E-dövlət mühitində terrorizmlə əlaqəli mətnlərin aşkarlanması metodu

Ramiz Alıquliyev[1], Günay Niftəliyeva[2]

AMEA İnformasiya Texnologiyaları İnstitutu

[1]r.aliguliyev@gmail.com; [2]gunay90@hotmail.com

***Xülasə*—Məqalədə e-dövlətin informasiya təhlükəsizliyinin təmin olunması üçün metod təklif olunmuşdur. Təklif edilmiş metod e-dövlətdə terrorizmi təbliğ edən mətnlərin aşkarlanmsına və analizinə imkan verir. Bu metod text mining texnologiyalarına əsaslanır. Metodun effektivliyini artırmaq üçün baxılan mövzu (terrorizm) üzrə lüğət bazası və sözlərin semantik şəbəkəsindən istifadə olunması təklif olunur. Mətnlər arasında yaxınlığı müəyyən etmək üçün hibrid metrika və yeni təsnifatlandırma metod təklif edilmişdir.**

Keywords—e-dövlət; e-dövlətin təhlükəsizliyi; terrorizm; text mining; təsnifatlandırma; hibrid yaxınlıq ölçüsü

I. GİRİŞ

Müasir dövrdə kriminal qruplar təkcə real aləmdə deyil, həm də virtual mühitdə (İnternet, e-dövlət) də dövlət və cəmiyyət əleyhinə öz bədniyyətli fəaliyyətlərini həyata keçirirlər. Bu fəaliyyət növləri müxtəlif məqsədli olur: dövlət əleyhinə təbliğat, mentalitetə uyğun gəlməyən, milli mənəvi dəyərlərin əsaslarını sarsıdan, terrorizmi təbliğ edən informasiyanın yayılması və s. [5-7, 9, 10, 17, 18].

E-dövlət mühitində bu məzmunda informasiyanın vaxtında aşkarlanması dövlətin və cəmiyyətin təhlükəsizliyinin təmin olunması baxımından mühüm əhəmiyyət kəsb edir və günümüzün ən aktual elmi-nəzəri və praktiki problemlərindən biridir [17, 18]. Heç də təsadüfi deyildir ki, e-dövlətin təhlükəsizliyi problemi Avropa Komissiyası tərəfindən qəbul edilmiş eGovRTD2020 layihəsində e-dövlət sahəsində araşdırılması vacib olan 13 ən aktual elmi-tədqiqat istiqamətindən biri kimi qeyd olunmuşdur [15].

E-dövlətin əsas funksiyalarından biri vətəndaşları ehtimal olunan zərər və zorakılıqlardan qorumaqdır. Linders [13] vətəndaş-dövlət münasibətlərinin təkamülünü araşdıraraq, belə qənaətə gəlmişdir ki, ehtimal olunan cinayətlər haqqında əvvəlcədən məlumat vermək, o cümlədən cəmiyyət üzvləri ilə hüquq-mühafizə orqanları arasındakı münasibətlərin yaxşılaşdırılması baxımından İnternet, xüsusi halda e-dövlət ən effektiv və əlverişli vasitədir. 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. Buna misal olaraq, 11 sentyabr 2011-ci il tarixində ABŞ-da həyata keçirilmiş terror hücumunu göstərmək olar. Terror hadisəsindən sonrakı təhlillər göstərdi ki, bu aktı həyata keçirən mütəşəkkil cinayətkar qrup bütün plan və fəaliyyətlərini İnternet şəbəkəsindən istifadə etməklə hazırlamış və koordinasiya etmişlər. Belə demək mümkünsə, virtual aləm cinayətkar qruplara öz əməllərini həyata keçirmək üçün çox əlverişli mühitdir.

Deməli, dövlətin mühüm vəzifələrindən biri də virtual mühitdə – İnternetdə və e-dövlətdə gizli fəaliyyət göstərən kriminal şəbəkələrin fəaliyyətini aşkarlamaq və analiz etməkdir. Bu mühit tez kommunikasiya yaratmaq və fəaliiyəti 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 edirlər. 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 [20]. Hal-hazırda biliklərin idarə olunmasında, müxtəlif mənbələrdə toplanmış mətnlərin intellektual analizində text mining ən qabaqcıl və effektiv texnologiyalardan biri hesab olunur [2].

Text mining texnologiyalarının belə populyar və tətbiq sahəsinin geniş olmasının digər səbəblərindən biri də hansı mühitdə (real və ya virtual) istehsal olunmasından asılı olmayaraq mətnlərin üstünlük təşkil etməsidir. Bütün istehsal olunan informasiyanın 80%-dən çoxunu mətnlər təşkil edir. Deməli, e-dövlətin təhlükəsizliyinin təmin olunması baxımından bü mühitdə dövr edən mətnlərin intellektual analizi çox aktual məsələdir

Beləliklə, problemin aktuallığını əsas tutaraq, məqalədə e-dövlətdə şübhəli (terrorizmlə əlaqəli) mətnlərin aşkarlanmsı üçün text mining texnologiyalarına əsaslanan metod təklif olunur. Bu metod [7]-də təklif olunmuş metoda oxşardır. Lakin təklif olunan metod bir neçə fərqli və üstün cəhətlərə malikdir:

- [7]-də təklif olunmuş metoddan fərqli olaraq, bu metodda sözlər arasındakı yaxınlığı hesablayarkən nəinki onlar arasındakı semantik yaxınlıq, həm də cümlənin sintaktik quruluşu, daha doğrusu sözlərin cümlədəki işlənmə ardıcıllığı nəzərə alınır;
- potensial təhlükəli sənədləri daha dəqiq aşkarlamaq üçün sənədlər arasındakı yaxınlıq yeni iterativ üsulla hesablanır: əvvəlcə sözlərin yaxınlığı təyin edilir; sonra sözlər arasındakı yaxınlıqdan istifadə etməklə cümlələrin yaxınlığı hesablanır; nəhayət, cümlələr arasındakı yaxınlıqdan istifadə olunmaqla sənədlər arasındakı yaxınlıq hesablanır.
- cümlələr arasında yaxınlığı hesablamaq üçün hibrid yaxınlıq ölçüsü daxil edilir;
- Təsnifatlandırma üçün yeni metod təklif olunmuşdur.

Məqalə aşağıdakı kimi strukturlaşdırılmışdır. Tədqiq olunan problemlə əlaqəli işlərin qısa icmalı ikinci bölmədə verilir. Üçüncü bölmədə təklif olunan metodun mərhələlərinin təsviri verilir. Yekun və gələcək tədqiqatlar barədə məlumat isə dördüncü bölmədə verilmişdir.

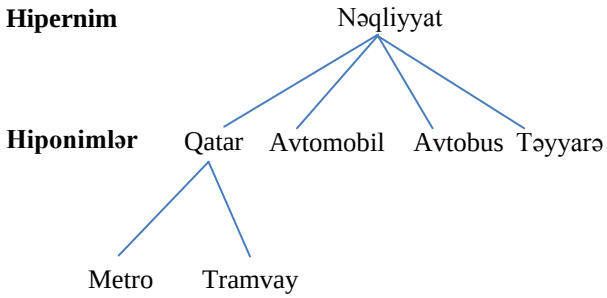
II. ƏLAQƏLİ İŞLƏRİN QISA İCMALI

Virtual mühitdə (İnternetdə, e-dövlətdə) kriminal və terrorizmlə bağlı informasiyanın aşkarlanması, identifikasiyası və izlənməsi üçün text mining texnologiyasına əsaslanan müxtəlif metodlar, alqoritmlər və modellər təklif edilmişdir. Məsələn, veb-də kriminal informasiyanın filtrasiyası və identifikasiyası məqsədilə sənədlər arasındakı oxşarlığı müəyyən etmək üçün [9, 10]-da yeni alqoritmlər təklif edilmişdir. Ərəb dilində kriminal sənədlərin identifikasiyası sistemi üçün [5]-də text mining texnologiyasının informasiyanın çıxarılması və klasterləşdirmə metodlarından istifadə olunmuşdur. İnformasiyanın çıxarılması üçün qaydalara əsaslanan yaxınlaşma, sənədlərin klasterləşdirilməsi üçün isə özü-özünə təşkil olunan neyron şəbəkə (Kohonen şəbəkəsi) tətbiq olunmuşdur. Kriminal sənədlərin tipinin identifikasiyası üçün [6]-də iki mərhələdən, sənədlərin aşkarlanması və onların klasterləşdirilməsindən ibarət metod təklif olunmuşdur. Birinci mərhələdə sənədlər əhəmiyyətsiz sözlərdən təmizlənir, sonra sənədləri əhəmiyyətli sözlərin vektoru kimi təsvir edib, onlar arsındakı yaxınlığı hesablamaq üçün metrika daxil edilir. İkinci mərhələdə klasterləşdirmə alqoritmini tətbiq etməklə sənədlər kriminal tiplərə görə qruplaşdırılır. İnternetdə terrorizmlə əlaqəli məqalələri aşkarlamaq üçün [7]-də mətnlərin analizinə əsaslanan yeni yanaşma təklif olunmuşdur. Bu yanaşma WordNet semantik şəbəkəsindən [14] istifadə etməklə terrorizmlə əlaqəli məqalələr çoxluğundan kontekst sözlərin (isimlərin) siyahısını yaradır. Sonra WUP [16] metrikasını tətbiq etməklə kontekst sözlərin əhəmiyyətlilik dərəcəsini hesablayır. Sonda isə biqramlardan və Keselj metrikasından [8] istifadə etməklə sənədləri təsnifatlandırır.

III. TƏKLİF OLUNAN METOD

Təklif olunan metod bir neçə mərhələdən ibarətdir: 1) tədqiq olunan mühitdə dövr edən sənədlərin (informasiyanın) dilindən asılı olaraq, həmin dil üçün terrorizmlə bağlı lüğət bazasının yaradılması; 2) baxılan dil üçün sözlərin semantik şəbəkəsinin yaradılması (metodun dəqiqliyi bu şəbəkədən çox asılıdır); 3) sözlərin morfoloji təhlili; 4) lügət bazasından istifadə etməklə sənədlərin ilkin filtrasiyası; 5) sözlər arasında semantik yaxınlığın hesablanması; 6) cümlələr arasında semantik yaxınlığın müəyyən edilməsi; 7) sənədlər arasında semantik yaxınlığın müəyyən edilməsi; 8) sənədin əvvəlcədən məlum olan hər hansı bir sinfə aid edilməsi (təsnifatlandırma).

Tutaq ki, tədqiq olunan mühitin dili üçün baxılan mövzu (terrorizm) ilə bağlı lüğət bazası (VBase) yaradılmış və sözlərin semantik şəbəkəsi qurulmuşdur (ingilis dilində yaradılmış şəbəkəyə oxşar olaraq bu şəbəkəni WordNet ilə işarə edək). Qeyd etmək lazımdır ki, bu biliklər bazası sözlər arasındakı semantik münasibətləri müəyyən etməyə imkan verir. Məsələn, bu şəbəkənin köməyilə sinonimləri, hipernimləri, hiponimləri və s. asanlıqla tapmaq mümkündür (Şəkil 1).



Şəkil 1. Hipernim və hiponimlər

Təklif olunan yanaşmanın hər bir mərhələsi aşağıda ətraflı izah edilir.

A. *Sənədlərin İlkin Filtrasiyası*

Sənədlərin ilkin filtrasiyası aşağıdakı qaydada həyata keçirilir. Əvvəlcə sənəddən terminlər çıxarılır, onlar morfoloji təhlil edilir (bu sözün başlanğıc formasını tapmaq üçündür, çünki eyni bir söz qəbul etdiyi şəkilçilərdən asılı olaraq müxtəlif formalarda olur) və sənəd terminlər çoxluğu kimi təsvir olunur, $D = (t_1, t_2, ..., t_m)$. Sonra Şimkeviç-Simpson ölçüsündən [21] istifadə edərək Vbase bazası ilə bu vektor arsındakı yaxınlıq ölçülür:

$$\text{sim}_{S-S}(D, \text{Vbase}) = \frac{|D \cap \text{Vbase}|}{|D|}, \quad (1)$$

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

Əgər $sim_{S-S}(D, \text{Vbase}) \geq \varepsilon$ olarsa, onda D sənədi şübhəli sənədlər çoxluğuna əlavə edilir və identifikasiya üçün növbəti mərhələyə keçid edilir. Burada ε eksperimental yolla müəyyən edilmiş limit qiymətidir.

B. *Sözlərin Semantik Yaxınlığı*

Sözlər arasındakı semantik yaxınlıq aşağıdakı ardıcıllıqla təyin edilir:

1. İki söz w_1 və w_2 götürülür.
2. WordNet semantik şəbəkəsindən bu sözlərin kökü tapılır.
3. WordNet leksik bazasındann hər bir sözün sinonimləri və onların sayı təyin edilir;
4. Sözlərin uzunluğu və WordNet şəbəkəsində istifadə etməklə onların ən yaxın ortaq (Least Common Subsume – LCS) kökü tapılır;
5. (2) və (3) düsturlarının köməyilə sözlər arasındakı semantik yaxınlıq hesablanır.

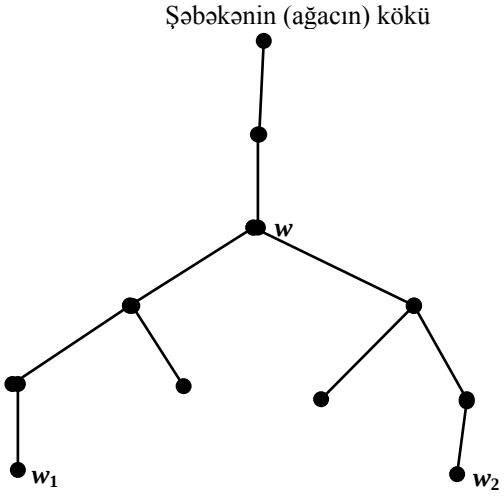
Sözlər arasındakı semantik yaxınlığı hesablamaq üçün əvvəlcə WordNet şəbəkəsindən istifadə etməklə, sözün informativ məzmunu (yükü) $IC(w)$ təyin edilir [1]:

$$IC(w) = 1 - \frac{\log(\text{synset}(w) + 1)}{\log(\max\_w)}. \quad (2)$$

Sonra (2) düsturundan istifadə edərək sözlər arasındakı semantik yaxınlıq hesablanır [1, 12]:

$$\text{sim}_{\text{IC}}(w_1, w_2) = \begin{cases} \dfrac{2 * \text{IC}(\text{LCS}(w_1, w_2))}{\text{IC}(w_1) + \text{IC}(w_2)}, & w_1 \neq w_2 \\ 1, & w_1 = w_2 \end{cases} \quad (3)$$

burada $\text{LCS}(w_1, w_2)$ – WordNet şəbəkəsində w_1 və w_2 sözlərinin ən yaxın olduğu ortaq söz (məsələn, şəkil 2-də göstərilən hal üçün $\text{LCS}(w_1, w_2) = w$), $\max\_w$ – WordNet semantik şəbəkəsindəki sözlərin ümumi sayı, $\text{synset}(w)$ – w sözünün sinonimlərinin sayıdır.



Şəkil 2. Sözlərin semantik şəbəkəsi

Sözlər arasındakı semantik yaxınlığı həm də WUP metrikasından [16] istifadə etməklə hesablayırıq:

$$\text{sim}_{\text{WUP}}(w_1, w_2) = \frac{2 * \text{depth}(w)}{\text{depth}(w_1) + \text{depth}(w_2) + 2 * \text{depth}(w)}, \quad (4)$$

burada $\text{depth}(w_1)$ – WordNet şəbəkəsində (ağacında) w_1-dən w-yə qədər olan qovşaqların sayı; $\text{depth}(w_2)$ – w_2-dən w-yə qədər olan qovşaqların sayı; $\text{depth}(w)$ – w-dən şəbəkənin kökünə qədər qovşaqların sayıdır. Şəkil 2-də göstərilən hal üçün $\text{depth}(w_1) = \text{depth}(w_2) = 3$ və $\text{depth}(w) = 2$. Onda

$$\text{sim}_{\text{WUP}}(w_1, w_2) = \frac{2 * 2}{3 + 3 + 2 * 2} = 0.4.$$

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

$$\begin{aligned} \text{sim}(w_1, w_2) = \alpha \times \text{sim}_{\text{IC}}(w_1, w_2) + \\ + (1 - \alpha) \times \text{sim}_{\text{WUP}}(w_1, w_2), \end{aligned} \quad (5)$$

burada $0 \le \alpha \le 1$ – çəki parametridir.

C. Cümlələrin Yaxınlıq Ölçüsü

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

D. Semantik Yaxınlıq

Cümlələr arasındakı semantik yaxınlıq sözlər arasındakı semantik yaxınlıqdan (5) istifadə edilərək hesablanır:

$$\text{sim}_{\text{semantic}}(S_1, S_2) = \frac{\sum_{w_1 \in S_1, w_2 \in S_2} sim(w_1, w_2)}{m_1 + m_2}, \quad (6)$$

burada m_1 və m_2 uyğun olaraq S_1 və S_2 cümlələrindəki sözlərin sayıdır.

E. Kosinus Metrikası

Kosinus metrikası vektor modelinə əsaslanan metrikadır. Vektor modelinə əsasən cümlələr arasındakı yaxınlığı hesablamaq üçün əvvəlcə onların hər biri vektor şəklində təsvir olunur, sonra isə iki vektor arasındakı məsafə (yaxınlıq) hesablanır. Tutaq ki, S_1 və S_2 cümlələri verilmişdir. Ənənəvi yanaşmalarda cümlələri vektor şəklində təsvir edərkən, vektorun uzunluğu sənəddə (yaxud sənədlər çoxluğunda) rast gəlinən sözlərin sayına bərabər götürülür. Aydındır ki, bu cür təsvir zamanı vektorun uzunluğu cümlənin uzunluğundan (cümlədəki sözlərin sayından) dəfələrlə böyük olur və vektorun elementlərinin böyük əksəriyyəti 0-a bərabər olur. Bu isə hesablama baxımından effektiv təsvir üsulu deyil. Ona görə də burada iki cümlə arasındakı yaxınlığı hesablayarkən, sözlər çoxluğu yalnız bu cümlələrdə rast gəlinən müxtəlif sözlərdən yaradılır. $\text{WS} = \{w_1, w_2, ..., w_N\}$ ilə bu sözlər çoxluğunu işarə edək, burada N müxtəlif sözlərin ümumi sayıdır. İki cümlədəki sözlər çoxluğu aşağıdakı ardıcıllıqla yaradılır [1, 19]:

1. İki cümlə götürülür, S_1 və S_2.
2. S_1 cümləsindən götürülmüş hər bir w sözü üçün aşağıdakı işlər aparılır:
   2.1. WordNet leksik bazasından istifadə etməklə onun kökü (RW) təyin edilir.
   2.2. Əgər RW sözü WS çoxluğunda iştirak edirsə, onda addım 2-yə getməli və S_1-dən götürülmüş növbəti söz üçün prosesi davam etdirməli, əks halda 2.3 addımına keçməli;
   2.3. Əgər sözün RW kökü sözlər çoxluğunda (WS) iştirak etmirsə, onda RW sözünü WS çoxluğuna əlavə edib, 2-ci addıma keçməli və prosesi S_1-dən götürülmüş növbəti söz üçün davam etdirməli. Proses S_1 cümləsindəki sözlər qurtarana kimi davam etdirilir.

Yuxarıdakı proses S_2 cümləsi üçün də təkrarlanır.

Cümlələr arasındakı yaxınlığı müəyyən etmək üçün semantik vektor modelindən istifadə edilir [3, 4]. Bunun üçün ilkin olaraq aşağıdakı əməliyyatlar yerinə yetirilir:

1. *Vektorun qurulması*. Vektorun hər bir elementi WS sözlər çoxluğundakı sözə uyğundur. Deməli, vektorun ölçüsü WS çoxluğundakı sözlərin sayına bərabərdir.
2. *Vektorun elementlərinin təyini*. Semantik vektorun hər bir elementi (sözün çəkisi) aşağıdakı qayda ilə təyin edilir:

2.4. Əgər WS sözlər çoxluğundan olan w sözü S_1 cümləsində iştirak edirsə, onda bu sözün vektordakı çəkisi 1 götürülür, əks halda növbəti addıma keçilir;
2.5. Əgər w sözü S_1 cümləsində iştirak etmirsə, onda (5) düsturunun köməyilə w sözü ilə S_1 cümləsindəki sözlər arasındakı yaxınlıq hesablanır.
2.6. Əgər sözlər arasında yaxınlıq varsa, onda w sözünün vektordakı çəkisi kimi bu qiymətlərdən ən böyüyü götürülür. Əks halda növbəti addıma keçid edilir;
2.7. Əgər sözlər arasında yaxınlıq yoxdursa, onda w sözünün vektordakı çəkisi 0 götürülür.

Beləliklə, kosinus metrikasından istifadə etməklə, iki vektor arasındakı yaxınlığı hesablamaq olar:

$$\mathrm{sim}_{\cos}(S_1,S_2)=\frac{\sum_{j=1}^{m}(w_{1j}\times w_{2j})}{\sqrt{\sum_{j=1}^{m}w_{1j}^2}\times\sqrt{\sum_{j=1}^{m}w_{2j}^2}}, \qquad (7)$$

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 j-ci sözün çəkisi; m isə sözlərin ümumi sayıdır.

F. Sintaktik Yaxınlıq

Cümlənin semantik yükü yalnız sözlərin semantik yükü ilə deyil, həm də sözlərin işlənmə arıcıllığından, yəni sözün cümlədəki mövqeyindən də birbaşa asılıdır. Məsələn, yuxarıdakı yaxınlıq ölçüsünə (semantik yaxınlıq) görə "Əli Həsənə zəng etdi" və "Həsən Əliyə zəng etdi" cümlələri oxşar cümlələr kimi qiymətləndiriliəcəkdir, çünki onlar eyni sözlərdən təşkil olunmuşdur. Buna görə də cümlələrin semantik yaxınlığını hesablayan zaman sözlərin cümlədəki işlənmə ardıcıllığı (onların cümlədəki mövqeyi) da mütləq nəzərə alınmalıdır. Beləliklə, cümlələrin sözlərin cümlədəki rast gəlmə mövqeyinə əsaslanan yaxınlığını hesablamaq üçün sintaktik-vektor yaxınlaşmasından istifadə olunur [11]. Bunun üçün əvvəlcə aşağıdakı əməliyyatlar yerinə yetirilir [1]:

1. Sintaktik vektor qurulur. Sintaktik vektorun qurulması üçün WS çoxluğundakı və sənəddəki sözlərdən istifadə edilir. Sintaktik-vektorun uzunluğu WS çoxluğundakı sözlərin sayına bərabərdir.
2. Sintaktik-vektorun hər bir elementi sözün çəkisini ifadə edir və o, sözün sənəddəki mövqeyinə bərabərdir. Bu çəki aşağıdakı kimi müəyyən edilir:

2.1. Əgər w sözü S_1 cümləsində rast gəlinirsə, onda onun vektorda çəkisi cümlədəki mövqeyinə bərabər götürülür, əks halda növbəti addıma keçilir;
2.2. Əgər w sözü S_1 cümləsində rast gəlinmirsə, onda (5) düsturunun köməyilə S_1 cümləsindəki sözlərlə w sözü arasındakı yaxınlıq hesablanır.
2.3. Əgər sözlər arasında yaxınlıq varsa, onda vektorda uyğun elementin qiyməti (sözün çəkisi) olaraq S_1 cümləsində ən böyük çəkiyə malik sözün mövqeyi götürülür.
2.4. Əgər sözlər arasında yaxınlıq yoxdursa, onda vektorda uyğun elementin qiyməti 0 götürülür.

Cümlələrin, sözlərin ardıcıllığına əsaslanan yaxınlığını hesablamaq üçün aşağıdakı düsturdan istifadə olunur [1, 11]:

$$\mathrm{sim}_{\text{wordorder}}(S_1,S_2)=1-\frac{\|O_1-O_2\|}{\|O_1+O_2\|}, \qquad (8)$$

burada $O_1=(d_{11},d_{12},...,d_{1m})$ və $O_2=(d_{21},d_{22},...,d_{2m})$ - S_1 və S_2 cümlələrinə uyğun sintaktik-vektorlar; d_{pj} isə O_p vektorunda j-ci sözün çəkisidir.

G. Xətti Kombinasiya

Cümlələr arasında yaxınlıq ölçüsünü hesablamaq üçün semantik, kosinus və sintaktik ölçülərinin xətti kombinasiyası istifadə olunur:

$$\begin{aligned}\mathrm{sim}_{\text{sentences}}(S_1,S_2)&=\beta_1\cdot\mathrm{sim}_{\text{semantic}}(S_1,S_2)+\\&+\beta_2\cdot\mathrm{sim}_{\text{wordorder}}(S_1,S_2)+\beta_3\cdot\mathrm{sim}_{\cos}(S_1,S_2)\end{aligned} \qquad (9)$$

burada $0\le\beta_i\le1,\ (i=1,2,3)$ çəki parametrləridir və onlar aşağıdakı şərti ödəyirlər:

$$\beta_1+\beta_2+\beta_3=1. \qquad (10)$$

H. Sənədlərin Yaxınlıq Ölçüsü

Sənədlər arasındakı yaxınlığı hesablamaq üçün cümlələr arasındakı yaxınlıqdan (9) istifadə olunur:

$$\mathrm{sim}_{\text{documents}}(\mathrm{D}_1,\mathrm{D}_2)=\frac{\sum_{S_1\in D_1,S_2\in D_2}\mathrm{sim}_{\text{sentences}}(S_1,S_2)}{n_1+n_2}, \qquad (11)$$

burada n_1 və n_2 uyğun olaraq D_1 və D_2 sənədlərindəki cümlələrin sayıdır.

I. Sənədlərin Təsnifatı

Sənədlərin təsnifatı dedikdə, yeni D sənədinin əvvəlcədən müəyyən edilmiş siniflərdən $C=(C_1,...,C_k)$ hər hansı birinə (yaxud bir neçəsinə) aid edilməsi prosesidir. Sənəd siniflərdən hansına daha çox yaxındırsa, həmin sinif(lər)ə aid edilir. Ədəbiyyatda kifayət qədər təsnifatlandırma metodları təklif edilmişdir. Burada D sənədinin C_q $(q=1,...,k)$ sinfinə aid olma dərəcəsini müəyyən etmək üçün aşağıdakı düstur təklif edilir:

$$\begin{aligned}\mathrm{score}(\mathrm{D}\mid C_q)&=\gamma\times\frac{\mathrm{sim}(O_D,O_{C_q})}{\sum_{d\in C}\mathrm{sim}(O_d,O_{C_q})}+\\&+(1-\gamma)\times\sum_{v\in C}\frac{\mathrm{sim}(O_d,O_v)}{\mathrm{sim}(O_d,O_v)}\end{aligned} \qquad (12)$$

burada $0\le\gamma\le1$ – çəki parametri; $\mathrm{score}(\mathrm{D}\mid C_q)$ – D sənədinin C_q sinfinə aid olma dərəcəsi; $\mathrm{sim}(O_D,O_{C_q})$ – D sənədinin O_D obrazı ilə C_q sinfinin O_{C_q} obrazı arasında yaxınlıq ölçüsüdür.

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

IV. YEKUN VƏ GƏLƏCƏK TƏDQİQATLAR

Məlumdur ki, text mining mətnlərin analizində və identifikasıyasında çox böyük imkanlara malik texnologiyadır. Tədqiqatlar göstərir ki, bu texnologiyanın tətbiq sahələri çox genişdir və hal-hazırda da böyük uğurla tətbiq edilməkdədir, Məqalədə e-dövlətin təhlükəsizliyinin təmin olunmasında bu texnologiyanın imkanları araşdırılmış və yeni yanaşma təklif olunmuşdur. Daha doğrusu e-dövlət mühitində terrorizmlə əlaqəli sənədlərin aşkarlanması üçün text mining texnologiyası metodlarına əsaslanan kompleks yanaşma təklif olunmuşdur. Lakin burada öz həllini gözləyən bir neçə problem var:

- ✓ tədqiq olunan dil üçün semantik şəbəkənin qurulması;
- ✓ kriminal qrupun kommunikasiyada xüsusi jarqon sözlərdən istifadə etməsi;
- ✓ sözlərin qrammatik cəhətdən düzgün yazılışı;
- ✓ mətnlərin müxtəlif dillərdə olması;
- ✓ mətndən terminlərin çıxarılması;
- ✓ təsnifat metodunun seçilməsi və dəqiqliyi.

Bütün bunlar təklif olunan metodun dəqiqliyinə birbaşa təsir edən amillərdir. Digər mühüm məsələ, burada fərz olundu ki, siniflər (mövzular) əvvəlcədən məlumdur və hər yeni sənəd bu sinflərdən birinə aid edilir. Lakin bu həmişə belə olmur və siniflər dinamikidir. Zaman keçdikcə yeni mövzular meydana çıxa bilər və bu mümkün haldır. Ona görə də burada ən düzgün yanaşma sənədləri avtomatik qruplaşdırıb, sonra identifikasiya etməkdir. Bütün bunlar onu göstərir ki, gələcək tədqiqatlar üçün kifayət qədər ciddi problemlər var. Bu problemlərin bir çoxu (məsələn, sözlərin avtomatik morfoloji təhlili üçün qaydaların yaradılması, sözlərin semantik şəbəkəsinin yaradılması) mutidissiplinar xarakterlidir.

ƏDƏBİYYAT

[1] Abdi A., Idris N., Alguliev R.M., Aliguliyev R.M. “Automatic summarization assessment through a combination of semantic and syntactic information for intelligent educational systems”, Information Processing & Management, 2015, vol.51, no.4, pp.340-358.

[2] Aggarwal C.C., Zhai C.X. Mining text data. Springer New York Dordrecht Heidelberg London. 2014.

[3] Alguliev R.M., Aliguliyev R.M., Mehdiyev C.A. “Sentence selection for generic document summarization using an adaptive differential evolution algorithm”, Swarm and Evolutionary Computation, 2011, vol.1, no.4, pp.213-222.

[4] Aliguliyev R.M. “A new sentence similarity measure and sentence based extractive technique for automatic text summarization”, Expert Systems with Applications, 2009, vol.36, no.4, pp.7764-7772.

[5] Alruily M., Ayesh A., Al-Marghilani A. “Using self organizing map to cluster arabic crime documents”, Proceedings of the International Multiconference on Computer Science and Information Technology, 2010. pp.357‑363.

[6] Bsoul Q., Salim J., Zakaria L.Q. “An intelligent document clustering approach to detect crime patterns”, Procedia Technology, 2013, vol.11, pp.1181-1187.

[7] Choi D., Ko B. Kim H., Kim H., “Text analysis for detecting terrorism-related articles on the web”, Journal of Network and Computer Applications, 2014, vol.38, pp.16-21.

[8] Keselj V., Peng F., Cercone N., Thomas C. “N-gram based author profiles for authorship attribution”, Proceedings of the Conference of the Pacific Association for Computational Linguistics, 2003, pp.255-264.

[9] Ku C.-H., Leroy G. “A crime reports analysis system to identify related crimes”, Journal of the American Society for Information Science and Technology, 2011, vol.62, no.8, pp.1533–1547.

[10] Ku C.-H., Leroy G. “A decision support system: automated crime report analysis and classification for e-government”, Government Information Quarterly, 2014, vol.31, no.4, pp.534-544.

[11] Li Y., McLean D., Bandar Z.A. O’shea J.D., Crockett K. “Sentence similarity based on semantic nets and corpus statistics”, IEEE Transactions on Knowledge and Data Engineering, 2006, vol.18, no.8, pp.1138-1150.

[12] Lin D. “An information-theoretic definition of similarity”, Proceedings of the Fifteenth International Conference on Machine Learning, 1998, pp.296-304.

[13] Linders D. “From e-government to we-government: defining a typology for citizen coproduction in the age of social media”, Government Information Quarterly, 2012, vol.29, no.4, pp.446-454.

[14] Miller G.A. “WordNet: a lexical database for English”, Communications on the ACM, 1995, vol.38, no.11, pp.39-41.

[15] Wimmer M., Codagnone C., Janssen M. “Future e-government research: 13 research themes identified in the eGovRTD2020 project”, Proceedings of the 41st Hawaii International Conference on System Sciences, Hawaii, USA, 7-10 January, 2008, pp.1-11.

[16] Wu Z., Palmer M. “Verb semantics and lexical selection”, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico, USA, 27-30 June, 1994, pp.133-138.

[17] Yildiz M. “E-government research: reviewing the literature, limitations, and ways forward”, Government Information Quarterly, 2007, vol.24, no.3, pp.646-665.

[18] Zhao J.J., Zhao S.Y., Zhao S.Y. “Opportunities and threats: security assessment of state e-government websites”, Government Information Quarterly, 2010, vol.27, no.1, pp.49-56.

[19] Zhao L., Wu L., Huang X. “Using query expansion in graph-based approach for query-focused multi-document summarization”, Information Processing & Management, 2009, vol.45, no.1, pp.35-41.

[20] Алыгулиев Р.М. “Роль технологии интеллектуального анализа текстов в обеспечении национальной безопасности”, Проблемы Информационных Технологий, 2013, № 1, с.38-43.

[21] ru.wikipedia.org/wiki/Коэффициент_Симпсона#cite_note-2