AZƏRBAYCAN RESPUBLİKASI

Əlyazması hüququnda

ELEKTRON DÖVLƏT-VƏTƏNDAŞ PLATFORMASINDA MƏLUMATLARIN İNTELLEKTUAL ANALİZİ ÜÇÜN METOD VƏ ALQORİTMLƏRİN İŞLƏNİLMƏSİ

İxtisas: 3338.01 – Sistemli analiz, idarəetmə və informasiyanın işlənməsi

Elm sahəsi: Texnika elmləri

İddiaçı: **Günay Yavər qızı İskəndərli**

Fəlsəfə doktoru elmi dərəcəsi almaq üçün
təqdim edilmiş dissertasiyanın

AVTOREFERATI

Bakı – 2021

Dissertasiya işi Azərbaycan Milli Elmlər Akademiyası (AMEA) İnformasiya Texnologiyaları İnstitutunda yerinə yetirilmişdir.

Elmi rəhbər: AMEA-nın müxbir üzvü, texnika elmləri doktoru
Ramiz Məhəmməd oğlu Alıquliyev

Rəsmi opponentlər: texnika elmləri doktoru, professor
Nadir Bafadin oğlu Ağayev

texnika üzrə fəlsəfə doktoru
Lalə Hekayət qızı Kərimova

texnika üzrə fəlsəfə doktoru
Vüqar Yadulla oğlu Musayev

Azərbaycan Respublikasının Prezidenti yanında Ali Attestasiya Komissiyasının Azərbaycan Milli elmlər Akademiyası İnformasiya Texnologiyaları İnstitutunun nəzdində fəaliyyət göstərən ED 1.35 Dissertasiya şurası

Dissertasiya şurasının sədri:

AMEA-nın həqiqi üzvü,
texnika elmləri doktoru, professor
Rasim Məhəmməd oğlu Əliquliyev

Dissertasiya şurasının elmi katibi:

texnika üzrə fəlsəfə doktoru, dosent
Fərhad Firudin oğlu Yusifov

Elmi seminarın sədri:

texnika elmləri doktoru
Mütəllim Mirzəəhməd oğlu Mütəllimov

İŞİN ÜMUMİ XARAKTERİSTİKASI

Mövzunun aktuallığı. İnformasiya cəmiyyətinin əsas elementlərindən olan e-dövlətin yaradılmasında əsas məqsəd dövlət idarələrinin vətəndaşlara göstərdikləri xidmətlərin səviyyəsini yüksəltməkdir. Belə ki, e-dövlət vasitəsilə informasiya resurslarına girişi sadələşdirmək və cəmiyyətin bütün təbəqələrinin dövlət idarəçiliyində aktiv iştirakını təmin etmək mümkündür. E-dövlət mühitində səmərəli qərarların qəbulu və dövlət təhlükəsizliyinin təmin olunması üçün bu mühitin tədqiqi mühüm əhəmiyyət kəsb edir.

Qeyd edək ki, e-dövlət mürəkkəb sosiotexnoloji mühitdir və bu mühitdə əsas hədəf hakimiyyət-vətəndaş arasında olan münasibətlər hesab olunur. E-dövlət mühitinin müxtəlif seqmentləri üzrə problemlər mövcuddur. Bu dissertasiya işində hədəfdə olan əsas məsələ hakimiyyət-vətəndaş münasibətlərində mövcud olan informasiya fəzasının analizindən ibarətdir. E-dövlət mühitində insanların maraq dairəsində olan məsələlərin vaxtında müəyyən olunması dövlət orqanlarına xidmətlərin keyfiyyətini yaxşılaşdırmağa, vətəndaş məmnuniyyətini artırmağa kömək edə bilər. E-dövlət xidmətlərinin əlçatanlığını və effektivliyini artırmaq üçün mütəmadi olaraq istifadəçi yönümlü qiymətləndirmə aparmaq lazımdır.

E-dövlətin əsas funksiyalarından biri də vətəndaşları ehtimal olunan zərər və zorakılıqlardan qorumaqdır. Təcrübə göstərir ki, bu əlverişli mühitdən cinayətkar qruplar da yaxşı "yararlanırlar" və onlar bu imkandan istifadə edərək dövlət və cəmiyyət üçün böyük təhlükə mənbəyinə çevrilirlər. Deməli, dövlətin mühüm vəzifələrindən biri də virtual mühitdə – İnternetdə və e-dövlətdə gizli fəaliyyət göstərən kriminal şəbəkələrin fəaliyyətini aşkarlamaq və analiz etməkdir. Bu mühit sürətli kommunikasiya yaratmaq və fəaliyyəti operativ koordinasiya etmək baxımından çox geniş imkanlara malikdir. Kriminal şəbəkənin üzvləri ünsiyyət qurmaq üçün veb-saytlardan, e-poçtdan, bloqlardan, onlayn çatdan və s. istifadə edir. Aydındır ki, belə kommunikasiya vasitələrində ötürülən informasiya növləri arasında mətnlər üstünlük təşkil edir. Ona görə də, mümkün ola biləcək terror aktlarının qarşısının alınması və dövlətin təhlükəsizliyinin təmin

olunması üçün virtual mühitdə, o cümlədən e-dövlətdə dövr edən mətnlərin analizi mühüm əhəmiyyət kəsb edir.

Yuxarıda qeyd olunduğu kimi, təhlükəsizlik çox mühüm məsələdir və təhlükəsizliyin təmin olunması üçün müxtəlif yanaşmalar, baxışlar mövcuddur. Təəssüflər olsun ki, e-dövlət mühitində vətəndaşların özünəmənsub olan şərhləri yetərincə analiz olunmayıb. Hal-hazırda biliklərin idarə olunmasında, müxtəlif mənbələrdə toplanmış mətnlərin intellektual analizində text mining ən qabaqcıl və effektiv texnologiyalardan biri hesab olunur. Dissertasiya işində bunları nəzərə alaraq, sosial şəbəkə və mətnlərin intellektual analizi (text mining) texnologiyalarının köməyilə e-dövlət-vətəndaş münasibətləri araşdırılmış, bu mühitdə səmərəli qərarların qəbulunu dəstəkləyən sistemlər üçün yeni yanaşmalar, metod və alqoritmlər təklif edilmişdir.

İşin məqsədi e-dövlətin təhlükəsizlik səviyyəsinin yüksəldilməsi və bu mühitdə göstərilən xidmətlərin keyfiyyətinin yüksəldilməsi məqsədilə sosial şəbəkə və mətnlərin intellektual analizi texnologiyalarının köməyilə dövlət-vətəndaş münasibətlərinin analizi üçün yeni yanaşmalar, metod və alqoritmlərin təklif edilməsidir.

Tədqiqatın metodları təbii dilin emalı, data mining, text mining, mövzu modelləşdirmə (topic modeling), qraflar nəzəriyyəsi, ehtimal nəzəriyyəsi, sosial şəbəkə analizi texnologiyalarına əsaslanır.

Müdafiəyə çıxarılan əsas müddəalar:

- e-dövlətdə terrorizmlə əlaqəli mətnlətin aşkarlanması üçün hibrid təsnifatlandırma metodu;
- e-dövlətdə terrorizmi təbliğ edən mətnlərin filtrasiyası üçün sentiment analiz texnologiyasına və Bayes klassifikatoruna əsaslanan metod;
- e-dövlətdə fəaliyyət göstərən gizli sosial şəbəkələrin aşkarlanması və analizi üçün sentiment analiz texnologiyasına əsaslanan metod və alqoritm;
- e-dövlət xidmətlərindən vətəndaş məmnuniyyətinin avtomatik qiymətləndirilməsi üçün metod;
- e-dövlətdə vətəndaşları (o cümlədən regionları) maraqlandıran aktual mövzuların müəyyən edilməsi üçün metod.

Dissertasiya işinin elmi yeniliyi aşağıdakı **nəticələrlə** təyin edilir:

- e-dövlətdə terrorizmlə əlaqəli mətnlərin aşkarlanması üçün hibrid təsnifatlandırma metodu işlənmişdir;
- e-dövlətdə terrorizmlə əlaqəli mətnlərin filtrasiyası üçün sentiment analiz texnologiyasına və Bayes klassifikatoruna əsaslanan metod işlənmişdir;
- e-dövlətdə gizli sosial şəbəkələrin aşkarlanması və analizi üçün sentiment analiz texnologiyasına əsaslanan metod və alqoritm işlənmişdir;
- e-dövlət xidmətlərindən vətəndaş məmnuniyyətinin avtomatik qiymətləndirilməsi üçün metod işlənmişdir;
- e-dövlətdə vətəndaşları (o cümlədən regionları) maraqlandıran aktual mövzuların müəyyən edilməsi üçün klasterləşmə və mövzu modelləşdirmə texnologiyalarına əsaslanan metod işlənmişdir.

İşin praktiki əhəmiyyəti. Əldə edilmiş elmi-nəzəri və praktiki nəticələr onlayn mühitlərdə fəaliyyət göstərən müxtəlif təbiətli sosial şəbəkələrin aşkarlanması və analizində, terrorizmlə əlaqəli fəaliyyətin aşkarlanması, terrorizmlə bağlı mətnlərin filtrasiyasında, e-xidmətlərin keyfiyyətinin yüksəldilməsində, vətəndaşların, regionların əsas maraq dairələrinin müəyyən olunmasında, istifadəçi şərhlərinin analizi əsasında aktual mövzuların çıxarılmasında və s. istifadə oluna bilər.

İşin aprobasiyası. Əsas elmi-nəzəri və praktiki nəticələr aşağıda adı çəkilən konfranslarda məruzə edilmiş və müzakirə olunmuşdur: "İnformasiya təhlükəsizliyinin multidissiplinar problemləri" II respublika elmi-praktiki konfransı (Bakı, 14 may 2015-ci il); "Big data: imkanları, multidissiplinar problemləri və perspektivləri" I respublika elmi-praktiki konfransı (Bakı, 25 fevral 2016-cı il); 10th International Conference on Application of Information and Communication Technologies – AICT 2016 (Baku, 12-14 October 2016); "İnformasiya təhlükəsizliyinin aktual multidissiplinar problemləri" IV respublika elmi-praktiki konfransı, (Bakı, 14 dekabr 2018-ci il); 2nd International Symposium on Applied Sciences and Engineering (Turkey, 7-9 April 2021).

Elmi nəşrlər: Dissertasiya mövzusu üzrə 14 elmi iş çap olunmuşdur. Onlardan 6 məqalə resenziya olunan jurnallarda, 7 tezis konfrans materiallarında və 1 ekspress-informasiya nəşr edilmişdir. Bu elmi işlərdən 3 məqalə Web of Science bazasında indeksləşən jurnallarda çap edilmişdir.

İşin strukturu və həcmi: Dissertasiya işi giriş, 4 fəsil, nəticə, 179 adda ədəbiyyat siyahısı və bir əlavədən, 11 şəkil və 14 cədvəldən ibarətdir.

İddiaçı AMEA-nın həqiqi üzvü, texnika elmləri doktoru, professor Rasim Əliquliyevə və elmi rəhbər, AMEA-nın müxbir üzvü, texnika elmləri doktoru Ramiz Alıquliyevə qiymətli məsləhətlərinə, dissertasiya işinin yerinə yetirilməsində göstərdikləri daimi diqqətə və hərtərəfli dəstəyə görə dərin minnətdarlığını bildirir.

İŞİN QISA İCMALI

Girişdə dissertasiya işinin aktuallığı əsaslandırılmış, tədqiqatın məqsədi və həll olunacaq məsələlər müəyyən edilmişdir. Əldə edilmiş nəticələrin elmi yeniliyi və praktiki əhəmiyyəti göstərilmişdir.

Birinci fəsildə ("E-dövlətin analizi texnologiyaları: text mining və sosial şəbəkələr") e-dövlət anlayışı, onun təkmil modelləri analiz olunmuş, e-dövlətin analizində text mining və sosial şəbəkə analizi texnologiyalarının rolu araşdırılmış, onun analizi sahəsində problemlərin müasir vəziyyəti şərh edilmişdir [4-6, 8].

İkinci fəsildə ("E-dövlətdə terrorizmi təbliğ edən mətnlərin aşkarlanması üçün metodlar") e-dövlət mühitində terrorizmlə əlaqəli məqalələrin aşkarlanması üçün text mining texnologiyasına əsaslanan metodlar təklif olunmuşdur.

Terrorizmlə əlaqəli sənədləri identifikasiya etmək üçün kNN, Bayes və yeni təklif olunan RG metodunun xətti kombinasiyasından ibarət hibrid təsnifatlandırma metodu **ikinci fəslin birinci bölməsində** verilmişdir [2, 3]:

Təklif olunan metod. Tutaq ki, tədqiq olunan mühitin dili üçün baxılan mövzu (terrorizm) ilə bağlı lüğət bazası (VBase) yaradılmış və sözlərin semantik şəbəkəsi (WordNet) qurulmuşdur.

Təklif olunan yanaşmanın hər bir mərhələsi aşağıda ətraflı izah edilmişdir:

1)Sənədlərin ilkin filtrasiyası: Əvvəlcə sənəddən terminlər çıxarılır, onlar morfoloji təhlil edilir və sənəd sözlər (terminlər) çoxluğu kimi təsvir olunur, $d = (\mathrm{t}_1, \mathrm{t}_2, ..., \mathrm{t}_m)$. Sonra Şimkeviç-Simpson ölçüsündən istifadə edərək VBase bazası ilə $d = (t_1, t_2, ..., t_m)$ çoxluğu arasındakı yaxınlıq hesablanır:

$$\mathrm{sim}_{\mathrm{S-S}}(d, \mathrm{VBase}) = \frac{|d \cap \mathrm{VBase}|}{|d|} \qquad (1)$$

burada $|A| - A$ çoxluğundakı elementlərin sayıdır.

Əgər $\mathrm{sim}_{\mathrm{S-S}}(d, \mathrm{VBase}) \geq \varepsilon$ olarsa, onda d sənədi şübhəli sənədlər

çoxluğuna əlavə edilir və identifikasiya üçün növbəti mərhələyə keçid edilir. Burada ε eksperimental yolla müəyyən edilmiş həddqiymətidir.

2)Sözlərin semantik yaxınlığı: Sözlər arasındakı semantik yaxınlığı hesablamaq üçün əvvəlcə WordNet şəbəkəsindən istifadə etməklə, sözün informativ məzmunu $\mathrm{IC}(t)$ təyin edilir:

$$\mathrm{IC}(t) = 1 - \frac{\log(\mathrm{synset}(t) + 1)}{\log(t_{max})} \tag{2}$$

Sonra (2) düsturundan istifadə edərək sözlər arasındakı semantik yaxınlıq hesablanır:

$$\mathrm{sim}_{\mathrm{IC}}(t_1, t_2) = \begin{cases} \dfrac{2 * \mathrm{IC}(\mathrm{LCS}(t_1, t_2))}{\mathrm{IC}(t_1) + \mathrm{IC}(t_2)}, & \text{əgər} \quad t_1 \neq t_2 \\ 1, & \text{əgər} \quad t_1 = t_2 \end{cases} \tag{3}$$

burada $\mathrm{LCS}(t_1, t_2)$ – WordNet şəbəkəsində t_1 və t_2 sözlərinin ən yaxın olduğu ortaq söz, t_{max} – WordNet semantik şəbəkəsindəki sözlərin ümumi sayı, $\mathrm{synset}(t)$ $-t$ sözünün sinonimlərinin sayıdır.

Sözlər arasındakı semantik yaxınlığı həm də WUP metrikasından istifadə etməklə hesablayırıq:

$$\mathrm{sim}_{\mathrm{WUP}}(t_1, t_2) = \frac{2 * \mathrm{depth}(t)}{\mathrm{depth}(t_1) + \mathrm{depth}(t_2) + 2 * \mathrm{depth}(t)} \tag{4}$$

burada $\mathrm{depth}(t_1)$ – WordNet semantik şəbəkəsində (ağacında) t_1-dən t -yə qədər olan qovşaqların sayı; $\mathrm{depth}(t_2)$ – t_2-dən t-yə qədər olan qovşaqların sayı; $\mathrm{depth}(t)$ – t-dən şəbəkənin kökünə qədər olan qovşaqların sayıdır.

Beləliklə, sözlər arasında semantik yaxınlıq (3) və (4) düsturları ilə verilən metrikaların xətti kombinasiyası kimi təyin olunur:

$$\mathrm{sim}(t_1, t_2) = \alpha \times \mathrm{sim}_{\mathrm{IC}}(t_1, t_2) + (1 - \alpha) \times \mathrm{sim}_{\mathrm{WUP}}(t_1, t_2) \tag{5}$$

burada $0 \leq \alpha \leq 1$ – çəki əmsalıdır.

3)Cümlələrin yaxınlıq ölçüsü: Cümlələr arasındakı yaxınlığı hesablamaq üçün 3 metrikadan istifadə olunacaqdır: semantik, kosinus və sintaktik.

Semantik yaxınlıq. Cümlələr arasındakı semantik yaxınlıq sözlər

arasındakı semantik yaxınlıqdan (5) istifadə edilərək hesablanır:

$$\text{sim}_{\text{semantic}}(s_1, s_2) = \frac{\sum_{t_1 \in s_1, t_2 \in s_2} \text{sim}(t_1, t_2)}{m_1 + m_2} \qquad (6)$$

burada m_1 və m_2 uyğun olaraq s_1 və s_2 cümlələrindəki sözlərin sayıdır.

Kosinus metrikası. Kosinus metrikasından istifadə etməklə, iki vektor arasındakı yaxınlıq aşağıdakı kimi hesablanır:

$$\text{sim}_{\cos}(s_1, s_2) = \frac{\sum_{j=1}^{m} (w_{1j} \times w_{2j})}{\sqrt{\sum_{j=1}^{m} w_{1j}^2} \times \sqrt{\sum_{j=1}^{m} w_{2j}^2}}$$

burada $s_1 = (w_{11}, w_{12}, ..., w_{1m})$ və $s_2 = (w_{21}, w_{22}, ..., w_{2m})$ – s_1 və s_2 cümlələrinə uyğun semantik vektorlar; w_{pj} – s_p vektorunda t_j sözünün çəkisi; m isə sözlərin ümumi sayıdır.

Sintaktik yaxınlıq. Cümlələrin, sözlərin cümlədəki mövqeyinə əsaslanan yaxınlığını, yəni sintaktik yaxınlığı hesablamaq üçün aşağıdakı düsturdan istifadə olunur:

$$\text{sim}_{\text{wordorder}}(s_1, s_2) = 1 - \frac{\|o_1 - o_2\|}{\|o_1 + o_2\|}$$

burada $o_1 = (w_{11}, w_{12}, ..., w_{1m})$ və $o_2 = (w_{21}, w_{22}, ..., w_{2m})$ – s_1 və s_2 cümlələrinə uyğun sintaktik-vektorlar; w_{pj} isə o_p vektorunda t_j sözünün çəkisi, $\|\cdot\|$-Evklid məsafəsidir.

Xətti kombinasiya. Cümlələr arasında yaxınlığı hesablamaq üçün semantik, kosinus və sintaktik ölçülərin xətti kombinasiyası istifadə olunur:

$$\text{sim}_{\text{sentences}}(s_1, s_2) = \beta_1 \cdot \text{sim}_{\text{semantic}}(s_1, s_2) + \beta_2 \cdot \text{sim}_{\text{wordorder}}(s_1, s_2) + \beta_3 \cdot \text{sim}_{\cos}(s_1, s_2) \qquad (7)$$

burada β_i $(0 \le \beta_i \le 1, \quad i = 1,2,3)$ çəki parametrləridir və

$\beta_1 + \beta_2 + \beta_3 = 1.$

4)Sənədlərin yaxınlıq ölçüsü: Sənədlər arasındakı yaxınlığı hesablamaq üçün cümlələr arasındakı yaxınlıqdan (7) istifadə olunur:

$$\text{sim}_{\text{documents}}(d_1, d_2) = \frac{\sum_{s_1 \in d_1, s_2 \in d_2} \text{sim}_{\text{sentences}}(s_1, s_2)}{n_1 + n_2}$$

burada n_1 və n_2 uyğun olaraq d_1 və d_2 sənədlərindəki cümlələrin sayıdır.

Sadəlik üçün aşağıda $\text{sim}_{\text{documents}}(d_1, d_2)$ əvəzinə $\text{sim}(d_1, d_2)$ yazılışından istifadə ediləcəkdir.

5)Sənədlərin təsnifatlandırılması: Tutaq ki, $\mathbf{C} = (C_1, ..., C_k)$ siniflər çoxluğu məlumdur. Burada d_i sənədinin C_q sinfinə aid olma dərəcəsini müəyyən etmək üçün kNN (k-Nearest Neighbor – k-ən yaxın qonşu), Bayes və yeni təklif olunan RG metodundan istifadə olunur.

kNN metodu. Bu metoda görə d_i sənədinin C_q sinfinə aid olması aşağıdakı düsturla hesablanmış kəmiyyətin qiyməti ilə müəyyən edilir:

$$\text{score}_{k\text{NN}}(d_i \mid C_q) = \sum_{d \in k\text{NN}_q(d_i)} \text{sim}(d_i, d), \quad i = 1, 2, ..., N; \quad q = 1, 2, ..., k \quad (8)$$

burada $k\text{NN}_q(d_i)$ – C_q sinfində d_i sənədinə ən yaxın olan k sayda sənədlər çoxluğudur.

d_i sənədi ən böyük $\text{score}_{k\text{NN}}(d_i \mid C_q)$ qiymətinə malik sinfə aid edilir, başqa sözlə $d_i \in C_{k^*}$, əgər $k^* = \arg\max_q \text{score}_{k\text{NN}}(d_i \mid C_q)$.

Modifikasiya olunmuş Bayes metodu. Bu metoda görə d_i sənədinin C_q sinfinə aid olma dərəcəsi aşağıdakı şərti ehtimalın qiyməti ilə müəyyən olunur:

$$\text{score}_{\text{M Bayes}}(C_q \mid d_i) = \text{P}(C_q \mid d_i) = \frac{\log \text{P}(C_q)}{w_i} + \sum_{j=1}^{m} \text{P}(\text{t}_j, d_i) \log \text{P}(t_j \mid C_q) \quad (9)$$

burada $P(t_j, d_i) = w_{ij} / w_i$ – t_j sözünün d_i sənədində işlənmə ehtimalıdır, $w_i = \sum_{j=1}^{m} w_{ij}$, $i = 1,...,n$; $q = 1,...,k$. w_{ij} – t_j sözünün d_i sənədində çəkisidir. $P(t_j | C_q)$ – t_j sözünün C_q sinfində olma ehtimalı, m – **D** sənədlər çoxluğundakı sözlərin sayıdır: $P(C_q)$ – sənədlərin C_q sinfində olma ehtimalıdır.

(9) düsturunda kNN metoduna oxşar olaraq $\text{score}_{\text{MBayes}}(C_q | d_i) = P(C_q | d_i)$ işarələməsi qəbul edilmişdir. Bu modelə görə d_i elə sinfə aid edilir ki, bu sinif üçün $P(C_q | d_i)$ ehtimalı ən böyük qiymətə malik olsun, $d_i \in C_{k^*}$, burada $k^* = \arg\max_{1 \le q \le k} \text{score}_{\text{MBayes}}(C_q | d_i)$

RG metodu. Bu metodun köməyilə d_i sənədinin C_q sinfinə aid olma dərəcəsi aşağıdakı düstürlə müəyyən edir:

$$\text{score}_{\text{RG}}(d_i \mid C_q) = \lambda \times \frac{\text{sim}(O_{d_i}, O_{C_q})}{\sum_{p=1}^{k} \text{sim}(O_{d_i}, O_{C_p})} + \\ + (1-\lambda) \times \frac{\sum_{v \in C_q} \text{sim}(O_{d_i}, O_v)}{\sum_{p=1}^{k} \sum_{d \in C_p} \text{sim}(O_{d_i}, O_d)} \qquad (10)$$

burada $\text{sim}(O_{d_i}, O_{C_q})$ – d_i sənədinin O_{d_i} obrazı ilə C_q sinfinin O_{C_q} obrazı arasında yaxınlıq ölçüsü; $\text{sim}(O_d, O_v)$ – d və v sənədlərinin O_d və O_v obrazları arasındakı yaxınlıq ölçüsü; λ isə çəki əmsalıdır, $0 \le \lambda \le 1$.

O_{C_q} obrazı C_q sinfinin mərkəzi kimi təyin edilir, $O_{C_q} = (w_1^q, w_2^q, ..., w_m^q)$:

$$w_j^q = \frac{1}{|C_q|} \sum_{d \in C_q} w_j^{q,d}, \; q = 1,...,k, \; j = 1,...,m$$

burada $|C_q|$ – C_q sinfindəki sənədlərin sayını, $w_j^{q,d}$ isə C_q sinfinə

daxil olan d sənədindəki j-ci sözün çəkisini göstərir.

Analoji qayda ilə O_d obrazı d sənədinin mərkəzi kimi təyin edilir, $O_d = (w_1^d, w_2^d, ..., w_m^d)$:

$$w_j^d = \frac{1}{|d|} \sum_{s \in d} w_j^{d,s}, \quad j = 1, ..., m$$

burada $|d| - d$ sənədindəki cümlələrin sayı, $w_j^{d,s}$ isə d sənədinə aid olan s cümləsindəki j-ci sözün çəkisidir.

Hibrid metod. Yekun təsnifatlandırma metodu kimi (8), (9) və (10) düsturlarının köməyilə alınmış nəticələrin xətti kombinasiyasından istifadə olunması təklif edilir:

$$\text{score}(d^{\text{new}} \mid C_q) = \gamma_1 \cdot \text{score}_{k\text{NN}}(d^{\text{new}} \mid C_q) + \gamma_2 \cdot \text{score}_{\text{Bayes}}(d^{\text{new}} \mid C_q) + \gamma_3 \cdot \text{score}_{\text{RG}}(d^{\text{new}} \mid C_q)$$

burada $0 \le \gamma_i \le 1$, $(i = 1,2,3)$ çəki əmsallarıdır və $\gamma_1 + \gamma_2 + \gamma_3 = 1$.

Beləliklə, d^{new} sənədi elə C_{k^*} sinfinə aid edilir ki, bu sinif üçün o ən böyük $\text{score}(d^{\text{new}} \mid C_q)$ qiymətinə malik olsun, $d^{\text{new}} \in C_{k^*}$, burada $k^* = \arg\max_q \text{score}(d^{\text{new}} \mid C_q)$.

6)Qiymətləndirmə: Təsnifatı qiymətləndirmək üçün dəqiqlik (accuracy), həssaslıq (precision), tamlıq (recall) və F-ölçü (F-measure) meyarlarından istifadə olunur:

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$\text{Precision} = \frac{T_p}{T_p + F_p}$$

$$\text{Recall} = \frac{T_p}{T_p + F_n}$$

$$\text{F} - \text{measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

burada T_p – doğru təsnif edilmiş terrorizmlə əlaqəli sənədlərin sayı; F_p –səhv təsnif edilmiş terrorizmlə əlaqəli sənədlərin sayı; T_n – doğru kimi təsnif edilmiş terrorizmlə əlaqəli olmayan sənədlərin sayı; F_p – səhv kimi təsnif edilmiş terrorizmlə əlaqəli olmayan sənədlərin sayıdır.

Məlum olduğu kimi e-dövlət interaktiv mühitdir və hər kəs bu mühitdən yararlana bilər. Belə ki, terror təşkilatının üzvləri dövləti hədələmək, siyasi mesaj göndərmək, özlərindən sonra iz qoymaq məqsədilə eyni vaxtda dövlətə hədələyici məktublar yaza bilərlər. Bunları nəzərə alaraq **ikinci fəslin ikinci bölməsində** terrorla bağlı mətnlərin filtrasiyası üçün yeni yanaşma təklif olunmuşdur:

Təklif olunmuş yanaşmada əsas məqsəd yazılan rəyləri avtomatik analiz etmək və potensial təhlükəli rəyləri aşkarlamaqdır [9,10].

Təklif olunan yanaşmanın hər bir mərhələsi aşağıda ətraflı şəkildə izah olunmuşdur:

İlkin emal: İlkin emal zamanı hər bir mətn (rəy) "." və "!" vasitəsilə cümlələrə ayrılır. Mətn ümumişlək sözlərdən təmizlənir. Hər bir söz müxtəlif formalarda şəkilçilər qəbul etdiyi üçün bütün sözlər ilkin variantına qaytarılır. Daha sonra hər bir sözün sinonimləri tapılır və o, çoxluq şəklində təsvir olunur.

Rəylərin polyarlığının müəyyən olunması: Əvvəlcə rəy $T = \{S_1, S_2, ..., S_N\}$ cümlələr çoxluğu şəklində təsvir olunur. Burada N – cümlələrin sayıdır. Daha sonra aşağıdakı düsturdan istifadə edərək hər bir rəyin polyarlığı müəyyən olunur:

$$score\,(T) = sign\,(\sum_{S \in T} score\,(S))$$

burada $score\,(S)$ – rəyə daxil olan S cümləsinin polyarlıq dərəcəsidir. $sign\,(x)$ işarə funksiyası aşağıdakı kimi təyin olunur:

$$sign\,(x) = \begin{cases} 1, & \text{əgər} \quad x > 0 \\ 0, & \text{əgər} \quad x = 0 \\ -1, & \text{əgər} \quad x < 0 \end{cases}$$

Rəyə daxil olan cümlələrin polyarlıq dərəcələrinin cəmi sıfırdan böyük olarsa, rəy pozitiv, sıfra bərabər olarsa, neytral, sıfırdan kiçik

olarsa, neqativ sinfə daxil olur.

$$score\,(S) = \sum_{w \in S} score\,(w)$$

w sözünün polyarlığı [10]-da göstərilən qayda ilə hesablanır.

Terrorla əlaqəli şübhəli rəylərin seçilməsi: Növbəti mərhələdə seçilmiş mənfi rəylərin terrorla əlaqəli olub-olmadığını müəyyən etmək üçün onların əvvəlcədən yaradılmış lüğət bazasında olan sözlərlə ilkin müqayisəsi aparılır. Əgər yaxınlıq müəyyən olunmuş həddən böyükdürsə, onda növbəti mərhələdə yazılmış rəy ilə lüğət bazası arasında daha detallı müqayisə aparılır.

Tutaq ki, V_{terror}–terrorla bağlı sözlər və onların genişlənməsindən ibarət lüğət bazasıdır. Burada genişlənmə dedikdə sözlərin sinonimləri çoxluğu nəzərdə tutulur. Seçilmiş mənfi rəylər sözlər çoxluğu $T = \{w_1, w_2, ..., w_m\}$ şəklində təsvir olunur. Hər bir söz isə sözlər çoxluğu $w_i \to synset\,(w_i),\ i = 1,2,...,m$ kimi təsvir olunur. Yazılmış rəyin lüğət bazasında olan sözlərlə yaxınlığını hesablamaq üçün aşağıdakı düsturdan istifadə olunması təklif olunur:

$$SW_i = synset\,(w_i) \cap V_{terror} = \{w_{i1}, w_{i2}, ..., w_{im_i}\},\ i = 1,...,m \quad (11)$$

Daha sonra (11) düsturu vasitəsilə tapılan sözlərin bir-birinə nə dərəcədə yaxın olduğu müəyyən olunur:

$$\theta_i = \frac{2}{m_i(m_i - 1)} \sum_{j=1}^{m_i - 1} \sum_{k=j+1}^{m_i} sim\,(w_{ij}, w_{ik}),\ i = 1,...,m \quad (12)$$

Sözlər arasındakı semantik yaxınlığı hesablamaq üçün [3]-də təklif olunan metoddan istifadə olunur.

Daha sonra (12) düsturu vasitəsilə hesablanan sözlərin bir-birinə yaxınlıq dərəcələri θ_i-lər toplanır. Hesablanan cəm əvvəlcədən təyin olunmuş t həddən böyük olarsa, rəyi yazan istifadəçi terrorla əlaqəli şübhəli hesab olunur və nəzarətə götürülür:

$$P(T) = \begin{cases} 1, & \text{əgər } \dfrac{1}{m} \sum_{i-1}^{m} \theta_i \geq t \\ 0, & \text{əks halda} \end{cases}$$

burada “1” yazılan rəyin terrorla bağlı şübhəli olduğunu, ”0” isə olmadığını göstərir.

Şübhəli bilinən istifadəçi rəyinin terrorla bağlılığının ehtimalını müəyyən etmək üçün Naive Bayes klassifikasiya metodundan istifadə olunması təklif olunur. Şübhəli bilinən rəyin terrorla bağlılığının ehtimalını hesablamaq üçün bu rəyə daxil olan hər bir sözün terrorla əlaqəli olma ehtimalı hesablanır:

$$P(T) = P(w_1, w_2, ..., w_n)$$

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i)$$

burada $P(T)$ -istifadəçi rəyinin ehtimalını, $P(w_i)$ isə bu rəyə daxil olan hər bir sözün terrorla əlaqəli olma ehtimalını göstərir.

Qeyd edək ki, burada hər bir sözün genişlənmiş sinonimləri çoxluğu da nəzərə alınır. Belə ki, bu rəyə daxil olan hər bir sözün və onun sinonimləri çoxluğunun əvvəlcədən yaradılmış V_{terror} lüğət bazasında işlənmə tezliyinə baxılır. Bunun əsasında şübhəli bilinən rəyin terrorla əlaqəli olma ehtimalı hesablanır:

$$P(V_{terror}|T) = \frac{P(T|V_{terror})P(V_{terror})}{P(T)}$$

$$P(T|V_{terror}) = P(w_1, w_2, ..., w_n|V_{terror}) = \prod_{i=1}^{n} P(w_i|V_{terror}) \qquad (13)$$

$$P(w_i|V_{terror}) = \frac{P(w_i \cap V_{terror})}{P(V_{terror})}$$

burada $P(V_{terror}|T)$ - şübhəli bilinən rəyin V_{terror} sinfindən olma ehtimalını göstərir.

Qeyd edək ki, (13)-də sözlərdən birinin lüğət bazasında olmaması halı nəticəyə birbaşa təsir edir (yəni ən azı bir söz V_{terror} lüğət bazasında olmadıqda istifadəçi rəyinin terrorla əlaqəli ehtimalı 0-a bərabər olur). Bu halda mətnin terrorla əlaqəli olma ehtimalını hesablamaq üçün (13) düsturunun əvəzinə aşağıdakı düsturdan istifadə edilməsi daha məqsədə uyğundur:

$$P(T|V_{terror}) = \frac{1}{n}\sum_{i=1}^{n} P(w_i|V_{terror})$$

Beləliklə, istifadəçi rəyləri əsasında terrorizmlə əlaqəli şübhəli bilinən istifadəçilər müəyyən olunur və onların fəaliyyəti müvafiq orqanlar tərəfindən nəzarətə götürülür.

Üçüncü fəsil ("E-dövlətdə gizli sosial şəbəkələrin aşkarlanması üçün metod və alqoritm") gizli sosial şəbəkələrin aşkarlanması məsələsinə həsr olunmuşdur. Burada müxtəlif mühitlərdə gizli sosial şəbəkələrin aşkarlanması üsulları araşdırılmış, e-dövlət mühitində istifadəçilərin yazdıqları şərhlərdən istifadə etməklə, mətnlərin intellektual analizi və sosial şəbəkə analizi texnologiyalarının köməyilə gizli sosial şəbəkələrin aşkarlanması üçün metod təklif olunmuşdur. [1,11].

Təklif olunan yanaşmada istifadəçilərin e-dövlət mühitində yazdığı şərhlər vasitəsilə gizli sosial şəbəkələrin aşkarlanması nəzərdə tutulmuşdur. Təklif olunan yanaşma aşağıdakı mərhələlərdən ibarətdir:

1. Verilənlərin toplanması və ilkin emalı;
2. Təsnifat;
3. Sosial şəbəkənin qurulması;
4. Sosial şəbəkənin analizi.

Təklif olunan yanaşmanın hər bir mərhələsi aşağıda ətraflı izah edilir.

1)Verilənlərin toplanması və ilkin emalı: Tutaq ki, e-dövlət mühitində n sayda məlumat (veb səhifə, informasiya) T_i, $(i = 1,2,...,n)$ yerləşdirilmişdir. i -ci məlumata yazılan şərhlər çoxluğu aşağıdakı kimi işarə olunur:

$$C_i = \{c_i^j\},\ i = 1,2,...,n,\ j = 1,2,...,m$$

burada c_i^j – i -ci məlumata j -ci istifadəçinin yazdığı şərhlər çoxluğu, m – istifadəçilərin sayıdır.

Şərhlər toplandıqdan sonra onlar üzərində ilkin emal həyata keçirilir.

2)Təsnifat: Növbəti mərhələdə hər bir məlumata yazılan şərhlər çoxluğu 3 sinifdə qruplaşdırılır, C_i^+ – pozitiv, C_i^- – neqativ və C_i^0 –

neytral:

$$C_i = C_i^+ \cup C_i^- \cup C_i^0,\ i = 1,...,n$$

Yazılan şərhləri bu üç sinifdə qruplaşdırmaq üçün mətnlərin analizində ən qabaqcıl texnologiyalardan biri olan sentiment analizdən istifadə olunması təklif olunur.

Yazılan şərhləri bu üç sinifdə qruplaşdırmaq üçün hər bir şərhin polyarlığı yuxarıda təklif olunan metoddan istifadə edərək təyin olunur:

$$C_i^- = \left\{ c_i^j \,\middle|\, score(c_i^j) = -1 \right\}$$

$$C_i^+ = \left\{ c_i^j \,\middle|\, score(c_i^j) = 1 \right\}$$

$$C_i^0 = \left\{ c_i^j \,\middle|\, score(c_i^j) = 0 \right\}$$

3)Sosial şəbəkənin qurulması: Bu mərhələdə sosial şəbəkənin aktorları və onlar arasında olan münasibətlər müəyyən olunur. İlkin olaraq neqativ sinfin ətrafında toplanan istifadəçilər qrupu müəyyən olunur. Hesab edirik ki, hər bir istifadəçi identifikasiya oluna bilir (ya qeydiyyatdan keçməklə, ya da IP ünvan vasitəsilə). Heç olmasa bir məlumata mənfi şərh yazan istifadəçilər qrupu aşağıdakı kimi təyin olunur:

$$U_\Sigma^- = \bigcup_{i=1}^{n} U_i^-$$

burada U_i^- $-i$-ci məlumata mənfi şərh yazan istifadəçilərdir.

Bütün məlumatlara mənfi şərh yazan istifadəçilər qrupu isə bu şəkildə müəyyən olunur:

$$U_\Pi^- = \bigcap_{i=1}^{n} U_i^-$$

U_Π^-- istifadəçilər qrupu quracağımız (aşkarlayacağımız) sosial şəbəkənin nüvəsini (əsas aktorlarını) təşkil edir.

Sosial şəbəkənin aktorları arasında münasibətlərin müəyyən olunmasında iki tip yanaşmadan istifadə olunur:

Birinci yanaşmada neqativ sinfə daxil olan istifadəçilərin birgə göründüyü məlumatlar və bu məlumatlara yazılan şərhlərin sayı əsasında sosial şəbəkənin aktorları arasında münasibətlərin müəyyən olunması nəzərdə tutulur:

$$w_1^{j_1 j_2} = \frac{n^{j_1 j_2}}{n^{j_1} + n^{j_2}}$$

burada $n^{j_1 j_2}$ – j_1 və j_2 istifadəçilərinin mənfi şərh yazdığı məlumatların sayı, n^{j_1} – j_1-ci istifadəçinin mənfi şərh yazdığı məlumatların sayı, n^{j_2} – j_2-ci istifadəçinin mənfi şərh yazdığı məlumatların sayıdır:

$$n^{j_1 j_2} = \sum_{i=1}^{n} I(c_i^{j_1}) \cdot I(c_i^{j_2})$$

$$n^{j_1} = \sum_{i=1}^{n} I(c_i^{j_1})$$

$$n^{j_2} = \sum_{i=1}^{n} I(c_i^{j_2})$$

burada $I(c_i^j)$ funksiyası aşağıdakı şəkildə təyin olunur:

$$I(c_i^j) = \begin{cases} 1, & \text{əgər} \quad c_i^j \neq \varnothing \\ 0, & \text{əks halda} \end{cases}$$

j-ci istifadəçi i-ci məlumata heç olmasa 1 dəfə şərh yazarsa, $I(c_i^j)$ funksiyası 1, əks halda, yəni j-ci istifadəçi i-ci məlumata şərh yazmayıbsa, 0 qiyməti alır.

Burada həmçinin istifadəçilərin eyni bir məlumata yazdığı mənfi şərhlərin sayı da nəzərə alına bilər. Bu halda istifadəçilər arasındakı əlaqənin çəkisi aşağıdakı düsturun köməyilə təyin olunur:

$$\tilde{w}_1^{j_1 j_2} = \frac{\sum_{i=1}^{n} \left(m_i^{j_1} + m_i^{j_2} \right) * I(c_i^{j_1}) \cdot I(c_i^{j_2})}{M^{j_1} + M^{j_2}}$$

$$M^j = \sum_{i=1}^{n} m_i^j$$

burada $m_i^{j_1}$ – j_1- ci istifadəçinin i-ci məlumata yazdığı şərhlərin sayı, $m_i^{j_2}$ – j_2-ci istifadəçinin i-ci məlumata yazdığı şərhlərin sayıdır.

$\sum_{i=1}^{n}\left(m_i^{j_1}+m_i^{j_2}\right)*I(c_i^{j_1})\cdot I(c_i^{j_2})-$ j_1 və j_2 istifadəçilərinin birgə göründükləri məlumatlara yazdıqları şərhlərin ümumi sayı, $M^{j}-$ j-ci istifadəçi tərəfindən yazılan şərhlərin ümumi sayıdır.

Təklif olunan ***ikinci yanaşmada*** neqativ sinfə daxil olan istifadəçilərin yazdığı şərhlərin semantik yaxınlığı əsasında sosial şəbəkənin aktorları arasında münasibətlərin müəyyən olunması nəzərdə tutulur. Aydındır ki, sosial şəbəkələrin analizində yaxınlıq ölçülərinin böyük rolu vardır. Yaxınlıq ölçüsü vasitəsilə istifadəçilər arasında əlaqənin gücü haqqında mühakimə yürütmək mümkündür. Bu məqalədə şərhlər arasındakı yaxınlığı hesablamaq üçün Jakkard ölçüsündən istifadə olunması təklif olunur:

$$w_2^{j_1 j_2}=sim(c^{j_1},c^{j_2})=\frac{\left|c^{j_1}\cap c^{j_2}\right|}{\left|c^{j_1}\cup c^{j_2}\right|}$$

burada $sim\,(c^{j_1},c^{j_2})-$ j_1 və j_2 istifadəçilərinin yazdığı şərhlər arasında semantik yaxınlıqdır.

Beləliklə, gizli sosial şəbəkənin aktorları arasında münasibətləri aşkarlamaq üçün yuxarıda təklif etdiyimiz üsulların xətti kombinasıyasından istifadə olunur:

$$w^{j_1 j_2}=\alpha\cdot\widehat{w}_1^{j_1 j_2}+(1-\alpha)\cdot w_2^{j_1 j_2}$$

burada $\alpha\left(0\le\alpha\le1\right)$ çəki parametridir.

4) Sosial şəbəkənin analizi: Qurulmuş sosial şəbəkədə əsas aktorların müəyyən olunması üçün nüvənin nə dərəcədə kompakt olduğunu göstərmək lazımdır. Bunun üçün qurulmuş sosial şəbəkədə istifadəçilər və onlar arasındakı əlaqələrin sayından istifadə olunması təklif olunur.

Sosial şəbəkəyə daxil olan istifadəçilər arasındakı əlaqələrin sayını müəyyən etmək üçün aşağıdakı düsturdan istifadə olunması təklif olunur:

$$M_{\Sigma}^{-}=\sum_{j_1,j_2\in U_{\Sigma}^{-}}I_1(w^{j_1 j_2})$$

$$M_{\Pi}^{-} = \sum_{j_1, j_2 \in U_{\Pi}^{-}} I_1(w^{j_1 j_2})$$

burada M_{Σ}^{-} – sosial şəbəkəyə daxil olan istifadəçilər arasında əlaqələrin ümumi sayı, M_{Π}^{-} – sosial şəbəkənin nüvəsinə daxil olan istifadəçilər arasında əlaqələrin sayıdır. $I_1(w^{j_1 j_2})$ funksiyası aşağıdakı şəkildə təyin olunur:

$$I_1(x) = \begin{cases} 1, & \text{əgər} \quad x > 0 \\ 0, & \text{əgər} \quad x = 0 \end{cases}$$

Daha sonra bütün şəbəkənin sıxlıq əmsalı aşağıdakı düsturun köməyilə təyin olunur:

$$\sigma_{\Sigma}^{-} = \frac{M_{\Sigma}^{-}}{\frac{N_{\Sigma}^{-}(N_{\Sigma}^{-} - 1)}{2}} \tag{14}$$

burada $N_{\Sigma}^{-} = \left| U_{\Sigma}^{-} \right|$ – sosial şəbəkənin istifadəçilərinin sayı, $\frac{N_{\Sigma}^{-}(N_{\Sigma}^{-} - 1)}{2}$ – sosial şəbəkənin aktorları arasında bütün mümkün əlaqələrin sayıdır.

Eyni qayda ilə nüvədə sıxlıq əmsalı təyin olunur:

$$\sigma_{\Pi}^{-} = \frac{M_{\Pi}^{-}}{\frac{N_{\Pi}^{-}(N_{\Pi}^{-} - 1)}{2}} \tag{15}$$

burada $N_{\Pi}^{-} = \left| U_{\Pi}^{-} \right|$ – sosial şəbəkənin nüvəsinə daxil olan istifadəçilərin sayı, $\frac{N_{\Pi}^{-}(N_{\Pi}^{-} - 1)}{2}$ – nüvənin aktorları arasındakı bütün mümkün əlaqələrin sayıdır.

(14) və (15) düsturlarının köməyilə nüvənin bütün sosial şəbəkədəki payı (çəkisi) aşağıdakı düstur vasitəsilə təyin olunur:

$$\sigma^{-} = \frac{\sigma_{\Pi}^{-}}{\sigma_{\Sigma}^{-}}$$

burada σ^- – nüvənin bütün sosial şəbəkədəki payı (çəkisi)-dir. Bunun əsasında nüvənin kompaktlığı müəyyən olunur.

Nüvənin kompaktlığı müəyyən olunduqdan sonra nüvəyə daxil olan aktorları ranqlaşdırmaq üçün aktorlar arasındakı münasibətlərin çəkisindən və əlaqələrin sayından istifadə olunur:

$$c_j^{wa} = k_j^{(1-\alpha)} \cdot s_j^{\alpha}, \quad 0 \le \alpha \le 1$$

$$k_j = \sum_{l \in U_\Sigma^-} I_1(w^{jl})$$

$$s_j = \sum_{l \in U_\Sigma^-} w^{jl}$$

burada $c_j^{w\alpha}$ – mərkəzilik dərəcəsinin ölçüsü, k_j – nüvənin j-ci aktoru ilə şəbəkəyə daxil olan digər aktorlar arasında əlaqələrin sayları cəmi, s_j – isə uyğun olaraq əlaqələrin çəkiləri cəmi və α – çəki parametridir.

Qeyd edək ki, ranqlaşdırma aktorların əhəmiyyətlilik dərəcəsinə görə azalan sıra ilə aparılır.

Dördüncü fəsildə ("E-dövlətdə əks-əlaqə mexanizmləri üçün metod və alqoritm") e-dövlət mühitində vətəndaşların əsas maraq dairəsində olan xidmətlərin aşkarlanması, xidmətlərdən vətəndaşların məmnunluq dərəcəsinin və regionların maraq dairələrinin müəyyən edilməsi, o cümlədən e-dövlət mühitində yazılan vətəndaş şərhlərinin hansı mövzulara həsr olunduğunu müəyyən etmək və e-xidmətlərin keyfiyyətinin yüksəldilməsi üçün metod təklif olunmuş və bu metodların yoxlanılması üçün eksperimentlər aparılmışdır [7,13].

Dördüncü fəslin birinci bölməsində təklif olunan metod aşağıdakı kimidir:

Tutaq ki, e-dövlət portalında təklif olunan xidmətlərin sayı m və bu xidmətlərdən istifadə edən vətəndaşların sayı n-dir. Bunları uyğun olaraq $(EG_1, EG_2, ..., EG_m)$ və $(U_1, U_2, ..., U_n)$ ilə işarə edək. Qeyd edək ki, hər bir xidməti müəyyən zaman müddətində qiymətləndirmək mümkündür. Bunun üçün vətəndaşların xidmətə müraciətlərinin sayı və məmnunluq dərəcəsindən istifadə oluna bilər. Təklif etdiyimiz metodda hər bir xidmət üzrə vətəndaşların məmnunluq dərəcəsini müəyyən etmək və hotspot xidmətləri tapmaq üçün müəyyən T – zaman periodu

ərzində əvvəlcə hər bir istifadəçinin xidmətdən istifadə vektoru qurulur:

$$U_i = \{ u_{i1}, u_{i2}, ..., u_{im} \}, \quad i = 1,2,...,n \tag{16}$$

burada u_{ij} $(i = 1,2,...,n,\ j = 1,2,...,m)$ – i-ci istifadəçinin j-ci xidmətdən istifadəsini ifadə edir:

$$u_{ij} = (u_{ij,1}, u_{ij,2}) \tag{17}$$

burada $u_{ij,1}$ – i-ci istifadəçinin j-ci xidmətə müraciətini, $u_{ij,2}$ – elementi isə xidmətdən məmnunluq dərəcəsini ifadə edir. Qeyd edək ki, burada iki variant mümkündür: 1) Xidmətdən istifadə sayı nəzərə alınmır; 2) Xidmətdən istifadə sayı nəzərə alınır.

Birinci variantda xidmətdən istifadə sayı nəzərə alınmır. Yəni, istifadəçinin xidmətə müraciətlərinin sayı nəzərə alınmır, yalnız onun xidmətə müraciət edib-etməməsinə baxılır. Əgər vətəndaş xidmətə müraciət edibsə, bu 1 ilə, etməyibsə, 0 ilə qiymətləndirilir:

$$u_{ij,1} = \begin{cases} 1, & \text{əgər } i\text{-ci istifadəçi } j\text{-ci xidmətdən istifadə edibsə,} \\ 0, & \text{əks halda.} \end{cases}$$

Vətəndaşların xidmətlərdən məmnunluq dərəcəsini qiymətləndirmək üçün [1,5] şkalasından istifadə olunması təklif olunur:

$$u_{ij,2} = \begin{cases} 1 & \text{çox pis,} \\ 2 & \text{pis,} \\ 3 & \text{orta,} \\ 4 & \text{yaxşı,} \\ 5 & \text{çox yaxşı,} \end{cases} \tag{18}$$

Qeyd edək ki, əgər istifadəçi xidmətə müraciət etməyibsə, onda xidmətdən istifadə vektoru $u_{ij} = (0,0)$ qəbul olunur.

(17)-ni nəzərə alsaq, onda (16) matrisini aşağıdakı şəkildə yaza bilərik:

$$U = \left\| \begin{matrix} u_{11} & \dots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{n1} & \dots & u_{nm} \end{matrix} \right\| = \left\| \begin{matrix} (u_{11,1}, u_{11,2}) & \dots & (u_{1m,1}, u_{1m,2}) \\ \vdots & \ddots & \vdots \\ (u_{n1,1}, u_{n1,2}) & \dots & (u_{nm,1}, u_{nm,2}) \end{matrix} \right\| \tag{19}$$

Hər bir istifadəçinin xidmətlər üzrə vektoru qurulduqdan sonra bu vektorların sətirlər üzrə uyğun elementlərini cəmləyə bilərik. Bu bizə hər bir xidmətə vətəndaşların nə dərəcədə tələbatı olduğunu və bu xidmətlərdən məmnunluğunu müəyyən etməyə kömək edə bilər. Bu halda hər bir e-xidmət $EG_j\ (j=1,2,...,m)$, iki ölçülü vektor şəklində ifadə oluna bilər:

$$EG_j=\sum_{i=1}^{n}u_{ij}=\sum_{i=1}^{n}\left(u_{ij,1},u_{ij,2}\right)=\left(u_{j,1},u_{j,2}\right)\quad j=1,2,...,m \qquad (20)$$

burada $u_{j,1}$ – j-ci xidmətdən istifadə sayı, $u_{j,2}$ – j-ci xidmətdən məmnunluq dərəcəsini ifadə edir. (20)-dən istifadə etməklə, j-ci xidmətdən orta məmnunluq dərəcəsini hesablaya bilərik:

$$EG_j^{savg}=\frac{u_{j,2}}{u_{j,1}}$$

burada EG_j^{savg} – j-ci xidmətdən orta məmnunluq dərəcəsini ifadə edir. EG_j^{savg}-a görə ranqlaşdırma (azalan sıra ilə) aparsaq, xidmətlərin məmnunluq reytinqini almış olarıq.

İkinci variantda xidmətdən istifadə sayı nəzərə alınır. Yəni, ola bilər ki, eyni vətəndaş xidmətdən bir neçə dəfə istifadə etsin və hər dəfə xidməti müxtəlif məmnunluq balları ilə qiymətləndirsin. Bu halda e-xidmətlərin orta məmnunluq reytinqini aşağıdakı qayda ilə aparmaq olar.

İstifadəçinin xidmətə müraciətlərinin sayını nəzərə alsaq, xidmətdən istifadə vektorunu aşağıdakı şəkildə yaza bilərik:

$$u_{ij}=\left(u_{ij,1},u_{ij,2}\right)=\left(N_{ij},\sum_{k=1}^{N_{ij}}u_{ij,2}^{k}\right)=\left(N_{ij},u_{ij,2}^{\Sigma}\right) \qquad (21)$$

burada N_{ij} – i-ci istifadəçinin j-ci xidmətə müraciətlərinin sayı, $u_{ij,2}^{k}$ – k-cı dəfə müraciət etdikdə verdiyi məmnunluq balı, $u_{ij,2}^{\Sigma}$ – isə i-ci istifadəçinin j-ci xidmətə verdiyi ümumi məmnunluq balıdır. (21)-dən istifadə etməklə həmçinin i-ci istifadəçinin j-ci xidmətdən orta

məmnunluq dərəcəsini tapa bilərik:

$$u_{ij}^{savg} = \frac{u_{ij,2}^{\Sigma}}{N_{ij}}$$

Bu halda (19) matrisi aşağıdakı şəkildə ifadə olunur:

$$U = \left\| \begin{matrix} u_{11} \dots u_{1m} \\ \vdots \ddots \vdots \\ u_{n1} \dots u_{nm} \end{matrix} \right\| = \left\| \begin{matrix} (N_{11}, u_{11}^{savg}) \dots (N_{1m}, u_{1m}^{savg}) \\ \vdots \qquad \ddots \qquad \vdots \\ (N_{n1}, u_{n1}^{savg}) \dots (N_{nm}, u_{nm}^{savg}) \end{matrix} \right\|$$

Bunları nəzərə alsaq, (20)-ni aşağıdakı şəkildə ifadə edə bilərik:

$$EG_j = \sum_{i=1}^{n} u_{ij} = \left(\sum_{i=1}^{n} N_{ij}, \sum_{i=1}^{n} u_{ij,2}^{\Sigma} \right) = \left(N_j, u_{j,2}^{\Sigma} \right) \qquad (22)$$

(22)-dən istifadə etməklə müraciətlərin sayını nəzərə almaqla, hər bir xidmətdən orta məmnunluq dərəcəsini tapa bilərik:

$$EG_j^{savg} = \frac{u_{j,2}^{\Sigma}}{N_j}$$

Burada da, EG_j^{savg}-ə görə ranqlaşdırma aparsaq, xidmətlərin məmnunluq reytinqini almış olarıq.

Burada həmçinin hər bir istifadəçinin bütün xidmətlərdən orta məmnunluq dərəcəsini tapa bilərik:

$$u_i^{savg} = \frac{1}{m} \sum_{j=1}^{m} u_{ij}^{savg}$$

burada $u_i^{savg} - i$-ci istifadəçinin bütün xidmətlərdən orta məmnunluq dərəcəsini ifadə edir.

Bu halda bütün istifadəçilərin e-xidmətlərdən orta məmnunluq dərəcəsi aşağıdakı şəkildə ifadə olunur:

$$U^{savg} = \frac{1}{n} \sum_{i=1}^{n} u_i^{savg} \qquad (23)$$

burada U^{savg} – istifadəçilərin e-xidmətlərdən orta məmnunluq dərəcəsini müəyyən edir.

(23) vasitəsilə e-sistem ümumi qiymətləndirilir. Burada e-sistem

dedikdə e-dövlət platforması nəzərdə tutulur.

Burada həmçinin hər bir xidmət üzrə regional qiymətləndirmə apara bilərik. Yəni, bu xidmətlərə daha çox hansı regionların müraciət etdiyi, o cümlədən razı qalıb-qalmadığını təyin edə bilərik. Bunun üçün u_{ij}^{savg}-dan istifadə oluna bilər. Belə ki, u_{ij}^{savg}-ə görə ranqlaşdırma aparsaq, istifadəçilərin j-ci xidmətdən məmnunluq reytinqini almış olarıq. Ranqlaşdırma aparıldıqdan sonra reytinq cədvəlini (18)-dən istifadə etməklə üç hissəyə bölə bilərik: U_j^- (məmnun deyil), U_j^0 (məmnundur), U_j^+ (çox məmnundur). Burada U_j^- – j-ci xidmətdən məmnunluq balları [1,2) intervalına düşən, U_j^0 – məmnunluq balları [2,4) intervalına düşən, U_j^+ – [4,5) intervalına düşən istifadəçilər qrupudur.

İstifadəçi qrupları təyin olunduqdan sonra xidmətlər üzrə bu qrupların kəsişməsinə baxa bilərik:

$$\begin{aligned} U^- &= \bigcap U_j^- \\ U^0 &= \bigcap U_j^0 \\ U^+ &= \bigcap U_j^+ \end{aligned} \qquad (24)$$

burada U^- – bütün xidmətlərdən narazı, U^0 – razı , U^+ – çox razı olan istifadəçilər qrupudur. Burada hər bir xidmətdən istifadə edən vətəndaşı regiona bağlaya bilərik. Belə ki, hər bir istifadəçinin arxasında IP ünvan olduğunu nəzərə alsaq, (24) düsturu vasitəsilə bütün xidmətlərdən narazı və razı olan vətəndaşların hansı regiona aid olduğu avtomatik olaraq müəyyən olunur. Ola bilsin ki, bu vətəndaşlar bir regiona toplansın və ya regionlar üzrə paylanmış olsunlar.

Həmçinin regionların ümumi maraq dairəsini təyin edə bilərik. Yəni, hər bir regionun istifadə etdiyi xidmətləri və bu xidmətlərin reytinqini müəyyənləşdirməklə, regionlar üçün hotspot xidmətləri təyin etmiş olarıq. Bunun üçün $(U_1, U_2, ..., U_n)$ istifadəçiləri yuxarıda qeyd olunduğu kimi IP vasitəsilə regionlara bölürük. (21)-dən istifadə etməklə hər bir istifadəçinin (regionun) istifadə etdiyi xidmətlər

müəyyən olunur. Eyni regiondan olan istifadəçilərin müraciət etdiyi xidmətləri ranqlaşdırsaq (azalma sırası ilə), xidmətlərin istifadə reytinqini alarıq. Bu reytinqə əsasən eyni regiondan olan istifadəçilərin ən çox müraciət etdiyi xidmətlər müəyyən olunur. Buradan isə avtomatik olaraq regionların e-xidmət maraqları təyin olunmuş olur.

Cədvəl 1

İstifadəçilərin xidmətlərə müraciət sayı və verdiyi ümumi bal

İstifadəçilər	***E-xidmətlər***				
	EG1	***EG2***	***EG3***	***EG4***	***EG5***
U_1	(41, 82)	(33, 44)	(22,80)	(38,76)	(17,34)
U_2	(46,153)	(1, 4)	(19,69)	(13,43)	(42,112)
U_3	(6,16)	(43,129)	(39,104)	(25,100)	(29,106)
U_4	(46,107)	(47,156)	(40,106)	(35,128)	(28,56)
U_5	(32,64)	(34,124)	(9,30)	(45,45)	(46,214)
U_6	(4,13)	(38,152)	(24,64)	(48,80)	(14,46)
U_7	(14,56)	(37,86)	(22,58)	(27,90)	(38,126)
U_8	(27,108)	(20,40)	(32,96)	(7,32)	(38,139)
U_9	(48,192)	(33,88)	(36,72)	(7,11)	(19,76)
U_{10}	(49,130)	(8,32)	(38,126)	(13,43)	(28,46)
U_{11}	(8,24)	(36,96)	(14,46)	(42,98)	(3,11)
U_{12}	(49,98)	(1,3)	(34,90)	(12,32)	(2,7)
U_{13}	(48,160)	(14,28)	(33,132)	(41,123)	(27,117)
U_{14}	(24,88)	(2,7)	(8,26)	(12,32)	(39,78)
U_{15}	(40,146)	(4,5)	(6,20)	(47,125)	(47,141)
U_{16}	(7,21)	(41,41)	(25,91)	(17,39)	(6,8)
U_{17}	(21,56)	(35,105)	(48,112)	(10,40)	(29,106)
U_{18}	(46,168)	(16,32)	(17,45)	(12,32)	(23,69)
U_{19}	(40,53)	(48,144)	(29,87)	(31,103)	(0,0)
U_{20}	(48,112)	(1,2)	(11,25)	(24,56)	(17,45)

Təklif olunan metodu qiymətləndirmək üçün hesablamalar *Matlab 2018* proqramında yerinə yetirilmiş, təklif olunan xidmətlərin sayı 5 və bu xidmətlərə müraciət edən vətəndaşların sayı 20 götürülmüşdür. Cədvəl 1-də müəyyən zaman intervalında e-xidmətlərə olan müraciətlərin sayı və bu xidmətlərdən məmnunluq dərəcəsini ifadə edən ballar, cədvəl 2-də isə bu xidmətlərə müraciət edən istifadəçilərin mənsub olduğu regionlar təsvir olunmuşdur.

Cədvəl 2

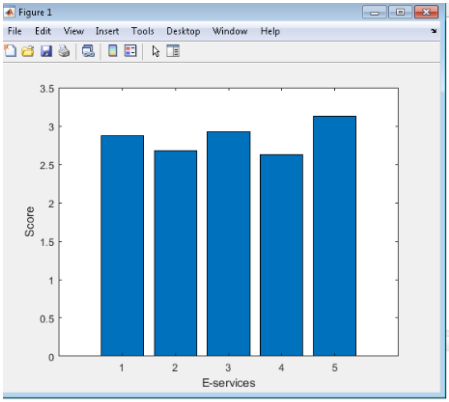
Regionlar üzrə istifadəçilərin paylanması

Regionlar	**İstifadəçilər**
R_1	U_2, U_5, U_7, U_{15}, U_{14}
R_2	U_1, U_{20}, U_9, U_{19}
R_3	U_8, U_{11}, U_{12}
R_4	U_3, U_4, U_{13}, U_{10}
R_5	U_{16}, U_{17}, U_6, U_{18}

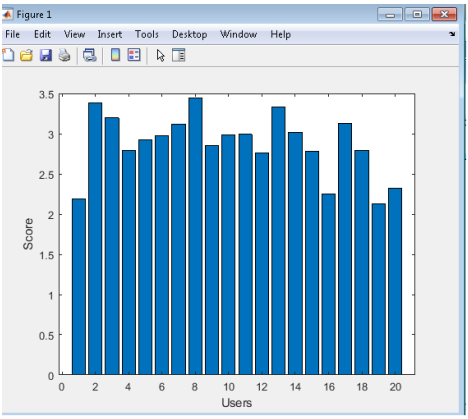
Bu verilənlər əsasında hesablamalar aparılmış, e-xidmətlərdən məmnunluq reytinqi və istifadəçilərin reytinqi müəyyən olunmuşdur (şəkil 1 və 2). Qeyd edək ki, burada istifadəçilərin reytinqi onların xidmətlərə olan müraciətlərinin sayı əsasında təyin olunur və bundan regionların xidmətlərə olan maraq dairələrinin təyin olunmasında istifadə olunur. Daha sonra bu hesablamalardan istifadə edərək e-sistem ümumi şəkildə qiymətləndirilmişdir:

$$U^{savg} = \frac{1}{20} * (2.1939 + 3.3864 + ... + 2.3173) = 2.8687$$

Sistemdən ümumi məmnunluq dərəcəsi $(U^{savg} = 2.8687)$ $[2,4)$ intervalına düşdüyündən sistemin fəaliyyətini "qənaətbəxş" hesab etmək olar.



Şəkil 1. E-xidmətlərdən məmnunluq reytinqi



Şəkil 2. İstifadəçilərin reytinqi

Eyni regiondan olan istifadəçilərin ən çox müraciət etdiyi xidmətləri təyin etdikdən sonra ranqlaşdırma vasitəsilə regionların əsas maraq dairəsində olan (hotspot) xidmətlər təyin olunmuşdur (cədvəl 3). Qeyd edək ki, cədvəldə *M.sayı* ilə müraciət sayı işarə olunmuşdur. Müraciət sayı dedikdə istifadəçilərin hər bir xidmətə olan müraciət sayından max olanı götürülmüşdür.

Beləliklə, xidmətlərin məmnunluq reytinqi, istifadəçilərin bütün xidmətlərdən ümumi məmnunluq reytinqi, bütün xidmətlərdən ancaq məmnun olan və olmayan regionlar, bu regionların maraq dairələri və e-sistem ümumi şəkildə qiymətləndirilmişdir.

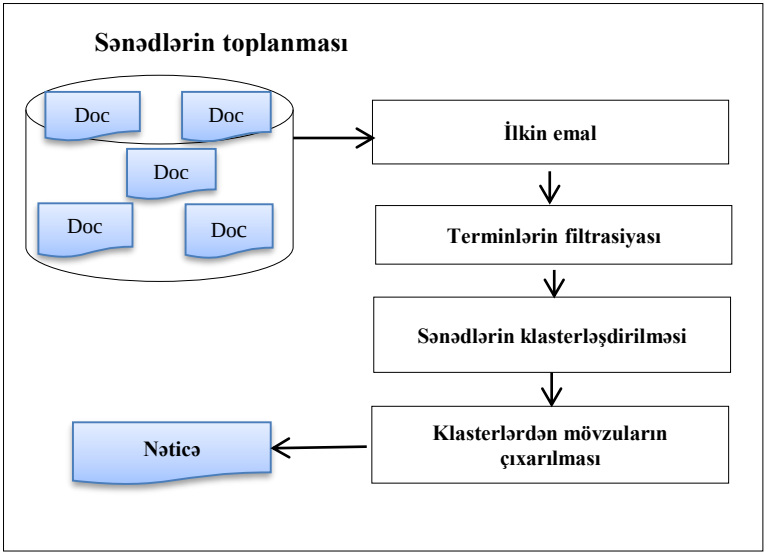
Cədvəl 3

Regionlar üzrə hotspot xidmətlər

E-xidmətlər	***Region 1***		***Region 2***		***Region 3***		***Region 4***		***Region 5***	
	M.sayı	***Ranq***	***M.sayı***	***Ranq***	***M.sayı***	***Ranq***	***M.sayı***	***Ranq***	***M.sayı***	***Ranq***
EG_1	46	*3*	48	*1*	49	*1*	49	*1*	46	*3*
EG_2	37	*4*	48	*2*	36	*4*	47	*2*	41	*4*
EG_3	22	*5*	36	*4*	34	*5*	40	*4*	48	*1*
EG_4	47	*2*	38	*3*	42	*2*	41	*3*	48	*2*
EG_5	47	*1*	19	*5*	38	*3*	29	*5*	29	*5*

Məlum olduğu kimi, e-dövlət mühitində vətəndaşlar hər hansı xidmətə şərhlər yazmaqla münasibət bildirə bilərlər. Bu şərhləri analiz

etməklə, onların narahat olduğu əsas məsələləri müəyyən etmək olar. Məlumdur ki, şərhlərin sayı artdıqca onları analiz etmək də çətinləşir. Vətəndaşların narahat olduğu əsas məqamları tez bir şəkildə müəyyən etmək üçün text mining metodlarının tətbiqi vacibdir. Bunları nəzərə alaraq, **dördüncü fəslin ikinci bölməsində** e-vətəndaşları maraqlandıran aktual mövzuların müəyyən olunması üçün yanaşma təklif olunmuşdur. Təklif olunan yanaşmada Gizli Dirixle Paylanması (GDP) və k-means alqoritmlərindən istifadə olunmuşdur [12, 14]. Təklif olunan yanaşmanın əsas addımları şəkil 3-də təsvir olunmuşdur:



Şəkil 3. Təklif olunan yanaşmanın sxemi

Təklif olunan yanaşmanın hər bir addımı aşağıda ətraflı şəkildə şərh olunmuşdur:

Addım 1. Əvvəlcə e-dövlət mühitində istifadəçi şərhləri toplanılır. Sadəlik üçün bu şərhlərə sənəd kimi baxılır:

$$D = \{d_1, d_2, ..., d_n\}$$

burada n – sənədlərin (şərhlərin) sayıdır.

Addım 2. Toplanan şərhlər ilkin emal olunur. İlkin emal zamanı sənədlərdən ümumişlək sözlər, rəqəmlər və durğu işarələri təmizlənir.

Hər bir söz müxtəlif formalarda şəkilçilər qəbul etdiyi üçün onlar ilkin variantına (kökünə) qaytarılır.

Addım 3. Şərhlərdən terminlər çıxarılır. Daha sonra sənədlər çoxluğu "Term Frequency-Inverse Document (TF-IDF)" sxeminin köməyilə vektor kimi təsvir olunur.

Sənədlər arasındakı məsafəni hesablamaq üçün Evklid məsafəsindən istifadə olunur.

Məlumdur ki, sənədlər çoxluğunda rast gəlinən terminlərin sayı həddən artıq çox olur və bu say bir sənəddə rast gəlinən terminlərin sayından çox-çox böyük olur. Onda sənədlərin TF-IDF sxemi ilə təsvir olunan vektorlarının elementlərinin çox hissəsi "0"-lar olacaq. Başqa sözlə vektorlar seyrək olacaq. Bu isə sənədlərin klasterləşməsində iki mühüm problem yaradır:

- "Lənətə gəlmiş" ölçü problemi;
- Klasterləşmənin keyfiyyəti.

Bu problemləri aradan qaldırmaq üçün vektordan seyrək terminlər əvvəlcədən təmizlənir. Seyrək terminlər təmizləndikdən sonra yuxarıda göstərilən problemlərə təsir edən digər bir amil ortaya çıxır. Bu da sənədlər çoxluğunda sinonim sözlərin olmasıdır. Belə ki, sənədlər çoxluğunda sinonim sözlər olduqda klasterləşmə zamanı oxşar məzmunlu sənədlər müxtəlif klasterlərə düşə bilər. Bu isə klasterləşmənin keyfiyyətinin aşağı düşməsinə gətirib çıxarır. Bu kimi halları aradan qaldırmaq üçün sənədlər çoxluğunda semantik yaxın sözlərin tapılması, onlardan birinin saxlanılması və digərlərinin kənarlaşdırılması təklif olunur. Sözlərin bir-birinə semantik yaxınlığını tapmaq üçün hər bir terminin genişlənmiş sinonimləri çoxluğundan istifadə olunması təklif olunur. Bunun üçün WordNet şəbəkəsindən istifadə etməklə hər bir terminin sinonimlər çoxluğu tapılır və $t_i \rightarrow synset\ (t_i)$ ilə işarə olunur.

Sözlər arasında semantik yaxınlıq aşağıdakı metrikadan istifadə etməklə hesablanır:

$$sim\,(t_g, t_s) = \frac{2\left|synset\ (t_g) \cap synset\ (t_s)\right|}{\left|synset\ (t_g) \cup synset\ (t_s)\right|} \geq \alpha, \quad g, s = 1,2,...,m$$

burada $\left|synset\ (t)\right|$ – t sözünün sinonimlərinin sayı, $0 \le \alpha \le 1$ – idarəolunan parametrdir. Əgər sözlər arasında yaxınlıq α ədəddindən böyükdürsə, bu sözlər bir termin kimi qəbul olunur. Belə ki, bu sözlərdən yalnız biri saxlanılır, digərləri kənarlaşdırılır. Beləliklə, d vektorunun ölçüsünü azaltmış oluruq. Bu halda d_i vektoru aşağıdakı vektora transformasiya olunur:

$$d_i \to d_i^* = \left\{\overline{w}_{i1}, \overline{w}_{i2}, ..., \overline{w}_{im_0}\right\}, \quad m_0 \le m$$

burada $\overline{w}_{ij}$ – sinonim sözlər kənarlaşdırıldıqdan sonra j-ci sözün i-ci sənəddəki TF-IDF çəkisidir.

Addım 4. Sənədlər vektor şəklində təsvir olunduqdan sonra klasterləşdirilir. Sənədləri klasterlərə ayırmaq üçün bir sıra metodlar mövcuddur. Bu məqalədə sənədləri klasterləşdirmək üçün k-means metodundan istifadə olunması təklif olunur. k-means böyük verilənlərin analizində icra müddətinin az və tətbiqinin asanlığı səbəbindən populyar alqoritmlərdən biri hesab olunur.

Addım 5. Sənədlər klasterlərə ayrıldıqdan sonra hər bir klaster üzrə mövzuları tapmaq üçün GDP-dən istifadə olunması təklif olunur. GDP vasitəsilə hər bir klaster üzrə sənədlərdən əsas mövzuların çıxarılması aşağıdakı şəkildə həyata keçirilir.

Tutaq ki, $\{C_1, C_2, ..., C_k\}$ klasterləri təyin olunmuşdur. Hər bir klasterə GDP alqoritmi tətbiq olunur və hər bir C_q klasteri üçün $T_q = \{T_{q1}, T_{q2}, ..., T_{qs}\}$ mövzuları təyin olunmuş olur. Burada s – mövzuların sayıdır.

Beləliklə, vətəndaşların yazdığı şərhlərin əsas mövzusunu müəyyən etmiş oluruq.

Təklif olunan metodu qiymətləndirmk üçün eksperiment *R* proqramlaşdırma dilində aparılmışdır. Eksperiment üçün BBC NEWS verilənlər çoxluğundan istifadə olunmuşdur. Bu verilənlər çoxluğu 2004-2005-ci illəri əhatə edən Biznes, Əyləncə, Siyasət, İdman və Texnologiya adlanan beş aktual sahəyə uyğun BBC xəbər veb saytından toplanmış 2225 sənəddən ibarətdir. Eksperimentdə Biznes, Əyləncə və İdman sahələrindən müxtəlif sayda sənədlər toplanmış və

analiz olunmuşdur. Klasterləşmə nəticələrini qiymətləndirmək üçün "təmizlik" (purity) əmsalından istifadə olunmuşdur.

İlkin emaldan əvvəl və sonra sənədlər çoxluğunda qalan sözlərin sayı cədvəl 4-də təsvir olunmuşdur.

Cədvəl 4

Sənədlər və sözlərin sayı

Sənədlərin sayı	**Sözlərin sayı**	
	İlkin emaldan əvvəl	**İlkin emaldan sonra**
100	8040	4421
300	18356	8851
500	25490	11766
800	33410	14750
1000	36346	15860

İlkin emaldan sonra sənədlər çoxluğuna təklif etdiyimiz metod tətbiq olunmuşdur. Sözlər arasında semantik yaxınlıq α –nın müxtəlif qiymətlərində (0.1, 0.2, 0.3, 0.4, 0.5) hesablanmışdır. Metoddan sonra sənədlər çoxluğunda qalan terminlərin sayı cədvəl 5-də təsvir olunmuşdur. Cədvəldə sənədlərin sayı 100 olan varianta baxdıqda, buradan aydın görünür ki, $\alpha = 0.1$ qiymətində daha çox sayda semantik yaxın sözlər aşkarlanmış və vektorun ölçüsü seyrək terminlər təmizləndikdən sonra qalan sözlərə nisbətən xeyli dərəcədə (26.42%) azalmışdır. α –nın qiyməti artdıqca, daha az sayda sözlər atılmışdır. Belə ki, $\alpha = 0.5$ qiymətinə baxdıqda, vektorun ölçüsünün daha az (1.29%) azaldığını görərik. Sənədlərin sayı artdıqca, semantik yaxın olan sözlərin sayı da uyğun olaraq artmış və vektorun ölçüsü xeyli dərəcədə azalmışdır. Məsələn, sənədlərin sayı 800 olan varianta baxsaq, $\alpha = 0.1$ qiymətində vektorun ölçüsünün 31.67% azaldığını görərik.

Daha sonra sənədlər çoxluğuna k-means klasterləşmə metodu tətbiq olunmuş və klasterləşmənin dəqiqliyi cədvəl 6-da təsvir olunmuşdur. Qeyd edək ki, burada $\alpha = 0$ qiyməti seyrək terminlər təmizləndikdən sonra qalan terminlər çoxluğunu ifadə edir. Cədvəldən göründüyü kimi terminlər çoxluğundan semantik yaxın sözlərin atılması klasterləşmənin keyfiyyətinə mənfi təsir göstərməmiş, əksinə təmizlik əmsalı kifayət qədər yüksək qiymət almışdır. Sənədlərin sayı artdıqca, təmizlik əmsalının qiyməti də kifayət qədər yüksək olmuşdur.

Cədvəl 5

Emaldan sonra sənədlərdəki sözlərin sayı

α =	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
Sənədlərin sayı	**Sözlərin sayı**					
	Seyrək terminlər təmizləndikdən sonra	**Semantik yaxın sözlər kənarlaşdırıldıqdan sonra**				
100	772	568 (26.4%)	691 (10.4%)	733 (5.05%)	756 (2.07%)	762 (1.29%)
300	878	633 (27.90%)	777 (11.50%)	838 (4.56%)	856 (2.51%)	865 (1.48%)
500	827	589 (28.77%)	725 (12.33%)	789 (4.59%)	809 (2.17%)	816 (1.33%)
800	821	561 (31.67%)	716 (12.79%)	778 (5.23%)	802 (2.31%)	810 (1.34%)
1000	801	561 (29.96%)	702 (12.36%)	765 (4.49%)	784 (2.12%)	791 (1.25%)

Cədvəl 6

α -nın müxtəlif qiymətlərində klasterləşmənin təmizlik (purity) əmsalı

α =	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
Sənədlərin sayı	**Purity**					
100	0.95	0.88	0.81	0.82	0.82	0.88
300	0.996	0.98	0.98	0.98	0.98	0.98
500	0.996	0.93	0.94	0.99	0.99	0.99
800	0.98	0.81	0.89	0.99	0.99	0.99
1000	0.982	0.97	0.98	0.97	0.98	0.98

Sənədlər klasterləşdirildikdən sonra hər bir klasterdən mövzuları çıxarmaq üçün GDP tətbiq olunmuşdur. Cədvəl 7 və 8-də hər bir klasterdən çıxarılan top 10 söz təsvir olunmuşdur.

Göründüyü kimi mövzular dəqiqliklə çıxarılmış, vaxtda isə uduş əldə olunmuşdur. Belə ki, cədvəl 9-da klasterləşməyə və hər bir klasterdən mövzuların çıxarılmasına sərf olunan zaman və onların müqayisəli analizi təsvir olunmuşdur.

Cədvəl 7

Hər bir klaster üzrə top 10 söz ($\alpha = 0$)

Klaster 1	**Klaster 2**	**Klaster 3**
music	ireland	cluster
award	england	growth
people	play	year
show	win	rate
year	wale	economi
won	side	bank
radio	game	econom
veto	beat	oil
years	nation	price
song	scotland	rise

Cədvəl 8

Hər bir klaster üzrə top 10 söz ($\alpha = 0.3$)

Klaster 1	**Klaster 2**	**Klaster 3**
film	england	cluster
star	play	team
year	cluster	year
award	year	rate
role	rugbi	economi
cluster	player	rise
includ	game	price
director	season	bank
play	cup	econom
bbc	week	month

Beləliklə, aparılan eksperiment və nəticələri onu göstərir ki, təklif olunmuş metod vasitəsilə böyük sənədlər çoxluğunun ölçüsünü xeyli dərəcədə azaldaraq, verilənlərin analizinə sərf olunan vaxta qənaət etmək, klasterləşmə və GDP alqoritminin keyfiyyətini yaxşılaşdırmaq olar.

Cədvəl 9

Sənədlərin klasterləşməsi və GDP-in tətbiqinə sərf olunan vaxt

α =	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
Sənədlərin sayı	**Sərf olunan zaman**					
100	11.6	9.3 (19.82%)	9.86 (15%)	10.32 (11.04%)	10.65 (8.19%)	11.06 (4.65%)
300	12.39	9.08 (26.71%)	9.74 (21.39%)	10.68 (13.80%)	11.26 (9.12%)	11.51 (7.10%)
500	18.7	11.35 (39.30%)	12.87 (31.18%)	13.28 (28.98%)	14.21 (24.01%)	14.98 (19.89%)
800	22.28	13.49 (39.45%)	14.59 (34.52%)	15.76 (29.26%)	16.57 (25.63%)	17.85 (20.33%)
1000	21.06	12.78 (39.31%)	14.09 (33.09%)	15.01 (28.72%)	16.01 (23.97%)	17.25 (18.09%)

NƏTİCƏLƏR

1. E-dövlətin analizində sosial şəbəkə və mətnlərin intellektual analizi texnologiyalarının rolu araşdırılmış, o cümlədən e-dövlətə qoyulan tələblər və onun inkişafına yanaşmalar, e-dövlətin inkişafının və analizinin əsas problemləri müəyyən olunmuşdur [4-6, 8];
2. E-dövlətdə terrorizmlə əlaqəli məqalələrin aşkarlanması üçün hibrid təsnifatlandırma metodu işlənmişdir [2, 3];
3. E-dövlətdə terrorizmlə bağlı mətnlərin filtrasiyası üçün sentiment analiz texnologiyasına və Bayes klassifikatoruna əsaslanan metod işlənmişdir [9, 10];
4. E-dövlətdə fəaliyyət göstərən gizli sosial şəbəkələrin aşkarlanması və analizi üçün sentiment analiz texnologiyasına əsaslanan metod və alqoritm işlənmişdir [1, 11];
5. E-dövlət xidmətlərindən vətəndaş məmnuniyyətinin avtomatik qiymətləndirilməsi üçün metod işlənmişdir [7, 12];
6. E-dövlətdə vətəndaşları (o cümlədən regionları) maraqlandıran aktual mövzuların müəyyən edilməsi üçün metod işlənmişdir [13, 14].

DİSSERTASİYA MATERİALLARI ÜZRƏ AŞAĞIDAKI ELMİ ƏSƏRLƏR ÇAP EDİLMİŞDİR:

1. Alıquliyev, R.M., Niftəliyeva, G.Y. E-dövlət mühitində gizli sosial şəbəkələrin aşkarlanmasi üçün yanaşma / "İnformasiya təhlükəsizliyinin multidissiplinar problemləri" üzrə II respublika elmi-praktiki konfransı, – Bakı, –14 may 2015, –s.116-118.
2. Alıquliyev, R.M., Niftəliyeva, G.Y. E-dövlət mühitində terrorizmlə əlaqəli mətnlərin aşkarlanması metodu / "İnformasiya təhlükəsizliyinin multidissiplinar problemləri" üzrə II respublika elmi-praktiki konfransı, – Bakı, – 14 may 2015, – s.111-115.
3. Aliguliyev, R.M., Niftaliyeva, G.Y. Detecting terrorism-related articles on the e-government using text-mining techniques // Problems of Information Technology,–2015, 6 (2), –p. 36-46.
4. Alıquliyev, R.M., Niftəliyeva, G.Y. E-dövlət sisteminin analizində Data mining texnologiyalarının tətbiq imkanları / "Big Data: imkanları, multidissiplinar problemləri və perspektivləri" I respublika elmi praktiki konfransı, –Bakı, 25 fevral 2016, – s. 81-84.
5. Alıquliyev, R.M., Niftəliyeva, G.Y. E-Dövlətin Big Data Mənbələri // "Big Data: imkanları, multidissiplinar problemləri və perspektivləri" I respublika elmi praktiki konfransı, – Bakı, 25 fevral 2016, – s. 78-80.
6. Alıquliyev, R.M. E-dövlətin analizi texnologiyları: text mining və sosial şəbəkələr. Ekspress-informasiya. "İnformasiya Texnologiyaları seriyası" / R.M. Alıquliyev, G.Y. Niftəliyeva – Bakı: –2016. –78 s.
7. Aliguliyev, R.M., Niftaliyeva, G.Y. Hotspot Information of Public Opinion in E-Government / 10th IEEE International Conference on Application of Information and Communication Technologies (AICT2016), –Baku, –12-14 October, 2016, – p.645-646.

8. Aliguliyev, R.M., Niftaliyeva, G.Y. The current state, problems and perspectives of e-government analysis technologies // Problems of Information Technology, – 2017, 8 (1), –p. 53-63.
9. İskəndərli, G.Y. E-dövlətə kiber hücumlar və onlarla mübarizə üsulları haqqında / "İnformasiya təhlükəsizliyinin aktual multidissiplinar problemləri" üzrə IV respublika elmi-praktiki konfransı, – Bakı, – 14 dekabr 2018, – s.158-160.
10. Alguliyev, R. M., Aliguliyev, R. M. Niftaliyeva, G. Y. Filtration of Terrorism-Related Texts in the E-government Environment // International Journal of Cyber Warfare and Terrorism, –2018, 8 (4), p.35-48. **(Web of Science)**
11. Alguliyev, R. M., Aliguliyev, R. M., Niftaliyeva, G. Y. A Method for Social Network Extraction From E-Government // International Journal of Information Systems in the Service Sector, –2019, 11 (3), p.37-55. **(Web of Science)**
12. Iskandarli G.Y. Using Hotspot Information to Evaluate Citizen Satisfaction in E-Government: Hotspot Information // International Journal of Public Administration in the Digital Age, –2020, 7 (1), –p. 47-62. **(Web of Science)**
13. Iskandarli G.Y. Applying Clustering and Topic Modeling to Automatic Analysis of Citizens' Comments in E-Government // International Journal of Information Technology and Computer Science, – 2020, 12 (6), –p.1-10.
14. Iskandarli G.Y. Detecting the Main Topics of Citizens' Comments in e-Government / 2nd International Symposium on Applied Sciences and Engineering,– Atatürk University, – Erzurum, –Turkey, –7-9 april, 2021, –pp.581-584.

Dissertasiyanın müdafiəsi **24 sentyabr 2021**-ci il tarixdə saat **14^{00}** – da AMEA İnformasiya Texnologiyaları İnstitutunun nəzdində fəaliyyət göstərən ED 1.35 Dissertasiya şurasının iclasında keçiriləcək.

Ünvan: Az 1141, Bakı şəhəri, B.Vahabzadə küçəsi, 9a

Dissertasiya ilə AMEA İnformasiya Texnologiyaları İnstitutunun kitabxanasında tanış olmaq mümkündür.

Dissertasiya və avtoreferatın elektron versiyaları AMEA İnformasiya Texnologiyaları İnstitutunun rəsmi internet saytında yerləşdirilmişdir.

Avtoreferat **22 iyul 2021**-ci il tarixində zəruri ünvanlara göndərilmişdir.

Çapa imzalanıb: 16.07.2021

Kağızın formatı: $60 \times 80^{1}/_{16}$

Həcm: 39098 işarə

Tiraj: 100 nüsxə