UOT 004.9

Alıquliyev R.M.[1]***, Niftəliyeva G.Y.***[2]
AMEA İnformasiya Texnologiyaları İnstitutu
[1] r.aliguliyev@gmail.com; [2] gunayniftali@gmail.com

TEXT MİNİNG METODLARININ KÖMƏYİLƏ E-DÖVLƏTDƏ TERRORİZMLƏ ƏLAQƏLİ MƏQALƏLƏRİN AŞKARLANMASI

Məqalədə e-dövlət mühitində terrorizmlə əlaqəli məqalələrin aşkarlanması üçün text mining texnologiyasına əsaslanan metod təklif olunmuşdur. Təklif olunmuş metod bir neçə mərhələdən ibarətdir: 1) terrorizmlə əlaqəli terminlərin lüğət bazasının yaradılması; 2) sözlərin semantik şəbəkəsinin yaradılması; 3) sözlərin morfoloji təhlili; 4) sənədlərin ilkin filtrasiyası; 5) sözlərin semantik şəbəkəsindən istifadə etməklə onlar arasında semantik yaxınlığın hesablanması; 6) cümlələr arasında semantik yaxınlığın müəyyən edilməsi; 7) sənədlər arasında semantik yaxınlığın müəyyən edilməsi; 8) sənədlərin təsnifatlandırılması. Sözlər, cümlələr və sənədlər arsında yaxınlığı hesablamaq üçün hibrid yaxınlıq ölçüləri daxil edilmişdir. Terrorizmlə əlaqəli sənədləri identifikasiya etmək üçün kNN, Bayes və yeni təklif olunan Ramiz-Günay metodlarının xətti kombinasiyasından ibarət hibrid təsnifatlandırma metodu təklif olunmuşdur.

***Açar sözlər**: e-dövlət; e-dövlətin təhlükəsizliyi; terrorizm; text mining; hibrid yaxınlıq ölçüsü; kNN metodu; modifikasiya olunmuş Bayes metodu; Ramiz-Günay metodu; hibrid təsnifatlandırma metodu.*

Giriş

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. [1–7].

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 [6, 7]. 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 [8].

E-dövlətin əsas funksiyalarından biri vətəndaşları ehtimal olunan zərər və zorakılıqlardan qorumaqdır. Linders [9] 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 [10]. Hal-hazırda biliklərin idarə olunmasında, müxtəlif mənbələrdə toplanmış mətnlərin intellektual analizində text mining ən qabaqcıl və effektiv texnologiyalardan biri hesab olunur [11].

Text mining texnologiyalarının belə populyar və tətbiq sahəsinin geniş olmasının digər səbəblərindən biri də real və ya virtual mühitdə istehsal olunmasından asılı olmayaraq informasiya növləri arasında mətnlərin üstünlük təşkil etməsidir. Beynəlxalq verilənlər korporasiyasının (International Data Corporation) analitiklərinin verdiyi məlumata görə istehsal olunan informasiyanın təqribən 80%-dən çoxunu mətnlər təşkil edir [12]. Deməli, e-dövlətin təhlükəsizliyinin təmin olunması baxımından bü mühitdə dövr edən mətnlərin intellektual analizi mühüm əhəmiyyət kəsb edir və elmi-tədqiqat nöqteyi-nəzərdən aktual məsələdir.

Beləliklə, problemin aktuallığını əsas tutaraq, məqalədə e-dövlətdə şübhəli (terrorizmlə əlaqəli) mətnlərin aşkarlanması üçün text mining texnologiyalarına əsaslanan metod təklif olunur. Bu metod [3]-də təklif olunmuş metoda oxşardır. Lakin təklif olunan metod bir neçə fərqli və üstün cəhətlərə malikdir:

- [3]-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 şübhəli 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 olunur.

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.

Əlaqəli işlərin qısa icmalı

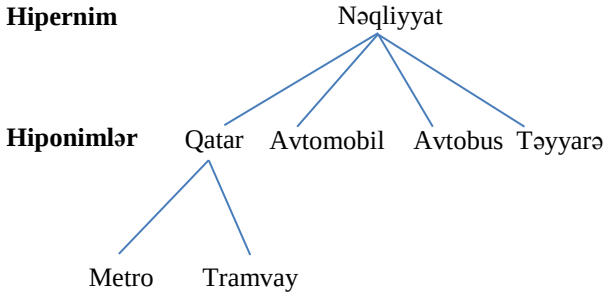
Virtual mühitdə (İnternetdə, e-dövlətdə) kriminal və terrorizmlə bağlı informasiyanın aşkarlanması, identifikasiyası və izlənməsi üçün text mining texnologiyasına əsaslanan müxtəlif metodlar, alqoritmlər və modellər təklif edilmişdir. Məsələn, veb-də kriminal informasiyanın filtrasiyası və identifikasiyası məqsədilə sənədlər arasındakı oxşarlığı müəyyən etmək üçün [4, 5]-da yeni alqoritmlər təklif edilmişdir. Ərəb dilində kriminal sənədlərin identifikasiyası sistemi üçün [1]-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 [2]-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 [3]-də mətnlərin analizinə əsaslanan yeni yanaşma təklif olunmuşdur. Bu yanaşma WordNet semantik şəbəkəsindən [13] istifadə etməklə terrorizmlə əlaqəli məqalələr çoxluğundan kontekst sözlərin (isimlərin) siyahısını yaradır. Sonra WUP [14] metrikasını tətbiq etməklə kontekst sözlərin əhəmiyyətlilik dərəcəsini hesablayır. Sonda isə biqramlardan və Keselj metrikasından [15]

istifadə etməklə sənədləri təsnifatlandırır. Çoxdilli terrorizmlə əlaqəli sənədlərin aşkarlanması üçün [16]-da təsnifatlandırma metoduna əsaslanan yeni yanaşma təklif olunmuşdur. Bu yanaşma veb sənədlərin qraf təsviri modeli ilə C4.5 təsnifatlandırma alqoritminin kombinasiyasından istifadə edir. [17]-də təklif olunmuş metod data mining alqoritmlərinin köməyilə veb saytlardakı mətnləri analiz etməklə terrorçuların fəaliyyətini (profilini) öyrənir. Kriminal məzmunlu mətnləri təsnifatlandırmaq üçün [18]-də qeyri-səlis qrammatikanın evolyusiyası (evolving fuzzy grammar) metodu təklif olunmuşdur. Bu metodda seçilmiş mətn fraqmentləri qeyri-səlis strukturda təsvir olunur.

Təklif olunan metod

Təklif olunan metod bir neçə mərhələdən ibarətdir: 1) tədqiq olunan mühitdə dövr edən sənədlərin (informasiyanın) dilindən asılı olaraq, həmin dil üçün terrorizmlə əlaqəli terminlərin 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 siniflərdən birinə 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.

1) Sənədlərin ilkin 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 sözlər (terminlər) çoxluğu kimi təsvir olunur, $d = (\mathrm{t}_1, \mathrm{t}_2, ..., \mathrm{t}_m)$. Sonra Şimkeviç-Simpson ölçüsündən [19] istifadə edərək VBase bazası ilə $d = (t_1, t_2, ..., t_m)$ çoxluğu arsındakı yaxınlıq hesablanır:

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

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

Əgər $\text{sim}_{S-S}(d, \text{VBase}) \geq \varepsilon$ olarsa, onda d sənədi şübhəli sənədlər çoxluğuna əlavə edilir və identifikasiya üçün növbəti mərhələyə keçid edilir. Burada ε eksperimental yolla müəyyən edilmiş sərhəd qiymətidir.

2) Sözlərin semantik yaxınlığı

Sözlər arasındakı semantik yaxınlıq aşağıdakı ardıcıllıqla təyin edilir:

1. İki söz t_1 və t_2 götürülür.
2. WordNet semantik şəbəkəsindən bu sözlərin kökü tapılır.
3. WordNet leksik bazasındann hər bir sözün sinonimləri və onların sayı təyin edilir;
4. WordNet şəbəkəsində istifadə etməklə, t_1 və t_2 sözlərinin ən yaxın ümumi (Least Common Subsume – LCS) kökü tapılır;
5. (2) və (3) düsturlarının köməyilə sözlər arasındakı semantik yaxınlıq hesablanır.

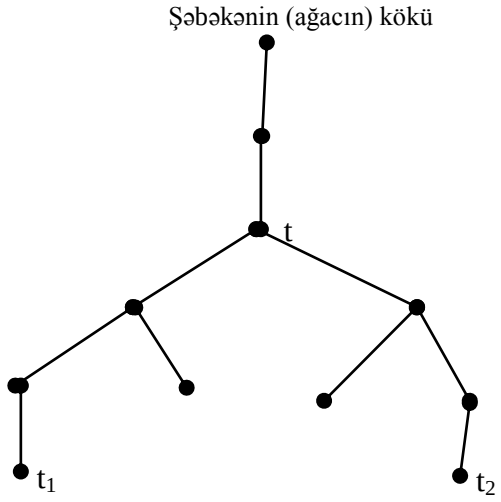
Sözlər arasındakı semantik yaxınlığı hesablamaq üçün əvvəlcə WordNet şəbəkəsindən istifadə etməklə, sözün informativ məzmunu (yükü) $\text{IC}(t)$ təyin edilir [20]:

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

Sonra (2) düsturundan istifadə edərək sözlər arasındakı semantik yaxınlıq hesablanır [20, 21]:

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

burada $\text{LCS}(t_1, t_2)$ – WordNet şəbəkəsində t_1 və t_2 sözlərinin ən yaxın olduğu ortaq söz (məsələn, şəkil 2-də göstərilən hal üçün $\text{LCS}(t_1, t_2) = t$), t_{max} – WordNet semantik şəbəkəsindəki sözlərin ümumi sayı, $\text{synset}(t) - t$ sözünün sinonimlərinin sayıdır.



Şəkil 2. Sözlərin semantik şəbəkəsi

Sözlər arasındakı semantik yaxınlığı həm də WUP metrikasından [14] istifadə etməklə hesablayırıq:

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

burada $depth(t_1)$ – WordNet semantik şəbəkəsində (ağacında) t_1-dən t-yə qədər olan qovşaqların sayı; $depth(t_2)$ – t_2-dən t-yə qədər olan qovşaqların sayı; $depth(t)$ – t-dən şəbəkənin kökünə qədər olan qovşaqların sayıdır. Məsələn, şəkil 2-də göstərilən hal üçün $depth(t_1) = depth(t_2) = 3$ və $depth(t) = 2$. Onda

$$sim_{WUP}(t_1, t_2) = \frac{2 * 2}{3 + 3 + 2 * 2} = 0{,}4.$$

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

$$sim(t_1, t_2) = \alpha \times sim_{IC}(t_1, t_2) + (1 - \alpha) \times sim_{WUP}(t_1, t_2), \quad (5)$$

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

3) Cümlələrin yaxınlıq ölçüsü

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

A) ***Semantik yaxınlıq.*** Cümlələr arasındakı semantik yaxınlıq sözlər arasındakı semantik yaxınlıqdan (5) istifadə edilərək hesablanır:

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

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

B) ***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 deməli, 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. $WS = \{t_1, t_2, ..., t_m\}$ ilə bu sözlər çoxluğunu işarə edək, burada m müxtəlif sözlərin ümumi sayıdır. İki cümlədəki sözlər çoxluğu aşağıdakı ardıcıllıqla yaradılır [20, 22]:

1. İki cümlə götürülür, s_1 və s_2.
2. s_1 cümləsindən götürülmüş hər bir t sözü üçün aşağıdakı işlər apaırılr:

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 keçməli 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 [23, 24]. 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:

2.1. Əgər WS sözlər çoxluğundan olan t sözü s_1 cümləsində iştirak edirsə, onda bu sözün vektordakı çəkisi 1 götürülür, əks halda növbəti addıma keçilir;

2.2. Əgər t sözü s_1 cümləsində iştirak etmirsə, onda (5) düsturunun köməyilə t sözü ilə s_1 cümləsindəki sözlər arasındakı yaxınlıq hesablanır.

2.3. Əgər sözlər arasında yaxınlıq sıfırdan fərqlidirsə, onda t sözünün vektordakı çəkisi kimi bu qiymətlərdən ən böyüyü götürülür. Əks halda növbəti addıma keçid edilir;

2.4. Əgər sözlər arasında yaxınlıq sıfra bərabərdirsə, onda t sözünün vektordakı çəkisi 0 götürülür.

Beləliklə, kosinus metrikasından istifadə etməklə, iki vektor arasındakı yaxınlıq aşağıdakı kimi hesablanır:

$$\mathrm{sim}_{\cos}(s_1, s_2) = \frac{\sum_{j=1}^{m}(w_{1j} \times w_{2j})}{\sqrt{\sum_{j=1}^{m} w_{1j}^2} \times \sqrt{\sum_{j=1}^{m} w_{2j}^2}}, \qquad (7)$$

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 t_j sözünün çəkisi; m isə sözlərin ümumi sayıdır.

C) ***Sintaktik yaxınlıq.*** Cümlənin semantik yükü yalnız sözlərin semantik yükü ilə deyil, həm də sözlərin işlənmə ardıcıllığından, 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 yanaşmasından istifadə olunur [25]. Bunun üçün əvvəlcə aşağıdakı əməliyyatlar yerinə yetirilir [20]:

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

2.1. Əgər t sözü s_1 cümləsində rast gəlinirsə, onda onun vektorda çəkisi cümlədəki mövqeyinə bərabər götürülür, əks halda növbəti addıma keçilir;

2.2. Əgər t sözü s_1 cümləsində rast gəlinmirsə, onda (5) düsturunun köməyilə s_1 cümləsindəki sözlərlə t sözü arasındakı yaxınlıq hesablanır.

2.3. Əgər sözlər arasında yaxınlıq sıfırdan fərqlidirsə, onda vektorda uyğun elementin qiyməti (sözün çəkisi) olaraq s_1 cümləsində ən böyük çəkiyə malik sözün mövqeyi götürülür.

2.4. Əgər sözlər arasında yaxınlıq sıfra bərabərdirsə, onda vektorda uyğun elementin qiyməti 0 götürülür.

Cümlələrin sözlərin cümlədəki mövqeyinə əsaslanan yaxınlığını hesablamaq üçün aşağıdakı düsturdan istifadə olunur [20, 25]:

$$\mathrm{sim}_{\mathrm{wordorder}}(s_1, s_2) = 1 - \frac{\|\mathrm{o}_1 - \mathrm{o}_2\|}{\|\mathrm{o}_1 + \mathrm{o}_2\|}, \quad (8)$$

burada $\mathrm{o}_1 = (w_{11}, w_{12}, ..., w_{1m})$ və $\mathrm{o}_2 = (w_{21}, w_{22}, ..., w_{2m})$ – s_1 və s_2 cümlələrinə uyğun sintaktik-vektorlar; w_{pj} isə o_p vektorunda t_j sözünün çəkisi, $\|\cdot\|$ – Evklid normasıdır.

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

$$\mathrm{sim}_{\mathrm{sentences}}(s_1, s_2) = \beta_1 \cdot \mathrm{sim}_{\mathrm{semantic}}(s_1, s_2) + \beta_2 \cdot \mathrm{sim}_{\mathrm{wordorder}}(s_1, s_2) + \beta_3 \cdot \mathrm{sim}_{\cos}(s_1, s_2), \quad (9)$$

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

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

4) Sənədlərin yaxınlıq ölçüsü

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

$$\mathrm{sim}_{\mathrm{documents}}(d_1, d_2) = \frac{\sum_{s_1 \in d_1, s_2 \in d_2} \mathrm{sim}_{\mathrm{sentences}}(s_1, s_2)}{n_1 + n_2}, \quad (11)$$

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

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

5) Sənədlərin təsnifatlandırılması

Tutaq ki, $\mathbf{C} = (C_1, ..., C_k)$ siniflər çoxluğu məlumdur. Verilmiş $\mathbf{D} = (d_1, ..., d_N)$ sənədlər çoxluğunun bu siniflərdən $\mathbf{C} = (C_1, ..., C_k)$ hər hansı birinə (yaxud bir neçəsinə) aid edilməsi prosesinə sənədlərin təsnifatı deyilir. Bu halda sənəd siniflərdən hansına daha çox yaxındırsa, onda o, həmin sinif(lər)ə aid edilir. Ədəbiyyatda kifayət qədər təsnifatlandırma metodları təklif edilmişdir. Burada d_i sənədinin C_q sinfinə aid olma dərəcəsini müəyyən etmək üçün *k*NN (k-Nearest Neighbor – k-ən yaxın qonşu) [26], Bayes [27] və yeni təklif olunan RG (Ramiz-Günay) metodundan istifadə olunur.

A) ***k*NN metodu.** Bu metoda görə d_i sənədinin C_q aid olması aşağıdakı düsturla hesablanmış kəmiyyətin qiyməti ilə müəyyən edilir:

$$\mathrm{score}_{k\mathrm{NN}}(d_i \mid C_q) = \sum_{d \in k\mathrm{NN}_q(d_i)} \mathrm{sim}(d_i, d),\ i = 1,2,...,N;\ q = 1,2,...,k, \quad (12)$$

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

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

B) **Modifikasiya olunmuş Bayes metodu.** Bayes metoduna görə d_i sənədinin C_q sinfinə aid olma dərəcəsi aşağıdakı şərti ehtimalın qiyməti ilə müəyyən olunur:

$$\mathrm{P}(C_q \mid d_i) = \frac{\mathrm{P}(C_q)\mathrm{P}(d_i \mid C_q)}{\mathrm{P}(d_i)}, \quad (13)$$

Təsnifatlandırma prosesində $\mathrm{P}(d_i)$ sabit qaldığından, (13) ifadəsində kəsrin məxrəcini nəzərə almamaq olar.

Burada $\mathrm{P}(C_q)$ – sənədlərin C_q sinfində olma ehtimalıdır. Bu aprior ehtimal aşağıdakı kimi təyin edilir:

$$\mathrm{P}(C_q) = \frac{\sum_{i=1}^{N} \mathrm{P}(C_q|d_i)}{N}. \tag{14}$$

$\mathrm{P}(d_i|C_q)$ kəmiyyətini hesablamaq üçün fərz olunur ki, sözlərin sənədlərdə işlənməsi bir-birindən asılı deyildir. Onda $\mathrm{P}(d_i|C_q)$ ehtimalı aşağıdakı düsturun köməyilə hesablanır:

$$\mathrm{P}(C_q|d_i) = \mathrm{P}(C_q)\prod_{j=1}^{m}\left(\mathrm{P}(\mathrm{t}_j|C_q)\right)^{w_{ij}}. \tag{15}$$

burada $\mathrm{P}(\mathrm{t}_j|C_q)$ – t_j sözünün C_q sinfində olma ehtimalı, m – **D** sənədlər çoxluğundakı sözlərin sayıdır:

$\mathrm{P}(\mathrm{t}_j|C_q)$ ehtimalı aşağıdakı kimi hesablanır:

$$\mathrm{P}(\mathrm{t}_j|C_q) = \frac{\sum_{d_i \in C_q} w_{ij}}{\sum_{j=1}^{m}\sum_{d_i \in C_q} w_{ij}},\ j = 1,...,m. \tag{16}$$

burada w_{ij} – t_j sözünün d_i sənədində çəkisidir.

Əgər t_j sözünün C_q sinfində çəkisi sıfra bərabər olarsa, onda (16) düsturundan alınır ki, $\mathrm{P}(\mathrm{t}_j|C_q)$ ehtimalı da sıfra bərabər olacaqdır. Beləliklə, (15) düsturundan asanlıqla alınır ki, bu hal üçün $\mathrm{P}(C_q|d_i)$ ehtimalı da sıfra bərabər olacaqdır. Ona görə də praktikada aşağıdakı düsturdan istifadə olunur:

$$\mathrm{P}(\mathrm{t}_j|C_q) = \frac{\frac{1}{N}\sum_{i=1}^{N} w_i + \sum_{d_i \in C_q} w_{ij}}{\sum_{i=1}^{N} w_i + \sum_{j=1}^{m}\sum_{d_i \in C_q} w_{ij}},\ j = 1,...,m. \tag{17}$$

(15) düsturunda loqarifmik miqyasa keçib, nəticəni d_i sənədindəki sözlərin ümumi çəkisinə (w_i) bölsək aşağıdakı bərabərliyi alarıq:

$$\mathrm{score}_{\mathrm{MBayes}}(C_q|d_i) = \mathrm{P}(C_q|d_i) = \frac{\log \mathrm{P}(C_q)}{w_i} + \sum_{j=1}^{m} \mathrm{P}(\mathrm{t}_j, d_i)\log \mathrm{P}(t_j|C_q), \tag{18}$$

burada $P(t_j, d_i) = w_{ij}/w_i$ – t_j sözünün d_i sənədində işlənmə ehtimalıdır, $w_i = \sum_{j=1}^{m} w_{ij}$, $i = 1,...,n;\ q = 1,...,k$.

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

C) Ramiz-Günay (RG) metodu. Bu metodun köməyilə d_i sənədinin C_q sinfinə aid olma dərəcəsi aşağıdakı düstürla müəyyən edir:

$$\text{score}_{\text{RG}}(d_i \mid C_q) = \lambda \times \frac{\text{sim}(\text{O}_{d_i}, \text{O}_{C_q})}{\sum_{p=1}^{k} \text{sim}(\text{O}_{d_i}, \text{O}_{C_p})} + (1-\lambda) \times \frac{\sum_{v \in C_q} \text{sim}(\text{O}_{d_i}, \text{O}_v)}{\sum_{p=1}^{k} \sum_{d \in C_p} \text{sim}(\text{O}_{d_i}, \text{O}_d)}, \quad (19)$$

burada $\text{sim}(\text{O}_{d_i}, \text{O}_{C_q})$ – d_i sənədinin O_{d_i} obrazı ilə C_q sinfinin O_{C_q} obrazı arasında yaxınlıq ölçüsü; $\text{sim}(\text{O}_d, \text{O}_v)$ – 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, $\text{O}_{C_q} = (w_1^q, w_2^q, ..., w_m^q)$:

$$w_j^q = \frac{1}{|C_q|} \sum_{d \in C_q} w_j^{q,d}, \; q = 1,...,k, \; j = 1,...,m, \quad (20)$$

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, $\text{O}_d = (w_1^d, w_2^d, ..., w_m^d)$:

$$w_j^d = \frac{1}{|d|} \sum_{s \in d} w_j^{d,s}, \; j = 1,...,m, \quad (21)$$

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

D) Hibrid metod. Yekun təsnifatlandırma metodu kimi (12), (18) və (19) düsturlarının köməyilə alınmış nəticələrin xətti kombinasiyasından istifadə olunması təklif edilir:

$$\text{score}(d^{\text{new}} \mid C_q) = \gamma_1 \cdot \text{score}_{k\text{NN}}(d^{\text{new}} \mid C_q) + \gamma_2 \cdot \text{score}_{\text{Bayes}}(d^{\text{new}} \mid C_q) + \gamma_3 \cdot \text{score}_{\text{RG}}(d^{\text{new}} \mid C_q), \quad (22)$$

burada $0 \le \gamma_i \le 1, \; (i = 1,2,3)$ çəki əmsallarıdır və aşağıdakı şərti ödəyirlər:

$$\gamma_1 + \gamma_2 + \gamma_3 = 1. \quad (23)$$

Beləlliklə, 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:

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}, \quad (24)$$

$$\text{Precision} = \frac{T_p}{T_p + F_p}, \quad (25)$$

$$\text{Precision} = \frac{T_p}{T_p + F_p}, \quad (26)$$

$$\text{F} - \text{measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (27)$$

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

Yekun və gələcək tədqiqatlar

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 qəbul olundu ki, siniflər (mövzular) əvvəlcədən məlumdur. Lakin bu həmişə belə olmur və siniflər dinamikdir. 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.

Ədəbiyyat

1. 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, Wisla, Poland, 18–20 October, 2010, pp.357–363.
2. Bsoul Q., Salim J., Zakaria L.Q. An intelligent document clustering approach to detect crime patterns // Procedia Technology, 2013, vol.11, pp.1181-1187.
3. 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.
4. 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.
5. 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.
6. Yildiz M. E-government research: reviewing the literature, limitations, and ways forward // Government Information Quarterly, 2007, vol.24, no.3, pp.646–665.
7. 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.
8. 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.

9. 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.
10. Алыгулиев Р.М. Роль технологии интеллектуального анализа текстов в обеспечении национальной безопасности // Проблемы Информационных Технологий, 2013, №1, с.38–43.
11. Aggarwal C.C., Zhai C.X. Mining text data. Springer New York Dordrecht Heidelberg London. 2014.
12. www.idc.com
13. Miller G.A. WordNet: a lexical database for English // Communications on the ACM, 1995, vol.38, no.11, pp.39-41.
14. 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.
15. 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, Nova Scotia, Canada, August 22–25, 2003, pp.255–264.
16. Last M., Markov A., Kandel A., Multi-lingual detection of terrorist content on the web // Lecture Notes in Computer Science, 2006, vol.3917, pp.16–30.
17. Shapira B., Last M., Elovici Y., Kandel A., Zaafrany O. Using data mining techniques for detecting terror-related activities on the web // Journal of Information Warfare, 2003, vol.3, no.1, pp.17–28.
18. Sharef N.M., Martin T. Evolving fuzzy grammar for crime texts categorization //Applied Soft Computing, 2015, vol.28, pp.175–187.
19. ru.wikipedia.org/wiki/Коэффициент_Симпсона#cite_note-2
20. 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.
21. Lin D. An information-theoretic definition of similarity / Proceedings of the Fifteenth International Conference on Machine Learning, 1998, pp.296–304.
22. 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.
23. 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.
24. 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.
25. 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.
26. Aliguliyev R.M. Effective summarization method of text documents / Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, France, September 19-22, 2005, pp.264–271.
27. Devroye L., Gyorfi L., Lugosi G. A probabilistic theory of pattern recognition, Springer, 1996.

UOT 004.9

Алыгулиев Рамиз М.[1]**, Нифталиева Гюнай Я.**[2]
Институт Информационных Технологий НАНА, Баку, Азербайджан
[1] r.aliguliyev@gmail.com; [2] gunay90@hotmail.com

Выявление связанных с терроризмом статей в электронном государстве с помощью методов text mining

В статье предложен метод, основанный на text mining-технологии, предназначенный для выявления статей, связанных с терроризмом в среде электронного государства. Предложенный метод состоит из нескольких этапов: 1) создание словаря, состоящего из терминов, связанных с терроризмом; 2) создание семантической сети слов; 3) морфологический анализ слов; 4) первичная фильтрация документов; 5) определение семантической близости между словами с использованием семантической сети слов; 6) определение семантической близости между предложениями; 7) определение семантической близости между документами; 8) классификация документов. Для определения связи между словами, предложениями и документами были введены гибридные меры близости. Для идентификации документов, связанных с терроризмом, был предложен гибридный метод классификации, состоящий из линейной комбинации методов *k*NN, Байеса и нового предложенного метода Рамиз-Гюнай.

Ключевые слова: *электронное государство, безопасность электронного государства, терроризм, text mining, гибридная мера близости, метод kNN, модифицированный метод Байеса, метод Рамиз-Гюнай, гибридный метод классификации.*

Ramiz M. Aliguliyev [1]**, Gunay Y. Niftaliyeva** [2]
Institute of Information Technology of ANAS, Baku, Azerbaijan
[1] r.aliguliyev@gmail.com; [2] gunay90@hotmail.com

Detecting terrorism-related articles on the e-government using text mining techniques

In this paper, a method based on text mining techniques for detecting terror-related articles on the e-government is proposed. The proposed method consists of several stages: 1) the creation of terror-related terms vocabulary; 2) the creation of semantic network of words; 3) morphological analysis of words; 4) the initial filtration of documents; 5) the calculation of the semantic similarity between words by using semantic network of words; 6) determination of semantic similarity between sentences; 7) determination of semantic similarity between documents; 8) classification of documents. Hybrid similarity measures to calculate the proximity between words, sentences and documents are introduced. Hybrid classification method combining the *k*NN, Bayes and new proposed Ramiz-Gunay methods for identification of terror-related articles is proposed.

Keywords*: e-government; e-government security; terrorism; text mining; hybrid similarity measure; kNN method; modified Bayes method; Ramiz-Gunay method; hybrid classification method.*