ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҲУЗУРИДАГИ ИЛМИЙ ДАРАЖАЛАР БЕРУВЧИ DSc.13/30.12.2019.T.07.02 РАҚАМЛИ ИЛМИЙ КЕНГАШ

ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ

АТАДЖАНОВ ЖАСУРБЕК АБДУШАРИБОВИЧ

КЎП ТИЛЛИ ТАҚСИМЛАНГАН АХБОРОТ-КУТУБХОНА ВА АРХИВ МАЪЛУМОТЛАР БАЗАЛАРИДАН ЎХШАШ МАТНЛАРНИ ҚИДИРИШ УСУЛ ВА АЛГОРИТМЛАРИ

05.01.09 – Ҳужжатшунослик. Архившунослик. Кутубхонашунослик

ТЕХНИКА ФАНЛАРИ ДОКТОРИ (DSc) ДИССЕРТАЦИЯСИ АВТОРЕФЕРАТИ

Тошкент – 2020

УДК 004.021

Докторлик (DSc) диссертацияси автореферати мундарижаси

Оглавление автореферата докторской (DSc) диссертации

Contents of the abstract of Doctoral (DSc) Dissertation

ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ ҲУЗУРИДАГИ ИЛМИЙ ДАРАЖАЛАР БЕРУВЧИ DSc.13/30.12.2019.T.07.02 РАҚАМЛИ ИЛМИЙ КЕНГАШ

ТОШКЕНТ АХБОРОТ ТЕХНОЛОГИЯЛАРИ УНИВЕРСИТЕТИ

АТАДЖАНОВ ЖАСУРБЕК АБДУШАРИБОВИЧ

КЎП ТИЛЛИ ТАҚСИМЛАНГАН АХБОРОТ-КУТУБХОНА ВА АРХИВ МАЪЛУМОТЛАР БАЗАЛАРИДАН ЎХШАШ МАТНЛАРНИ ҚИДИРИШ УСУЛ ВА АЛГОРИТМЛАРИ

05.01.09 – Ҳужжатшунослик. Архившунослик. Кутубхонашунослик

ТЕХНИКА ФАНЛАРИ ДОКТОРИ (DSc) ДИССЕРТАЦИЯСИ АВТОРЕФЕРАТИ

Тошкент – 2020

Фан доктори (DSc) диссертацияси мавзуси Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Олий аттестация комиссиясида B2020.2.DSc/T169 рақам билан рўйхатга олинган.

Диссертация Тошкент ахборот технологиялари университетида бажарилган.

Диссертация автореферати уч тилда (ўзбек, рус, инглиз (резюме)) Илмий кенгаш веб саҳифасида (www.tuit.uz) ва «Ziyonet» ахборот-таълим порталида (www.ziyonet.uz) жойлаштирилган.

Илмий маслаҳатчи:	**Рахматуллаев Марат Алимович** техника фанлар доктори, профессор
Расмий оппонентлар:	**Умаров Абдусалом Одилович** ижтимой фанлар доктори, профессор
	Нуралиев Фахриддин Муродиллаевич техника фанлари доктори, доцент
	Кучкаров Тахир Сафарович иқтисод фанлари доктори, профессор
Етакчи ташкилот:	**Ислом Каримов номидаги Тошкент давлат техника университети**

Диссертация ҳимояси Тошкент ахборот технологиялари университети ҳузуридаги DSc.13/30.12.2019.T.07.02 рақамли Илмий кенгашнинг 2020 йил «28» июня да соат 10⁰⁰ даги мажлисида бўлиб ўтади. (Манзил: 100084, Тошкент шаҳри, Амир Темур кўчаси, 108-уй. Тел.: (99871) 238-64-43; факс: (99871) 238-65-52; e-mail: tuit@tuit.uz).

Диссертация билан Тошкент ахборот технологиялари университети Ахборот-ресурс марказида танишиш мумкин (156 рақам билан рўйхатга олинган). (Манзил: 100084, Тошкент, Амир Темур кўчаси, 108-уй. Тел.: (+99871) 238-65-44).

Диссертация автореферати 2020 йил «16» июня куни тарқатилди.

(2020 йил «15» июня даги 3 рақамли реестр баённомаси.)

И.Х.Сиддиков
Илмий даражалар берувчи
илмий кенгаш раиси,
т.ф.д., профессор

Х.Э. Хужаматов
Илмий даражалар берувчи
илмий кенгаш илмий котиби,
техника фанлари бўйича фалсафа доктори (PhD)

Т.С. Кучкаров
Илмий даражалар берувчи
илмий кенгаш ҳузуридаги илмий семинар раиси,
и.ф.д., профессор

КИРИШ (фан доктори (DSc) диссертацияси аннотацияси)

Диссертация мавзусининг долзарблиги ва зарурати. Жаҳонда ахборот моддий бойликлардан бири сифатида тан олиниб, ҳозирда уларни излаш, сақлаш ҳамда улардан фойдаланиш жараёнларини ривожлантириш борасида жуда катта ишлар амалга оширилмоқда. Ахборот технологиялари ривожланиши натижасида маълумотларни сақлаш, узатиш ва ундан фойдаланиш жараёни бир неча баробар осонлашди. Бу эса ўз навбатида маълумотлар ҳажмини кескин равишда ортишига замин яратди. Айни пайтда MARC (Machine-Readable Cataloging-машина ўқий оладиган каталоглаштириш) ва Dublin Core форматлари асосида ишлайдиган тизимларнинг ривожланиши натижасида жуда кўплаб электрон каталоглар ва электрон кутубхоналар ташкил этилмоқда. Бу эса ўз-ўзидан мазкур соҳага тўлиқ матнлардан маълумот қидириш, электрон каталогни шакллантириш жараёнини автоматлаштириш, тўлиқ матнларни солиштириш ва уларни ўзаро ўхшашлик даражасини аниқлаш каби бир қатор вазифаларни қўймоқда. Бу борада ривожланган мамлакатларда, жумладан Австралия, Австрия, Буюк Британия, Германия, АҚШ, Канада, Гонконг, Россия Федерацияси каби давлатларда тўлиқ матнлар асосида ахборот қидириш ҳамда уларнинг ўзаро ўхшашлигини аниқлаш воситаларини ишлаб чиқиш муҳим вазифалардан бири ҳисобланади.

Жаҳонда ахборот-кутубхона ва архив маълумотлар базасидан тўлиқ матнли маълумотларни қидириш, матн таркибидаги сўзларни морфологик таҳлил қилиш ҳамда маъно англатувчи сўзлар тўпламини аниқлаш, матнларни ўзаро ўхшашлигини аниқлаш ҳамда уларни машина ёрдамида рукнини аниқловчи моделлар, алгоритмлар ва дастурий таъминот ишлаб чиқишга йўналтирилган илмий тадқиқотлар олиб борилмоқда. Бу йўналишда кўп тилли маълумотлар базасидан ўзаро ўхшаш матнларни аниқлашни таъминлайдиган моделлар, алгоритмлар ва дастурий воситалар ишлаб чиқиш долзарб муаммолардан бири ҳисобланади.

Республикамизда ахборот-кутубхона фондлари электрон каталогини ҳамда архив маълумотлар базасини шакллантириш борасида кенг қамровли ишлар амалга оширилмоқда. 2017-2021 йилларда Ўзбекистон Республикасини янада ривожлантириш бўйича Ҳаракатлар стратегиясида, жумладан «... сифатли таълим хизматлари имкониятларини ошириш, ... таълим ва ўқитиш сифатини баҳолашнинг халқаро стандартларини жорий этиш асосида олий таълим муассасалари фаолиятининг сифати ҳамда самарадорлигини ошириш»[1] вазифалари белгиланган. Мазкур вазифаларни амалга оширишда кўп тилли тақсимланган ахборот-кутубхона ва архив маълумот базаларидан тўлиқ матнлардан ахборот қидириш, уларни ўзаро ўхшашлигини аниқлаш ҳамда рефератив маълумотлар асосида электрон

[1] Ўзбекистон Республикаси Президентининг 2017 йил 7 февралдаги ПФ-4947-сон «Ўзбекистон Республикасини янада ривожлантириш бўйича Ҳаракатлар стратегияси тўғрисида»ги Фармони

каталогни автоматик тарзда шакллантириш моделлари, алгоритмлари ва дастурий тизимини ишлаб чиқиш муҳим вазифалардан бири ҳисобланади.

Ўзбекистон Республикаси Президентининг 2017 йил 7 февралдаги ПФ-4947-сон «Ўзбекистон Республикасини янада ривожлантириш бўйича Ҳаракатлар стратегияси тўғрисида»ги Фармони, 2019 йил 20 сентябрдаги ПҚ-3107-сон «Ўзбекистон Республикаси «Ўзархив» агентлиги фаолиятини такомиллаштириш тўғрисида»ги Қарори, 2019 йил 7 июндаги ПҚ-4354-сонли «Ўзбекистон Республикаси аҳолисига ахборот-кутубхона хизмати кўрсатишни янада такомиллаштириш тўғрисида»ги Қарори ҳамда мазкур фаолиятга тегишли меъёрий-ҳуқуқий ҳужжатларда белгиланган вазифаларни амалга оширишда ушбу диссертация тадқиқоти муайян даражада хизмат қилади.

Тадқиқотнинг республика фан ва технологияларини ривожланишининг устувор йўналишларига мослиги. Мазкур тадқиқот республика фан ва технологиялар ривожланишининг IV. «Ахборотлаштириш ва ахборот-коммуникация технологияларини ривожлантириш» устувор йўналиши доирасида бажарилган.

Диссертация мавзуси бўйича хорижий илмий тадқиқотлар шарҳи. Тўлиқ матнлардан маълумот қидириш ва уларни ўзаро ўхшашлигини аниқлаш борасида моделлар, алгоритмлар ва дастурий таъминотлар ишлаб чиқиш бўйича тадқиқотлар жаҳон илмий марказлари ва олий таълим муассалари, жумладан Curtin University of Technology Perth (Австралия), Graz University of Technology (Австрия), South Bank University's School of Computing (Буюк Британия), Bauhaus University Weimar (Германия), Computer Science Stanford University (АҚШ), School of Computing Queen’s University at Kingston Ontario (Канада), University of Applied Sciences Berlin (Германия), The Hong Kong Polytechnic University Kowloon (Гонконг)да тадқиқотлар олиб борилган ва олиб борилмоқда. Шу билан бирга Google (АҚШ), Яндекс, Антиплагиат (Россия Федерацияси) ташкилотларида ҳам бу борада бир қатор изланишлар олиб борилмоқда.

Жаҳонда ахборот-кутубхона ва архив маълумотлар базасини шакллантириш ҳамда улардан тўлиқ матнли маълумотларни излаш бўйича Katipo Communications (Янги Зерландия), Washington Library Network (АҚШ)да изланишлар шунингдек матнларни бир тилдан иккинчи тилга нейрон тармоқлар асосида таржима қилиш ҳамда уларни плагиатга текшириш борасида ҳам илмий изланишлар олиб борилмоқда.

Матнларнинг ўзаро ўхшашлигини аниқловчи усуллар, моделлар ва алгоритмларни ишлаб чиқишга йўналтирилган илмий изланишлар жаҳоннинг етакчи илмий марказлари ва олий таълим муассасалари, жумладан Bank University's School of Computing (Буюк Британия), Graz University of Technology (Австрия), The Hong Kong Polytechnic University Kowloon (Гонконг)да илмий тадқиқот ишлари олиб борилмоқда.

Муаммонинг ўрганилганлик даражаси. Матнларни ўзаро солиштириш ҳамда уларни ўхшашлигини аниқлаш борасида ишлаб чиқилган

моделлар, алгоритмлар U. Manber, N. Heintze, V.L. Hong, A. Broder, D. Fetterly, A. Chowdhury, W. Pugh, C. Ильинский, M. Hermann, Z. Bilal, Ю.Г. Зеленков, И.В. Сегалович ва бошқа олимларнинг илмий ишларида кўриб чиқилган. Электрон кутубхона ва архив тизимларини шакллантириш жараёнини автоматлаштириш ҳамда улардан тўлиқ матнли маълумотларни излаш борасида эса Я.Л.Шрайберг, Ф.С. Воройский, А.С. Карауш, А. И. Бродовский, Е.В. Линдеман каби олимлар тадқиқот ишлари олиб боришган.

Ўзбекистон Республикасида ахборот-кутубхона ва архив соҳасини автоматлаштириш, уларда маълумот қидириш жараёни модели ва алгоритмлари тадқиқига М. А. Раҳматуллаев, У. Ф. Каримов, А. Ш. Мухаммадиев ва бошқа олимларнинг илмий ишларида ўрганилган. Матнларни морфологик ва лексик таҳлил қилиш жараёнлари эса А. К. Пўлатов, С. Ризаев, С. Муҳамедова ва бошқа олимлар ишларида келтириб ўтилган.

Бу соҳадаги тадқиқотларнинг таҳлили шуни кўрсатдики, санаб ўтилган муаллифларнинг илмий тадқиқот ишларида кўп тилли тўлиқ матнли маълумотлар базасида ўзаро ўхшаш матнларни аниқлаш масалалари етарли даражада ўрганилмаган.

Диссертация тадқиқотининг диссертация бажарилган олий таълим муассасасининг илмий-тадқиқот ишлари режалари билан боғлиқлиги. Диссертация тадқиқоти Муҳаммад ал-Хоразмий номидаги Тошкент ахборот технологиялари университети илмий тадқиқот режасининг ЕФ4-003-«Корпоратив ҳисоблаш тармоқлари ишончлилигини баҳолаш модел ва алгоритмлари тадқиқоти» (2012-2013), А5-ФА-А012-«Ўзбекистон Республикаси Олий аттестация комиссиясининг ахборот тизими» (2012-2014) ҳамда А-Ф-А015-«Даврий нашрлар электрон каталоги ва электрон кутубхонасини яратиш» (2015-2016) мавзусидаги илмий лойиҳаси доирасида бажарилган.

Тадқиқотнинг мақсади. Илмий-таълимий фаолиятда ахборот таъминотини такомиллаштириш учун, кўп тилли тўлиқ матнли кутубхона ва архив маълумотлар базаларидан ахборот қидириш усули, алгоритмлари ва дастурий таъминоти ишлаб чиқишдан иборат.

Тадқиқотнинг вазифалари:

ахборот-кутубхона ва архив тизимларида тўлиқ матнларни лексик, морфологик таҳлил қилувчи усуллар, алгоритмлар ҳамда уларнинг ўзаро ўхшашлигини аниқлаш воситаларини ретроспектив таҳлил этиш;

корпоратив ахборот-кутубхона, архив тармоқларида тўлиқ матнларни лексик, морфологик таҳлил қилиш алгоритмларини такомиллаштириш;

корпоратив ахборот-кутубхона, архив тармоқларида тўлиқ матнлар асосида маълумот қидириш жараёни модели ва алгоритмларини ишлаб чиқиш;

кўп тилли илмий-таълимий тўлиқ матнли библиографик ахборотлар рукнини аниқловчи модел ва алгоритм ишлаб чиқиш;

кўп тилли маълумотлар базаси, интернет тармоғи ҳамда катта маълумотлар (big data) асосида маълумот қидириш ва матнларни ўзаро ўхшашини аниқлаш усулини ишлаб чиқиш;

таклиф қилинган модел ва алгоритмлар асосида, кўп тилли маълумотлар базасида ўзаро ўхшаш матнларни аниқлаш тизимини ишлаб чиқиш.

Тадқиқотнинг объекти сифатида ахборот-кутубхона ва архив маълумотлар базалари ҳамда тўлиқ матнларни қидириш жараёнлари олинган.

Тадқиқотнинг предметини ахборот-кутубхона ва архив тармоқларида кўп тилли маълумотлар базасида матнларни ўхшашлигини аниқловчи моделлар, алгоритмлар ва дастурий воситалари ташкил этади.

Тадқиқотнинг усуллари. Тизимли таҳлил, тўпламлар назарияси, чекли автоматлар, машина ёрдамида таржима қилиш, математик статистика, нейрон тармоқлар ҳамда дастурий тизимларни лойиҳалаш усулларидан фойдаланилган.

Тадқиқотнинг илмий янгилиги қуйидагилардан иборат:

ахборот-кутубхона ва архив тизимларида кўп тилли тўлиқ матнли маълумотлар асосида ахборот қидириш моделлари ва алгоритмлари ишлаб чиқилган;

матн таркибидаги сўзларни морфологик таҳлил қилиш ҳамда маъно англатувчи сўзлар тўпламини аниқлаш бўйича модел ва алгоритмлар ишлаб чиқилган;

сўзларни синоним шаклларини қўллаш асосида юзага келадиган яширин плагиатни аниқлашга имкон берувчи модел ва алгоритм ишлаб чиқилган;

сўзларни синоним шаклларини инобатга олган ҳолда илмий-таълимий ахборотларни машина ёрдамида рукнини аниқловчи алгоритм ишлаб чиқилган;

кўп тилли архив ва кутубхона маълумотлар базаларида матнларни ўзаро ўхшашлигини аниқловчи CLAD (Cross Language Analog Detector – кўп тилли матнларнинг ўхшашлигини аниқлаш) усули ишлаб чиқилган.

Тадқиқотнинг амалий натижалари қуйидагилардан иборат:

интернет тармоғида матндаги сўзларнинг синоним шаклларини инобатга олган ҳолда ўхшашлигини аниқлаш дастурий таъминоти ишлаб чиқилган;

тўлиқ матн таркибидаги терминлар асосида илмий-таълимий рукнини аниқлаш дастурий воситаси ишлаб чиқилган;

рефератив матн асосида электрон каталогни автоматик тарзда шакллантириш дастурий воситаси ишлаб чиқилган;

матнни бир тилли ва кўп тилли тўлиқ матнлар базасидан плагиатга текшириш дастурий воситаси ишлаб чиқилган;

кўп тилли ахборот-кутубхона ва архив маълумотлар базасидан ўхшаш матнларни аниқловчи дастурий воситаси ишлаб чиқилган.

Тадқиқот натижаларининг ишончлилиги. Тадқиқот натижаларининг ишончлилиги тадқиқот давомида олинган натижаларнинг солиштирма таққосланганлиги, матнларни ўзаро ўхшашлигини аниқлашда фойдаланилган моделлар ва алгоритмларнинг қатъийлиги, амалий ва ҳисоблаш

математикасининг синовдан ўтган усулларининг қўлланилиши ҳамда мос гувоҳномалар ва жорий қилинганлик тўғрисидаги далолатномалар билан тасдиқланган натижаларнинг сифат ва миқдор жиҳатдан баҳоланганлиги билан изоҳланади.

Тадқиқот натижаларининг илмий ва амалий аҳамияти. Тадқиқот натижаларининг илмий аҳамияти ўзбек тилидаги сўзларни морфологик, ўзбек, рус ва инглиз тилларидаги матнларни лексик таҳлил қилувчи, матнларни бир ва кўп тилли ўзаро ўхшашлигини аниқловчи моделлар, алгоритмлар ва усул ишлаб чиқилганлиги билан изоҳланади.

Тадқиқот натижаларининг амалий аҳамияти матнларни кўп тилли ўхшашлигини текшириш, корпоратив электрон кутубхоналар ва тўлиқ матнли маълумотлар базасидан маълумот қидиришга ҳамда реферативе маълумотлар асосида электрон каталогни автоматик тарзда шакллантиришга имкон берувчи дастурий восита ишлаб чиқиш билан изоҳланади.

Тадқиқот натижаларининг жорий қилиниши. Кўп тилли тақсимланган ахборот-кутубхона ва архив маълумотлар базаларидан ўхшаш матнларни қидириш усул ва алгоритмлари бўйича олинган натижалар асосида:

сўзларнинг синоним шаклларини қўллаш асосида юзага келадиган яширин плагиатни аниқлашга имкон берувчи модел, алгоритм ҳамда кўп тилли архив ва кутубхона маълумотлар базаларида матнларни ўзаро ўхшашлигини аниқловчи CLAD усули Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Олий аттестация комиссияси ахборот тизимига дастурий модул сифатида тадбиқ қилинган (Ахборот технологиялари ва коммуникацияларини ривожлантириш вазирлигининг 2019 йил 28 декабрдаги 33-8/9210-сонли маълумотномаси). Илмий тадқиқот натижасида, илмий-таълимий маълумотлари асосида электрон кутубхонани ташкил қилиш ҳамда муассаса ходимлари учун матнларни ўзбек, рус ва инглиз тилларидаги ўхшашларини аниқлаш имконияти яратилган;

матн таркибидаги сўзларни морфологик таҳлил қилиш ҳамда маъно англатувчи сўзлар тўпламини аниқлаш бўйича модел ва алгоритмлар ҳамда матнларни машина ёрдамида рукнини аниқлаш алгоритмлари «Ўзбектелеком» акционерлик жамияти симли телефон абонентларига хизмат кўрсатиш учун мўлжалланган КАСР ва tBilling тизимларига тадбиқ қилинган (Ахборот технологиялари ва коммуникацияларини ривожлантириш вазирлигининг 2019 йил 18 октябрдаги 33-8/7394-сонли маълумотномаси). Натижада, телефон станцияларидан келган CDR файлларни биллинг тизимига юклаш жараёни тўлиқ автоматлаштирилган. Биллинг тизимига сўзлашувларни такрор киритилишининг олди олинди ҳамда мазкур жараёнда фаолият олиб бораётган ходимларнинг кунига ўртача 12% иш вақти тежалган ҳамда ўртача йиллик 36 млн. сўмлик хизмат кўрсатиш таннархининг камайишига эришилган;

ахборот-кутубхона ва архив тизимларида кўп тилли тўлиқ матнли маълумотлари асосида ахборот қидириш модели ва алгоритми асосида

Хоразм вилояти Қўшкўпир туман халқ таълими бўлими ва унга қарашли 52 та ўрта мактаблар учун марказлашган электрон кутубхона ишлаб чиқилган (Ахборот технологиялари ва коммуникацияларини ривожлантириш вазирлигининг 2019 йил 28 декабрдаги 33-8/9210-сонли маълумотномаси). Мазкур электрон кутубхонада тўлиқ матнли файллардан маълумот қидириш жараёни йўлга қўйилиши натижасида керакли маълумотларни излаб топиш учун сарфланадиган вақтни 20-30 баробарга қисқаришига олиб келган.

Тадқиқотлар натижаларининг апробацияси. Мазкур назарий ва амалий тадқиқот натижалари 5 та халқаро илмий-техник конференцияларида маъруза қилинган ва муҳокамадан ўтказилган.

Тадқиқот натижаларининг эълон қилинганлиги. Тадқиқотлар мавзуси бўйича жами 26 та илмий иш чоп этилган, жумладан Ўзбекистон Республикаси Олий аттестация комиссиясининг докторлик диссертациялари асосий илмий натижаларини чоп этишга тавсия этилган илмий нашрларда 14 та мақола, 7 таси республика журналларида ва 7 таси хорижий журналларда нашр қилинган, ҳамда 2 та ЭҲМ учун яратилган дастурий воситаларни қайд қилиш гувоҳномалари олинган.

Диссертациянинг тузилиши ва ҳажми. Диссертация кириш, бешта боб, хулоса, фойдаланилган адабиётлар рўйхати, иловалардан иборат. Диссертациянинг ҳажми 166 саҳифани ташкил қилади.

ДИССЕРТАЦИЯ ИШИНИНГ АСОСИЙ МАЗМУНИ

Киришда диссертация иши мавзусининг долзарблиги ва зарурати асосланган, тадқиқотнинг мақсади ва вазифалари, объект ва предмети тавсифланган, республика фан ва технологиялари ривожланишининг устувор йўналишларига мослиги кўрсатилган, тадқиқотнинг илмий янгилиги ва амалий натижалари баён қилинган, олинган натижаларнинг ишончлилиги, илмий ва амалий аҳамияти очиб берилган, тадқиқот натижаларини амалиётга жорий қилиш, ишнинг апробацияси, нашр этилган ишлар ва диссертация тузилиши бўйича маълумотлар келтирилган.

Биринчи «**Ахборот-кутубхона ва архив тизимларида илмий-таълимий маълумотлар базасидан тўлиқ матнли маълумотларни излаш усулларини тизимли таҳлили**» бобида тўлиқ матнли маълумотлар базасидан маълумот қидириш усуллари, моделлари ва алгоритмлари таҳлил қилинган. Матнларни лексик ва морфологик таҳлил қилувчи кенг тарқалган модел ва алгоритмлар ўрганилган ҳамда уларнинг қўлланилиш жараёнлари таҳлил қилинган. Матнларни машина ёрдамида таржима қилиш усуллари, уларнинг ўзаро фарқлари ҳамда қўлланилиш соҳалари келтириб ўтилган. Матнларнинг ўзаро ўхшашлигини аниқлаш шакллари ва босқичлари ҳамда бу борада бажариладиган вазифалар кўриб чиқилган. Шу билан бирга, матнларни лексик ва морфологик таҳлил қилувчи модел ва алгоритмлар ўрганилган.

Кўп тилли тўлиқ матнли маълумотлар базасидан маълумот қидириш ҳамда ўзаро ўхшаш матнларни аниқлаш нафақат матнларни ўзаро плагиатга

текшириш балки, ахборот-кутубхона ва архив тизимларига тегишли тўлиқ матнларни рукнини аниқлаш, рефератив матн ёки мақола асосида тўлиқ матнлар тўпламини аниқлаш жараёнларини автоматлаштиришда фойдаланилади.

Иккинчи «**Кўп тилли маълумотлар базасида тўлиқ матнларни ўзаро ўхшашлигини аниқлаш босқичлари, моделлари ва алгоритмлари**» бобида кўп тилли тўлиқ матнли маълумотлар базасидан маълумот қидириш жараёни босқичлари ишлаб чиқилган бўлиб, у қуйидаги қисмлардан иборат:

- тўлиқ матнни формаллаштириш – ҳар хил форматдаги рақамли маълумотларни (PDF, DOC, HTML ва ҳакозо) оддий матн шаклига ўтказиш;
- матнни сўзлар тўплами шаклига ўтказиш ва маъно берувчи сўзларни аниқлаш;
- тўпламдаги ҳар бир сўзнинг техник ўзак қисмини аниқлаш – табиий тил морфологик қоидалари асосида ишлаб чиқилган алгоритмларга кўра ҳар бир сўзнинг техник ўзак шаклини аниқлаш;
- терминларни бир тилдан бошқа тилга таржима қилиш – луғат асосида машинавий таржима қилиш усули ёрдамида амалга ошириш. Мазкур босқичда фақат ҳар хил табиий тилда тавсифланган матнлар учун амалга ошириш;
- ҳар бир термин ва унинг таржимавий шаклини лексик таҳлил қилиш – терминнинг синоним шаклларини ва улар орасидан умумий шаклини аниқлаш;
- терминларни маълумотлар базасига сақлаш – терминлар тўплами шаклига ўтказилган матнни, кўп тилли маълумотлар базасига махсус шаклда киритиш;
- матнларни ўзаро ўхшашлигини аниқлаш ва уларни маъно жиҳатидан гуруҳлаш.

Матнларни кўп тилли маълумотлар базасига киритиш ёки уларни ўзаро солиштириш жараёнида, дастлаб ҳар бир матн қуйидаги шаклга ўтказилади:

$$D = ((d_1, d_1^{'}, n_1),\ (d_2, d_2^{'}, n_2), .., (d_k, d_k^{'}, n_k)), \qquad (1)$$

бу ерда D - берилган матн, k - матн таркибидаги маъно англатувчи терминлар сони, d_i - D матннинг i терминнинг умумий синоним шакли, $d_i^{'}$ - d_i терминнинг таржима шакли, n_i - d_i терминнинг D матн таркибидаги қатнашишлар сони. Бундан ташқари матнларни ўзаро ўхшашлигини аниқлаш жараёнида, ҳар бир матнни оғирлиги ҳам инобатга олинади.

$$N = \sum_{i=1}^{k} n_i\ , \qquad (2)$$

бу ерда N - берилган D матн оғирлиги.

Маълумки, табиий тилда шакли ҳар хил, аммо маъно жиҳатдан бир хил сўзлар мавжуд. Матн таркибида ҳам бундай сўзлар фаол иштирок қилади. Тилшуносликда бундай сўзлар ўзаро синоним сўзлар деб юритилади. Қуйида

мазкур шакл жиҳатдан ҳар хил, маъно жиҳатдан эса бир хил иккита гап берилган.

q) Ўғлим, афтингни артиб ол *t) Ўғлим, юзингни артиб ол*

Агар уларнинг қисмлари ўзаро ўхшаш бўлса, берилган q ва t матнлар ўзаро ўхшаш ҳисобланади. Терминларни ўзаро ўхшашлик даражасини ифодалаш учун:

$$\varphi(q,t) = (q \approx t), \tag{3}$$

$$\varphi(q,t) \in R,\ R = [0..1]$$

бу ерда $\approx$ - символи матн таркибидаги сўзларнинг маъно жиҳатдан ўзаро ўхшашлигини ифодалайди. R-ўзаро яқинлик мезонини ифодаловчи тўплам бўлиб, 0...1 орасидаги қийматга эга ҳақиқий сонлардан иборат элементлардан ташкил топган.

$$\varphi(q,t) = \begin{cases} 1,\ q \approx t \\ 0, q \cong t \end{cases}$$

бу ерда $\cong$- символи матн таркибидаги сўзлар ўзаро ўхшаш бўлмаслигини, ёки белгиланган миқдордан кам бўлганлигини ифодалайди.

Ҳозирги кунда, матн таркибидаги сўзларнинг синоним шаклларини инобатга олган ҳолда уларни ўзаро солиштириш жараёнида, ҳар бир сўзнинг барча синоним шакллари текшириб ўтилади.

$$s = \sum_{i=1}^{l} c_i, \tag{4}$$

бу ерда c_i - i-термин синонимлари сони бўлиб, ихтиёрий c_i учун, $c_i \geq 1$ шарти ўринли, s - умумий солиштиришлар сони. Таклиф қилинаётган CLAD усулида эса бу борада қуйидагича ёндашув илгари сурилади.

$$\omega(w_{i1}) = \omega(w_{i2}) = \omega(w_{i3}) .. = \omega(w_{in}) = w_i, \tag{5}$$

$$w_i \in [w_{i1}, w_{i2}, w_{i3}, .., w_{in}],\ \forall w_{ij} \forall w_{ik} \varphi(w_{ij}, w_{ik}) \approx 1$$

бу ерда w_{ij} - шакли ҳар хил аммо маъноси бир хил бўлган сўзлар яъни ўзаро синоним сўзлар/терминлар тўплами, $\omega(w)$ - w сўзни асосий шаклини аниқловчи функция бўлиб, юқорида кўсатилгандек функцияга ўзаро синоним сўзлар параметр сифатида берилса натижа бир хил қийматга эга бўлади.

Терминни лексик таҳлил қилиш деб белгиланган мазкур жараённи матннинг барча терминлари учун қўлланилади. Натижада, матнларни ўхшашлигини текшириш жараёнида сўзни синоним шаклларини ҳам инобатга олган ҳолда алгоритмнинг **ишлаш вақти** $V(l)$ қуйидагича бўлади.

$$V(l) \in O(l), \tag{6}$$

бу ерда $O(l)$- чизиқли вақт, яъни алгоритмнинг умумий ишлаш вақтига сўзларнинг синоним шакллари сони таъсир қилмайди. Шу ўринда, $\omega(w)$ функция учун қуйидаги шартлар бажарилиши лозим.

• $[w_{i1}, w_{i2},.., w_{in}]$ тўпламга янги w_{in+1} элемент қўшилганда ҳам $\omega(w_{ij})$ ни натижасига таъсир қилмаслиги лозим, бу ерда $j \in [1..n+1]$;

• $[w_{i1}, w_{i2},.., w_{in}]$ тўпламдан ихтиёрий w_{ij} элементи олиб ташланган тақдирда ҳам $\omega(w_{ik})$ қиймати ўзгармаслиги лозим, бу ерда $k \in [1.., j-1, j+1, n]$ бўлиб, қуйида мазкур шартларни математик модели келтирилган.

$$W = [w_{i1},\ w_{i2},\ w_{i3},.., w_{in}] \tag{7}$$

$$\forall w_{ik} \in W\ \omega(w_{ik}) = w_i,\ \ w_i \in W,\ V = W \cap [w_{in}] = [w_{i1},.., w_{in-1}]$$

$$Z = W \cup [w_{in+1}] = [w_{i1},.., w_{in}, w_{in+1}],\ \forall v \in V\ \omega(v) = \forall z \in Z\ \omega(z) = \forall w \in W\ \omega(w) = w_i$$

Тадқиқот давомида ўзбек тилидаги гап таркибидаги сўзларнинг техник ўзак қисмини аниқлашда қўлланиладиган чекли автомат ишлаб чиқилди. Қуйидаги жадвалда ўзбек тилидаги қўшимчалар турлари келтириб ўтилган.

1-жадвал: Ўзбек тилидаги қўшимчалар турлари

Коди	Қўшимчалар тури	Қайси қўшимчалардан кейин келиши	Мисол
0	ўзак		
Сўз ясовчи қўшимчалар			
1	От ясовчи	0	-чи, -увчи, -ла
2	Феъл ясовчи	0	-ла
3	Сифат ясовчи	0	-севар
Шакл ясовчи қўшимча			
11	Кўплик қўшимчаси	0, 1, 3, 14	-лар
12	Бўлишсизлик шакли	0, 2	-ма (-мас)
13	Сифат даражалари	0, 3	-роқ
14	Феъл замонлари	0, 2, 12, 15	-ди, -моқчи
15	Феъл нисбатлари	0, 2	-(и)ш, -тир
Сўз ўзгартирувчи қўшимча			
101	Келишик қўшимча	0, 1, 11, 103	-ни, -га, -нинг
102	Шахс-сон	0, 1, 14	-(и)к, -(и)м
103	Эгалик қўшимча	0, 1, 11	-(и)м, -(и)миз
104	Ҳаракат номи	0, 2	-(и)ш, -у(в)
105	Сифатдош	0, 2, 15	-ган (-кан, -қан)

Мазкур жадвалда ҳар бир тур қўшимчалари маълум бир код билан белгиланиб, ўз навбатида у қайси турдаги қўшимчадан кейин ишлатилиши келтириб ўтилган. Сўзни морфологик таҳлил қилиш босқичи чекли автомат асосида амалга оширилган бўлиб, қуйида ушбу жараён таркиби келтирилган.

- ҳолатлар тўплами – Q (мазкур тўплам чекли бўлади);
- берилган символлар тўплами – E (мазкур тўплам чекли бўлади);
- ўтиш функцияси – δ (бир ҳолатдан иккинчи бир ҳолатга ўтказувчи функция);
- дастлабки ҳолат $q_0 \in Q$.
- натижавий ҳолатлар тўплами F (мазкур тўплам Q нинг қисм тўпламидир).

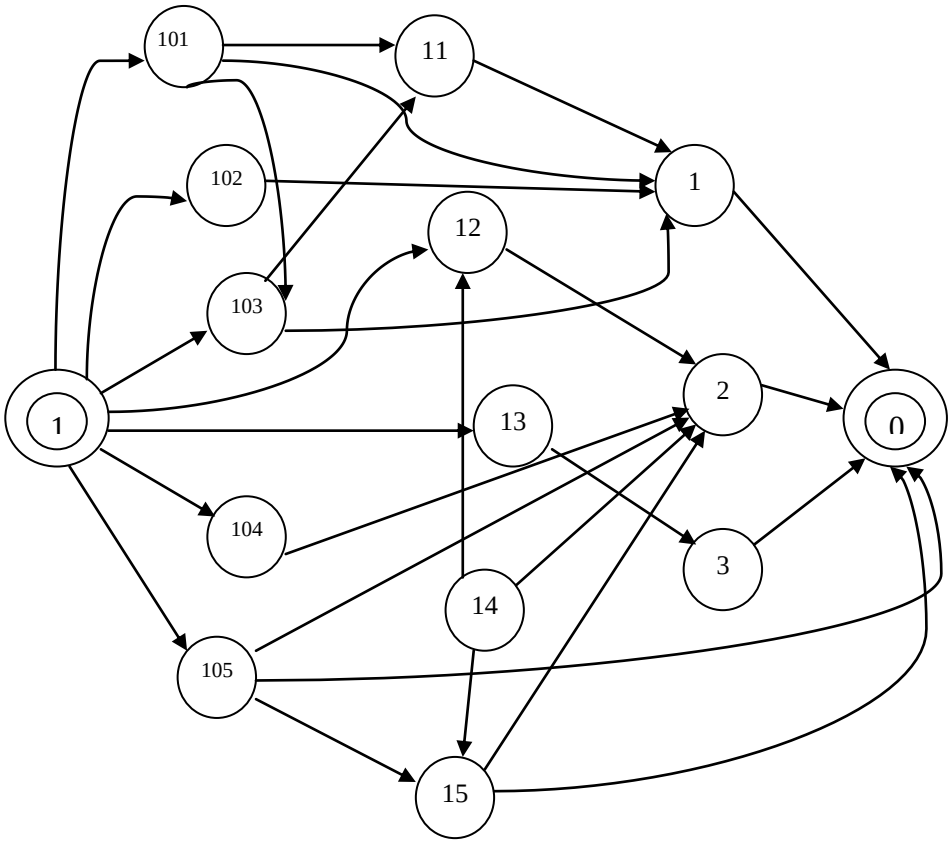
Юқорида келтирилган чекли автоматнинг ҳар бир қисмини «*ишламаганларни*» сўзи мисолида кўриб чиқсак:

1. Ҳолатлар тўплами - берилган сўзнинг таркибидаги ҳар бир қўшимчани кетма-кет олиб ташлашда ҳосил бўлган сўзлар тўпламидир. **Q={ишламаганларни, ишламаганлар, ишламаган, ишлама, ишла, иш} Q_i** ўзбек тилида қўшимчалар **асос (ўзак) + сўз ясовчи + шакл ясовчи + сўз ўзгартирувчи** шаклда келишини инобатга олган ҳолда берилган сўзнинг ўзагини аниқлаш жараёни қуйидаги ҳолатларда бўлади.

a) Негиз – берилган сўз «*ишламаганларни*»;
b) Сўз ўзгартирувчи қўшимчаларсиз ҳолат – «*ишламаганлар*»;
c) шакл ясовчи қўшимчаларсиз ҳолат – «*ишла*»;
d) сўз ясовчи қўшимчасиз ҳолат – «*иш*»;
e) ўзак – сўзнинг ўзгармас қисми.

2. Берилган символлар тўплами – сўзни бир ҳолатдан иккинчи ҳолатга ўтказувчи символлардир. Бизнинг ҳолатда эса бу ўзбек тилидаги қўшимчалардир.

Σ={ни, лар, ган, ма, ла} W_i



1-расм. Ўзбек тилидаги сўзларни ўзагини аниқлаш модели

3. Ўтиш функцияси – сўзни бир ҳолатдан бошқа ҳолатга берилган қўшимча ёрдамида ўтказувчи функция.

q_1-ишла, q_0 – иш, кирувчи қўшимча –*ла*

$q_1=\delta(q_0, ла)$ = ишла

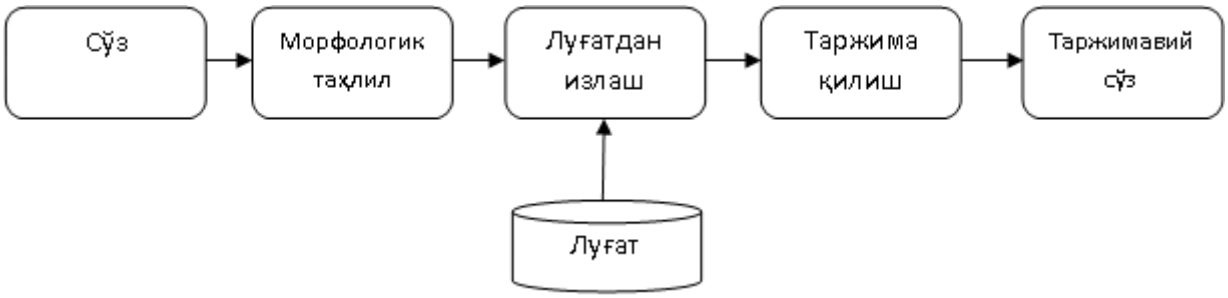
Дастлаб, ЧАни чапдан ўнга (ўзак+қўшимча1+қўшимча2+...) шаклида ишлаб чиқилади. Уни ўнгдан чапга ишлайдиган қилиш учун $\boldsymbol{\delta}$ ни тескари тартибда ишлайди, яъни $\boldsymbol{Q_0}$ = ишла , $\boldsymbol{Q_1}$-ишла

$$\mathbf{Q_1 = \delta\ (Q_0\ ,\ ла)}$$

4. Натижавий негиз тўплам – бу сўзнинг ўзак қисмидан ташкил топган бўлиб, мазкур жараёнда у фақат битта элементга эга бўлади.

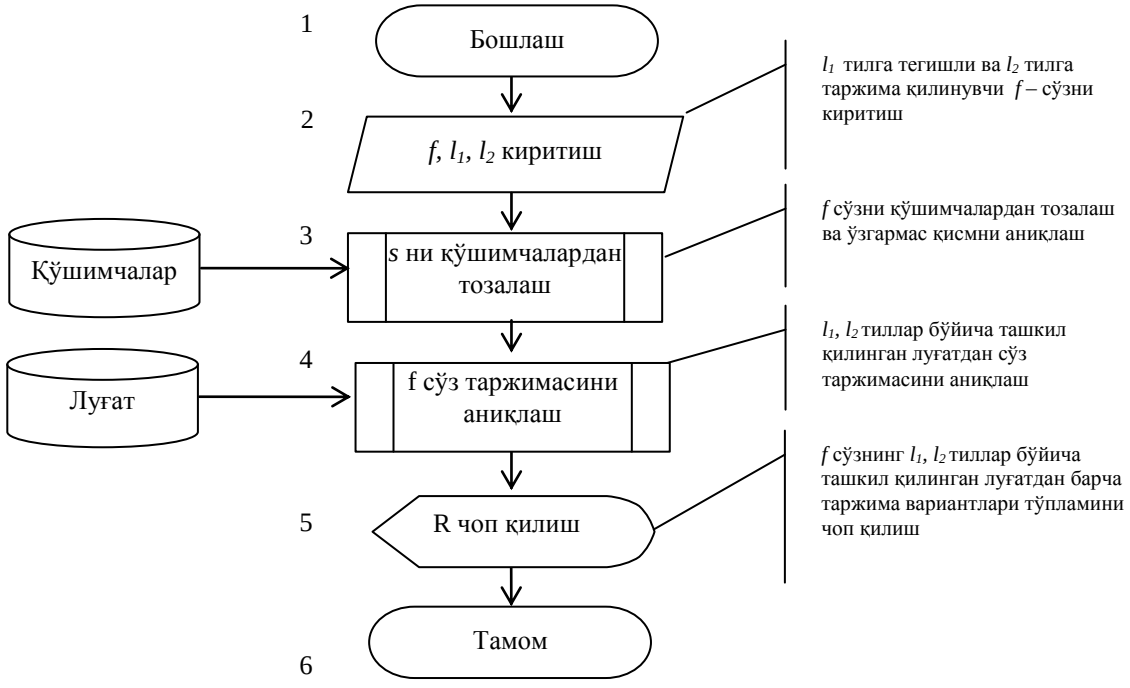
Мазкур чекли автомат асосида ўзбек тилидаги сўзларнинг техник ўзак қисмини аниқлаш алгоритми ва дастурий таъминотини ишлаб чиқиш мумкин.

Ҳар хил тилда тавсифланган матнларни ўзаро солиштириш жараёнида дастлаб, улар битта табиий тил кўринишига олиб келинади. Таъкидлаб ўтилганидек, таклиф қилинаётган CLAD усулида терминларни таржима қилиш луғат асосида ташкил қилиниб, у қуйидаги босқичлардан ташкил топган.



2-расм. Сўзни таржима қилиш босқичлари

Қуйида f сўзни l_1 тилдан l_2 тилга машина ёрдамида таржима қилиш алгоритми келтириб ўтилган.



3-расм. Луғат асосида сўзларни таржима қилиш алгоритми

Маълумки, терминларни таржима қилиш жараёни матнни машина ёрдамида таржима қилиш йўналишига тегишли бўлиб, ҳозирда бу борада жуда кўп натижаларга эришилган. CLAD усули асосида келтирилган терминларни таржима қилиш жараёни қуйидагилар билан мавжуд усуллардан фарқ қилади:

- таржима қилинган термин олдинги морфологик шаклига келтирилмайди;
- солиштириш жараёнида терминнинг барча таржима шакллари инобатга олинади.

Шуни таъкидлаш лозимки, агарда гап мазмуни ўзгармаса гапни таржима қилиш жараёнида сўзнинг синоним шаклларидан ихтиёрий биттасини қўллаш мумкин. Яъни, бу ерда сўз шакли эмас гап мазмуни эътиборга олинади. Матнларни солиштириш жараёнида эса нафақат гап мазмуни балки, сўз шакли ҳам муҳим аҳамият касб этади. Яъни, ҳар хил тилда бўлган матнларни таққослаш жараёнида қўлланиладиган терминларни таржима қилиш жараёни матнни таржима қилиш жараёнидан мураккаброқ ҳисобланади.

Мазкур масалани ечишда, ҳар бир терминнинг таржима шакллари лексик таҳлилдан ўтказилади. Бу эса матнларни солиштириш жараёнида терминнинг барча синоним шаклларини инобатга олган ҳолда солиштиришга имкон беради.

Матнларнинг ўзаро ўхшашлигини текшириш жараёнини иккита D ва T матнлар асосида келтириб ўтсак. Дастлаб, (1) шаклга мос равишда T матн ҳам қуйидаги шаклга олиб келинади.

$$T = ((t_1, t_1^{'}, m_1), (t_2, t_2^{'}, m_2), ..., (t_p, t_p^{'}, m_p)), \quad (8)$$

бу ерда T - берилган матн, p - матн таркибидаги маъно англатувчи терминлар сони, t_i - T матннинг i терминнинг умумий синоним шакли, $t_i^{'}$ - t_i терминнинг таржима шакли, m_i - t_i терминнинг T матн таркибидаги қатнашишлар сони. Агар T матн CLAD усулида асос қилиб олинган табиий тилда ёзилган бўлса, у ҳолда $\forall t_i \in [t_1, t_2, .., t_p] = t_i^{'}$ яъни, T матн учун терминлари таржима қилинмасдан, тўғридан-тўғри ўз шаклига олинади. Шу ўринда, (2) формулага мос равишда, T матннинг оғирлиги M ҳисобланади.

$$M = \sum_{i=1}^{p} m_i \quad (9)$$

Матнларни ўзаро ўхшашлигини текшириш жараёни матнлар тавсифланган тилга қараб икки хил босқичда амалга оширилади. Агар $lang(D) = lang(T)$ бўлса, у ҳолда маълумотларни ўзаро солиштириш жараёнида D матн учун $[d_1, d_2, .., d_k]$ терминлар тўплами асосида, T матн учун эса мос равишда $[t_1, t_2, .., t_p]$ асосида амалга оширилади. Агар $lang(D) \neq lang(T)$ бўлса, у ҳолда маълумотларни ўзаро солиштириш жараёнида D матн учун $[d_1^{`}, d_2^{`}, .., d_k^{`}]$

терминлар тўплами асосида, T матн учун эса мос равишда $[t_1^{'}, t_2^{'}, .., t_p^{'}]$ асосида амалга оширилади. Бу ерда $lang(X)$ - X матн ёзилган табиий тил ҳисобланади.

Шу ўринда кейинги тавсифлашларни соддалаштириш мақсадида қуйидаги белгилашларни киритсак. Юқоридаги (1) шаклида берилган D ҳамда (8) шаклида берилган T матнларни қуйидагича белгиланган

$$D = ((r_1, n_1), (r_2, n_2), ..., (r_k, n_k)) \quad (10)$$

$$T = ((q_1, m_1), (q_2, m_2), ..., (q_p, m_k)), \quad (11)$$

бўлиб, бу ердаги r_i ва q_j қуйидагича ҳисобланади.

$$\forall r_i = \begin{cases} d_i, & lang(D) = lang(T) \\ d_i^{'} \end{cases}, \; R = ((r_1, n_1), ..., (r_k, n_k))$$

$$\forall q_j = \begin{cases} t_j, & lang(D) = lang(T) \\ t_j^{'} \end{cases}, \; Q = ((q_1, m_1), ..., (q_p, m_k))$$

Кейинги босқичларда айнан солиштириш жараёнида R ва Q объектлар асосида амалга оширилади. Чунки, D ҳамда T учун қуйидаги шарт ўринли бўлади.

$$sim(D, T) = sim(R, Q) \quad (12)$$

Кейинги қадам эса, $[r_1, r_2, .., r_k]$ ва $[q_1, q_2, .., q_p]$ тўпламларнинг ўзаро кесишмаси асосида янги $[x_1, x_2, ..., x_l]$ тўплам ҳосил қилинади. Агар мазкур тўплам бирорта элементга эга бўлмаса, яъни $[r_1, r_2, .., r_k] \cap [q_1, q_2, .., q_p] = \theta$ матнларнинг ўхшашлик даражаси 0 га тенг деб топилса, солиштириш жараёнини шу қадамда тугатамиз, акс ҳолда солиштириш жараёни давом эттирилади.

$$[r_1, r_2, .., r_k] \cap [q_1, q_2, .., q_p] = [x_1, x_2, .., x_l] \quad (13)$$

Ҳосил бўлган $\forall x_i$ ва уларни мос равишда R ҳамда Q объектлар таркибида қатнашиш сонлари асосида қуйидаги X объект шакллантирилади.

$$X = ((x_1, n_1, m_1), (x_2, n_2, m_2), ..., (x_l, n_l, m_l)), \quad (14)$$

бу ерда n_i - x_i терминни D матнда такрорланишлар сони яъни оғирлиги, m_i - x_i терминни T матнда такрорланишлар сони яъни оғирлиги. Берилган x_i нуқтаи назаридан D матнни T матнга тегишлилик даражаси $belong(x_i, D, T)$ сифатида белгиланади ва у қуйидагича ифодаланади.

$$belong(x_i, D, T) = \frac{n_i \cdot m_i}{N^2} \quad (15)$$

Мос равишда, берилган x_i нуқтаи назаридан T матнни D матнга тегишлилик даражаси $belong(x_i, T, D)$ сифатида белгиланади ва у қуйидагича ифодаланади.

$$belong(x_i, T, D) = \frac{n_i \cdot m_i}{M^2} \quad (16)$$

Шундан келиб чиққан ҳолда, D матнни T матнга тегишлилик даражаси қуйидагича ҳисобланади.

$$belong(D, T) = \sum_{i=1}^{l} \frac{n_i \cdot m_i}{N^2} \quad (17)$$

Мос равишда, T матнни D матнга тегишлилик даражаси қуйидагича ҳисобланади.

$$belong(T, D) = \sum_{i=1}^{l} \frac{n_i \cdot m_i}{M^2} \quad (18)$$

Шу ўринда, (17) ва (18) натижаларига кенгроқ тўхталиб ўтсак. Маълумки, гоҳида бир матн иккинчи бир матнни таркиби бўлиши мумкин. Масалан, D - матн бирон мақола ёки қисса бўлса, у ўз навбатида, T - мақолалар тўплами ёки асар таркиби бўлиши мумкин. Мазкур ҳолда T матнни D - матнга ўхшашлик даражаси кам бўлиши табиий, аксинча D - матн T матнга 100% ўхшаш ҳисобланади. Мазкур ҳолларда тўғри натижага (17) ва (18) ёрдамида эга бўлиш мумкин. Бундан ташқари, берилган матнларнинг ўзаро ўхшашлик даражасини ушбу функция натижаларининг ўрта арифметиги асосида аниқланади.

$$sim(D, T) = sim(T, D) = \frac{belong(D, T) + belong(T, D)}{2}.$$

Юқорида келтирилган функция CLAD усули асосида матнларни ўзаро ўхшашлигини ифодалайди. Таъкидлаб ўтилганидек, CLAD усули нафақат матнларни бир-бирига ўхшашлигини балки, уларни ўзаро тегишлилигини ҳам аниқлайди.

Учинчи **«Корпоратив ва глобал тармоқда илмий-таълимий ва архив маълумотлар асосида тўлиқ матнли ахборот қидириш модели ва алгоритмлари»** бобида матнни интернет тармоғидан ўхшашларини аниқлаш, кўп тилли тўлиқ матнлар базасида матнларни илмий-таълимий ахборотлар рукни бўйича автоматик таснифлаш ҳамда таклиф қилинаётган CLAD усули асосида ечилиши мумкин бўлган масалаларга тўхталиб ўтилган.

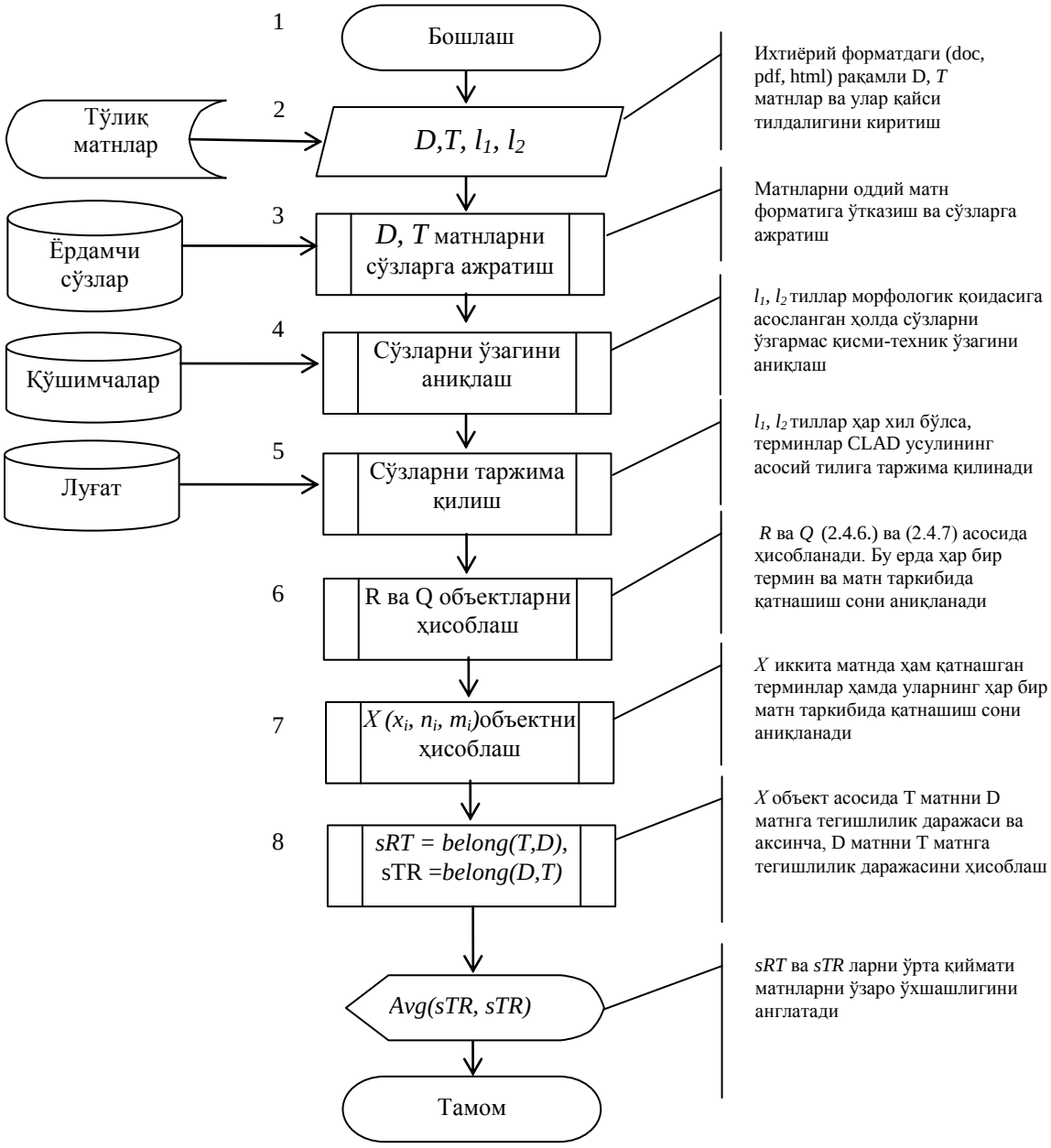
Матнни интернет тармоғидан ўхшашини топиш жараёни қуйидаги қисмлардан ташкил топган:

- ҳар хил форматда тавсифланган матнни оддий матн шаклига ўтказиш;
- матнни терминларга ажратиш;
- терминлар асосида интернет тармоғидан терминларга мос Web-саҳифалар рўйхатини аниқлаш;
- ҳужжатларни солиштириш ва ўхшашлик даражасига қараб Web-саҳифаларни саралаш.

Сўзлар тўплами асосида Web саҳифаларни аниқлашда Yahoo, Google, Yandex XML каби тизимлардан фойдаланиш мумкин. Таклиф қилинаётган алгоритмнинг мавжуд алгоритмлардан асосий фарқи, матнларни

солиштириш жараёнида матн таркибидаги терминларнинг синоним шакллари ҳам инобатга олинган. Қуйида берилган D матн ва T Web саҳифа матнларини сўзнинг синоним шакли асосида солиштириш алгоритми берилган.

1. D матннинг $[x_1, x_2,..x_k]$ тўпламига оид бўлмаган барча терминлари $[d_{11}, d_{21},.., d_{r1}]$ аниқланади. Бу ерда r - $[x_1, x_2,..x_k]$ тўпламга тегишли бўлмаган D матн терминлари сони.



4-расм. Матнларни ўзаро ўхшашлигини текшириш алгоритми

2. d_{i1} терминнинг синоним шакллари $[d_{i2}, d_{i3},.., d_{it}]$ аниқланади. Агар бирорта ҳам синоним шакл топилмаса, у ҳолда 4-қадамга ўтилади.

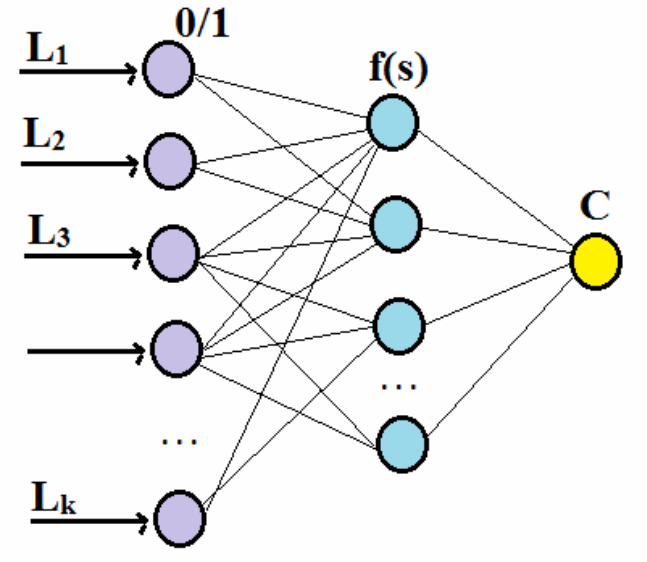
3. $[d_{i2}, d_{i3}, .., d_{it}]$ терминлар $[t_1, t_2, ..., t_s]$ тўпламдан изланади. Биринчи топилган d_{ij} термин ва уни T матн таркибида қатнашиши сони асосида X объектга янги $(x_{k+1}, n_{k+1}, m_{k+1})$ элемент шаклида қўшилади. n_{k+1} - d_{i1} терминни D матн таркибида қатнашиш сони, m_{k+1} - d_{ij} терминни T матн таркибида қатнашиш сони.

4. Навбатдаги термин учун 2 ва 3 қадамлар бажарилади. Барча терминлар учун 2 ва 3 қадамлар бажарилгач, алгоритмнинг иш жараёни ўз ниҳоясига этади.

Ҳосил бўлган X объект асосида D ва T матнларининг ўзаро ўхшашлиги аниқланади. Мазкур жараён 4-расмда келтирилган алгоритм асосида амалга оширилади.

Берилган матнни интернет тармоғидан ўхшашини аниқлаш, нафақат ўзаро плагиат матнларни, қолаверса реферативматн ёрдамида электрон каталогни автоматик тарзда интернет ресурсларидан фойдаланиб шакллантиришга имкон беради. Бу эса ўз навбатида илмий тадқиқот билан шуғулланиш жараёнининг таҳлил қисми учун зарур бўлган манбаларни автоматик тарзда аниқлашни таъминлайди. Тадқиқотчи ўзига керакли соҳа бўйича икки ёки уч саҳифадан иборат матн тайёрлайди. Ушбу матнга ўхшаш бўлган манбалар тўплами юқоридаги алгоритм асосида қурилган дастурий воситадан фойдаланган ҳолда интернет тармоғидан аниқланади.

Матнларнинг илмий-таълимий ахборот рукнини аниқлаш жараёни нейрон тармоқлар асосида амалга оширилган. Тармоқни ўргатиш усули **ўқитувчи ёрдамида,** активация функцияси **сигмоид,** кириш сигнали **луғатдаги термин коди,** архитектураси **3 қатламдан иборат,** чиқиш қиймати **матн тегишли бўлган гуруҳ коди,** хатоликни баҳолаш **ўрта квадрат усули** асосида амалга оширилган. Матн таркибидаги термин оғирлиги эса TF-IDF формуласи асосида аниқланган. Қуйидаги 5-расмда мазкур нейрон тармоқнинг архитектураси келтирилган.



5-расм. Матнларни рукнларга ажратиш нейрон тармоғи архитектураси

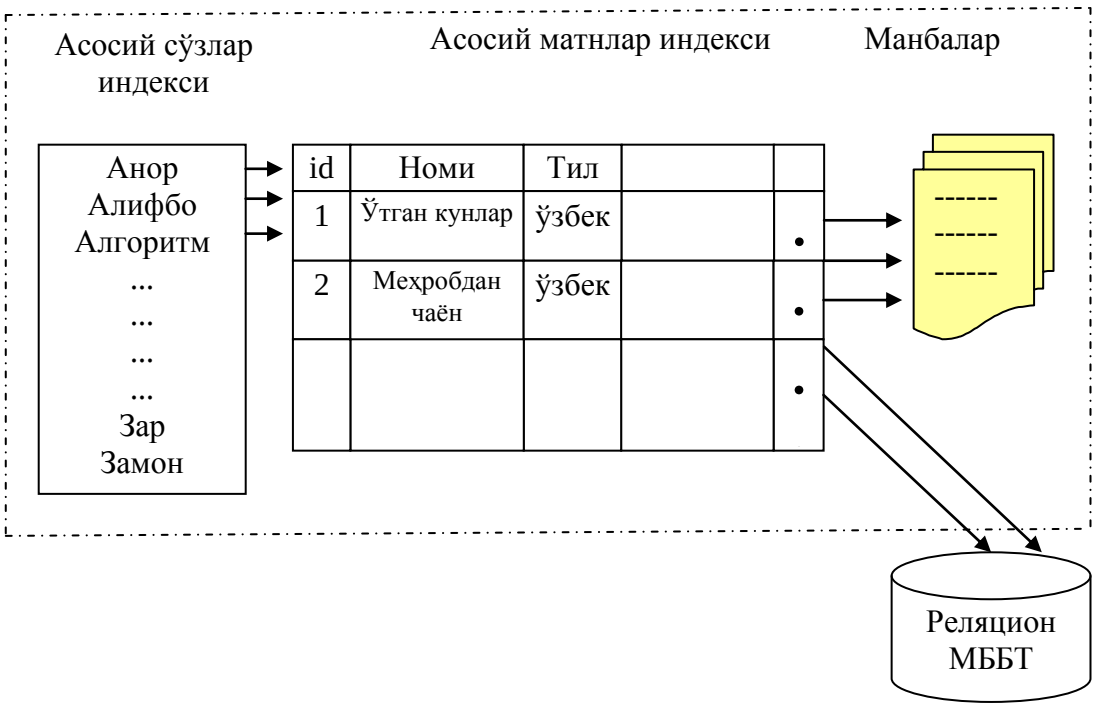
Ҳозирда мавжуд бўлган матнларни автоматик таснифлаш усулларидан фарқли равишда, ҳар бир терминнинг дастлаб синоним шакллари ҳам инобатга олинади, яъни матн таркибидаги терминнинг оғирлигини аниқлаш жараёнида терминнинг барча синоним шакллари ҳам қаралади. Мазкур ҳолат электрон журнал ва тўпламларнинг рукнини аниқлаш жараёнини янада такомиллаштиради.

Таклиф қилинаётган CLAD усули асосида ҳал этилиши лозим бўлган масалалар келтирилиб ўтилган.

3-жадвал. CLAD усулини амалий жиҳатдан аҳамияти

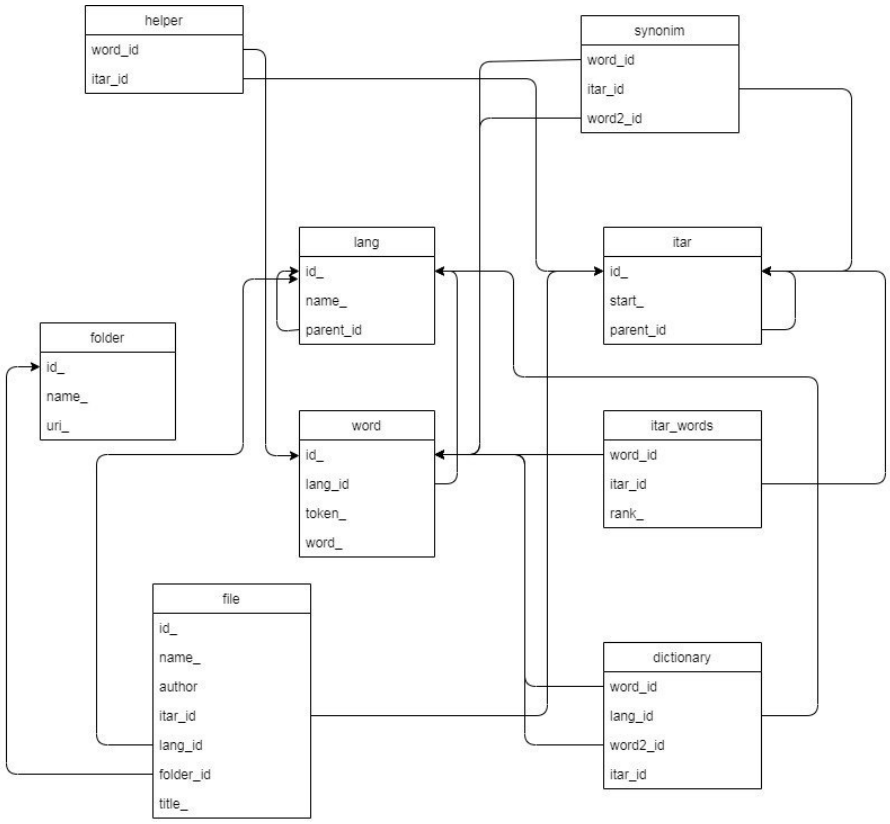
№	Соҳа	Ҳал этилиши кўзда тутилган масалалар
1.	Электрон ҳужжат алмашинуви тизимларида	- Матнларни соҳаларга ажратиш; - матнлардан маълумот излаш; - матнларни тақсимланган ҳолда сақлаш.
2.	Электрон кутубхона ва автоматлаштирилган кутубхона тизимларида	- Электрон китоблар базасини ташкил қилиш ва ундан маълумот излаш; - матн тегишли бўлган илмий-таълимий ахборот рубрикасини автоматик тарзда аниқлаш; - матнга тегишли калит сўзлар тўпламини автоматик шакллантириш.
3.	Нашриёт соҳасида	- Нашр қилинган матнлар базасини шакллантириш, у асосида янги келган матнларни ўхшашлигини текшириш; - матн тегишли бўлган соҳани автоматик тарзда аниқлаш.
4.	Электрон архивларда	- Электрон матнлар архивини ташкил қилиш; - архивга янги келган маълумот турини автоматик аниқлаш; - архив маълумотлардан маълумот қидириш.
5.	Антиплагиат тизимларида	Матнларнинг синоним сўзларини инобатга олган ҳолда қуйидаги турдаги ўхшашликларини аниқлаш: - Матнни кўчириб ташлаш (Copy&Paste); - синонимлар асосида яширилган плагиат; - таржима плагиатни аниқлаш.

Тўртинчи **«Катта маълумотлар (big data) асосида маълумотлар ахборот қидириш тизими таркиби»** бобида CLAD усули асосида ишлаб чиқилган дастурий воситанинг NoSQL (Not Only SQL - нафақат SQL) тамойили асосида қурилган маълумотлар базалари таркиби келтириб ўтилган. Жумладан, кўп тилли тўлиқ матнларни индекслаш жараёнида нореляцион маълумотлар базасидан фойдаланилган. Тизим иш жараёни учун керак бўлган маълумотномалар (сўзнинг синоним шакллари, луғат ва бошқа маълумотлар) эса реляцион маълумотлар базасида сақланади. Тадқиқод асосида ишлаб чиқилган тизимда, реляцион маълумотлар базасини бошқариш тизими сифатида MySQL тизими, нореляцион маълумотлар базасини бошқариш тизими сифатида эса Apache Lucene тизими, тўлиқ матнлар учун файл сервердан фойдаланилган. Қуйидаги 6-расмда тизим маълумотлар базаси таркиби келтирилган.



6-расм. Матнларни нореляцион маълумотлар базасида сақлаш жараёни

Реляцион маълумотлар базаси синоним сўзлар базаси, таржима жараёнида қўлланиладиган луғатлар тўплами, табиий тил таркибидаги ёрдамчи сўзлар ва бошқа маълумотлардан ташкил топади. Қуйидаги 7-расмда мазкур маълумотлар базаси таркиби келтирилган.



7-расм. Тизим реляцион МБ таркиби

Тизим Z39.50 ва HTTP протоколлари асосида автоматлаштирилган ахборот-кутубхона тизимлари маълумотлар базасидан маълумот излаш ҳамда тўлиқ матнларни ўз базасига кўчириб ўтказиш имкониятига эга. Шу билан бирга, тизимга киритилган ҳар хил тўлиқ матн хеч қандай қайта ишловсиз махсус FTP (File Transfer Protocol – файл узатиш протоколи) серверларда сақланади. Юқоридаги расмда мазкур сервер манбалар сифатида берилган.

Бешинчи **«CLAD усули асосида корпоратив ва глобал тармокда кутубхона ва архив тўлиқ матнли маълумотларини масофадан излаш тизими»** бобида тадқиқот асосида ишлаб чиқилган «jComporator» тизимининг функционал таркиби, унинг ахборот-ҳужжат оқими жараёнида бошқа тизимлар билан электрон ҳужжатларни алмашинуви, корпоратив ахборот-кутубхоналар тармоғида тўлиқ матнлардан маълумот қидириш жараёнлари келтириб ўтилган. Шу билан бирга, мазкур тизим асосида интернет тармоғидан ва кўп тилли маълумотлар базасидан ахборот қидириш жараёнининг таҳлили кўриб чиқилган.

Мазкур тизим нафақат ўз маълумотлар базаси балки, корпоратив электрон кутубхоналар негизида ишлаш учун мўлжалланган. Маълумки, электрон кутубхоналарда библиографик ёзувлар MARC формат ёки Dublin Core форматида сақланади. Мазкур форматлар қатъий белгиланган таркибга эга бўлиб, уларда библиографик ёзувга боғланган файл, уни муаллифи, матн ёзилган табиий тил ҳақидаги маълумотлар машина ўқий оладиган шаклда сақланади. Бу эса бизга электрон кутубхона таркибидаги матнлардан қўшимча воситаларсиз фойдаланиш имконини беради. Шу сабабдан, тизимда электрон каталог таркибидаги библиографик ёзувга боғланган матнларни индекслаш имконияти ишлаб чиқилган.

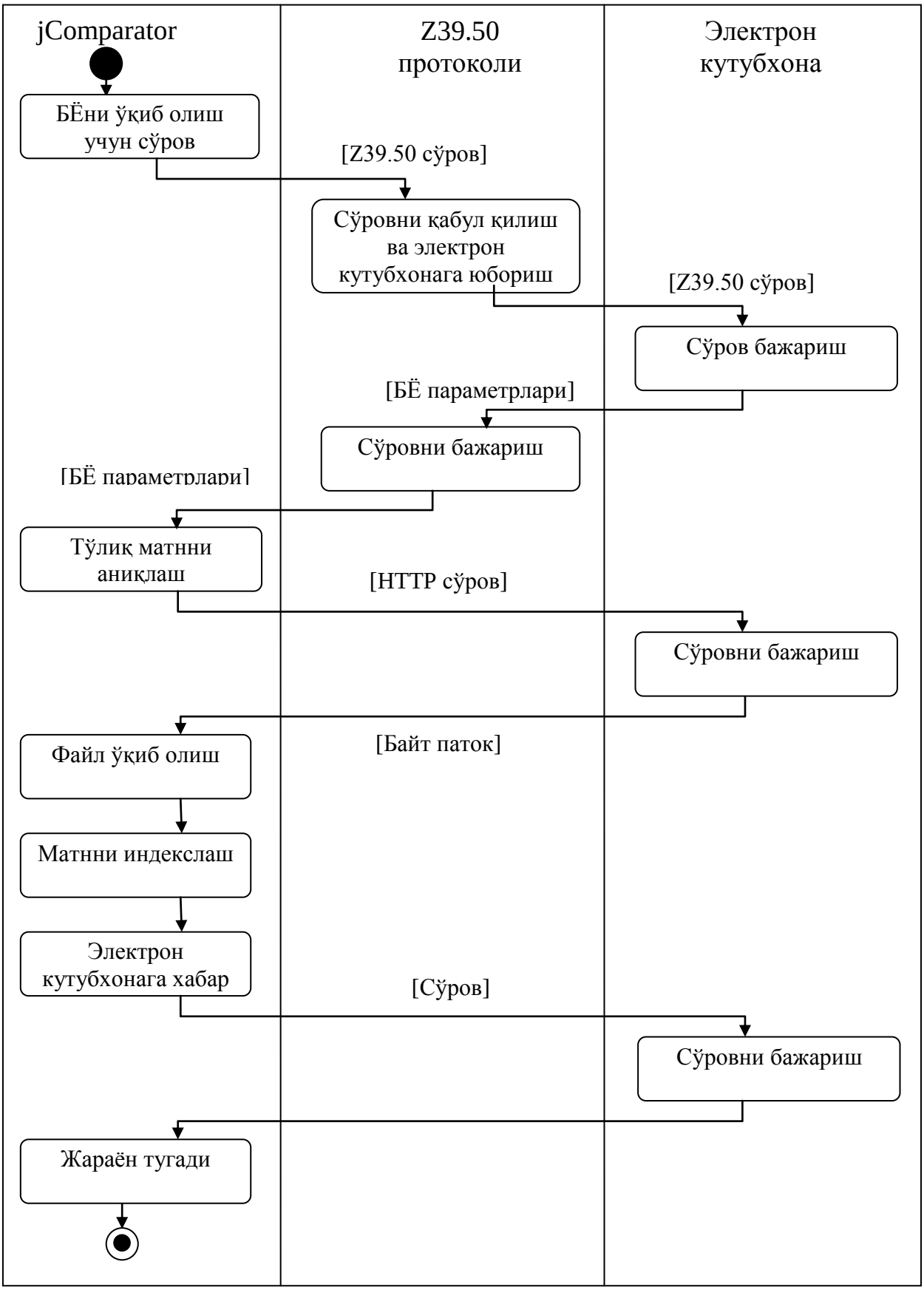
Автоматлаштирилган ахборот - кутубхона ва архив тизимлари билан ахборот алмашиш жараёни қуйидаги шаклларда амалга оширилади:

- Z39.50 – протоколи орқали;
- JDBC протокол асосида электрон кутубхона маълумотлар базасига боғланиш орқали.

Санаб ўтилган шакллар асосида тизим электрон кутубхона библиографик ёзувларига боғлана олади. Аммо, мазкур библиографик ёзувга боғланган тўлиқ матнга ҳам мурожаат этиш ҳуқуқи таъминланиши керак.

Автоматлаштирилган ахборот-кутубхона тизимларида библиографик ёзувлар алмашиниш учун махсус Z39.50 протоколи ишлаб чиқилган бўлиб, фойдаланиладиган маълумотлар базаси ҳамда қўлланиладиган MARC форматдан қатъий назар мазкур протокол асосида ҳар хил тизимлар ўзаро маълумот алмашишлари мумкин. Агар автоматлаштирилган ахборот-кутубхона тизимлари Z39.50 протоколи асосида ишлай олса, jComporator тизими мазкур электрон кутубхона маълумотлар базасига тегишли библиографик ёзувга боғланган тўлиқ матнларни индекслаш имкониятига эга. Бу эса электрон кутубхона ва архив тизимларида тўлиқ матнлар асосида маълумот қидириш, уларнинг ўзаро ўхшашлигини аниқлаш каби масалаларни таъминлайди.

Қуйидаги 8-расмда тизимнинг электрон кутубхона тизимлари билан Z39.50 протоколи асосида маълумот алмашиниш жараёни келтирилган.

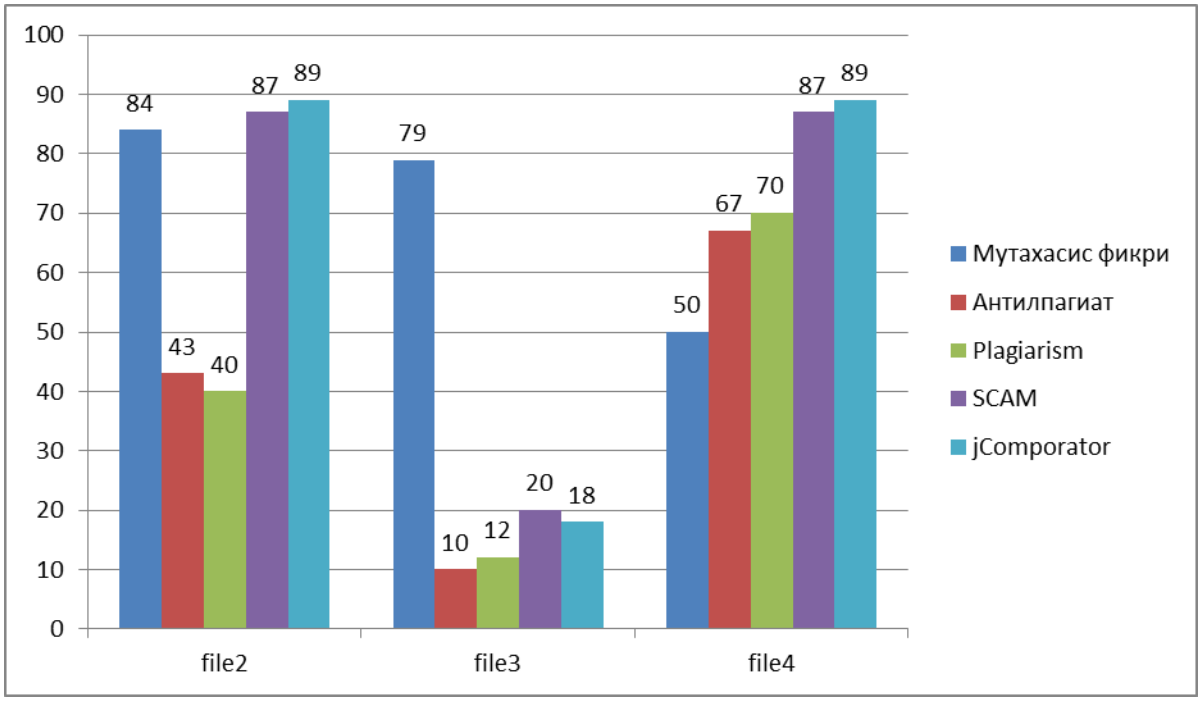


8-расм. Z39.50 протоколи асосида электрон кутубхона билан ишлаш

Матнни интернет тармоғидан ўхшашини аниқлаш алгоритмининг таҳлил жараёни учун **file1.html** матн олинди ҳамда у қуйидаги учта шаклда тавсифланди:

- матн таркибидаги сўзлар синоним шаклларига алмаштирилди ва у file2.html деб номланди;
- матн кўпчилик гапларнинг маъносини ўзгартирмаган ҳолда сўз бирикмалари ҳамда омоним сўзлар тадбиқ қилинди ва мазкур матн file3.html деб номланди;
- кўпчилик гап таркбидаги сўзлар ҳамда матн таркибидига гаплар ўринлари алмаштирилди ва ҳосил бўлган матн file4.html деб номланди.

Таҳлил жараёнида 10 кишидан иборат мутахассислар гуруҳи танлаб олинди ва улардан ҳосил қилинган матнларни дастлабки file1.html матнига ўхшашлик даражасини аниқлаш сўралди. Матнларни ўзаро ўхшашлигини [0..10] орасидаги қийматлар билан белгилаш сўралди. Шу билан бирга мазкур матнлар Антиплагиат, Plagiarism тизимлари ҳамда SCAM алгоритми асосида тузилган дастурий восита асосида ўхшашлиги аниқланди. Қуйидаги расмда таҳлил жараёнидан олинган натижаларнинг кўриниши диаграмма шаклида келтирилган.



9-расм. Сўзларнинг синоним шаклларини инобатга олган ҳолда ўхшашлигини текшириш жараёни таҳлили

Таҳлил жараёни акс этган юқоридаги расм шуни кўрсатадики, тадқиқот асосида ишлаб чиқилган матнларни ўзаро ўхшашлигини аниқловчи алгоритм мавжуд алгоритмларга қараганда 2% яхшироқ натижага эришилган. Аммо **file3.html** – матн таркибида омоним сўзларни ҳамда сўз бирикмаларини қўлланилиши борасида ўтказилган таҳлил жараёни қониқарсиз бўлганлигини

кўрсатади. Бу эса мазкур жараёнда кўпроқ изланишлар олиб боришни тақозо қилади.

Тадқиқот асосида ишлаб чиқилган дастурий тизим модулли принцип асосига қурилган бўлиб, у қуйидаги қисмлардан ташкил топган:

- тўлиқ матнни морфологик ва лексик таҳлил қилиш ҳамда терминларни луғат асосида таржима қилишга мўлжалланган модул;
- тўлиқ матнни Big Data тамойили асосида қурилган маълумотлар базасида сақлаш ҳамда ундан маълумот қидиришга асосланган модул;
- бошқа тизимлар билан биргаликда ишлашни таъминловчи Z39.50 ва HTTP протоколларига асосланган модул;
- терминлар тўпламининг ўхшашлигини аниқловчи модул;
- тизим маълумотномалари ва маълумотлар базаларини бошқарувчи модул.

ХУЛОСА

«Кўп тилли тақсимланган ахборот-кутубхона ва архив маълумотлар базаларидан ўхшаш матнларни қидириш усул ва алгоритмлари» мавзусидаги диссертация иши бўйича олиб борилган тадқиқотлар натижасида қуйидаги хулосалар тақдим этилди:

1. Тўлиқ матнли маълумотлар базасида ҳужжатларни ўхшашлигини текшириш усули, алгоритмлари ва дастурий воситалари таҳлил қилиниб, ҳозирги кунда ҳар хил табиий тил асосида ёзилган матнларни ўзаро ўхшашлигини аниқлаш моделлари, алгоритмлари ва дастурий тизимларини ишлаб чиқиш ўз ечимини кутаётган масалалардан бири эканлиги асослаб берилди.
2. Кўп тилли матнларни ўзаро ўхшашлигини аниқлаш жараёни учун, матнларни сўзларга ажратиш, сўзларнинг ўзгармас қисмини аниқлаш, терминларни машина ёрдамида таржима қилиш, уларни умумий синоним шаклларини аниқлаш модел ва алгоритмлари ишлаб чиқилди. Мазкур модел ва алгоритмлар ҳар хил форматдаги тўлиқ матнларни маълумотлар базасида сақлаш ҳамда улардан маълумот қидиришга имкон берди.
3. Сўзларнинг матн таркибида қатнашиш сони асосида, матндаги калит сўзларни аниқлаш алгоритми ишлаб чиқилди. Ушбу алгоритм асосида тўлиқ матнни каталоглаштиришда калит сўзларни аниқлаш жараёнини тўлиқ автоматлаштиришга олиб келди.
4. Кўп тилли матнлар таркибидаги сўзлар тўплами асосида уларни ўзаро ўхшашлигини аниқлаш ҳамда ўхшаш қисмларни ажратиб бериш модел ва алгоритми ишлаб чиқилди. Мазкур жараён кўп тилли ва яширин плагиат матнларни аниқлашни таъминлади.
5. Илмий-таълимий ва архив маълумотларни интернет тармоғида терминларни синоним шаклларини инобатга олган ҳолда ўхшашини аниқловчи модел ва алгоритм ишлаб чиқилди. Натижада, рефератив маълумотлар асосида электрон каталогни автоматик шакллантириш жараёни йўлга қўйилди.

6. Кўп тилли матнларнинг илмий-таълимий ахборот рукнини аниқлаш модел ва алгоритмлари ишлаб чиқилди. Натижада, матнларни электрон каталог таркибидаги библиографик тавсифга боғлаш жараёнида илмий-таълимий ахборот рукнини аниқлаш босқичи тўлиқ автоматлаштирилди.

7. NoSQL ёндашуви асосида тўлиқ матнларни сақлаш ҳамда улардан маълумот излашни таъминлайдиган реляцион ва нореляцион маълумотлар базаси таркиби ишлаб чиқилган бўлиб, катта ҳажмли матнларни тақсимланган маълумотлар базаларида сақлашни самарали таъминлайди.

8. Z39.50 ва HTTP протоколлари асосида автоматлаштирилган ахборот-кутубхона тизимлари таркибидаги тўлиқ матнлардан маълумот қидириш модел ва алгоритмлари ишлаб чиқилди. Натижада нафақат библиографик ёзув балки унга боғланган тўлиқ матнлардан маълумот излаш жараёни йўлга қўйилди.

9. Кўп тилли тақсимланган ахборот-кутубхона ва архив маълумотлар базаларидан ўхшаш матнларни аниқловчи CLAD усули ишлаб чиқилди. Мазкур усулда кўп тилли матнларни ўзаро ўхшашлигини аниқлаш жараёни тизимли ташкил қилинган.

10.Диссертация тадқиқоти давомида ишлаб чиқилган моделлар, алгоритмлар, CLAD усули ва дастурий воситалар Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Олий аттестация комиссияси ахборот тизимига, «Ўзбектелеком» акциядорлик жамияти биллинг тизимларига, Хоразм вилояти Қўшкўпир туман халқ таълими бўлимига жорий қилинди. Биллинг тизимларига қўллаш жараёнида иш жараёни 12% тезлашишига эришилди.

НАУЧНЫЙ СОВЕТ DSc.13/30.12.2019.T.07.02 ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ ТАШКЕНТСКОМ УНИВЕРСИТЕТЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

ТАШКЕНТСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

АТАДЖАНОВ ЖАСУРБЕК АБДУШАРИБОВИЧ

МЕТОД И АЛГОРИТМЫ ПОИСКА АНАЛОГОВ ТЕКСТОВ ИЗ МНОГОЯЗЫЧНЫХ РАСПРЕДЕЛЕННЫХ ИНФОРМАЦИОННО-БИБЛИОТЕЧНЫХ И АРХИВНЫХ БАЗ ДАННЫХ

05.01.09 - Документалистика. Документоведение. Архивоведение

АВТОРЕФЕРАТ ДИССЕРТАЦИИ ДОКТОРА ТЕХНИЧЕСКИХ НАУК (DSc)

Ташкент – 2020

Тема докторской диссертации (DSc) зарегистрирована в Высшей аттестационной комиссии при Кабинете Министров Республики Узбекистан за № B2020.2.DSc/T169.

Диссертация выполнена в Ташкентском университете информационных технологий.

Автореферат диссертации на трех языках (узбекский, русский, английский (резюме)) размещен на веб-странице Научного совета (www.tuit.uz) и на Информационно-образовательном портале «Ziyonet» (www.ziyonet.uz).

Научный консультант:	**Рахматуллаев Марат Алимович** доктор технических наук, профессор
Официальные оппоненты:	**Умаров Абдусалом Одилович** доктор социальных наук, профессор
	Нуралиев Фахриддин Муродиллаевич доктор технических наук, доцент
	Кучкаров Тахир Сафарович доктор экономических наук, профессор
Ведущая организация:	**Ташкентский государственный технический университет имени Ислама Каримова**

Защита диссертации состоится «23» июня 2020 г. в 10[00] часов на заседании научного совета DSc.13/30.12.2019.T.07.02 при Ташкентском университете информационных технологий (Адрес: 100084, г. Ташкент, ул. Амира Темура, 108. Тел.: (99871) 238-64-43; факс: (99871) 238-65-52; e-mail: tuit@tuit.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Ташкентского университета информационных технологий (регистрационный номер №156). (Адрес: 100202, г. Ташкент, ул. Амира Темура, 108. Тел.: (99871) 238-65-44).

Автореферат диссертации разослан «16» июня 2020 года.
(протокол рассылки №3 от «15» июня 2020 г.).

И.Х.Сиддиков
Председатель научного совета по присуждению ученых степеней, д.т.н., профессор

Х.Э. Хужаматов
Ученый секретарь научного совета по присуждению ученых степеней, доктор философии (PhD) по техническим наукам

Т.С. Кучкаров
Председатель научного семинара при научном совете по присуждению учёных степеней, д.э.н., профессор

ВВЕДЕНИЕ (Аннотация диссертации доктора наук (DSc))

Актуальность и востребованность темы диссертации. Информация признана одним из важных ценностей в мире для развития общества и в настоящее время проводится большая работа по развитию методов и средств их поиска, хранения и использования. В результате развития информационных технологий процессы хранения, передачи и использования данных стали намного эффективнее. Но это, в свою очередь, привело к резкому увеличению объема данных. В настоящее время многие электронные каталоги и электронные библиотеки создаются в результате развития систем, работающих на основе форматов MARC (Machine-Readable Cataloging – машиночитаемая каталогизация) и Dublin Core. И это, в свою очередь, ставит ряд задач в этой области, таких как: поиск информации из полнотекстовых баз данных, автоматизация процесса создания электронного каталога, сравнение полных текстов и определение степени их сходства. В этой связи, одной из важных задач в развитых странах, включая Австралию, Австрию, Великобританию, Германию, США, Канаду, Гонконг, Российскую Федерацию, является разработка инструментов для поиска информации на основе полных текстов и выявления их сходств.

В развитых странах проводятся масштабные научные исследования в целях разработки моделей, алгоритмов и программного обеспечения для поиска полнотекстовой информации из информационно-библиотечных и архивных баз данных, морфологического анализа слов в тексте и определения набора значимых слов, выявления сходств между текстами и машинного выявления их рубрики. Одной из актуальных проблем в этой сфере является разработка моделей, алгоритмов и программного обеспечения, которые позволяют идентифицировать схожие тексты из многоязычной базы данных.

В нашей стране также осуществляются работы по формированию электронного каталога информационно-библиотечных фондов и архивной базы данных. В Стратегии действий по дальнейшему развитию Республики Узбекистан в 2017-2021 годах определены такие задачи, как «... повышение доступности качественных образовательных услуг, ... повышение качества и эффективности деятельности высших образовательных учреждений на основе внедрения международных стандартов обучения и оценки качества преподавания»[1]. Одной из важных задач при реализации данных целей является разработка моделей, алгоритмов и программных обеспечений для поиска информации в полных текстах из многоязычных распределенных информационно-библиотечных и архивных баз данных, выявления их сходств и автоматического формирования электронных каталогов на основе реферативных данных.

[1] Указ Президента Республики Узбекистан от 7 февраля 2017 года № УП-4947 «О Стратегии действий по дальнейшему развитию Республики Узбекистан»

Настоящее диссертационное исследование в определенной степени служит осуществлению задач, определенных в Указе Президента Республики Узбекистан № УП-4947 от 7 февраля 2017 года «О Стратегии действий по дальнейшему развитию Республики Узбекистан», Постановлении № ПП-3107 от 20 сентября 2019 года «О совершенствовании деятельности агентства «Узархив» Республики Узбекистан», Постановлении № ПП-4354 от 7 июня 2019 года «О дальнейшем совершенствовании информационно-библиотечного обслуживания населения Республики Узбекистан», а также задачам, определенным в нормативно-правовых актах в данной сфере.

Соответствие исследования приоритетным направлениям развития науки и технологий республики. Данное исследование выполнено в рамках приоритетного направления развития науки и технологий республики IV - «Информатизация и развитие информационно-коммуникационных технологий».

Обзор зарубежных научных исследований по теме диссертации. Исследования по разработке моделей, алгоритмов и программного обеспечения для поиска информации в полных текстах и выявления их сходств проводились и проводятся международными научными центрами и высшими учебными заведениями, такими как Curtin University of Technology Perth (Австралия), Graz University of Technology (Австрия), South Bank University's School of Computing (Великобритания), Bauhaus University Weimar (Германия), Computer Science Stanford University (США), School of Computing Queen’s University at Kingston Ontario (Канада), University of Applied Sciences Berlin (Германия), The Hong Kong Polytechnic University Kowloon (Гонконг). В то же время Google (США), Yandex, Антиплагиат (Российская Федерация) проводят ряд исследований в этой сфере.

Исследования по формированию информационно-библиотечной и архивной базы данных, поиску из них полнотекстовой информации, а также научные исследования по переводу текстов с одного языка на другой и их проверка на плагиат на основе нейронных сетей проводятся в Katipo Communications (Новая Зеландия) и Washington Library Network (США).

Научно-исследовательские работы, направленные на разработку методов, моделей и алгоритмов определения сходства текстов, проводятся такими ведущими научными центрами и высшими учебными заведениями по всему миру, как Bank University's School of Computing (Великобритания), Graz University of Technology (Австрия), The Hong Kong Polytechnic University Kowloon (Гонконг).

Степень изученности проблемы. Модели и алгоритмы, разработанные для сравнения текстов и определения их сходства, рассмотрены в научных работах таких ученых, как U. Manber, N. Heintze, V.L. Hong, A. Broder, D. Fetterly, A. Chowdhury, W. Pugh, С. Ильинский, M. Hermann, Z. Bilal, Ю.Г. Зеленков, И.В. Сегалович, а также других ученых. Такие же ученые как Я.Л.Шрайберг, Ф.С. Воройский, А.С. Карауш, А. И. Бродовский, Е.В. Линдеман проводили исследования в сфере автоматизации процесса

формирования электронных библиотек и архивных систем и поиска в них полнотекстовой информации.

Модели и алгоритмы автоматизации информационно-библиотечной и архивной сферы в Республике Узбекистан, а также процесса поиска информации в них изучены в научных работах М. А. Рахматуллаева, У. Ф. Каримова, А. Ш. Мухаммадиева и других ученых. Процессы морфологического и лексического анализа текстов рассматриваются в работах А. К. Пулатова, С. Ризаева, С. Мухамедова и других ученых.

Анализ исследований в данной сфере показывает, что в научных исследовательских работах вышеуказанных авторов вопросы идентификации взаимно схожих текстов в многоязычной полнотекстовой базе данных изучены не в полном объеме.

Связь диссертационного исследования с планами научно-исследовательских работ высшего учебного заведения, в котором выполнена диссертация. Диссертационное исследование выполнено в соответствии с научно-исследовательским планом Ташкентского университета информационных технологий имени Мухаммада аль-Хорезми ЕФ4-003 - «Исследование моделей и алгоритмов оценки надежности корпоративных вычислительных сетей» (2012-2013), А5-ФА-А012 - «Информационная система Высшей аттестационной комиссии Республики Узбекистан» (2012-2014), а также в рамках научного проекта А-Ф-А015 - «Создание электронного каталога и электронной библиотеки периодических изданий» (2015-2016).

Цель исследования. Разработка методов, алгоритмов и программного обеспечения для поиска информации из многоязычных полнотекстовых библиотечных и архивных баз данных в целях совершенствования информационного обеспечения научной и образовательной деятельности.

Задачи исследования:

ретроспективный анализ методов лексического, морфологического анализа полных текстов, алгоритмов и средств определения их сходства в информационно-библиотечных и архивных системах;

совершенствование алгоритмов лексического, морфологического анализа полных текстов в корпоративных информационно-библиотечных, архивных сетях;

разработка моделей и алгоритмов полнотекстового поиска информации в корпоративных информационно-библиотечных, архивных сетях;

разработка модели и алгоритма определения рубрики многоязычных научно-образовательных полнотекстовых библиографических данных;

разработка метода поиска информации и определения сходства текстов в многоязычных базах данных, сети Интернет и в системах “больших данных” (big data);

разработка программного комплекса системы идентификации схожих текстов в многоязычной базе данных на основе предложенных моделей и алгоритмов.

Объектом исследования являются информационно-библиотечные и архивные базы данных, а также процессы полнотекстового поиска

Предметом исследования являются модели, алгоритмы и программные средства, определяющие сходство текстов в многоязычных базах данных в информационно-библиотечных и архивных сетях.

Методы исследования. В работе использовались такие методы, как системный анализ, теория множеств, конечные автоматы, машинный перевод, математическая статистика, проектирование нейронных сетей и программных систем.

Научная новизна исследования состоит в следующем:

разработаны модели и алгоритмы поиска информации в информационно-библиотечных и архивных системах на основе многоязычных полнотекстовых данных;

разработаны модели и алгоритмы морфологического анализа слов в содержании текста и определения набора значимых слов;

разработаны модель и алгоритм, позволяющие обнаруживать скрытый плагиат на основе использования синонимичных форм слов;

разработан алгоритм машинного определения рубрики научной и образовательной информации;

разработан метод CLAD (Cross Language Analog Detector – определение сходства многоязычных текстов) для выявления сходства текстов в многоязычных архивных и библиотечных базах данных.

Практическими результатами исследования являются следующие:

разработано программное обеспечение для определения сходства слов в тексте в сети Интернет с учетом синонимичных форм;

разработано программное средство для определения научно-образовательной рубрики на основе терминов в полном тексте;

разработано программное средство для автоматического формирования электронного каталога на основе реферативного текста;

разработано программное средство для проверки текста в одноязычной и многоязычной полнотекстовой базе данных на плагиат;

разработано программное средство, который выявляет схожие тексты в многоязычной информационной-библиотечной и архивной базе данных.

Достоверность результатов исследования. Достоверность результатов исследования обеспечивается осуществлением сравнения результатов, результатами статистического анализа, полученных в ходе исследования, надежностью моделей и алгоритмов, используемых для определения сходства текстов, применением проверенных методов прикладной и вычислительной математики, а также качественной и количественной оценкой результатов, подтвержденных соответствующими сертификатами и актами внедрения.

Научная и практическая значимость результатов исследований. Научная значимость результатов исследования объясняется разработкой моделей, алгоритмов и метода морфологического анализа слов на узбекском

языке, лексического анализа текстов на узбекском, русском и английском языках, выявления одноязычной и многоязычной схожести текстов.

Практическая значимость результатов исследования объясняется разработкой программного средства, позволяющего проверять сходство текстов, представленных на разных языках, осуществлять поиск информации в корпоративных электронных библиотеках и полнотекстовых базах данных, а также автоматически формировать электронные каталоги на основе реферативных данных.

Внедрение результатов исследований. На основе результатов, полученных из алгоритмов и методов поиска схожих текстов из многоязычных распределенных информационно-библиотечных и архивных баз данных:

модель, алгоритм, позволяющие выявлять скрытый плагиат, возникающий на основе использования синонимичных форм слов, а также метод CLAD для определения сходства текстов в многоязычных архивных и библиотечных базах данных внедрены в качестве программного модуля в информационную систему Высшей аттестационной комиссии при Кабинете Министров Республики Узбекистан (справка Министерства развития информационных технологий и коммуникаций от 28 декабря 2019 г. № 33-8/9210). В результате исследования создана возможность формирования электронной библиотеки на основе научно-образовательных данных, а также выявления сходства текстов на узбекском, русском и английском языках;

модели и алгоритмы морфологического анализа слов в тексте и определения набора значимых слов, а также алгоритмы машинного определения рубрики текста внедрены в системы KACP и tBilling, предназначенные для обслуживания абонентов проводной телефонной связи Акционерного Общества «Узбектелеком» (справка Министерства развития информационных технологий и коммуникаций от 18 октября 2019 г. № 33-8/7394). В результате процесс загрузки файлов CDR с телефонных станций в биллинговую систему полностью автоматизирован. Повторное введение разговоров в биллинговую систему предотвращено, и для сотрудников, работающих в этом процессе, сэкономлено в среднем 12% рабочего времени в день, а также ежегодная себестоимость услуг снижена в среднем на 36 млн. сумов;

с помощью модели и алгоритма поиска информации на основе многоязычной полнотекстовой информации в информационно-библиотечных и архивных системах разработана централизованная электронная библиотека для отделения народного образования Кушкупырского района Хорезмской области и 52 средних школ, находящихся в его системе (справка Министерства развития информационных технологий и коммуникаций от 28 декабря 2019 г. № 33-8/9210). В результате того, что процесс поиска информации в полнотекстовых файлах в данной электронной библиотеке

был надлежащим образом налажен, время, используемое на поиск необходимой информации, сократилось в 20-30 раз.

Апробация результатов исследования. Результаты данного теоретического и практического исследования были представлены и обсуждены в 5 международных научно-практических конференциях.

Публикация результатов исследования. По тематике исследований всего было опубликовано 26 научных работ, среди которых 14 статей в научных журналах, рекомендованных для публикации основных научных результатов докторских диссертаций Высшей аттестационной комиссией Республики Узбекистан, 7 - в республиканских журналах и 7 - в зарубежных журналах, а также получены 2 свидетельства о регистрации программного обеспечения для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, пяти глав, заключения, списка использованной литературы, приложений. Объем диссертации составляет 166 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИОННОЙ РАБОТЫ

Во введении обоснованы актуальность и востребованность темы диссертационной работы, описаны цели и задачи, объект и предмет исследования, их соответствие приоритетным направлениям развития науки и технологий в республике, отмечены научная новизна и практические результаты исследования, раскрыты достоверность, научная и практическая значимость полученных результатов, предоставлена информация о внедрении результатов на практике, апробации работы, опубликованных работах, а также о структуре диссертации.

В первой главе **«Систематический анализ методов поиска полнотекстовых данных из научно-образовательных баз данных в информационно-библиотечных и архивных системах»** анализируются методы, модели и алгоритмы поиска информации в полнотекстовых базах данных. Изучены широко распространенные модели и алгоритмы лексического и морфологического анализа текстов, а также проанализированы процессы их применения. Приведены методы машинного перевода текстов, их отличия и области применения. Рассмотрены формы и этапы определения сходства текстов, а также задачи, которые необходимо выполнить в этом направлении. Вместе с тем, изучены модели и алгоритмы, осуществляющие лексический и морфологический анализ текстов.

Поиск информации из многоязычной полнотекстовой базы данных и выявление схожих текстов используется не только для проверки плагиата текстов, но и для автоматизации процессов определения рубрики полных текстов соответствующих информационно-библиотечных и архивных систем, идентификации набора полных текстов на основе абстрактного текста или статьи.

Во второй главе **«Этапы, модели и алгоритмы определения сходства полных текстов в многоязычной базе данных»** разработаны этапы

процесса поиска информации из многоязычной полнотекстовой базы данных, который состоит из следующих частей:

- форматирование полного текста – преобразование цифровых данных в различных форматах (PDF, DOC, HTML и т. д.) в простой текст;
- преобразование текста в набор слов и определение значимых слов;
- определение технического корня каждого слова в наборе – определение технического корня каждого слова в соответствии с алгоритмами, разработанными на основе морфологических правил естественного языка;
- перевод терминов с одного языка на другой – с использованием метода машинного перевода на основе словаря. На данном этапе – осуществлять только для текстов, описанных различными естественными языками;
- лексический анализ каждого термина и его формы перевода – для определения синонимичных форм термина и общей формы среди них;
- сохранение терминов в базу данных – ввод текста, преобразованного в набор терминов, в многоязычную базу данных в специальной форме;
- выявление сходства между текстами и их группировка по смыслу.

При вводе текстов в многоязычную базу данных или их сравнении каждый текст первоначально преобразовывается в следующий формат:

$$D = ((d_1, d_1^{'}, n_1),\ (d_2, d_2^{'}, n_2), .., (d_k, d_k^{'}, n_k)), \qquad (1)$$

где D - заданный текст, k - количество значимых терминов в составе текста, d_i - общая синонимичная форма термина i текста D, $d_i^{'}$ - форма термина d_i после перевода, n_i - количество наличия термина d_i в составе текста D. Кроме того, в процессе определения сходства текстов также учитывается объем каждого текста.

$$N = \sum_{i=1}^{k} n_i , \qquad (2)$$

где N - объем заданного текста D.

Как известно, в естественном языке существуют слова различные по форме, но имеющие одинаковое значение. Такие слова также принимают активное участие в тексте. В лингвистике такие слова называются синонимами. Ниже приведены два предложения, которые отличаются по форме, но имеют одинаковое значение.

q) Сынок, вытри личико *t) Сынок, вытри лицо*

Если их части схожи друг с другом, заданные тексты q и t считаются взаимно схожими. Чтобы выразить степень сходства терминов:

$$\varphi(q,t) = (q \approx t), \qquad (3)$$

$$\varphi(q,t) \in R,\ R = [0..1]$$

где $\approx$ - символ, выражающий сходство слов в тексте по значению. R - это набор, выражающий критерий взаимной близости, который образован из элементов, состоящих из действительных чисел со значением от 0 до 1.

$$\varphi(q,t)=\begin{cases}1,\ q\approx t\\ 0,q\cong t\end{cases}$$

где $\cong$- символ, если слова в тексте не являются взаимно схожими или они меньше указанного количества.

В настоящее время в процессе сравнения слов в тексте, учитывая их синонимичные формы, проверяются все синонимичные формы каждого слова.

$$s=\sum_{i=1}^{l}c_i\,, \tag{4}$$

где c_i - количество синонимов термина-i, для произвольного c_i, соответствует условие $c_i\geq 1$, s - количество общих сравнений. В предлагаемом методе CLAD в этом направлении предлагается следующий подход.

$$\omega(w_{i1})=\omega(w_{i2})=\omega(w_{i3})..=\omega(w_{in})=w_i\,, \tag{5}$$

$$w_i\in[w_{i1},w_{i2},w_{i3},..,w_{in}],\ \forall w_{ij}\forall w_{ik}\varphi(w_{ij},w_{ik})\approx 1$$

где w_{ij} - слова разные по форме, но одинаковые по значению, т.е. набор синонимичных слов/терминов, $\omega(w)$ - является функцией, которая определяет базовую форму слова w, и если функции заданы взаимно синонимичных слов в качестве параметра, как показано выше, результат будет иметь то же значение.

Этот процесс, определенный как лексический анализ термина, применяется ко всем терминам текста. В результате в процессе проверки сходства текстов, учитывая также синонимичные формы слова, **время работы** алгоритма $V(l)$ выглядит следующим образом.

$$V(l)\in O(l)\,, \tag{6}$$

где $O(l)$ - линейное время, то есть общее время работы алгоритма не зависит от количества синонимичных форм слов. Вместе с тем, для функции $\omega(w)$ должны быть выполнены следующие условия.

- даже в случае добавления в набор $[w_{i1},w_{i2},..,w_{in}]$ нового элемента w_{in+1} это не должно влиять на результат $\omega(w_{ij})$, где $j\in[1..n+1]$;
- даже если произвольный элемент w_{ij} будет исключен из набора $[w_{i1},w_{i2},..,w_{in}]$ значение $\omega(w_{ik})$ не должно меняться, где $k\in[1..,j-1,j+1,n]$, математическая модель этих условий приведена ниже.

$$W=[w_{i1},\ w_{i2},\ w_{i3},..,w_{in}] \tag{7}$$

$$\forall w_{ik}\in W\ \omega(w_{ik})=w_i,\ \ w_i\in W,\ V=W\cap[w_{in}]=[w_{i1},..,w_{in-1}]$$

$$Z=W\cup[w_{in+1}]=[w_{i1},..,w_{in},w_{in+1}],\ \forall v\in V\ \omega(v)=\forall z\in Z\ \omega(z)=\forall w\in W\ \omega(w)=w_i$$

В ходе исследования разработан конечный автомат, используемый для определения технического корня слов в составе предложений на узбекском языке. В нижеуказанной таблице перечислены виды суффиксов в узбекском языке.

Таблица 1: Виды суффиксов в узбекском языке

Код	Виды суффиксов	После каких суффиксов добавляются	Пример
0	Корень		
Словообразующие суффиксы			
1	Образующие существительное	0	-чи, -увчи, -ла
2	Образующие глагол	0	-ла
3	Образующие прилагательное	0	-севар
Формообразующие суффиксы			
11	Суффикс множественного числа	0, 1, 3, 14	-лар
12	Отрицательная форма	0, 2	-ма (-мас)
13	Степени прилагательного	0, 3	-роқ
14	Времена глаголов	0, 2, 12, 15	-ди, -моқчи
15	Соотношение глаголов	0, 2	-(и)ш, -тир
Суффиксы, изменяющие слово			
101	Суффиксы падежей	0, 1, 11, 103	-ни, -га, -нинг
102	Местоимение-числительное	0, 1, 14	-(и)к, -(и)м
103	Притяжательные суффиксы	0, 1, 11	-(и)м, -(и)миз
104	Название действия	0, 2	-(и)ш, -у(в)
105	Причастие	0, 2, 15	-ган (-кан, -қан)

В данной таблице каждый вид суффиксов отмечен определенным кодом, который, в свою очередь, указывает, после какого вида суффикса используется каждый из них. Этап морфологического анализа слова выполнен на основе конечного автомата, и структура этого процесса приведена ниже.

- набор состояний - Q (этот набор является конечным);
- заданный набор символов - E (этот набор является конечным);
- функция перехода - δ (функция перехода из одного состояния в другое);
- начальное состояние $q_0 \in Q$.
- Результирующий набор состояний - F (данный набор является частью набора Q).

Рассмотрим каждую часть вышеуказанного конечного автомата на примере слова «*ишламаганларни*»:

1. Набор состояний – это набор слов, образованный последовательным удалением каждого суффикса в данном слове. **Q={ишламаганларни,**

ишламаганлар, ишламаган, ишлама, ишла, иш} Q_i в узбекском языке процессе определения корня (основы) заданного слова, учитывая тот факт, что суффиксы используются в следующей последовательной форме **основа (корень) + словообразующий + формообразующий + словоизменяющий**, представляется следующим образом.

f) Корень – заданное слово «*ишламаганларни*»;
g) Состояние без словоизменяющих суффиксов – «*ишламаганлар*»;
h) Состояние без формообразующих суффиксов – «*ишла*»;
i) Состояние без словообразующего суффикса – «*иш*»;
j) Основа – неизменяемая часть слова.

2. Заданный набор символов – это символы, которые перемещают слово из одного состояния в другое. В нашем случае это суффиксы узбекского языка.

Σ={ни, лар, ган, ма, ла} W_i

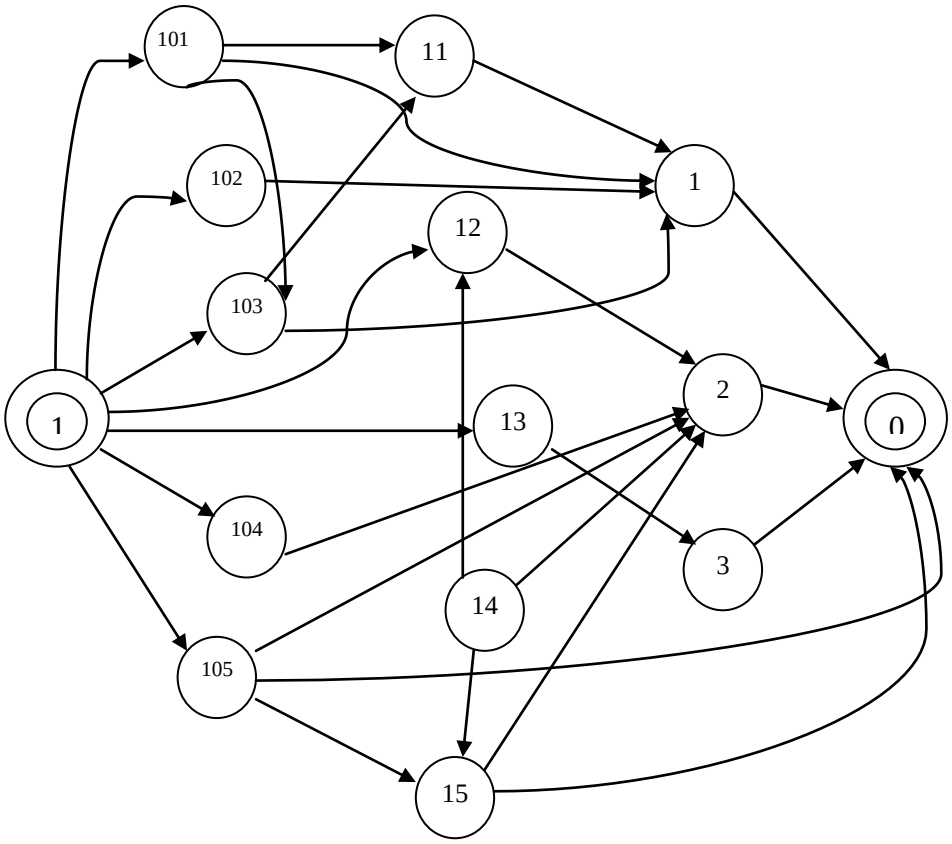


Рисунок 1. Модель определения корня слов в узбекском языке

3. Функция перехода – это функция, которая переводит слово из одного состояния в другое с использованием заданного суффикса.

q_1-ишла, q_0 – иш, добавленный суффикс – ***ла***

$q_1=\delta\ (q_0, ла\ )$ = ишла

Первоначально КА разрабатывается в форме слева направо (корень + суффикс1 + суффикс2 + ...). Чтобы запустить его справа налево нужно запустить $\boldsymbol{\delta}$ в обратном порядке, то есть $\boldsymbol{Q_0}$ = ишла , $\boldsymbol{Q_1}$-ишла

$$\mathbf{Q_1 = \delta\ (Q_0\ ,\ ла)}$$

4. Результирующий корневой набор состоит из основной части (основы) слова, при этом в настоящем процессе он имеет только один элемент.

На основе данного конечного автомата можно разработать алгоритм и программное обеспечение для определения технического корня (основы) слов в узбекском языке.

В процессе сравнения текстов между собой, описанных на разных языках, первоначально они преобразовываются в форме одного естественного языка. Как уже отмечалось, перевод терминов с помощью предлагаемого метода CLAD организован на основе словаря, который состоит из следующих этапов.

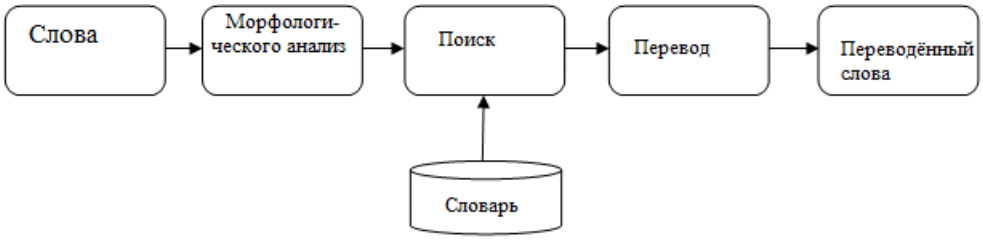


Рисунок 2. Этапы перевода слова

Ниже приведен алгоритм машинного перевода слова f с языка l_1 на язык l_2.

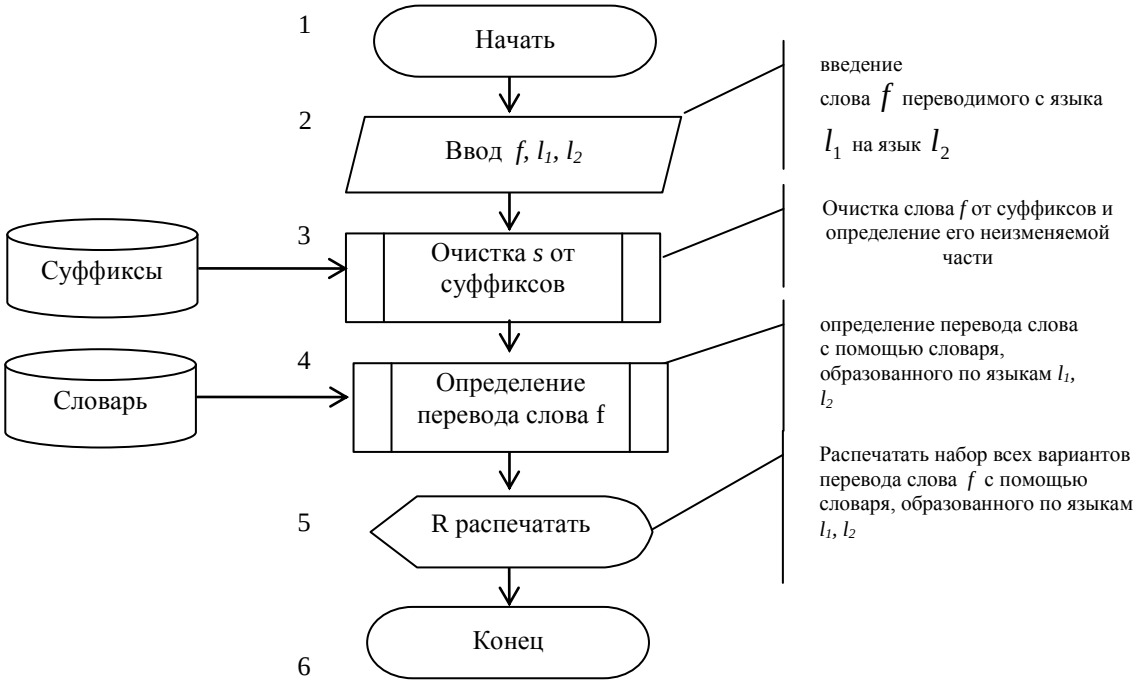


Рисунок 3. Алгоритм перевода слов на основе словаря

Как известно, процесс перевода терминов относится к направлению машинного перевода текста, и в настоящее время достигнуты большие результаты в этом отношении. Процесс перевода терминов на основе метода CLAD отличается от существующих методов следующим:

- переведенный термин не преобразовывается в предыдущую морфологическую форму;

- в процессе сравнения учитываются все формы перевода термина.

Следует отметить, что если содержание предложения не меняется, одна из произвольных синонимичных форм слова может использоваться в процессе перевода предложения. То есть здесь принимается во внимание содержание предложения, а не форма слова. А в процессе сравнения текстов важно не только содержание предложения, но и форма слова. То есть процесс перевода терминов, используемых в процессе сравнения текстов на разных языках, является более сложным, чем процесс перевода текста.

При решении этого вопроса формы перевода каждого термина подвергаются лексическому анализу. Это, в свою очередь, позволяет сравнивать тексты, принимая во внимание все синонимичные формы термина.

Рассмотрим процесс проверки сходства текстов на основе двух D и T текстов. Первоначально текст T также преобразовывается в следующую форму в соответствии с формой (1).

$$T = ((t_1, t_1^{'}, m_1), (t_2, t_2^{'}, m_2), ..., (t_p, t_p^{'}, m_p)), \quad (8)$$

где T - заданный текст, p - количество значимых терминов в тексте, t_i - общая синонимичная форма термина i в тексте T, $t_i^{'}$ - форма перевода термина t_i, m_i - количество использования термина t_i в составе текста T. Если текст T написан на естественном языке, основанного на методе CLAD, в этом случае $\forall t_i \in [t_1, t_2, .., t_p] = t_i^{'}$ то есть термины для текста T принимаются непосредственно в их собственной форме без перевода. Вместе с тем, объемом текста T является M в соответствии с формулой (2).

$$M = \sum_{i=1}^{p} m_i . \quad (9)$$

Процесс проверки сходства текстов выполняется в два вида этапов в зависимости от языка, на котором описаны тексты. В случае, если $lang(D) = lang(T)$, процесс сравнения данных для текста D осуществляется на основе набора терминов $[d_1, d_2, .., d_k]$, а для текста T - соответственно на основе $[t_1, t_2, .., t_p]$. В случае же, если $lang(D) \neq lang(T)$, процесс сравнения данных для текста D осуществляется на основе набора терминов $[d_1^{`}, d_2^{`}, .., d_k^{`}]$, а для текста T - соответственно на основе $[t_1^{'}, t_2^{'}, .., t_p^{'}]$. $lang(X)$ здесь является естественным языком, на котором написан текст X.

Вместе с тем, чтобы упростить следующие описания, введем следующие обозначения. Текст D, заданный в вышеуказанной форме (1), а также текст T, заданный в форме (8) обозначим следующим образом.

$$D = ((r_1, n_1), (r_2, n_2), ..., (r_k, n_k)) \tag{10}$$

$$T = ((q_1, m_1), (q_2, m_2), ..., (q_p, m_k)), \tag{11}$$

где r_i и q_j считаются как:

$$\forall r_i = \begin{cases} d_i, & lang(D) = lang(T) \\ d_i^{'} \end{cases}, \; R = ((r_1, n_1), ..., (r_k, n_k))$$

В псоледующих этапах именно процесс сравнения будет осуществляться на основе объектов R и Q, т.к. для D и T целесообразным считается следующее условие.

$$sim(D,T) = sim(R,Q) \tag{12}$$

Следующим шагом является создание нового набора $[x_1, x_2, ..., x_l]$ на основе взаимного пересечения наборов $[r_1, r_2, .., r_k]$ и $[q_1, q_2, .., q_p]$. Если данный набор не содержит ни одного элемента, то есть $[r_1, r_2, .., r_k] \cap [q_1, q_2, .., q_p] = \theta$ если будет установлено, что степень сходства текстов равна 0, мы завершаем процесс сравнения на этом шаге, в противном случае процесс сравнения будет продолжен.

$$[r_1, r_2, .., r_k] \cap [q_1, q_2, .., q_p] = [x_1, x_2, .., x_l] \tag{13}$$

На основе полученного $\forall x_i$ и количества их использования соответственно в составе объектов R и Q образуется следующий объект X.

$$X = ((x_1, n_1, m_1), (x_2, n_2, m_2), ..., (x_l, n_l, m_l)), \tag{14}$$

где n_i - количество повторений термина x_i в тексте D, т.е. объем, m_i - количество повторений термина x_i в тексте T, т.е. объем. С точки зрения заданного x_i соответствие текста D тексту T определяется как $belong(x_i, D, T)$ и выражается следующим образом.

$$belong(x_i, D, T) = \frac{n_i \cdot m_i}{N^2} \tag{15}$$

Соответственно, с точки зрения заданного x_i соответствие текста T тексту D определяется как $belong(x_i, T, D)$ и выражается следующим образом.

$$belong(x_i, T, D) = \frac{n_i \cdot m_i}{M^2} \tag{16}$$

Исходя их этого, степень соответствия текста D тексту T считается следующим.

$$belong(D,T)=\sum_{i=1}^{l}\frac{n_i\cdot m_i}{N^2} \quad (17)$$

Соответственно, степень соответствия текста T тексту D считается следующим.

$$belong(T,D)=\sum_{i=1}^{l}\frac{n_i\cdot m_i}{M^2} \quad (18)$$

Теперь рассмотрим результаты (17) и (18) более подробно. Как известно, иногда один текст может быть содержимым другого текста. Например, если текст D - представляет собой какую-нибудь статью или историю, он, в свою очередь, может быть содержимым сборника статей или произведения - T . В данном случае естественно, что степень сходства текста T с текстом D будет низкой, напротив текст D считается на 100% схожим с текстом T **.** Именно в этих случаях правильный результат получается на основе (17) и (18). Кроме того, степень сходства данных текстов определяется на основе средней арифметики результатов этой функции.

$$sim(D,T)=sim(T,D)=\frac{belong(D,T)+belong(T,D)}{2}$$

Вышеуказанная функция выражает сходство текстов на основе метода CLAD. Как уже отмечалось, метод CLAD не только определяет схожесть текстов между собой, но также и их принадлежность друг другу.

В третьей главе **«Модели и алгоритмы поиска полнотекстовой информации на основе научно-образовательных и архивных данных в корпоративной и глобальной сети»** рассматриваются вопросы выявления сходства текста с стекстами в Интернет сети, автоматической классификации текстов по рубрике научно-образовательных данных, а также вопросы, которые могут быть решены на основе предлагаемого метода CLAD.

Процесс поиска аналогичного текста в Интернете состоит из следующих частей:

- конвертация текста, описанного в различных форматах, в обычную форму текста;
- разделение текста на термины;
- определение списка подходящих Web-страниц в сети Интернет на основе терминов;
- сравнение документов и, на основе уровня сходства, сортировка Web-страниц.

Такие системы, как Yahoo, Google, Yandex XML могут быть использованы для идентификации Web-страниц на основе набора слов. Основное отличие предлагаемого алгоритма от существующих алгоритмов состоит в том, что в процессе сравнения текстов также учитываются синонимичные формы терминов в тексте. Ниже приведен алгоритм сравнения заданного текста D с текстом T из Web-страницы на основе синонимичной формы слова.

1. Определяются все термины $[d_{11}, d_{21},.., d_{r1}]$, не относящиеся к набору $[x_1, x_2,..x_k]$ текста D, где r - количество терминов текста D, не относящиеся к набору $[x_1, x_2,..x_k]$.

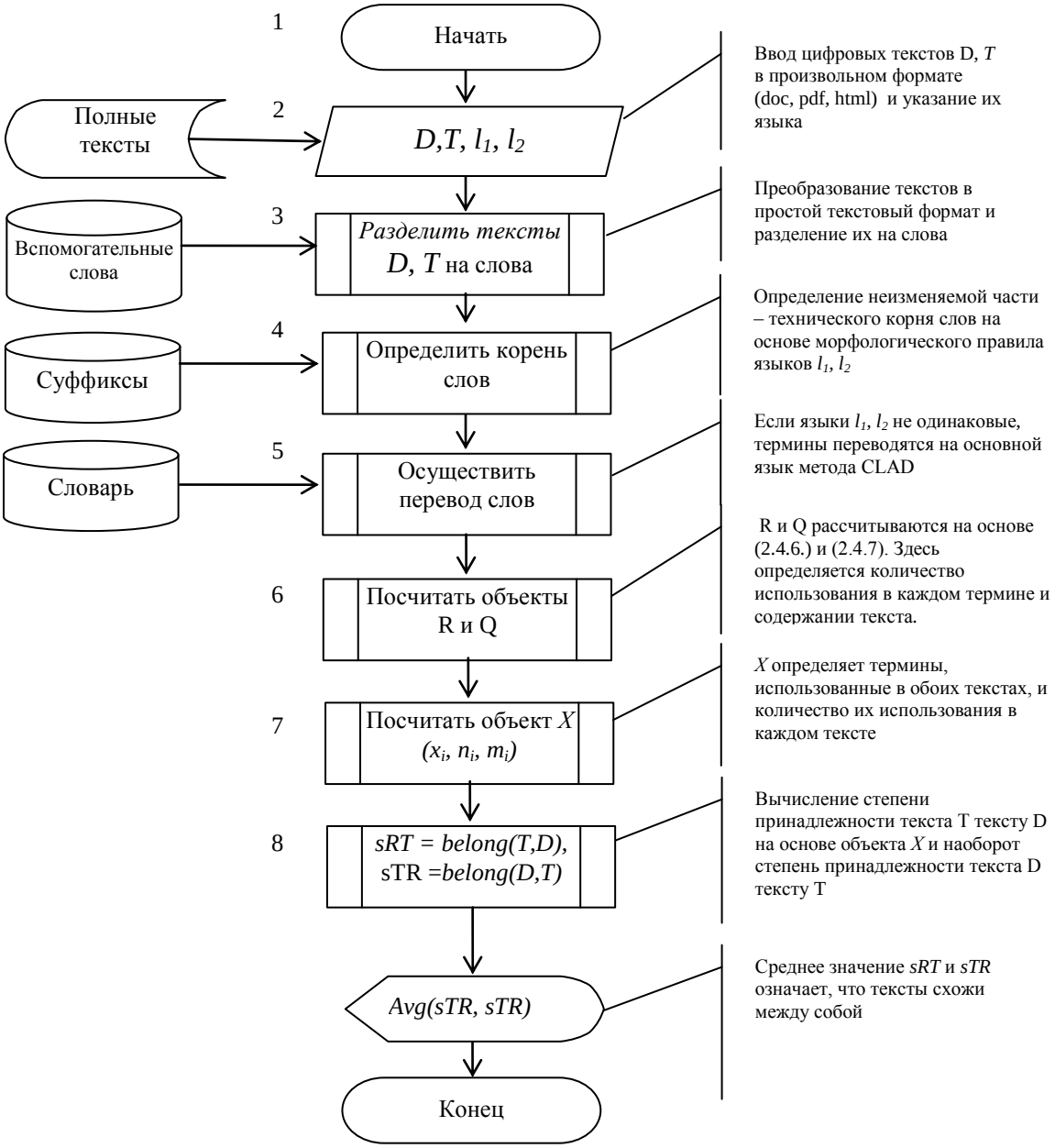


Рисунок 4. Алгоритм проверки сходства текстов

2. Определяются синонимичные формы $[d_{i2}, d_{i3},.., d_{it}]$ термина d_{i1}. Если ни одна синонимичная форма не найдена, переходим к шагу 4.

3. Осуществляется поиск терминов $[d_{i2}, d_{i3},.., d_{it}]$ из набора $[t_1, t_2,..., t_s]$.

Первый найденный термин d_{ij} на основе количества его использования в тексте T, добавляется к объекту X в форме нового $(x_{k+1}, n_{k+1}, m_{k+1})$ элемента. n_{k+1}- количество использования термина d_{i1} в тексте D, m_{k+1}- количество использования термина d_{ij} в тексте T.

4. Выполняются шаги 2 и 3 для следующего термина. После выполнения шагов 2 и 3 для всех терминов рабочий процесс алгоритма завершается.

На основании полученного объекта X определяется сходство текстов D и T. Данный процесс осуществляется на основе алгоритма, приведенного в рисунке 4.

Определение сходства заданного текста с текстами в сети Интернет позволяет не только идентифицировать плагиатные тексты, но и автоматически создавать электронный каталог на основе реферативного текста с использованием интернет-ресурсов. Это, в свою очередь, обеспечивает автоматическую идентификацию ресурсов, необходимых для аналитической части процесса научной исследовательской работы. Исследователь может подготовить текст, состоящий из двух или трех страниц, в необходимой для него сфере. На основании этого текста набор подходящих ресурсов из сети Интернет определяется с использованием программного средства, созданного на основе вышеупомянутого алгоритма.

Процесс определения научно-образовательной информационной рубрики текстов основан на нейронных сетях. Метод обучения сети осуществлен с помощью преподавателя, функция активации осуществлена на основе сигмоида, входной сигнал – на основе кода термина в словаре, архитектура состоит из 3 слоев, выходное значение осуществлено на основе группового кода по принадлежности текста, оценка ошибки осуществлена на основе метода среднего квадрата. Объем термина в тексте определен на основе формулы TF-IDF. В нижеследующем 5-рисунке приведена архитектура данной нейронной сети.

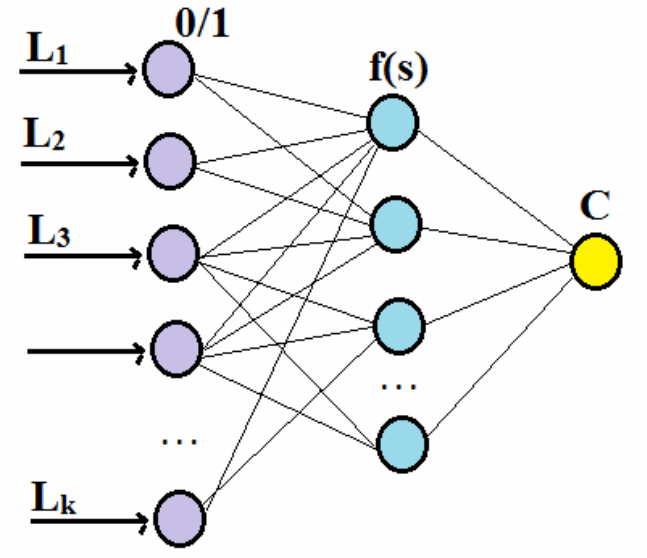


Рисунок 5. Архитектура нейронной сети, разделяющей тексты на рубрики

В отличие от доступных в настоящее время методов автоматической классификации текстов, первоначально учитываются также синонимичные формы каждого термина, то есть все синонимичные формы термина также учитываются в процессе определения объема термина в тексте. Данное

обстоятельство приводит к дальнейшему совершенствованию процесса определения рубрики электронных журналов и сборников.

Приводятся вопросы, которые необходимо решить на основе предлагаемого метода CLAD.

Таблица 3 Практическая значимость метода CLAD

№	Сфера применения	Решаемые задачи
1.	В системах электронного обмена документами	- разделение (разбивка) текстов по сферам; - поиск информации в текстах; - хранение текстов на основе распределения.
2.	В электронной библиотеке и автоматизированных библиотечных системах	- Создание базы электронных книг и поиск информации в ней; - автоматическая идентификация соответствующей научно-образовательной информационной рубрики текста; - автоматическое формирование набора ключевых слов, относящихся к тексту.
3.	В области издательского дела	- Формирование базы данных опубликованных текстов, на основе которой проверяется сходство новых текстов; - Автоматическое определение сферы, к которой принадлежит текст.
4.	В электронных архивах	- формирование архива электронных текстов; - автоматическое определение вида новой информации, поступившей в архив; - Поиск информации из архивных данных.
5.	В системах антиплагиата	Определение следующих видов сходства с учетом синонимов в тексте: - копирование и вставление текста (Copy&Paste); - плагиат, скрытый на основе синонимов; - Обнаружение плагиатного перевода.

В четвертой главе **«Структура информационно-поисковой системы на основе больших данных (big data)»** описывается структура баз данных, сформированных на основе принципа NoSQL (Not Only SQL - не только SQL-SQL) программного средства, разработанного на основе метода CLAD. В частности, в процессе индексации многоязычных полных текстов использовалась нереляционная база данных. Справки (синонимичные формы слова, словарь и другие данные), необходимые для рабочего процесса системы, хранятся в реляционной базе данных. В системе, разработанной на основе исследований, использовалась система MySQL в качестве системы управления реляционными базами данных, а система Apache Lucene использовалась в качестве системы управления нереляционными базами данных, файловый сервер использовался для полнотекстовых документов. В рисунке-6 ниже приведена структура базы данных системы.

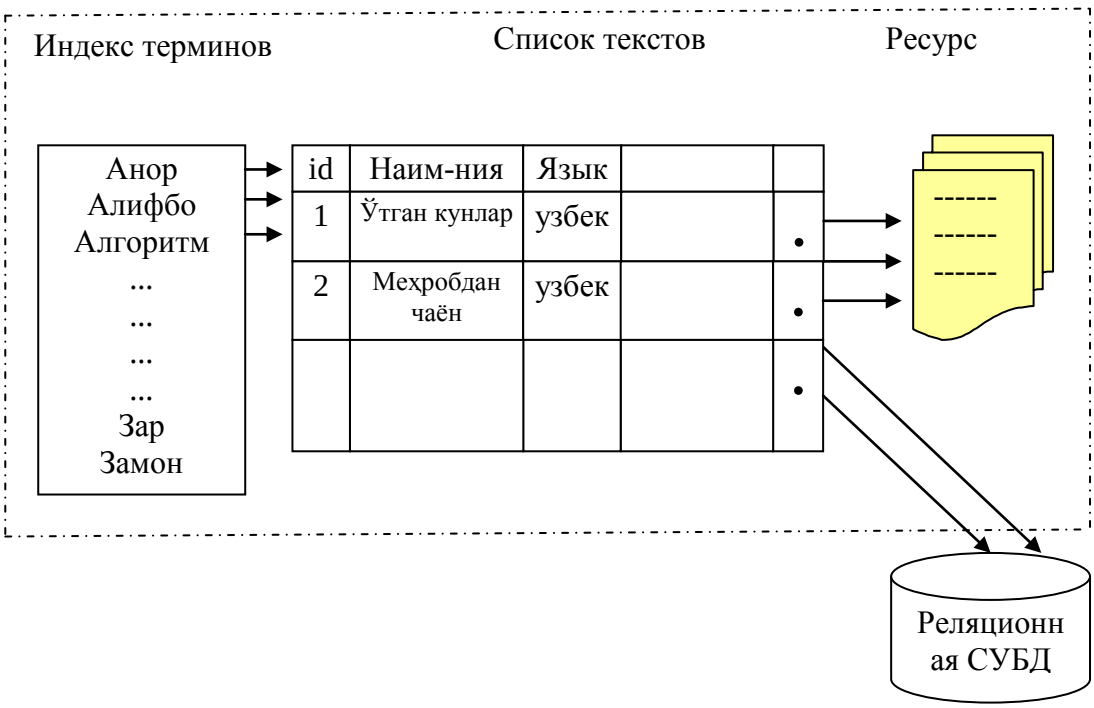


Рисунок 6. Процесс хранения текстов в нереляционной базе данных

Реляционная база данных состоит из базы данных синонимичных слов, набора словарей, используемых в процессе перевода, вспомогательных слов в естественном языке и других данных. В рисунке-7 ниже показана структура этой базы данных.

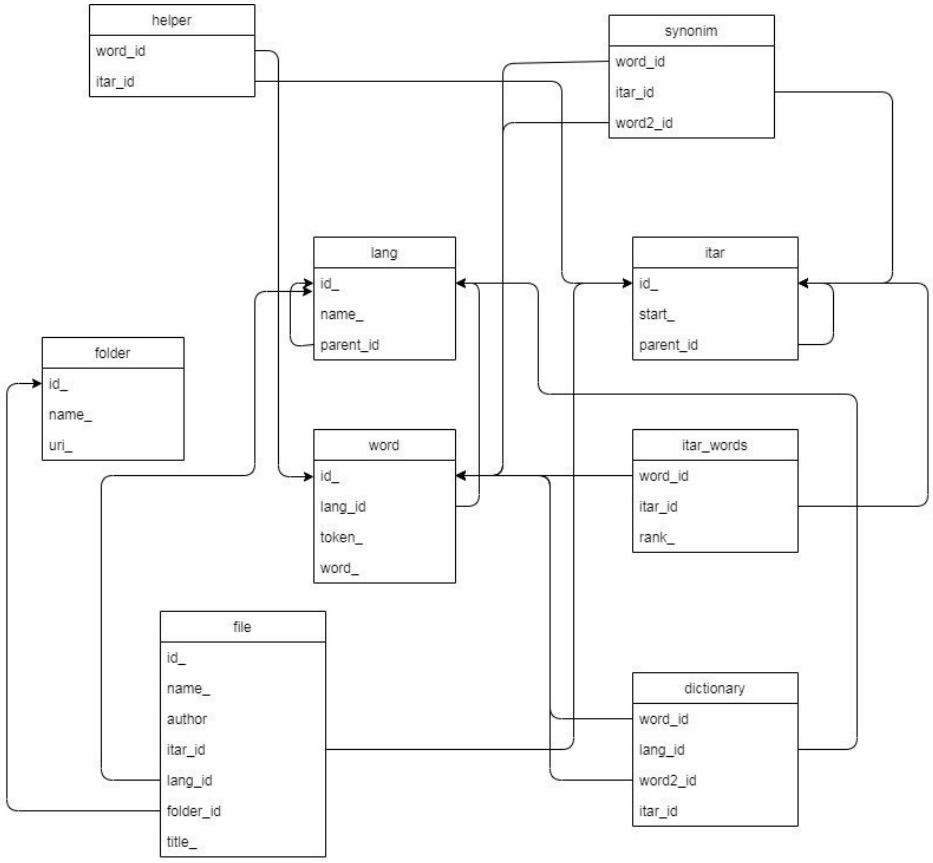


Рисунок 7. Структура реляционной БД системы

Система способна осуществлять поиск информации из базы данных автоматизированных информационно-библиотечных систем на основе протоколов Z39.50 и HTTP, а также копировать полные тексты в свою базу данных. Вместе с тем, введенный в систему разного рода полный текст хранится на специальных серверах FTP (File Transfer Protocol – протокол трансфера файла) без какой-либо обработки. В вышеуказанном рисунке этот сервер указан в качестве ресурсов.

В пятой главе **«Система удаленного поиска полнотекстовых библиотечных и архивных данных в корпоративных и глобальных сетях на основе метода CLAD»** описывается функциональная структура системы «jComporator», разработанная на основе исследований. Представлены функции обмена электронными документами с другими системами в процессе документооборотного потока, приводятся процессы поиска информации из полных текстов в корпоративной информационно-библиотечной сети. В то же время, рассматривается анализ процесса поиска информации из Интернета и многоязычных баз данных на основе этой системы.

Данная система предназначена для работы не только с локальными базами данных, но и на основе корпоративных электронных библиотек. Как известно, библиографические записи в электронных библиотеках хранятся в формате MARC или Dublin Core. Эти форматы имеют строго определенную структуру, в которых файл, связанный с библиографической записью, данные о его авторе, естественном языке, на котором написан текст, хранятся в машиночитаемой форме. Это позволяет использовать тексты в электронной библиотеке без дополнительных инструментов. Для этого в системе разработана возможность индексации текстов, связанных с библиографической записью в структуре электронного каталога.

Автоматизированный процесс обмена информацией с информационно-библиотечными и архивными системами осуществляется в следующих формах:

- через протокол Z39.50;
- путем подключения к базе данных электронной библиотеки на основе протокола JDBC.

На основании перечисленных форм система может связаться с библиографическими записями электронной библиотеки. Однако также должно быть обеспечено право обращаться к полному тексту, связанному с данной библиографической записью.

Для обмена библиографическими записями в автоматизированных информационно-библиотечных системах используется специальный протокол Z39.50, который позволяет различным системам обмениваться информацией на основе этого протокола независимо от используемой базы данных и применяемого формата MARC. Если автоматизированные информационно-библиотечные системы могут работать на основе протокола Z39.50, то система jComporator может индексировать полные тексты,

связанные с соответствующей библиографической записью в базе данных электронной библиотеки. Это обеспечивает решение таких вопросов, как поиск информации на основе полных текстов в электронных библиотеках и архивных системах, выявление их сходства.

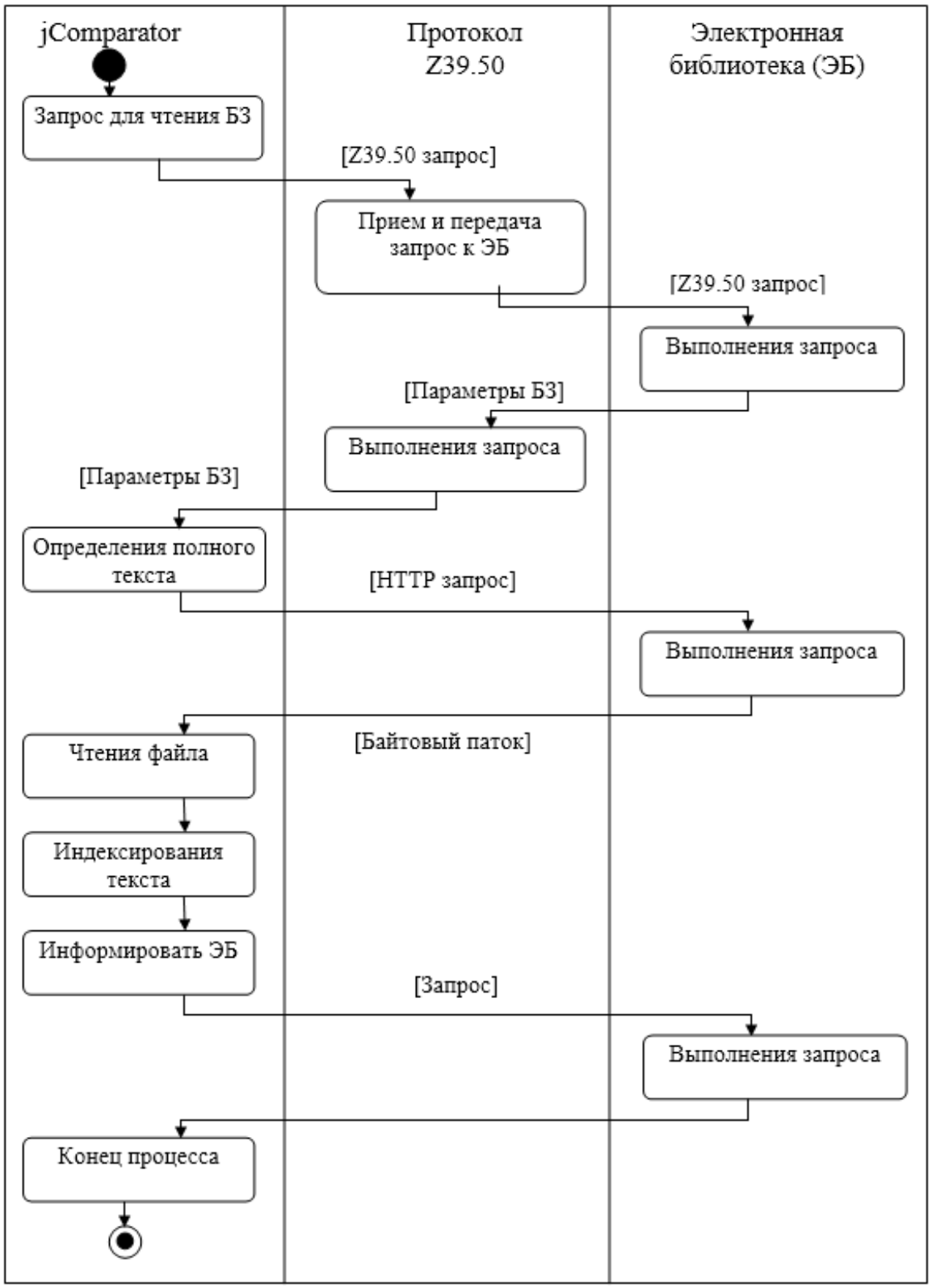


Рисунок 8. Работа с электронной библиотекой на основе протокола Z39.50

На рисунке-8 приведен процесс обмена данными jComparator на основе протокола Z39.50 с электронными библиотечными системами данной системы.

Для процесса анализа алгоритма определения сходства текста с текстом из сети Интернет был взят текст **file1.html**, который был приведен в следующих трех формах:

- слова в тексте были заменены синонимичными формами и текст был назван file2.html;
- в текст введены фразы и омонимы без изменения значения большинства предложений, и текст был назван file3.html;
- осуществлена замена мест большинства слов в предложениях и большинства предложений в тексте, и полученный текст был назван file4.html.

В ходе анализа была выбрана группа из 10 специалистов, которых попросили определить степень сходства сгенерированных текстов с исходным текстом file1.html. Вместе с тем, была просьба отметить сходство текстов со значениями между [0..10]. Также, на основе программного средства, основанного на системах Антиплагиат, Plagiarism и алгоритме SCAM, было выявлено, что эти тексты были схожи. На следующем рисунке показаны полученные в процессе анализа результаты в виде диаграммы.

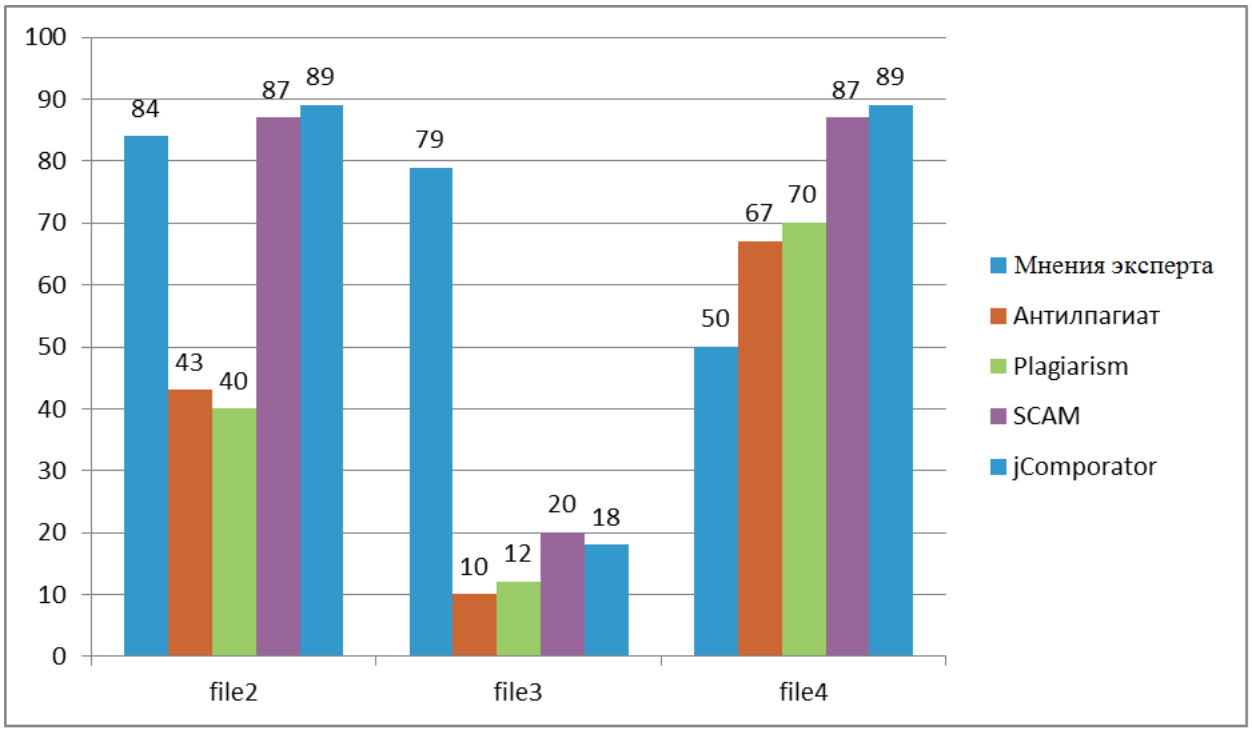


Рисунок 9. Анализ процесса проверки сходства слов с учетом их синонимичных форм.

Анализа, показывает, что алгоритм для определения сходства текстов, разработанных на основе исследования, достиг на 2% более хорошего результата, чем существующие алгоритмы. Однако file3.html показывает, что процесс анализа использования в тексте омонимичных слов и фраз является

неудовлетворительным. Это, в свою очередь, требует дополнительных исследований данного процесса.

Программная система, разработанная на основе исследований, основана на модульном принципе и состоит из следующих частей:

- модуль, направленный на морфологический и лексический анализ полного текста и перевода терминов на основе словаря;
- модуль, основанный на хранении полного текста в базе данных, сформированных по принципу Big Data, и поиске информации в ней;
- модуль, основанный на протоколах Z39.50 и HTTP, позволяющий работать с другими системами;
- модуль определения сходства набора терминов;
- модуль, управляющий данными системы и базами данных.

ЗАКЛЮЧЕНИЕ

По результатам исследования, проведенного в рамках диссертационной работы, посвященной теме: «Методы и алгоритмы поиска схожих текстов в многоязычных распределенных информационно-библиотечных и архивных базах данных», были сделаны следующие выводы:

1. Проведен анализ методов, алгоритмов и программных средств проверки сходства документов в полнотекстовой базе данных, и обосновано, что разработка моделей, алгоритмов и программных систем для определения сходства текстов, написанных на основе различных естественных языков, является одним из вопросов, ожидающих свое решение.
2. Для процесса определения сходства многоязычных текстов разработаны модели и алгоритмы, позволяющие разделять тексты на слова, определять неизменяемую часть слов, осуществлять машинный перевод терминов, определять их общие синонимичные формы. Эти модели и алгоритмы позволяют хранить полные тексты в различных форматах в базах данных и искать информацию в них.
3. Разработан алгоритм определения ключевых слов в тексте на основе количества использования слов в тексте. Этот алгоритм дает возможность автоматизировать процесс выявления ключевых слов при каталогизации полного текста.
4. На основе набора слов в многоязычных текстах разработаны модель и алгоритм определения их сходства и выделения схожих частей. Данный процесс обеспечил обнаружение многоязычных и скрытых плагиатных текстов.
5. Разработаны модель и алгоритм определения сходства научно-образовательных и архивных данных с учетом синонимичных форм терминов в сети Интернет. В результате был налажен процесс автоматического формирования электронного каталога на основе реферативных данных.

6. Разработаны модель и алгоритмы определения научно-образовательной информационной рубрики многоязычных текстов. В результате этого этап определения научно-образовательной информационной рубрики в процессе связывания текстов с библиографическим описанием, имеющимся в электронном каталоге, полностью автоматизирован.

7. На основе подхода NoSQL разработана структура реляционных и нереляционных баз данных, которая обеспечивает хранение полных текстов и поиска информации в них, а также эффективно обеспечивает хранение больших объемов текста в распределенных базах данных.

8. Разработаны модели и алгоритмы поиска информации в полных текстах, содержащихся в информационно-библиотечных системах, автоматизированных на основе протоколов Z39.50 и HTTP. В результате был реализована функция поиска информации не только в библиографических записях, но и в связанных с ними полных текстах.

9. Разработан метод CLAD для идентификации схожих текстов в многоязычных распределенных информационно-библиотечных и архивных базах данных. На основе данного метода организован процесс определения сходства многоязычных текстов.

10. Модели, алгоритмы, метод CLAD и программные средства, разработанные в ходе диссертационного исследования, в той или иной мере были внедрены в информационную систему Высшей аттестационной комиссии при Кабинете Министров Республики Узбекистан, биллинговые системы АК «Узбектелеком», Кушкупырское районное отделение народного образования Хорезмской области. При применении к биллинговым системам было достигнуто ускорение рабочего процесса на 12%. Акты внедрения прилагаются в диссертационную работу.

SCIENTIFIC COUNCIL AWARDING SCIENTIFIC DEGREES DSc.13/30.12.2019.T.07.02 AT TASHKENT UNIVERSITY OF INFORMATION TECHNOLOGIES

TASHKENT UNIVERSITY OF INFORMATION TECHOLOGIES

ATADJANOV JASURBEK ABDUSHARIBOVICH

METHOD AND ALGORITHMS FOR SEARCHING FOR ANALOGUES OF TEXTS FROM MULTILINGUAL DISTRIBUTED INFORMATION LIBRARY AND ARCHIVE DATABASES

05.01.09 – Documentary. Documentation. Archival studies

ABSTRACT OF THE DOCTORAL (DSc) DISSERTATION OF TECHNICAL SCIENCES

Tashkent – 2020

The theme of doctoral (DSc) dissertation was registered at the Supreme Attestation Commission at the Cabinet of Ministers of the Republic of Uzbekistan under number B2020.2.DSc/T169.

The dissertation has been prepared at Tashkent University of Information Techologies.

The abstract of the dissertation is posted in three languages (Uzbek, Russian, English (resume)) on the Scientific Council website www.tuit.uz and on the website of «ZiyoNet» Information and Educational portal www.ziyonet.uz.

Scientific adviser:	**Rakhmatullaev Marat Alimovich** Doctor of Technical Sciences, Professor
Official opponents:	**Umarov Abdusalom Odilovich** Doctor of Social Sciences, Professor
	Nuraliev Faxriddin Murodillaevich Doctor of Technical Sciences, Docent
	Kuchkarov Taxir Safarovich Doctor of Economic Sciences, Professor
Leading organization:	Tashkent State Technical University named after Islam Karimov

The defense of the dissertation will be held «28» July 2020 at 10⁰⁰ at the meeting of Scientific council No. DSc.13/30.12.2019.T.07.02 at Tashkent University of Information Technologies (Address: 100084, Tashkent city, Amir Temur street, 108. Ph.: (+99871) 238-64-43; fax: (+99871) 238-65-52, e-mail: tuit@tuit.uz).

The dissertation can be reviewed at the Information Resource Centre of Tashkent University of Information Technologies (is registered under No. 156). (Address: 100084, Tashkent city, Amir Temur street, 108. Ph.: (+99871) 238-64-43; fax: (+99871) 238-65-52, e-mail: tuit@tuit.uz).

The abstract of dissertation was sent out on «16» July 2020 y.
(mailing protocol No. 3 on «15» July 2020 y.).

I.Kh.Siddikov
Chairman of the Scientific Council awarding scientific degrees, Doctor of Technical Sciences, Professor

H.E.Khujamatov
Scientific Secretary of the Scientific Council awarding scientific degrees, PhD on Technical Sciences

T.S. Kuchkarov
Chairman of the Acadenic Seminar at the Scientific Council awarding scientific degrees, Doctor of Economic Sciences, Professor

INTRODUCTION (abstract of the dissertation of doctor of science (DSc))

The aim of the research work. To improve the information support in scientific and educational activities, it is necessary to develop methods, algorithms and software for information retrieval from multilingual full-text libraries and archival databases.

The object of the research work are information-library and archival databases, as well as the process of searching for full texts.

The scientific novelty of the research work:

information retrieval models and algorithms based on multilingual full-text data in information-library and archival systems;

developed models and algorithms for morphological analysis of words in the text and the definition of a set of meaningful words;

a model and algorithm have been developed that allow the detection of hidden plagiarism based on the use of synonymous forms of words;

an algorithm was developed to determine the column of scientific and educational information using a machine, taking into account the synonymous forms of words;

a CLAD (Cross Language Analog Detector) method has been developed to detect the similarity of texts in multilingual archive and library databases.

Implementation of research results. Based on the results of methods and algorithms for searching for similar texts from multilingual distributed information-library and archival databases:

a model, algorithm for detecting hidden plagiarism based on the use of synonymous forms of words and the clad method for determining the similarity of texts in multilingual archive and library databases have been introduced as a software module in the information system of the higher attestation commission under the cabinet of ministers. ministry reference no. 33-8 / 9210 of 28 december 2019). as a result of the research, it is possible to create an electronic library based on scientific and educational data, as well as to identify similarities of texts in uzbek, russian and english;

models and algorithms for morphological analysis of words in the text and the definition of semantic words, as well as algorithms for determining the position of the text by machine were applied to KASR and tBilling systems for wired telephone subscribers of Uzbektelecom JSC (Ministry of Information Technologies and Communications October 18, 2019 Reference No. 33-8 / 7394). As a result, the process of uploading CDR files from telephone exchanges to the billing system is fully automated. The reintroduction of conversations into the billing system was prevented, and an average of 12% of working time per day was saved for employees working in this process, and an average of 36 mln. UZS reduction of service cost;

based on the model and algorithm of information retrieval on the basis of multilingual full-text information in information-library and archival systems, a

centralized electronic library was developed for the Department of Public Education of Koshkopir district of Khorezm region and 52 secondary schools (Ministry of Information Technologies and Communications Reference No. -8 / 9210). The process of searching for information in full-text files in this e-library has reduced the time spent searching for the required information by 20-30 times.

The structure of the dissertation. The dissertation consists of an Introduction, five Chapters, Conclusion, References and Appendices.

The volume of the thesis is 166 pages.

ЭЪЛОН ҚИЛИНГАН ИШЛАР РЎЙХАТИ
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
LIST OF PUBLISHED WORKS

I бўлим (I часть; I part)

1. Atadjanov J.A. - Models of Morphological Analysis of Uzbek Words // Кибернетика и программирование. - 2016. - № 6. - С. 70 - 73. DOI: 10.7256/2306-4196.2016.6.20945. (05.00.00; №45).

2. Атаджанов Ж.А., Атаджанов Б.А. Ахборот-ресурс марказларининг автоматлаштирилган тизимида тўлиқ матнли бибилографик ёзувлардан ахборот қидириш // «ТАТУ хабарлари» ilmiy-texnika va axborot-tahliliy jurnali. №3(43) сон 2017 йил, 35-40 бетлар. (05.00.00; №10).

3. Атаджанов Ж.А. Ахборот-кутубхона ресурслари орасидан ўхшаш матнларни аниқлаш алгоритми // «ТАТУ хабарлари» ilmiy-texnika va axborot-tahliliy jurnali. №4(44) сон 2017 йил, 8-14 бетлар. (05.00.00; №10).

4. Атаджанов Ж.А. Ахборот-ресурс марказларининг автоматлаштирилган тизими: янги имкониятлар // «ТАТУ хабарлари» ilmiy-texnika va axborot-tahliliy jurnali. №4(44) сон 2017 йил, 24-29 бетлар. (05.00.00; №10).

5. Atadjanov J. The analysis of algorithms outlined for finding words roots // «Electronic journal of actual problems of modern science, education and training». Urgench., №1.- 2019. – P. 558-566. – ISSN 2181-9750. (05.00.00; №26).

6. Атаджанов Ж.А. Матнларни сўзларга ажратувчи тизимлар таҳлили // Хоразм Маъмун академияси ахборотномаси. №1 сон 2019 йил, 92-94 бетлар. (10.00.00; №21).

7. Атаджанов Ж.А. Матнларни ўҳшашликка текшириш усуллари ва дастурий воситалари таҳлили // Хоразм Маъмун академияси ахборотномаси. №1 сон 2019 йил, 94-97 бетлар. (10.00.00; №21).

8. Атаджанов Ж.А. Электрон кутубхона ва архивларда тўлиқ матнли ахборот қидириш жараёнини ташкил қилиш // «Kutubxona.uz» илмий-услубий журнали. №4(44), 2019 йил, 32-34 бетлар. (05.00.00; №30).

9. Atadjanov J. Document Plagiarism Detection On The Internet With Accounting Synonym Forms Of Words // International Journal of Scientific & Technology Research. ISSN: 2277-8616, Volume-8, Issue-11. December 2019 Edition.- P. 1517-1519.

10. Atajanov J., Atadjanov B., Abdulazizov Sh., Makhmanov O. Method for Measuring the Semantic-similarity of Textual Document and Web-pages // International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075. Volume-9, Issue-2. December 2019.- P. 1601-1606. DOI: 10.35940/ijiteee.B7314.129219.

11. Atajanov J., Atadjanov B. Cross-Language Plagiarism Detection Based on CLAD Method // International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075. Volume-9, Issue-2. December 2019.- P. 4903-4909. DOI: 10.35940/ijiteee.B7404.129219.

12. Rakhmatullayev M.A., Atadjanov J.A., Gandjaeva L., Matyakubova Y., Allabergenova K. Cross-Language Plagiarism Detection Steps // International Journal of Scientific & Technology Research ISSN: 2277-8616, Volume-9 Issue-01, January 2020.- P. 3303-3308.

13. Atadjanov J.A., Atadjanov B.A. Cross Language Analog Detecting Process // International Journal of Advanced Research in Science, Engineering and Technology ISSN: 2350-0328, Volume-7 Issue-6, June 2020.- P. 14005-14012, (05.00.00; №8).

14. Atadjanov J.A. Format for Bibliographic Information Exchange // International Journal of Advanced Research in Science, Engineering and Technology ISSN: 2350-0328, Volume-7 Issue-6, June 2020.- P. 14054-14057, (05.00.00; №8).

II бўлим (II часть; II part)

15. Каримов У.Ф., Атаджанов Ж.А. КАРМАТ тизимида даврий нашрлар (газета ва журналлар) электрон каталогини шакллантириш технологияси // «Электрон кутубхона тармоқларида илмий-таълимий ахборотлар яратиш ва улардан фойдаланиш технологиялари» тўплами. Алишер Навоий номидаги Ўзбекистан Миллий кутубхонаси нашриёти. Тошкент-2011.- Б. 56-68.

16. Атаджанов Ж.А. Ахборот-ресурс марказларининг автоматлаштирилган тизимида библиографик маълумотларнинг хавфсизлигини таъминланиши // «Электрон кутубхона тармоқларида илмий-таълимий ахборотлар яратиш ва улардан фойдаланиш технологиялари» тўплами. ELINE – PRESS. Тошкент-2013.- Б.27-30.

17. Каримов У.Ф., Атаджанов Ж.А., Каримов Ў.У. Ахборот-ресурс марказларининг автоматлаштирилган тизими (ARMAT) // «Электрон кутубхона тармоқларида илмий-таълимий ахборотлар яратиш ва улардан фойдаланиш технологиялари» тўплами. ELINE – PRESS. Тошкент-2013.- Б.37-51.

18. Атаджанов Ж.А., Исломова Х.Э. Таълимда электрон кутубхоналар // «Билимлар жамияти сари: ўзгараётган дунёда кутубхоначиларнинг янги роли» халқаро анжуман материаллари. Тошкент-2011.- Б. 88-92.

19. Atadjanov J.A. Antiplagiat dastur yaratishda o‘zbekcha so‘zlarning morfologik tahlili // «infoCOM.UZ» O`zbekiston axborot kommunikatsiya texnologiyalari. №5(185), 2017 й. -Б.52-54.

20. Атаджанов Ж.А. Интернет тармоғида ўхшаш матнларни излаш модели ва алгоритми // «Инновацион ривожланиш учун илмий ахборот ресурслар» 11-халқаро семинар ва презентациялар тўплами. Тошкент-2019 й.- Б. 79-83.

21. Атаджанов Ж.А. Кўп тиллик антиплагиат тизими // «Хорижий тилларни ўргатишнинг инновацион технологиялари» халқаро илмий-амалий анжуман материаллари. Самарқанд-2019. - Б.44-46.

22. Атаджанов Ж.А. Таълим жараёнларига jcomporator тизимини татбиқ қилиш // «Амалий математика ва информацион технологияларнинг долзарб

муаммолари» халқаро анжуман тезислари тўплами. Тошкент-2019. - Б. 224-225.

23. Атаджанов Ж.А. Ўзбек тилидаги матнларни морфологик таҳлил қилиш усули // «Замонавий филология тараққиётида инновацияларнинг роли» халқаро илмий конференцияси. Тошкент-2019.- Б.387-389.

24. Atadjanov J.A. Ilmiy-ta`limiy axborotlar rubrikasini aniqlash algoritmi // «infoCOM.UZ» O`zbekiston axborot kommunikatsiya texnologiyalari. №9.- 2019 й.

25. Махманов О.К., Таджиходжаев З.А., Рахматуллаев М.А., Атаджанов Ж.А., Хакимов З.Т. Component // O'zbekiston Respublikasi Adliya vazirligi huzuridagi intellektual mulk agentligi. EHM uchun yaratilgan dasturning rasmiy ro'yxatdan o'tkazilganligi to'g'risidagi guvohnoma. № DGU 03666. Toshkent, 21.04.2016.

26. Атаджанов Ж.А., Муртазин И.В. «Yagona konvergent avtomatlashtirilgan hisob tizimi» ЭҲМ учун дастур // O'zbekiston Respublikasi Adliya vazirligi huzuridagi intellektual mulk agentligi. EHM uchun yaratilgan dasturning rasmiy ro'yxatdan o'tkazilganligi to'g'risidagi guvohnoma. № DGU 05707. Toshkent, 18.10.2018.

Автореферат «Муҳаммад ал-Хоразмий авлодлари» илмий журнали таҳририятида ўзбек, рус ва инглиз тилларидаги матнларининг мослиги текширилди (14.07.2020).

Бичими 60x84$^{1}/_{16}$. Рақамли босма усули. Times гарнитураси.
Шартли босма табоқи: 3. Адади 100. Буюртма № 85.

Гувохнома reestr № 10-3719
«Тошкент кимё технология институти» босмахонасида чоп этилган.
Босмахона манзили: 100011, Тошкент ш., Навоий кўчаси, 32-уй.