UOT 004.02

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

MƏTNLƏRİN AVTOMATİK REFERATLAŞMASI ÜÇÜN OPTİMALLAŞMA MODELİ

Məqalədə mətnlərin avtomatik referatlaşdırılması üçün öyrədilməyən yanaşma təklif edilmişdir. Bu yanaşma mətndən cümlələrin seçilməsinə əsaslanır. Təklif olunanan yanaşmada cümlələrin seçilməsi optimallaşma məsələsi kimi modelləşdirilmişdir. Bu model üç xüsusiyyəti optimallaşdırmağa imkan verir: relevantlıq – referat mətnin əsas məzmununu daşıyan cümlələri özündə saxlamalıdır; izafilik – referatda eyni məzmun daşıyan cümlələr olmamalıdır; ölçü – referatın həcmi məhduddur.

Açar sözlər: *informasiya yükü, mətn analizi, mətn referatlaşdırması, izafilik, əhatəlilik, optimallaşma modeli.*

Giriş

İnformasiya – kommunikasiya texnologiyalarının sürətli inkişafı ilə informasiyanın emalı və ötürülməsi, informasiya daşıyıcıları və onlarda informasiyanın saxlanılması formaları əhəmiyyətli dərəcədə dəyişmişdir. İnformasiyaya hüquqi sənəd statusu verən əsas rekvizitlər (*imza və s.*) də dəyişikliklərə məruz qalmışdır. İnformasiyanın (*sənədin*) yeni forması – elektron sənəd (*e-sənəd*) meydana çıxmışdır. Bu tip sənədlər biznes, vətəndaş və hakimiyyət orqanları arasında informasiya mübadiləsinin əsas formasına, onların idarə edilməsində bir alət rolunu oynayan elektron sənəd dövriyyəsi sistemləri (ESDS) isə yaradılmaqda və inkişaf etməkdə olan elektron dövlətin (*e-dövlət*) ən vacib komponentinə çevrilmişdir. Bu onunla izah olunur ki, dövlət qurumlarının kargüzarlıq fəaliyyəti elektronlaşdırılır, dövlət idarələri arasında əlaqə şəbəkə mühitində həyata keçirilir. Dövlətin vətəndaşlara, biznes sektora göstərdiyi xidmətlər məkan və zamandan asılı olmayaraq İnternet üzərindən onlayn yerinə yetirilir [1, 2]. Dövlətin göstərdiyi elektron xidmətlərin populyarlaşması ilə bu sistemlərdə çoxlu sayda e-sənədlər (*vətəndaşların müraciətləri, ərizə və şikayətləri, biznes sektordan daxil olan sənədlər, dövlət qurumları arasındakı xidməti yazışmalar və s.*) dövr edir, emal olunur. Sənəd axınının çox böyük, həm də dinamik olduğunu nəzərə alaraq, dövlət hakimiyyət orqanlarında ESDS-in tətbiqi zamanı insan əməyini yüngülləşdirən, sistemin işinin səmərəliliyini artıran, ümumiyyətlə, sistemin intellektuallaşdırılmasına imkan verən elementlərin olması çox vacibdir. E-dövlətin əsas funksiyalarını effektiv həyata keçirmək üçün bu sənədlər müxtəlif məqsədlər üzrə analiz edilməlidir. Nəzərə alsaq ki, əsas informasiya daşıyıcısı kimi mətnlər əksəriyyət *(80–90%)* təşkil edir, o zaman bu tip sənədlərin avtomatik təsnifatlandırılması, rəhbər və məmurlar tərəfindən tez oxunması, məzmunu barəsində müəyyən fikir əldə olunması və düzgün qərar verilməsində ciddi problem yaranır. Aydındır ki, mətn tipli e-sənədləri hazırda kargüzarlıqda tətbiq olunan ESDS və ya verilənlərin idarə edilməsi sistemləri vasitəsi ilə analiz etmək qeyri-mümkündür. İnformasiyanın həddindən artıq çoxalması ilə yaranan "informasiya yükü" şəraitində böyük həcmdə informasiyalarla işləmək üçün mətn sənədlərin sıxılmış formasının – referatının alınması daha effektiv görünür [1, 2]. Çünki, böyük həcmdə sənədin əvəzində, onun qısaldılmış referatını təqdim etməklə, tez bir zamanda insanları sənədin əsas məzmunu ilə tanış etmək olar.

Referatlaşdırma əsas məzmunu saxlamaqla sənədin qısaldılmış variantının yaradılması prosesidir. Onun əsas problemləri:

- ölçü;
- informativ cümlələrin və tematik bölmələrin aşkarlanması;
- izafilik – referatda eyni məna daşıyan cümlələrin təkrarlanmaması;
- yaxınlıq ölçüsünün seçilməsidir.

Referatlaşdırma sənəddən (*və ya sənədlərdən*) daha xarakterik informasiyanın – cümlələrin

seçilməsini və onların ardıcıl birləşdirilməsini yerinə yetirir. Referata alternativ cümlələrin daxil edilməsi, referatda eyni məzmun daşıyan cümlələrin təkrarlanmasına – izafiliyə səbəb olur və izafiliyi aradan qaldırmaq üçün mexanizmə ehtiyac yaranır. 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 işdə mətnlərin avtomatik referatlaşdırılması üçün geniş əhatəliyi və minimum izafiliyi idarə edən optimallaşma modeli təklif olunur.

Mövcud referatlaşdırma metod və alqoritmlərin analizi

İnternetdə informasiyaya əlyetərliyin artması mətnlərin avtomatik referatlaşdırılması sahəsində tədqiqatların aparılmasını bir qədər də intensivləşdirmişdir. [4–9]-da avtomatik referatlaşdırma, onun problemləri və müasir vəziyyəti geniş şərh olunmuşdur. Referatlaşmada məqsəd, sənədin (və ya sənədlərin) məzmununu əks etdirən informativ hissələrin (cümlə, abzas) müəyyən edilməsidir. Mövqe, açar sözlərin tezliyi, başlıq-açar sözlər, sintaktik meyarlar, göstərici ifadələr kimi cümlə xüsusiyyətləri hər bir cümlənin relevantlığını bildirir. Bu günə kimi təklif olunmuş müxtəlif ekstraktiv metodlar cümlələri çəkilərinə görə ranqlaşdırır və referata daxil etmək üçün onlardan yalnız ən yüksək çəkiyə malik olanını seçir. Ouyang və digərləri tədqiqatlarında cümlə-ranqlaşdırma məsələsinə reqressiya modelini SVR (Support Vector Regression) tətbiq etməklə referatlaşdırmanı həyata keçirmişlər [10]. Sənəd çoxluğunun ekstraktiv metodla referatlaşdırılması üçün ilk dəfə çoxdilli MEAD platforması yaradılmışdır *(http://www.summarization.com/mead/)*. MEAD cümlələri çıxarmaq üçün "*centroid*" xüsusiyyətlərdən, yəni klasterlərin ağırlıq mərkəzlərindəki informasiyadan istifadə edir. Klasterdəki hər cümlə üçün üç xüsusiyyəti (ağırlıq mərkəz qiyməti – *centroid value*; mövqe qiyməti – *positional value*; ust-ustə düşən ilkin cümlə – *first-sentence overlap*) hesablayır və bu xüsusiyyətlərin xətti kombinasiyasını istifadə etməklə hansı cümlənin daha vacib olduğunu təyin edir. Cümlələrin seçimi referatın üzunluğu və yeni cümlələrin seçilmiş cümlələrlə kosinus yaxınlığı yoxlanılmaqla məhdudlaşdırılır [11]. Huang və digərləri cümlələrin relevantlıq ölçüsünü təyin etmək üçün qeyri-səlis hibrid sxem tətbiq edirlər [12].

Avtomatik mətn referatlaşdırmada, sənədlərə daxil olan fərqli fikirlərin seçim prosesi *müxtəliflik* adlanır. Müxtəliflik referatlaşdırılmış mətndə izafiliyə nəzarət etmək və daha yaxşı referat yaratmaq üçün çox əhəmiyyətlidir. [13]-də müxtəlifliyə əsaslanan referatlaşdırma modeli – MMR (Maximal Marginal Relevance) təqdim olunmuşdur. Bu yanaşmada "acgöz" alqoritmlər (*greedy algorithm*) daha relevant cümlələri seçə və eyni zamanda əvvəldən seçilmiş cümlələr arasında daha yaxın olanları silə bilir, bununla da izafilikdən nisbətən azad olmaq olur. MMR-in əsas problemi isə qeyri-optimal olmasıdır. [14]-də qeyri-mənfi matris faktorizasiyadan (*NMF – Non-negative Matrix Factorization*) istifadə etməklə, təlimsiz ümumi sənəd referatlaşdırma modeli verilmişdir. NMF gizli semantik təhlilə əsaslanan (*LSA – Latent Semantic Analysis*) metodlara nisbətən daha məzmunlu cümlələri seçir, daha yaxşı interpretasiya olunmuş semantik xüsusiyyətləri istifadə edə və sənədlərin təbii strukturunu tuta bilir. LSA metodlar cümlələri semantik xüsusiyyətlərin xətti birləşməsi ilə təqdim edirlər [15]. [16]-da cümlələrin semantik təhlilinə (*SLSS – sentence-level semantic analysis*) və simmetrik qeyri-mənfi matris faktorizasiyaya (*SNMF – symmetric non-negative matrix factorization*) əsaslanan model təklif edilir. SLSS əvvəlcə semantik təhlildən istifadə etməklə cümlələr arasındakı münasibətləri qurur və oxşarlıq matrisi yaradır. Sonra SNMF cümlələri klasterlərdə qruplaşdırmaq üçün istifadə olunur. Nəhayət, referat yaratmaq üçün hər klasterdən ən informativ cümlələr seçilir.

[17–20]-də referatlaşdırma üçün qraflar nəzəriyyəsinə əsaslanan müxtəlif metodlar təklif olunmuşdur. Bu metodlar sənədləri qraflar şəklində təqdim edir. LexRank əvvəlcə kosinus yaxınlığına əsaslanan cümlə birləşməsi qrafını yaradır və sonra məxsusi vektor (*eigenvector*) mərkəzləşmə konsepsiyasına əsaslanan vacib cümlələri seçir [18]. [21]-də abzasların çıxarılması üçün TRM (Text Relationship Map) metodikası tətbiq olunur. Qrafın təpələri cümlələri, tillərin çəkisi

isə onların yaxınlıq dərəcəsini göstərir. Təklif olunmuş ekstraktiv metodlar sənədləri abzaslar (*cümlələr*) çoxluğuna ayırır, kosinus ölçüsünü istifadə etməklə onlar arasındakı yaxınlığı hesablayır.

Referatın keyfiyyətini artırmaq üçün klasterləşmə metodlarına əsaslanan yanaşmalar da geniş yayılmışdır [4,22,23]. Yanaşmalar iki addımdan ibarətdir: klasterləşdirmə və ranqlaşdırma. Əvvəlcə tematik bölmələri müəyyən etmək üçün cümlələr məzmuna görə qruplaşdırılır və klasterdəki bir cümlə ilə qalan cümlələr arasında orta kosinus yaxınlığına əsaslanan cümləyə mərkəz qiyməti təyin olunur. Cümlələr qiymətlərinə görə ranqlaşdırılır və referat yaratmaq üçün namizəd cümlə kimi hər klasterdəki ən yüksək ranqlı cümlələr seçilir. Klasterləşmə cümlələr arasındakı müxtəlifliyi tapmaq üçün ən effektiv alət kimi istifadə olunur [22].

Ekstraktiv sənəd referatlaşmada optimal referatın tapılmasına optimallaşma problemi kimi baxıla bilər. Çünki sənədlərdə informativ cümlələrin identifikasiyasının özü mahiyyətcə optimallaşma məsələsidir. Son on ildə sənəd referatlaşmada optimallaşmaya əsaslanan yanaşmalar daha intensiv tədqiq edilməkdədir [2,3,7,24–30]. Optimal referatın yaradılması ideyası ilk dəfə 2004-cü ildə Filatova və Hatzivassiloglou tərəfindən irəli sürülmüşdür [26]. Onlar sənədləri iki ölçülü fəzada mətn və konseptual vahidlər kimi təqdim etmişlər, eyni zamanda əsas mətn vahidlərini seçə bilən və informasiyanın üst-üstə düşmələrini minimallaşdıran formal model təklif etmişlər [26]. Takamura və Okamura mətn referatlaşdırmasını maksimum əhatəlilik məsələsi kimi təqdim etmişlər (*MCKP – maximum coverage knapsack problem*) [27]. Huang və başqaları referatlaşdırmaya dörd məqsəd funksiyasını (*informasiya əhatəliliyi, vacibliyi, izafiliyi və mətn ardıcıllığı*) daxil etməklə, çoxkriteriyalı optimallaşdırma məsələsi kimi baxmışlar [3]. İzafiliyi aradan qaldırmaq üçün onlar spektral klasterləşdirmədən istifadə edir və hər cümləni semantik əlaqəli cümlələrdən ibarət olan qruplarda təsnif edirlər. Sənəd daxilindəki cümlələrin vacibliyi Markov modelini istifadə etməklə təyin edilir. Optimal referatın yaradılması sahəsində R.M.Əliquliyev və R.M.Alıquliyevin rəhbərliyi ilə aparılan tədqiqatlar isə daha cəlbedicidir. [2,7,24,29–31]-də müəlliflər cümlələrin seçilməsini optimallaşma məsələsi kimi formalizə etmiş və təkamül alqoritmlərin köməyi ilə həll etmişlər. [30]-da referatlaşdırma p-median metoduna əsaslanan çoxkriteriyalı (*relevantlıq, əhatəlilik və fərqlilik*) optimallaşdırma məsələsi kimi modelləşdirilmiş və qoyulmuş məsələ adaptiv qarışqa alqoritminin köməyi ilə həll edilmişdir. [2]-də maksimum əhatəliliyi və minimum izafiliyi təmin edən referatlaşdırma modeli kvadratik bul proqramlaşdırma məsələsi kimi formalizə edilmişdir. Modeldə məqsəd funksiyası məzmun əhatəliliyi və izafiliyin çəkili kombinasiyası şəklində təqdim olunmuş və optimallaşma məsələsinin həlli üçün binar diferensial təkamül alqoritmi işlənmişdir. Modelin üstünlüyü cümlələri seçərkən üst-üstə düşən cümlələri silməklə referatda yüksək müxtəlifliyi təmin etmək, yəni izafiliyi minimallaşdırmaqdadır. [31]-də referatlaşdırma tamədədli kvadratik xətti proqramlaşdırma məsələsi kimi modelləşdirilmiş və sürü intellektinə əsaslanan diskret alqoritm vasitəsilə həll edilmişdir. [29]-da müəlliflər çoxsənədli ümumi sənəd referatlaşdırılması üçün riyazi model təklif edirlər. Bu yanaşmada onlar vacib cümlələri çıxarmaq və izafiliyi azaltmaq üçün cümlə-sənəd kolleksiyası, referat-sənəd kolleksiyası və cümlə-cümlə münasibətlərindən istifadə edirlər. Optimallaşma problemini həll etmək üçün təkmilləşdirilmiş diferensial təkamül alqoritmi işlənmişdir.

Təklif edilən referatlaşdırma modeli

Tədqiqatlardan da göründüyü kimi, əhatəlilik və izafilik referatın keyfiyyətini həll edən əsas meyarlardır. Bu məqsədlə məqalədə mətnlərin avtomatik referatlaşdırılması üçün izafiliyi idarə edən optimallaşma modeli təklif olunur.

Cümlələrin təsviri və yaxınlıq ölçüsü. Ümumiyyətlə, mətn referatlaşdırmanın məqsədi mətndən elə cümlələr çoxluğunu tapmaqdır ki, o sənədin əsas məzmununu özündə əks etdirsin. Başqa sözlə, elə referat yaratmaq lazımdır ki, sənəd kolleksiyası ilə referat arasındakı oxşarlıq maksimum həddə olsun. Modeli təqdim etməzdən əvvəl sənədi cümlələr toplusu kimi, $D=\{s_i, i=\overline{1,n}\}$, təqdim edək. Burada s_i – D-dəki i-ci cümləni göstərir, n – sənəddəki cümlələrin sayıdır. $T=\{t_1,t_2,...,t_m\}$ isə D sənəddə olan bütün sözləri göstərir. Cümlələr məlum

vektor modelini istifadə etməklə təqdim olunur. Bu modelə əsasən hər bir s_i cümləsi m ölçülü fəzada sözlərdən ibarət xarakteristik vektor şəklində təsvir edilir, $s_i = \{w_{i1}, w_{i2}, ..., w_{im}\} = \{w_{ij}, i = \overline{1,n}, j = \overline{1,m}\}$. Burada m sənədlərdəki sözlərin sayı, w_{ij} isə i-ci cümlədəki j-ci sözün çəkisidir və o, t*f*isf* (*term frequency-inverse sentence frequency*) modelini istifadə etməklə hesablanır:

$$w_{ij} = f_{ij} \times \log(n / n_j),$$

burada f_{ij} – s_i cümləsində t_j sözünün işlənmə tezliyidir, *isf* isə bütün cümlələrin sayının t_j sözünün işləndiyi cümlələrin sayına nisbətinin loqarifmidir. n_j – t_j terminləri rast gəlinən cümlələrin sayıdır $i = \overline{1,n}$, $j = \overline{1,m}$. Bu modelin əsas problemi böyük ölçü – əlamətlər fəzasının həddindən artıq böyük olmasıdır. Əlamətlər fəzasının kiçildilməsində ən çox istifadə edilən yanaşmalar: sənəddəki bəzi sözləri *(stop words)* ixtisar etmək və sözün kökünün (*stemming)* təyin edilməsidir.

Referatlaşdırmada əsas məsələlərdən biri cümlələr arasındakı yaxınlıq ölçüsünün təyin edilməsidir. Adətən mətn sənədlər bir-birinə o vaxt yaxın hesab olunur ki, onların terminoloji tərkibi oxşar olsun. Mətn vahidləri arasındakı "yaxınlığı" (*similarity*) təyin etmək üçün *evklid məsafəsi, kosinus, Jaccard, Pearson, Kullback-Leibler divergence* ölçülərindən istifadə edilir. Kosinus ölçüsü mətn vektorlarının ən populyar yaxınlıq ölçülərindəndir. Mətn analizində tez-tez istifadə olunan kosinus ölçüsü iki vektor arasındakı bucağın kosinusunu hesablayır. $s_i = \{w_{i1}, w_{i2}, ..., w_{im}\}$ və $s_j = \{w_{j1}, w_{j2}, ..., w_{jm}\}$ cümlələri arasındakı kosinus yaxınlığı aşağıdakı kimi hesablanır:

$$sim(s_i, s_j) = \cos(s_i, s_j) = \frac{\sum_{k=1}^{m} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \cdot \sum_{k=1}^{m} w_{jk}^2}}, \; i, j = 1, ..., n$$

Məsələnin riyazi formalizasiyası. Referat yaradarkən adətən üç xüsusiyyət optimallaşdırılır:

- *relevantlıq*: referat istifadəçi üçün relevant olan informativ mətn hissələrindən ibarət olmalıdır.
- *izafilik*: referatlar eyni informasiyanı daşıyan mətn hissələrindən ibarət olmamalıdırlar.
- *ölçü*: referat uzunluğa görə məhdudlaşdırılmalıdır. Adətən, referat tələbatdan asılı olaraq ilkin mətnin 5 – 30%-ni təşkil etməlidir.

Bu üç xüsusiyyətin birgə optimallaşdırılması çətin və qlobal referatlaşdırma məsələsidir. Relevant mətn hissələrin daxil edilməsi təkcə onların öz xüsusiyyətlərinə deyil, həm də referatdakı digər mətn hissələrinin xüsusiyyətlərinə əsaslanır [25].

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

$$x_i = \begin{cases} 1, əgər\ s_i\ referata\ daxil\ edilmişdirsə \\ 0, əks\ halda \end{cases}$$

$$x_i = \begin{cases} 1, əgər\ s_i\ və\ s_j\ referata\ daxil\ edilmişdirsə \\ 0, əks\ halda \end{cases}$$

Onda mətn referatlaşdırılması məsələsi aşağıdakı kimi formalizə oluna bilər:

$$\sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (w_i + w_j) sim_{ij} . x_{ij} \to \max \qquad (1)$$

$$\sum_{i=1}^{n} l_i x_i \le L \tag{2}$$

$$x_i \in \{0,1\} \tag{3}$$

$$x_{ij} \in \{0,1\} \tag{4}$$

burada $w_i - s_i$ cümləsinin əhəmiyyətlilik dərəcəsi (*çəkisi*), $sim_{ij} - s_i$ və s_j sümlələri arasında yaxınlıq ölçüsüdür. w_i çəkisini aşağıdakı kimi təyin etmək olar:

$$w_i = \sum_{j=1}^{n} m_j . e^{-\left(-\frac{1}{sim_{ij}}\right)^2} \quad , \; i = 1,...,n$$

m_j aşağıdakı kimi hesablanır:

$$m_j = \frac{\sum_{k=1}^{m} w_{jk}}{\sum_{k=1}^{m} \overline{w_k}} \quad , \; j = 1,...,n \qquad ,$$

burada $\overline{w_k} = \frac{1}{n}\sum_{i=1}^{n} w_{ik}$, $k = 1,...,m$ – cümlələr çoxluğunun mərkəzinin koordinatlarıdır.

(2)-də l_i - s_i cümləsinin, L-isə yaradılacaq referatın uzunluğudur. Uzunluq sözlərin sayı və ya həcm (*baytla*) ola bilər.

Nəticə

E-dövlət inkişaf etdikcə onun informasiya mühitində emal edilən informasiyanın həcmi də böyük sürətlə artır və dövr edən sənədlərin əksəriyyətini strukturlaşdırılmamış mətnlər təşkil edir. Mətn sənədləri kiçik zaman ərzində analiz etmək, operativ qərar qəbul etmək çox ciddi problem yaradır. Problemin həllində avtomatik mətn referatlaşdırmadan istifadə daha məqsədəuyğundur. Referatlaşmanın əsas problemi isə məzmun əhatəliliyi, izafilikdir. Bu məqsədlə yaradılacaq referatda izafiliyi minimallaşdırmaq üçün sənəd referatlaşdırılması xətti optimallaşdırma məsələsi kimi formalizə olunmuşdur. Təklif olunan model dövlət qurumlarında tətbiq olunan ESDS-lərin intellektuallaşdırılmasına imkan verməklə, məmurların qərar qəbul etməsinə dəstək ola bilər. Çətin həll olunan optimallaşma məsələsinin həllində təbii alqoritmlərin (*qarışqa, arı və s.*) tətbiqi daha məqsədəuyğundur. Növbəti tədqiqatımızın hədəfi məhz (1-4) optimallaşma məsələsinin həlli üçün alqoritmin işlənməsidir.

Ədəbiyyat

1. Hacırəhimova M.Ş. Elektron dövlət mühitində sənəd dövriyyəsi sistemlərinin aktual problemləri və həll yolları // İnformasiya cəmiyyəti problemləri, 2010, №2, s.21–29.
2. 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, no.16, pp.12460–12473.
3. Huang L., He Y., Wei F., Li W. Modeling document summarization as multi-objective optimization / Proceedings of the Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, China, 2010, april 02–04, pp.382–386.
4. Aliguliyev R.M. Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization // Computational Intelligence, 2010, vol.26, no.4, pp.420–448.
5. Jones K.S. Automatic summarizing: the state of the art // Information Processing and

Management, 2007, vol.43, no.6, pp.1449–1481.
6. Das D., Martins A. F.T. A Survey on Automatic Text Summarization // Language, 2007, no.4, pp.1–31. http://www.cs.cmu.edu/~nasmith/LS2/das-martins.07.pdf
7. Alguliev R.M., Aliguliyev R.M., Isazade N.R. MR&MR-SUM: maximum relevance and minimum redundancy document summarization model // İnternational Journal of İnformation Technology & Decision Making, 2013, vol.12, no.3, pp.361–393
8. Tucker R. Automatic summarizing and the CLASP system, PhD thesis, University of Cambridge, UK, 1999, 190 p.
9. Zajic D.M. Mutipe alternative sentence compressions as a tool for automatik summarization task, PhD Thesis, University of Maryland College park,. 2007, 229 p. www.umiacs.umd.edu
10. Ouyang Y., Li W., Li S., Lu Q. Applying regression models to query-focused multi-document summarization // Information Processing & Management, 2011, vol.47, no.2, pp.227–237.
11. Radev D., Jing H., Stys M., Tam D. Centroid-based summarization of multiple documents // Information Processing and Management, 2004, vol.40, no.6, pp.919–938.
12. Huang H.H., Yang H.C., Kuo Y.H. A fuzzy-rough hybrid approach to multi-document extractive summarization / Proceedings of the Ninth International Conference on Hybrid Intelligent Systems, Shenyang, China, 2009, august 12–14, pp.168–173.
13. Carbonell J.G., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries / Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998, august 24–28, pp.335–336.
14. Lee J.H., Park S., Ahn C.M., Kim D. Automatic generic document summarization based on non-negative matrix factorization // Information Processing and Management, 2009, vol.45, no.1, pp.20–34.
15. Gong Y., Liu X. Generic text summarization using relevance measure and latent semantic analysis / Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval, New Orleans, USA, 2001, september 9–12, pp.19–25.
16. Wang D., Li T., Zhu S., Ding C. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization / Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval, Singapore, 2008, july 20–24, pp.307–314.
17. Wan X., Xiao J. Graph-based multi-modality learning for topic-focused multi-document summarization / Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09), Pasadena, USA, 2009, july 11–17, pp.1586–1591.
18. Erkan G., Radev D. Lexrank: graph-based centrality as salience in text summarization // Journal of Artificial Intelligence Research, 2004, vol. 22, pp.457–479.
19. Zhang J., Xu H., Cheng X. GSPSummary: a graph-based sub-topic partition algorithm for summarization / Proceedings of the Asia Information Retrieval Symposium, Harbin, China, 2008, january 15–18, pp.321–334.
20. Zhao L., Wu L., Huang X. Using query expansion in graph-based approach for query-focused multi-document summarization // Information Processing and Management, 2009, vol.45, no.1, pp.35–41.
21. Mitra M., Singhal A., Buckley C. Automatic text summarization by paragraph extraction / Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, 1997, pp.39–46.
22. Binwahlan M.S., Salim N., Suanmali L. Fuzzy swarm diversity hybrid model for text summarization // Information Processing and Management, 2010, vol.46, no.5, pp.571–588.
23. Nomoto T., Matsumoto Y. The diversity-based approach to open-domain text summarization // Information Processing and Management, 2003, vol.39, no.3, pp.363–389.
24. Alguliev R., Aliguliyev R., Hajirahimova M. Multi-document summarization model based

on integer linear programming // Intelligent Control and Automation, 2010, vol.1, no.1, pp.105–111.
25. McDonald R. A study of global inference algorithms in multi-document summarization / Proceedings of the 29th European Conference on IR Research, Rome, Italy, Springer-Verlag, LNCS, 2007, april 2–5, no.25, pp.557–564.
26. Filatova E., Hatzivassiloglou V. A formal model for information selection in multi-sentence text extraction / Proceedings of the 20th International Conference on Computational Linguistics (COLING'04), Geneva, Switzerland, 2004, august 23–27, pp.397–403.
27. Takamura H., Okumura M. Text summarization model based on maximum coverage problem and its variant / Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece, 2009, march 30–april 3, pp.781–789.
28. Lin J., Madnani N., Dorr B. Putting the user in the loop: interactive maximal marginal relevance for query-focused summarization / Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, USA, 2010, june 1–6, pp.305–308.
29. Alguliev R.M., Aliguliyev R.M., Isazade N.R. Multiple documents summarization based on evolutionary optimization algorithm // Expert Systems with Applications, 2013, vol.40, no.5, pp.1675–1689.
30. Алыгулиев Р.М., Мехтиев Ч.А. Моделирование реферирования документов как модифицированная задача о р-медиане и адаптивный муравьиный алгоритм для решения задачи оптимизации // Информационные технологии, 2011, №9, стр.9–17.
31. Alguliev R.M., Aliguliyev R.M., Isazade N.R. CDDS: Constraint-driven document summarization models // Expert Systems with Applications, 2013, vol.40, no.2, pp.458–465.

УДК 004.02

Алыгулиев Рамиз М.[1], Гаджирагимова Макруфа Ш.[2]
Институт Информационных Технологий НАНА, Баку, Азербайджан
[1]a.ramiz@science.az; [2]makrufa@science.az

Оптимизационная модель для автоматического реферирования текстов

В статье предложен необучаемый подход к автоматическому реферированию документов. Этот подход основан на выборе предложений. В предлагаемом подходе, выбор предложений моделирован как задача оптимизации. Эта модель дает возможность оптимизации трех свойств: релевантности – реферат должен содержать информативные предложения, несущие основные темы исходного текста; избыточности – реферат не должен содержать предложения, передающие ту же информацию; размерности – реферат ограничен в длину.

Ключевые слова: *информационная перегрузка, текст анализ, текст резюмирование, избыточность, охват, оптимизационная модель.*

Ramiz Aliguliyev M.[1], Makrufa Hajirahimova S.[2]
Institute of Information Technology of ANAS, Baku, Azerbaijan
[1]a.ramiz@science.az; [2]makrufa@science.az

An optimization model for automatic text summarization

In this paper, an unsupervised approach to automatic document summarization is proposed. This approach is based on sentence selection. In the proposed approach, sentence selection is modeled as an optimization problem. This model generally attempts to optimize three properties: relevance – summary should contain informative sentences that carry the main topics of the source text; redundancy – summaries should not contain multiple sentences that convey the same information; length – summary is bounded in length.

Key words: *information overload, text mining, text summarization, redundance, coverage, optimization model.*