АҢДАТПА

6D070400 – «Есептеу техникасы және бағдарламалық қамтамасыз ету» мамандығы бойынша «философия докторы» (Ph.D) дәрежесі бойынша диссертация

Якунин Кирилл Олегович

“Табиғи тілдерді өңдеу әдістеріне негізделген бұқаралық ақпарат құралдарын жинау, талдау және сыныптау модельдері мен әдістерін әзірлеу”

Диссертациялық жұмыста медиа-мониторинг жөніндегі тапсырмаларын шешуге арналған табиғи тілдегі мәтіндерді жіктеу әдістері ұсынылды. Ұсынылған әдістер мәтін корпусының жасырын құрылымдары туралы білімді алу үшін үлкен көлемді гетерогенді деректерді (Big Data) қолдану идеясына негізделген, бұл мәтіндердің тиімді векторлық көрінісін құруға мүмкіндік береді, жіктеу мәселелерін шешеді. Сонымен қатар, ұсынылған әдістердің есептеу күрделілігі төмен және қолмен таңбалаудың минималды мөлшерімен ерікті бағалау критерийі бойынша немесе кейбір жағдайларда автоматты түрде (таңбасыз) мәтін жіктеуіштерін құруға мүмкіндік береді.

Жұмыс мақсаты.

Жұмыстың мақсаты таратылған ақпараттық жүйе шеңберінде бұқаралық ақпарат құралдарнан және әлеуметтік желілерден мәтіндік ақпаратты автоматты түрде көп критериалды бағалаудың үлгілері мен әдістерін әзірлеу болып табылады.

Зерттеу мақсаттары.

Қойылған мақсатқа жету үшін жұмыста келесі міндеттер қойылды:

- мәтіндер үшін тиімді (өнімділік және талап етілетін сараптамалық баға бойынша) жіктеу үлгілерін құру мүмкіндігін шектейтін негізгі факторларды анықтау;
- таңбаланбаған деректер негізінде мәтіндерді тиімді векторлау әдісін әзірлеу (Үлкен деректер-Big Data);
- құжаттардың жасырын параметрлерін тану үшін объективті айқын көрсеткіштерге негізделген жіктеу моделін өздігінен үйрену әдісін әзірлеу;
- құжаттар мен бұқаралық ақпарат құралдары көздерін көп критериалды бағалау әдістемесін әзірлеу;
- бұқаралық ақпарат құралдары мен әлеуметтік желілердегі мәтіндік ақпаратты жинау, сақтау, өңдеу және жіктеу үшін таратылған ақпараттық жүйені әзірлеу және ақпараттық жүйе нәтижелерінің дұрыстығын тексеру.

Зерттеудің өзектілігі. Медиа мониторинг тапсырмалары әлемдік нарықта, ең алдымен, беделді басқару өнімдері түрінде өте танымал. Алайда,

мемлекеттік деңгейдегі медиа кеңістіктің деректері негізінде шешім қабылдауға байланысты жоғары деңгейдегі міндеттер қазіргі уақытта іс жүзінде автоматты түрде шешілмейді.

Мәтінді жіктеу тақырыбының елеулі дамуына қарамастан, қазіргі заманғы SoTA (state of the art – қазіргі кездегі ең жақсысы) модельдері берілген критерий бойынша жіктеу үшін үлкен мөлшерде деректерді белгілеуді, сонымен қатар жоғары өнімділікті (оқытуға арналған жүздеген және мыңдаған процессорларды) қажет етеді.

Зерттеу объектісі - мәтіндік құжаттарды жіктеу әдістері.

Зерттеу пәні - мәтіндік құжаттарды тақырыптық модельдеуге негізделген көп критерийлі бағалау әдістері.

Диссертациялық зерттеудің мақсаты - таратылған ақпараттық жүйе шеңберінде медиа-көздері мен әлеуметтік желілерден алынған мәтіндік ақпаратты автоматты түрде көп критерийлі бағалау модельдері мен әдістерін әзірлеу.

Зерттеу әдістері. Диссертациялық жұмыста келесі зерттеу әдістері қолданылады: жіктеу әдістері, мәтіндік құжаттарды векторизациялау әдістері, шешім қабылдау үшін көп өлшемді талдау әдістері, ақпараттық жүйелерді жобалау мен әзірлеу технологиялары.

Ғылыми жаңалық

1. BigARTM тақырыптық моделін пайдаланып мәтіндік құжаттарды векторизациялау әдісі ұсынылған.
2. Жіктеу моделін өздігінен үйрену үшін тақырыптық денеаралық тепе-теңдікті бағалау әдісі ұсынылған.
3. Басылымның әлеуметтік маңыздылығын көрсететін көп факторлы бағалау әдісі ұсынылады.
4. Байес агрегаттау жүйесіне негізделген массалық ақпарат құралдарын ММҚ көп критериалды бағалаудың ұсынылған әдістемесі, иерархияны талдау әдісі (AHP) және тақырыптық модельдеу.

Зерттеудің негізгі нәтижелері мыналар болып табылады: мәтіндік құжаттар мен бұқаралық ақпарат құралдары көздерін көп критериалды бағалау әдістемесін және сәйкес үлгілерді өз бетінше оқу мен оқытудың соған байланысты әдістерін әзірлеу; ұсынылатын әдістерді бағдарламалық іске асыру болып табылатын бұқаралық ақпарат құралдары мен әлеуметтік желілерден деректерді жинау, сақтау, өңдеу және жіктеу үшін таратылған ақпараттық жүйені әзірлеу; бұқаралық ақпарат құралдары мен әлеуметтік желілердегі жарияланымдардың әлеуметтік маңыздылығын бағалаудың көп факторлы әдіснамасы. Диссертациялық жұмысты орындау барысында алынған келесі практикалық қорытындылар мен ұсыныстарды бөліп көрсетуге болады:

1. Бұқаралық ақпарат құралдар мониторингі қызметін ұсынатын жүйелер нарығына талдау, сондай-ақ нормативтік-құқықтық база мен техникалық ерекшеліктерді талдау жүргізілді. Қолданыстағы шешімдердің әлсіз жақтары ашылады, ұсыныстар қалыптастырылады;

2. Ашық ақпарат көздерінің қоғамға әсері туралы мәселе зерттелді, әсер етудің негізгі бағыттары анықталды, ақпараттық белгілердің тізімі қалыптастырылды, олардың негізінде бұл әсерді бағалауға болады;

3. Құжаттарды жіктеуге және мәтінді векторлауға қатысты қолданыстағы тәсілдер зерттелді, ағымдағы шешімдердің проблемалары мен әлсіз жақтары анықталды, ұсыныстар қалыптастырылды;

4. Тақырыптық үлгі негізінде мәтіндерді векторлау тәсілі әзірленді;

5. Берілген атрибутқа қатысты тақырып салмағын автоматты немесе жартылай автоматты түрде алуға мүмкіндік беретін корпус аралық тақырыптық теңгерімсіздікті бағалау әдісі әзірленді;

6. Агрегацияның Байес үлгісіне негізделген бұқаралық ақпарат құралдары көздерін көп критериалды бағалау әдісі әзірленді;

7. Мәтіндік ақпаратты жинауға (сырлауға), сақтауға, өңдеуге, сондай-ақ нәтижелерді визуализациялау мүмкіндігімен тақырыптық модельдер мен классификаторларды құруға мүмкіндік беретін Open Source шешімдеріне негізделген таратылған ақпараттық жүйе әзірленді;

8. Қазақстандық және ресейлік дереккөздерден 6 миллионнан астам жарияланымдар, соның ішінде басылым мәтіндері де, метадеректер де жинақталған;

9. Ұсынылған үлгілер мен әдістер расталды. Есептеу эксперименттерінің негізгі сериясы қауіпті деп аталатын жаңалықтарды анықтаумен байланысты болды (әлеуметтік маңызды, резонансты жағымсыз басылымдар). BERT терең оқыту моделімен салыстыру жүргізілді, ұсынылған модельдердің сапа көрсеткіштері салыстырмалы, есептеу күрделілігі аз және қолмен сарапшылық бағалау көлеміне аз талаптар қойылады.

Ғылыми нәтижелер тақырыптардың немесе мақала метадеректерінің жоғары деңгейлі белгілеуін пайдалана отырып, мәтіндік құжаттарды бірқатар мүмкіндіктер (критерийлер) бойынша қолмен белгілеудің ең аз мөлшерімен жіктеуге мүмкіндік беретін үлгілерді, әдістерді және бағдарламалық құралдарды әзірлеуге мүмкіндік берді. Зерттеу шеңберінде әзірленген ақпараттық жүйе оңай масштабталады (функционалдық жағынан да, өнімділігі жағынан да) және деректер мәтіндерінің үлкен көлемін сақтауға, өңдеуге, біріктіруге және визуализациялауға мүмкіндік береді.

Нәтижелерді нақты пайдалану үшін ұсыныстар мен бастапқы деректер

Ғылыми зерттеулердің нәтижелері, атап айтқанда, әзірленген ақпараттық жүйе бірқатар мақсаттарда пайдаланылуы мүмкін:

1. Зерттеушілер мен ғалымдардың пайдалануы. Жұмыстың 5.3.1 бөлімінде көрсетілгендей, әртүрлі гуманитарлық зерттеулер үшін әзірленген ақпараттық жүйені пайдаланудың әлеуеті зор.

2. Шешім қабылдауды қолдау үшін ірі компаниялар мен мемлекеттік органдардың пайдалануы.

3. Беделді басқару мәселесін шешу үшін ірі компаниялардың пайдалануы.

4. Қарапайым пайдаланушылардың қызығушылық тудыратын деректерді барлау іздеу үшін пайдалануы.

Сонымен қатар, жүйенің әмбебаптығын атап өту керек - оны мәтіндік деректер корпусының кең ауқымына қолдануға болады, мысалы, ұйымдардың ішкі құжаттарына, ғылыми жарияланымдарға, жеке хаттарға және т.б.

Қорғаныс ережелері:

1. Мәтіндік құжаттарды тақырыптық векторлауды пайдалана отырып, бұқаралық ақпарат құралдарды бағалаудың көп өлшемді әдістемесі негізінде әлеуметтік маңыздылықты мультифакторлық бағалау әдістемесі әзірленді.

2. Media Analytics бұқаралық ақпарат құралдарының және әлеуметтік желілердің мәтіндік деректерін жинау, өңдеу, бағалау және визуализациялау үшін бөлінген ақпараттық жүйенің архитектурасы мен бағдарламалық қамтамасыз етілуі әзірленді, оның ішінде ұсынылған бағалау әдістемесі, сондай-ақ оны бағалау және тексеру.

Тақырыптың ғылыми-зерттеу бағдарламаларының жоспарларымен байланысы. Ұсынылған нәтижелер ҚР БҒМ БҒМ ИИВТ (Қаржыландыру көзі ҚР БҒМ Ғылым комитеті): бағдарламалық-мақсатты қаржыландыру (НҚҚ) ҚР БҒМ BR05236839 «Жеке тұлғаның тұрақты дамуын ынталандыру үшін ақпараттық технологиялар мен жүйелерді дамыту» жобасын іске асыру барысында алынды. цифрлық Қазақстанды дамыту негіздерінің бірі ретінде» 2018-2020 жж.

Зерттеу нәтижелерін апробациялау. Зерттеудің негізгі ережелері мен нәтижелері: Procedia 9-шы халықаралық жас ғалымдар конференциясында (Scopus, 2020), Сандық трансформация және жаһандық қоғам туралы халықаралық конференцияда (Springer, 2019, 2020) және басқа конференцияларда баяндалған. Зерттеудің негізгі нәтижелері импакт-факторы 2,51 (Q1) болатын Symmetry журналында жарияланды. Әзірленген ақпараттық жүйе Қазақстан Республикасы Білім және ғылым министрлігінде енгізілген (Диссертацияның А қосымшасы).

Диссертация тақырыбы бойынша 16 жұмыс жарияланды, оның ішінде Scopus және Thomson Reuters мәліметтер қорына кіретін журналдарда 6 мақала жарияланды (2 мақала - Q1, 2 мақала - Q2, 2 мақала - Q3), олардың бірі ҚР БҒМ білім және ғылым саласындағы бақылау комитеті ұсынған жарияланымдар, тағайындалған квартилсіз Scopus-та CiteScore бар журналдарда 3 мақала (2 мақала - CiteScore 17%, 1 мақала - CiteScore 69%), 2 мақала РИНЦ-ке енгізілген

журналдарда (импакт-фактор 0,482 және 1,385), 6 ғылыми мақалалар халықаралық ғылыми-практикалық конференциялар жинақтарында жарияланды. Жұмыс нәтижесі бойынша 4 авторлық куәлік берілді.

Жұмыс кіріспеден, бес бөлімнен, қорытындыдан және пайдаланылған әдебиеттер тізімінен тұрады. Диссертацияның жалпы көлемі 130 бет.