкил қилишдан иборат. Аммо ушбу формат асосида БЁга бириктирилган тўлиқ матндан маълумот қидириш имкони мавжуд эмас. Электрон кутубхона ёки архивларда матнлар тўлиқ хам сақланишини инобатга оладиган бўлсак, мазкур масала долзарб муаммолардан бири ҳисобланади.

Ҳозирги кунда мамлакатимиз ахборот-кутубхона муассаларида кенг тарқалган тизимлар – ИРБИС, MarcSQL, АРМАТ ва КАДАТАларнинг бирортасида ҳам тўлиқ матнга боғланган маълумотлардан ахборот қидириш имконияти мавжуд эмас. Бу борада биз jComporator тизимидан фойдаланишимиз мумкин.

Мазкур тизим ҳар хил форматдаги (PDF, MsWord, HTML) матнларни ўзаро ўхшашликка текшириш мақсадида ишлаб чиқилган. Ҳозирги кунда мазкур тизим ва АРМАТ тизимини биргаликда ишлашини таъминловчи Z39.50 протоколи асосида ишловчи махсус модул яратилган бўлиб, мазкур модулнинг иш жараёни қуйидаги манбада кўрсатилган [2].

Шу ўринда Z39.50 протоколига қисқача тўхталиб ўтсак. Мазкур протокол клиент-сервер оиласига тегишли протокол ҳисобланиб, дастурий воситалар орасида маълумот алмашиниш учун 1970 йилда ишлаб чиқилган [2, 4]. Мазкур протокол библиографик маълумотларни алмашиниш учун қулай бўлганлиги боис, соҳада ҳам кенг қўлланила бошланди.

Матнларни кўп тиллик ўхшашликка текшириш jComparator тизими фойдаланувчи интерфейси, функционал қисм ва маълумотлар базаларидан ташкил топиб, қуйидаги автоматлаштирилган иш жойларидан (АИЖ) иборат:

· тизим маълумотлар базаси(МБ)ни шакллантирувчи АИЖ;

· матнларни маълумотлар базасига киритувчининг АИЖ;

· матнларни ўхшашликка текшириш АИЖ.

Тизим МБсини шакллантирувчининг АИЖ асосий ҳар хил турдаги маълумотномаларни шакллантириш ва уларни бошқаришдан иборат. Мазкур маълумотномалар қуйидагилардан иборат:

· тизим ишлайдиган табиий тиллар маълумотномаси;

· табиий тилдаги маъно англатмайдиган ёрдамчи сўзлар маълумотномаси;

· илмий-таълимий ахборот рубрикаси маълумотномалари;

· илмий-таълимий ахборот рубрикасига боғланган умумий сўзлар маълумотномалари;

· сўзларни таржима қилиш жараёнида қўлланилувчи луғат;

· синоним сўзлар маълумотномалари.

Матнларни маълумотлар базасига киритувчи АИЖ. Мазкур АИЖ тизим МБдаги матнлар билан ишлаш учун мўлжалланган бўлиб қуйидаги асосий вазифаларни бажаришга йўналтирилган:

· тизимдаги матнлар рўйхати ва уларнинг таркиби билан танишиш;

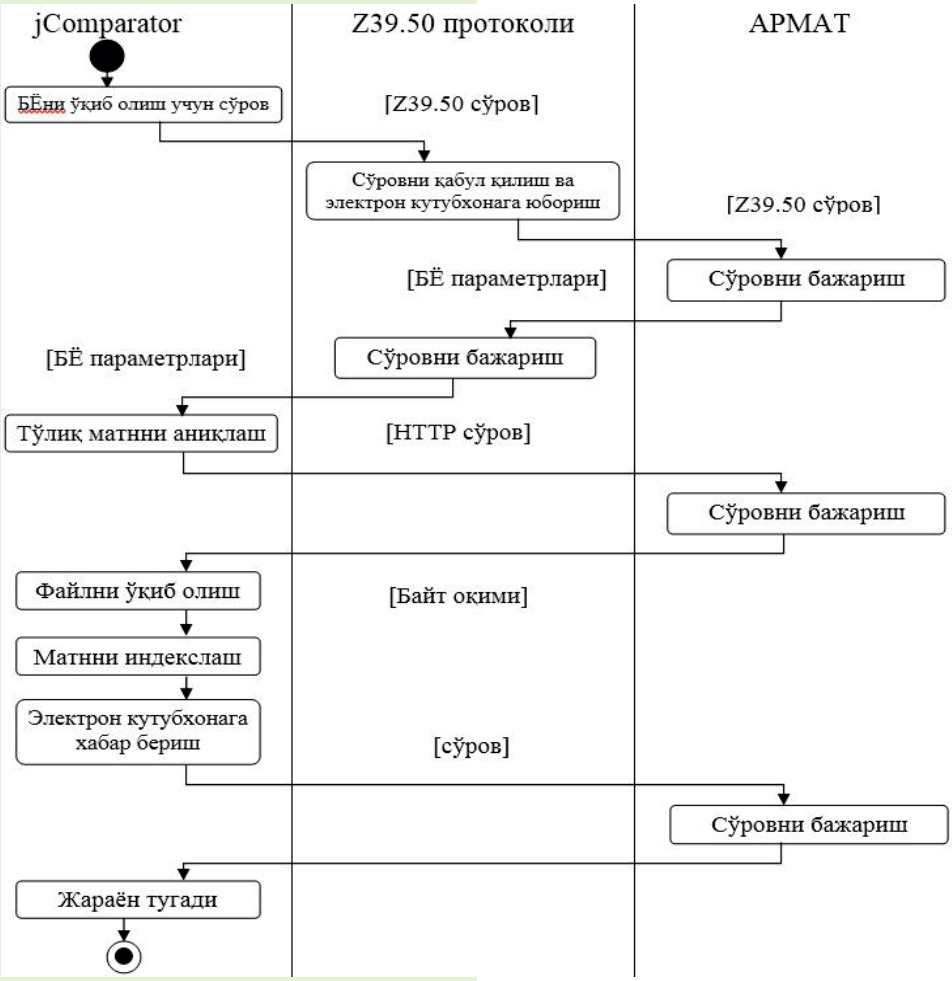
· МБ индексларини қайта қуриш – мазкур қисм одатда матнга маълумот киритилгандан сўнг тизим маълумотномаларида (ёрдамчи сўзлар, синоним ёки луғатларда) ўзгариш юзага келса қўлланилади. Иш жараёни давомийлиги тизим таркибидаги маълумотлар ҳажмига боғлиқ;

· МБга янги матн қўшиш;

· тизим таркибидаги мавжуд матнни таҳрирлаш;

· тизимдан киритилган матнни олиб ташлаш.

Мазкур АИЖда тизим МБ шакллантирилиб, маълумотлар ҳам реляцион, ҳам нореляцион маълумотлар базасини бошқариш тизимида сақланади. Бундан ташқари матннинг тўлиқ шакли тизимга ажратилган файл тизимда ҳам сақланади. МБ индексларини қайта қуриш жараёнида айнан мазкур файл тизимдаги матнлар асосида қайта киритилади. Бу жараён дастлаб тизим МБсидан тўлиқ матн сўзлари асосида шаклланган индексларни тозалаб, матнлари қайта индекслашдан иборат.



1-расм. Z39.50 протоколи асосида электрон кутубхонанинг МБ билан ишлаши

Матнларни ўхшашликка текшириш АИЖ берилган матнни тизимдаги матнлар орасидан ўхшашини аниқлаш учун мўлжалланган. Ўхшашликка текшириш жараёнида фойдаланувчидан матн жойлашган жой, матн ёзилган табиий тилни кўрсатиш талаб қилинади.

Ҳозирги кунга келиб йирик кутубхоналар ахборот тизимлари мазкур форматда маълумот алмашиниш имкониятига эга. Шулардан бири АҚШ Конгресс кутубхонасидир [4].

Ишнинг дастлабки босқичи электрон каталогга (ЭК) киритилган тўлиқ матнларни тизим базасига киритишдан иборат бўлиб, мазкур жараён ҳар бир БЁ учун кетма-кет тарзда амалга оширилади. Z39.50 протоколи орқали БЁни ўқиш жараёнида тизим матн ёзилган тил, унинг жойлашган манзили, муаллифи каби маълумотларга эга бўлади. Айнан мазкур маълумотлар асосида БЁ боғланган матн jComporator тизими базасига киритилади. Шу ўринда юқорида 1-расмда келтириб ўтилган алгоритмнинг иш жараёнига кенгроқ тўхталиш зарур.

Дастлаб, jComporator тизими АРМАТ тизими ЭКдан тўлиқ матнга эга БЁни MARC21 форматида кўчириб олади. Маълумки, MARC21 форматида БЁга боғланган тўлиқ матн ҳақидаги маълумотлар (файл номи, уни жойланган манзили, файл формати) учун махсус майдонлар ажратилган. Мазкур майдлонлардан фойдаланган ҳолда биз БЁга бириктирилган тўлиқ матнни кўчириб олишимиз мумкин. Юқоридаги 1-расмда jComporator тизимининг тўлиқ матнга мурожаат ҳуқуқи мавжудлиги кўриб чиқилди. Агар тўлиқ матн ва jComporator тизим орасида тўғридан-тўғри боғланиш мавжуд бўлмаса, у ҳолда мазкур ишни SOAP (Simple Object Access Protocol — объектга содда мурожаат протоколи) протоколи асосида хам амалга ошириши мумкин [5].

Шундан сўнг навбатдаги босқич, ҳар хил форматдаги тўлиқ матнни оддий матн шаклига ўтказиш, матн таркибидаги сўзларни ажратиб олиш, ҳосил бўлган сўзлар тўпламидан маъно англатмайдиган сўзларни олиб ташлаш каби жараёнлар бажарилади. Ҳосил бўлган сўзлар тўплами тизимнинг нореляцион МБсига киритилади. МБда сўз қуйидаги шаклда сақланади:

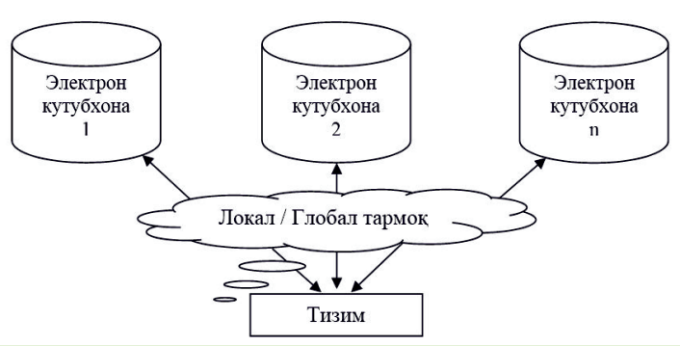
· сўзнинг ўзгармас шакли – матн таркибидаги сўзлар таҳлил жараёнида ўзак қисмга ажратилади ва МБда сўзнинг фақат ўзак қисми сақланади. Матнларни ўзаро таққослаш жараёнида хам айнан сўзнинг ўзак шакллари иштирок қилади;

· сўзнинг матн таркибида қатнашган сони – МБга киритиладиган ҳар бир сўз матн таркибида неча марта такрорланганлиги ҳам сақланади. Ушбу маълумот матнларнинг ўзаро ўхшашлик даражасини аниқлашда қўлланилганлиги сабабли МБга киритилади;

· сўзнинг матн таркибида жойлашган ўрни – мазкур маълумот матнларнинг ўхшаш қисмларини ажратиб кўрсатиш учун қўлланилади.

Матнларни ўзаро ўхшашликка текшириш жараёнида айнан мазкур МБ асосида амалга оширилади. Матнларнинг бир-бирига ўхшашлик даражаси улар таркибидаги бир хил сўзларнинг такрорланиш сони нисбатлари асосида ишлайдиган махсус алгоритм асосида аниқланади.

Мазкур тизим асосида нафақат битта, балки бир нечта электрон кутубхоналарнинг ЭК асосида тўлиқ матнли маълумот қидириш жараёнини ташкил қилиш мумкин. Қуйида мазкур турдаги боғланишниг ташкилий тузилиши берилган.



2-расм. Тизимнинг корпоратив электрон кутубхоналар билан боғланиши

Хулоса ўрнида шуни айтиш мумкинки, мазкур тизим асосида нафақат электрон кутубхоналар балки, автоматлаштирилган ахборот-кутубхона тизимлари, архивлар ва бошқа тўлиқ матнли маълумотлар билан ишловчи тизимларда матнларни ўхшашликка текшириш жараёнини амалга ошириш мумкин. Мазкур jComporator тизими нафақат Z39.50 протоколи, балки RabbitMC модули асосида ҳам маълумот алмашиниш имкониятига эга.

Фойдаланилган адабиётлар

1. Раҳматуллаев М.А., Каримов У.Ф., Муҳаммадиев А.Ш., Атаджанов Ж.А. Корпоратив ахборот-ресурс марказларининг автоматлаштирилган тизими. – Алишер Навоий номидаги Ўзбекистон Миллий кутубхонаси нашриёти, Тошкент: 2008.

2. Племнек А.И., Усманов Р.Т., Сова Д.Н. Использование протоколов Z39.50 и HTTP в современных библиотечных информационных системах. http://www.unilib.neva.ru/rus/olsc/publications/z39_01.html

3. Атаджанов Ж.А. Кўп тиллик антиплагиат тизими // Чет тилларини ўргатишда инновацион технологиялар. Материалы международной конференции, 10-11 октября. Самарканд, 2019.

4. https://ru.wikipedia.org/wiki/Z39.50

5. Atadjanov J.A. Exchanging bibliographic data with its full text by SOAP // 2010 4th International Conference on Application of Information and Communication Technologies. – 12-14 October 2010. – Tashkent: 2010.