Mətn referatlaşmada izafiliyi idarə edən optimallaşma modeli

Ramiz Alıquliyev[1], Məkrufə Hacırəhimova[2]
[1,2] AMEA İnformasiya Texnologiyaları İnstitutu, Azərbaycan, Bakı
[1]*a.ramiz@science.az*, [2]*makrufa@science.az*

***Xülasə*— Məqalə avtomatik mətn referatlaşdırılması məsələsinə həsr olunur. Məqalədə mətnlərin avtomatik referatlaşdırılması xətti proqramlaşdırma məsələsi kimi formalizə edilir. Təklif olunan model mətnlərin optimal referatının yaradılmasına imkan verir.**

Açar sözlər - mətn referatlaşdırılması, izafilik, əhatəlilik, kiçik qalıq, optimallaşma modeli, text mining, təkamül alqoritmi

I. GİRİŞ

İnformasiya Cəmiyyətinə keçidlə əlaqədar olaraq dünyada informasiyanın əhəmiyyəti vacib resurs kimi artmış, informasiyanın emalı və ötürülməsi formaları, informasiya daşıyıcıları və onlarda informasiyanın saxlanılması, informasiyaya hüquqi sənəd statusu verən əsas rekvizitlər əhəmiyyətli dərəcədə dəyişmişdir. İnformasiyanın (*sənədin*) yeni forması – elektron sənəd (*e-sənəd*) meydana çıxmışdır. Bu tip sənədlər biznes, vətəndaşlar və hakimiyyət orqanları arasında informasiya mübadiləsinin əsas formasına, onların idarə edilməsində bir alət rolunu oynayan elektron sənədlərin idarə edilməsi sistemləri (ESİS) isə yaradılmaqda və inkişaf etməkdə olan e-dövlətin ən vacib komponentinə [1] çevrilmişdir. Əgər əvəllər hakimiyyət orqanlarında informasiya sistemləri təşkilat daxilində istifadə üçün yaradılırdısa, hazırda bu sistemlər dövlət idarələri ilə vətəndaşların (G2C -Government to Citizen) və qeyri-dövlət sektorunun (G2B - Government to Business), eyni zamanda dövlət idarələrinin bir-biri ilə (G2G - Government to Government) həm şaquli, həm də üfüqi istiqamətdə elektron qarşılıqlı əlaqəsini təmin edir [1].

İnformasiya kommunikasiya texnologiyalarının sürətli inkişafı ilə külli miqdarda elektron sənədlər (*e-sənədlər*) yaradılır, WWW-də və rəqəmsal kitabxanalarda, elektron arxivlərdə toplanır. Onların içərisində strukturlaşdırılmamış mətnlər əksəriyyət (80-90%) təşkil edir və 2020-ci ilə qədər rəqəmsal informasiyanın həcminin 44 zeta bayt (*44 trilyon qiqabayt*) olacağı qeyd olunur [1]. Bu vəziyyət hazırda dövlət qurumlarında tətbiq olunan *ESDS* üçün də xarakterikdir. Belə ki, e-hökumət xidmətlərinin populyarlaşması ilə bu sistemlərdə çoxlu sayda e-sənədlər (*vətəndaşların müraciətləri, ərizə və şikayətləri, biznes sektordan daxil olan sənədlər, dövlət qurumlarının arasındakı xidməti yazışmalar və s.)* emal olunur və sənədlərin əksəriyyəti mətn tiplidir. Bu tip sənədlərin avtomatik sistemləşdirilməsi, rəhbər və məmurlar tərəfindən operativ oxunması, məzmunu barəsində müəyyən fikir əldə olunmasında və düzgün qərar verilməsində ciddi problem yaradır [2]. E-dövlətin əsas funksiyalarını effektiv həyata keçirmək üçün bu sənədlər müxtəlif məqsədlər üzrə analiz edilməlidir. Aydındır ki, bu tip e-sənədləri hazırda kargüzarlıqda tətbiq olunan kifayət qədər funksional-texniki imkanlara malik mövcud ESİS və ya verilənlərin idarə edilməsi sistemləri vasitəsi ilə analiz etmək qeyri-mümkündür.

II. "TEXT MINING" VƏ ONUN ƏSAS MƏSƏLƏLƏRİ

Problemin həllində mətn sənədlər üzərində məzmuna görə avtomatik klassifikasiya *(classification) və* klasterləşmə *(clustering),* referatlaşdırma *(summarization)* kimi intellektual əməliyyatların aparılmasına ehtiyac yaranır. Hazırda *text mining* bu əməliyyatların yerinə yetirilməsində ən mükəmməl texnologiyadır. İnformasiyanın həddindən artıq çoxalması ilə yaranan "informasiya yükü" şəraitində böyük həcmdə informasiyalarla işləmək üçün "text mining"in həll etdiyi mətn sənədlərin sıxılmış formasının – avtomatik referatının alınması daha effektiv görünür [2,3]. Referatlaşdırma mətnin əsas məzmununu saxlamaqla sənədin qısaldılmış variantının yaradılması prosesidir və onun əsas problemi: ölçü, informativ cümlələrin və tematik bölmələrin aşkarlanması; izafilik – referatda eyni məna daşıyan cümlələrin təkrarlanmaması; yaxınlıq ölçüsünün seçilməsi və həll alqoritmidir. Referatda seçilmiş cümlələrin hər biri fərdi şəkildə vacib olmalıdır və məzmunca bir-birini təkrarlamamalıdır. Ona görə də namizəd cümlələrin əksəriyyəti verilmiş referatın uzunluğunu nəzərə alaraq yararlıdır və ən yaxşı referatın seçim strategiyası ən yaxşı cümlələrin seçilməsinə nisbətən daha vacib olur. Ən yaxşı cümlələrin seçim proseduru ilə müqayisədə ən yaxşı referatın seçilməsi qlobal optimallaşma problemidir [3]. Problemin həlli ilə əlaqədar olaraq mətnlərin avtomatik referatlaşdırılması üçün geniş əhatəliliyi və minimum izafiliyi idarə edən optimallaşma modeli təklif olunur.

III. MƏSƏLƏNİN RİYAZİ FORMALİZASİYASI VƏ HƏLLİ

Tutaq ki, D sənədlər çoxluğu verilmiş və o, cümlələr çoxluğu $D = \{s_1, s_2, \dots, s_n\}$, kimi təsvir edilmişdir. Burada s_i D -dəki i -ci cümləni bildirir, n isə sənəd çoxluğundakı cümlələrin sayıdır.

Fərz edək ki, D coxluğunda hər bir cümlə referata daxil edilmək şansına malikdir. Bunun üçün belə dəyişən daxil edək

$$x_{ij} = \begin{cases} 1, əgər\ s_i\ və\ s_j\ cümlələri\ referata\ daxil\ edilmişdir \\ 0, əks\ halda \end{cases}$$

Onda mətn referatlaşdırılması məsələsi aşağıdakı kimi formalizə oluna bilər:

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(sim(s_i,O)+sim(s_j,O))\,x_{ij} \rightarrow max$$

$$sim(s_i,s_j) < \varepsilon$$

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(l_i+l_j)x_{ij} \le L$$

$$x_{ij} \in \{0,1\}$$

ε - referatda izafiliyin səviyyəsini təyin etmək üçün parametrdir. Bu parametrin qiymətindən asılı olaraq yaradılacaq referatda izafiliyin (*təkrarçılığın*) səviyyəsini idarə etmək olar. Bu parametrin qiymətini aşağıdakı kimi təyin etmək olar:

$$\varepsilon = \frac{1}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} sim(s_i,s_j)$$

l_i -s_i cümləsinin, L-isə yaradılacaq referatın uzunluğudur.

Uzunluq sözlərin sayı və ya həcm (baytla) ola bilər. Optimallaşma məsələsinin həllində son zamanlar təbii fenomenlərin sosial davranışları ilə təlqin edilən alqoritmlərin tətbiqi daha məqsədə uyğundur [3,4].

IV. NƏTİCƏ

Təklif olunan modelin sənədlə bağlı informasiya sistemlərində tətbiqi, xüsusən də dövlət qurumlarında tətbiq olunan ESDS-lərin intellektuallaşdırılmasına imkan verməklə, məmurların qərar qəbul etməsində operativliyi təmin etmiş olacaqdır.

ƏDƏBİYYAT

[1] The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. www.emc.com/leadership/digital-universe/

[2] Hacırəhimova M.Ş. Elektron dövlət mühitində sənəd dövriyyəsi sistemlərinin aktual problemləri və həll yolları // İnformasiya cəmiyyəti problemləri, 2010, №2, s. 21-29.

[3] Alguliev R.M., Aliguliyev R.M., Hajirahimova M.S. GenDocSum + MCLR: Generic document summarization based on maximum coverage and less redundancy // Expert Systems with Application, 2012, vol. 39, №16, pp. 12460-12473.

[4] Alguliev R.M., Aliguliyev R.M. Evolutionary algorithm for extractive text summarization // Intelligent Information Management, 2009, vol. 1, №2, pp. 128–138.