BAKI UNİVERSİTETİNİN XƏBƏRLƏRİ

№1 ***Fizika-riyaziyyat elmləri seriyası*** ***2013***

İNFORMATİKA

UOT 621.39

MƏTN TİPLİ İNFORMASİYALARIN ANALİZİ ÜSULU VƏ MODELİ

S.Y.ƏLİZADƏ
Azərbaycan Texniki Universiteti
akif@inbox.ru

Məqalə mətn sənədlərinin avtomatlaşdırılmış məzmunca intellektual emalının təminatı nəzəriyyəsinin inkişafına, informasiya şəbəkələrində mətn tipli sənədlərinin emalının intellektuallaşdırılması sistemlərinin yaradılması zamanı modellərin və elmi-nəzəri, metodoloji əsasların təminatı üsullarının analizi, elektron fondların yaradılması və informasiya-axtarış sistemlərinin effektivliyinin artırılmasına həsr edilmişdir.

Açar sözlər: axtarış sistemləri, semantik analiz, kontekst, bilik.

Son zamanlar elektron şəklində olan mətn tipli informasiyaların miqdarı o qədər artmışdır ki, mövcud informasiya resursları arasında tələb olunan məlumatların axtarışının çətinliyi ilə əlaqədar, informasiyanın biliyə çevrilməməsi və qiymətdən düşməsi təhlükəsi yaranmışdır. Yaranmış şəraitdə, sənədlərin analizi və axtarışı sistemlərində intellektuallaşmaya tələbat kəskin artmışdır. Qlobal informasiya fəzasının indiki həcmində (1000 ekzabaytdan çox) informasiya axtarışı məsələlərinin həlli, lazımi informasiyaya vaxtında müraciəti təmin etmək üçün yalnız prioritet deyil, həm də vacib sayılır. Hazırda toplanmış nəhəng informasiya toplusu, onun həcminin kəsilməz olaraq artması informasiya analizi və axtarışı sahəsində məqalənin mövzusunun aktuallığını və tədqiqatların əhəmiyyətini müəyyənləşdirir.

Tədqiqatların nəticələri göstərir ki, biliklərin əldə edilməsi və çıxarılması üçün istifadə edilən mövcud axtarış sistemlərinin effektivliyi, daha çox birbaşa analiz üsullarından asılıdır. Belə ki, onlar istənilən axtarış sisteminin əsasını təşkil edir və əksər hallarda bu sistemin imkanlarını və məhdudiyyətlərini təyin edir [1].

Təhlillər göstərir ki, intellektual sistemlər üzrə tədqiqatlar oblastı cox genişdir; genetik alqoritmlər; koqnitiv modelləmə; intellektual interfeyslər; nitqin tanınması və sintezi; deduktiv modellər, çoxagentli sistemlər; ontologiyalar; biliklər menecmenti; yumşaq hesablamalar və digərləri.

İnternet bilik mənbəyinə çevrilməklə bərabər qeyri-formal kommunikasiyaların imkanını inqilabi şəkildə dəyişmişdir (dünya elmi cəmiyyətlərində informasiyaya müraciət sürəti). Elektron poçt, elanlar taxtası, internet "konfransları" və s.

Qeyd etmək lazımdır ki, informasiya axtarış dilinin seçilməsi və sənədlərin təbii dildə emalı bilik texnologiyasının vacib məsələlərindəndir [2]. Qarşıya qoyulmuş məqsədlərin mürəkkəbliyi və bu metodların reallaşdırılması üçün işlənmiş nəzəriyyələrin kifayət dərəcədə olmaması göstərir ki, interpretasiya prosesi (yəni bir anlama sistemindən digərinə tərcümə) ünsiyyət sisteminin (birgə fəaliyyətin) mürəkkəbliyi və universallığının artması ilə əlaqədar mövcud bir sıra çətinliklərlə bağlıdır. Lakin yenə də aydın görünür ki, daha çevik və təbii ünsiyyət üçün, təbii dil daha adekvatdır, belə ki, onun tətbiqi, insanın maşına sərbəst müraciətini, insanın və kompüterin kompleks adaptasiyasını, onların birgə fəaliyyətinin etibarlılığını təmin etməyə imkan verir. Ona görə də təbii dildə ünsiyyət sisteminin effektivliyi, əksər hallarda hesablama vasitələrinin işinin məhsuldarlığından və keyfiyyətindən asılıdır.

Formal olaraq məsələnin translyasiyasını belə təsvir etmək olar: tutaq ki, təbii dil altçoxluğu və hesablama sisteminin formal dili verilir, qısa olaraq M-dil adlandıraq. M-dildə problemli mühitin modeli təsvir edilir. Burada translyasiya məsələsi hər hansı bir mətnin $t_i \in T$ (burada T-peşəkar yönlü bütün mətinlər çoxluğundan) hər hansı bir mətnə $m_i \in M$ (burada M- M-dilinin problemli redaksiyasının bütün mətinlər çoxluğudur) cevrilməsindən və ya T mətnlər çoxluğunda, verilmiş $t_i \in T$ mətni üzrə ona adekvat $m_i \in M$ mətnini qurmağa imkan verən ψ :T→M təsviri təyin olunur.

Adekvakt sözünün mənası belə izah olunur. Tutaq ki, ψ^{-1}:T→M təsviri və hər hansı bir $\{\Pi_i\}$ ekvivalent linqvistik çevrilmə çoxluğu mövçuddur. ψ^{-1} təsviri bəzi $m_i \in M$ mətnləri üçün $t_i \in T$ mətnlərini qurmağa imkan verir. Onda $t_2 \in T$ üçün ψ üzrə ona adekvat olan $m_i \in M$ mətnini, yalnız o zaman qurmaq olar ki, $m_i \in M$ mətninə ψ^{-1} köməyilə müxtəlif $t_i^1, t_i^2, ..., t_i^n (n \geq 1)$ çoxluğu almaq olar ki, hər biri $t_i^j \in T$ mətni $\{\Pi_{ij}\}$ çoxluqları qədər dəqiqliklə ekvivalentdir.

Başqa sözlə desək, əgər trasnslyator təbii dil (TD) mətnin bir neçə dəfə dəyişərsə və alınan dəyişmiş mətndə ilkin verilmiş mətnə adekvat cümlələr olmazsa translyator TD- mətnini "başa düşər".

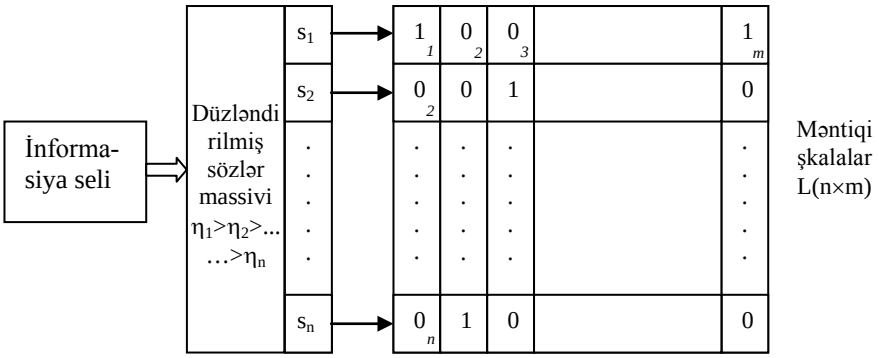
Verilmiş işdə əsas diqqət, sənədlərin tematik yaxınlığı məsələlərinə və onunla bağlı məsələlərin həllinə yönəlmişdir. Müəlliflərin şəxsi tədqiqatları nəticəsində mətn tipli informasiyaların analizi üçün onların bağlılığını nəzərə alan - verilənlərin sıxılması üsulundan istifadə edərək mətnin semantik modeli işlənmişdir (şəkil 1) [3].

Bu üsulun mahiyyəti ondan ibarətdir ki, mətnlər, bu mətnlərin məzmununu təşkil edən sözlər lüğətinin (S) rastgəlmə tezliyinə görə nizamlanmış şəkildə və mətndə hər bir sözün mövqeyini kodlaşdıran məntiqi şkalalar (L) şəklində təsvir olunur. Bunun üçün əvvəlcə mətnlərin statistik analizi aparılır və hər bir sözün rastgəlmə tezliyi (S-informasiya elementi) təyin edilir və hər bir mətn, informasiya axtarışı zamanı axtarış tezliyinin artırılması üçün assosiativ yaddaşda saxlanılır. L məntiqi şkalasının uzunluğu mətndəki bütün sözlərin bitlərlə sayına bərabərdir. Strukturun belə təsviri zamanı məntiqi şkalaların hamısı bir yerdə (mətnin "semantik kartı") L (nxm) ölçülü matrisasını təşkil edir. Burada n- mətni təşkil edən unikal sözlərin sayı, m – mətndəki bütün sözlərin sayıdır.

Hər sözün mövqeyi uyğun məntiqi şkalalarda vahidlərlə qeyd edilir. Hər şkaladakı vahidlərin sayı, mətndə uyğun sözün rastgəlmə tezliyinin sayına bərabərdir. Bu cür matrisanın köməyilə açar sözlər və onların kontekstlərini asanlıqla təyin etmək olar. Kontekst, açar sözlərin qonşu sütunlarının köməyilə təyin olunur. Kontekstin dərinliyi ekspert tərəfindən verilən sütunların sayı ilə müəyyən olunur. Unikal sözlər yaddaşda rastgəlmə tezliyinin azalması istiqamətində (η), yerləşdiyi üçün, açar sözlərin sayının sərhəd qiymətini təyin etmək olar.

Deməli, bu yanaşmanın mahiyyəti, mətnin strukturunun informasiya seli ilə modellənməsi və bu selin vasitəsilə mətnin strukturunun nxm ölçülü matris şəklində formalaşmasından ibarətdir. Mətnin struktur təsviri modelinə keçid aşağıdakı kimi həyata keçirilir.

Mətnə, informasiya elementlərindən – sözlərdən təşkil olunmuş informasiya seli kimi baxılır. Əgər mətndən, birincidən başlayaraq sonuncuya kimi sözləri ardıcıl götürsək, onda bu elə P informasiya seli olacaq. Bu zaman mətndə bütün sözlər toplusunu sonlu informasiya elementləri çoxluğuna ayırmaq olar:



Şək. 1. Mətnin semantik modeli

$$S = \{s_1, s_2, \ldots, s_n\},$$

Burada S_i - mətndə müəyyən sözə uyğun informasiya elementidir.

İnformasiya strukturu D(S, C) – informasiya elementləri çoxluğu S (məntiqi şkalaların vahid mərtəbələri) və bu elementlər arasında C əlaqələr toplusunun cəmidir.

$$D(S, C) \equiv P$$

C – informasiya elementləri cütləri arasında əlaqələr toplusu olub, P informasiya selinin eyni elementlər cütündən dəfələrlə keçdiyi halda, təkrar olunan əlaqələrdən ibarət ola bilər.

$$C = (c_1, c_2, \ldots, c_{n-1}),$$

burada $c_i = (s_i, s_{i+1})$ – iki informasiya elementi arasında əlaqə olub, P selində s_i, s_{i+1} informasiya elementləri ardıcıllığını göstərir

$$\forall s \in S \; \exists \, C(s) = ((c_i, c_{i+1})_1, \ldots, (c_j, c_{j+1})_n).$$

D(S, C) strukturuna daxil olan S çoxluğundakı hər bir informasiya elementi üçün əlaqə cütləri mövcuddur, harada c_i, c_j – giriş əlaqələr (özündən əvvəlki sütunlar), c_{i+1}, c_{j+1} – çıxış əlaqə, n – əlaqə cütlərinin sayıdır.

Giriş əlaqə, verilmiş informasiya elementindən keçən seli təsvir edən əlaqələr toplusunda çıxış əlaqədən əvvəlki əlaqəni göstərir.

n(C(s)) - C(s) toplusunda verilmiş informasiya elementinin D(S, C) strukturundakı digər informasiya elementləri ilə əlaqələrinin sayını xarakterizə edir. n(C(s)) mətndə sözlərin sayına bərabərdir. Əlaqə cütlərinin sayını d(s) – informasiya elementinin dərəcəsi kimi işarələndirək:

$$d(s) = n(C(s)),$$

$$\forall s \in S, 0 < d(s) \le n(S) - 1.$$

$d(D(S, C))_{max}$ – D(S,C) informasiya strukturu üçün inforamsiya elementinin (daha tez-tez rastgəlinən) maksimal dərəcəsi:

$$d(D(S, C))_{max} = \max d(s), \; s \in S.$$

$d(D(S, C))_{min}$ – D(S,C) informasiya strukturu üçün informasiya elementinin minimal dərəcəsi:

$$d(D(S, C))_{min} = \min d(s), \qquad s \in S.$$

Təqdim olunan model əsasında aşağıdakı iki məsələnin həlli üçün mətnin tematik analizinin üsul və alqoritmlərinin işlənməsi yerinə yetirilə bilər:

1) mətn informasiyasının tematik təsnifatı;
2) mətnin verilmiş sinifə tematik aidiyyat dərəcəsinin hesablanması.

Mətnin tematikasının düzgün və adekvat təsvirinə yalnız açar sözləri deyil, həm də bu sözlərin konteksti daxildir, belə ki, istənilən sözün mənası, birmənalı olaraq, mətndə onunla birlikdə, yaxın və yanaşı işlənən sözlərin kontekstində müəyyən olunur [4]. Açar sözlər, kontekstdən ayrılıqda, mətnin tematik istiqamətliyini tam olaraq əks etdirmir. Mövcud psixolinqvistik tədqiqatlar verilmiş tezisi təsdiq edir.

Bu halda, açar sözlərin kontekstlərlə tamamlanması, axtarış sisteminin effektivliyini artırır (şəkil 2).

Bu üsülün ümumi ardıcıllığı aşağıdakı kimidir:

1) Mətnin modelləşməsi və onun informasiya strukturunun formalaşması.

2) Mətndə təkrarlanma sayına görə düzləndirilmiş, bütün informasiya elementləri çoxluğunun seçilməsi.

3) M_p açar elementlər çoxluğunun seçilməsi.

Bütün informasiya elementləri çoxluğundan ilk *n*-i götürək (*n*-sərhəd kəmiyyəti əsasında təyin olunur), hansı ki, $M_p = \{k_1s_1, k_2s_2, \ldots, k_ns_n\}$ ilk açar elementləri çoxluğudur, $k_1, k_2, \ldots, k_n$ çəki əmsalları verilmiş tematikada bu və ya digər informasiya elementinin çəkisini (qiymətini) təyin edir.

4) M_p informasiya elementləri çoxluğunun kontekst analizi əsasında M_k dəqiqləşdirici çoxluğun formalaşması. Kontekst analiz, informasiya elementlərinin ətrafının, əvvəlcədən formalaşdırılmış informasiya strukturu üzrə analizinə əsaslanır.

5) Ümumi açar elementləri çoxluğunun alınması. Bu da mətnin tematikasını müəyyənləşdirir: $M = M_p + M_k$. Bu üsulun nəticəsi isə $M= \{k_1s_1, k_2s_2, \ldots, k_ns_n\}$ çoxluğudur.

Mətnin verilmiş sinifə tematik aidiyyat dərəcəsinin hesablanması. Tutaq ki, $M = \{k_1s_1, k_2s_2, \ldots, k_ns_n\}$ – nümumə-mətnin açar elementləri çoxluğudur; $M_f = \{k_{f1}s_1, k_{f2}s_2, \ldots, k_{fn}s_n\}$ – mətnin axtarışı zamanı tapılan (informasiya-axtarış sistemi vasitəsilə tapılan sənəd) hər hansı bir açar elementləri çoxluğudur ki, biz mətn-nümunəyə nəzərən tematik yaxınlığı araşdırmalıyıq.

Hər bir informasiya elementi üzrə ω_i tematik yaxınlığı belə hesablamaq lazımdır:

$$\omega_i = \frac{k_{i\min}}{k_{i\max}} \cdot k_i,$$

$$k_{i\min} = k_i, k_{i\max} = k_{fi} \Leftrightarrow k_i < k_{fi},$$

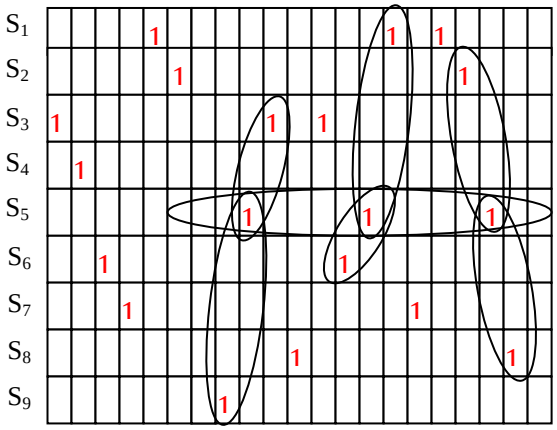
$$k_{i\min} = k_{fi}, k_{i\max} = k_i \Leftrightarrow k_{fi} < k_i.$$

Bütün mətn üzrə ümumi tematik yaxınlıq əmsalı, bütün ω_i-lərin cəmi kimi hesablamaq lazımdır:

$$\omega = \sum_{i=1}^{n} \omega_i$$

Hər bir tapılmış mətn (sənəd) üçün ω hesablasaq, bu sənədlərin tematik yaxınlığa görə seçilməsini yerinə yetirmək olar.

Yuxarıda nəzərdən keçirilən tematik yaxınlığın hesablanması üsulunu vacib bir fikirlə tamamlamaq lazımdır. Sözün mənası onun kontekstinə görə, onunla birlikdə işlənən sözlərə görə təyin olunur. Xüsusilə bu, tematik yaxınlıq hesablanarkən əhəmiyyətlidir. S və S_f – də iştirak edən eyni bir söz özündə tamamilə müxtəlif məna



Şək. 2. Söz birləşmələrinin (kontekstlərin) seçilməsi.

çaları, məna yükü daşıya bilər. Elə buna görə də çəki əmsallarının sadə müqayisəsi, tematik yaxınlığın dəqiq hesablanması üçün kifayət deyil, sözlərin kontekstini də nəzərə almaq lazımdır.

Nəticə. Məzmuna görə tammətnli analiz müasir elektron fondların yaradılmasının avtomatlaşdırılması sisteminin vacib hissəsi hesab olunur. Mətnlərin məzmuna görə analizi və təsnifatlaşdırılması məsələləri zəif formalizə olunan məsələlər sinifinə aiddir, evristik alqoritmlərin köməyi və linqvistlərin iştirakı ilə həll edilməlidir. Məqalədə əsasən aşağıdakılar təklif edilir:

1. Mətnin semantik əlaqəliliyini və ardıcıllığını struktur şəklində təsvir etməyə imkan verən, sözlərin sıxılmasına əsaslanan ixtiyari məzmunlu mətnin təsvir üsulu;

2. Mətnin mövzusunu çəkili açar sözlərin yığımı şəklində təsvir etməyə və bu sözlərin verilmiş tematikada qiymətliliyini xarakterizə etməyə imkan verən tezlik-kontekst təsnifatlaşdırma üsulu;

3. Mətnlərin mövzu yaxınlığının kəmiyyətcə qiymətləndirilməsini əldə etməyə imkan verən və bu mətnlərin istifadəçi sorğusuna tematik yaxınlığının dərəcəsini hesablayan alqoritm.

Təklif edilmiş model, metod və alqoritmlər tematik analizin dəqiqliyini və adekvatlığını əhəmiyyətli dərəcədə artırmağa imkan verir.

ƏDƏBİYYAT

1. Аббасов А.М., Сулейманов А.Ш. Модульный принцип организации нейронных сетей для распознавания смысла слов и текста / Доклады международной конференции «Информационные средства и технологии». М., 2003, с. 54-57.
2. Suleymanov A.Sh. Calculation algorithm of the semantic distance between texts // Die Fachzeitschrift Erscheint unter Dachherrschaft der Deutschen Gesellschaft für Informatik, Bochum, Deutschland, Germany. Fachzeitschrift "Spektrum", 2005, №3, p. 35-39.

3. Сулейманов А.Ш., Челабиева С.Ю. Интегрирование методов статистического и лингвистического анализа текстовых документов / Международная конференция «Информационные средства и технологии». М., 2005, с.33.
4. Сулейманов А.Ш. Метод определение контекстных слов при анализе текста // Информационные технологии, теоретический и прикладной научно-технический журнал, М.,2009, 7(155), с.46-49.

МОДЕЛЬ И МЕТОД АНАЛИЗА ТЕКСТОВОЙ ИНФОРМАЦИИ

С.Ю.АЛИЗАДЕ

РЕЗЮМЕ

Статья посвящена развитию теоретического обеспечения автоматизированной интеллектуальной обработки текстовой информации, анализу методов обеспечения модели, научно-практического и методологического основ при разработке системы интеллектуальной обработки текстовых документов в информационных сетях, созданию электронных фондов и повышению эффективности информационно- поисковых систем.

Ключевые слова: поисковые системы, семантический анализ, контекст, знание.

ANALYSIS METHODS AND MODELS OF TEXTUAL TYPES OF INFORMATION

S.Y.ALIZADEH

SUMMARY

The article deals with the development of support theory of notional intellectual processing of textual documents, analysis of the models and methods of scientific-theoretical and methodical bases at the processing of the intellectualization of textual documents in information networks, establishment of electronic foundation and the efficiency of information retrieval systems.

Key words: search systems, semantic analysis, context, knowledge

Принято в редакцию:06.02.2013 г.
Подписано к печати: 06.03.2013 г.