Elektron Kitabxana Sistemlərində Mətnlərin Avtomatik Təsnifatı üçün Metod

Nigar İsmayılova[1], Sevinc Mərdanova[2], Nərgiz İsmayılova[3]
[1,2,3]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*nigar@iit.ab.az,* [2]*mardanovasevinj@gmail.com,* [3]*nargiz.ni21@gmail.com*

***Xülasə* — Məqalədə e-kitabxana sistemlərində mətnlərin avtomatik təsnifatı üçün bir sıra metod və alqoritmlərə baxılmışdır. Mövzunun aktuallığı əsaslandırılmış, mətnlərin avtomatik təsnif olunmasının tətbiq sahələri göstərilmiş, tətbiq olunan üsulların üstünlükləri və çatışmazlıqları şərh olunmuş, açar sözlər əsasında mətnlərin təsnif olunması üçün qeyri-səlis neyron şəbəkələrdən istifadə təklif olunmuşdur.**

Açar sözlər — avtomatik təsnifat, təsnifat üsulları, açar sözlərin avtomatik təyin olunması, qeyri-səlis neyron şəbəkə

I. GİRİŞ

İnformasiya texnologiyalarının sürətli inkişafı, verilənlər bazalarının imkanlarının artması nəticəsində elektron resursların həcminin çoxalması bu resursların avtomatik təsnif olunmasını zəruriyyətə çevirir. Mətnlərin müəllifinin, aid olduğu istiqamətin avtomatik təyin olunması müasir e-kitabxanaların əsas problemlərindəndir. Hal-hazırda fəaliyyət göstərən ən məşhur bibliometrik bazalarda (Thomson Reuters, Scopus) məqalələr jurnalların aid olduğu istiqamətlərə görə təsnifat olunur, lakin bəzi jurnallarda müxtəlif istiqamətli məqalələr çap olunur, nəticədə məqalələrin bu cür təsnifatını adekvat hesab etmək olmaz.

Beləliklə, mətnlərin avtomatik təsnif olunması nəticəsində e-kitabxana və bibliometrik bazalarda elektron resursların avtomatik kateqoriyalaşması, xəbər saytlarında xəbərlərin qruplaşdırılması, korporativ şəbəkələrdə məktubların avtomatik yönləndirilməsi, internetdə məlumatların müəyyən məhdudiyyətlərə (yaş, dini mənsubiyyət və s.) görə kateqoriyalaşdırılması, cinayət faktlarının aşkarlanması və s. kimi məsələlər asanlıqla həll oluna bilər.

II. İSTİFADƏ OLUNAN ÜSULLAR

Hal-hazırda dünya alimləri tərəfindən müxtəlif təsnifat metod və alqoritmləri, məsələn, ən yaxın qonşu klassifikatoru (Nearest neighbour classifier), Bayes üsulu (Bayesian clasification), Dayaq vektorlar üsulu (Support vector machine), Assosiasiyalara (assosiativ qaydalara) əsaslanan təsnifat (Association based clasification), Terminlərin qraf modeli (Term Graph Model), Neyron şəbəkələr vasitəsilə təsnifatlaşdırma üsulu (Classification using neural network) mətnlərin təsnifatı üçün tətbiq olunmuşdur. Bu üsullar vasitəsilə müxtəlif eksperimentlər aparılmış, onların üstünlükləri və çatışmazlıqları aşkarlanmışdır [1,2]:

- Ən yaxın qonşu alqoritmi öyrədici nümunələrin yaxınlıq dərəcəsinə əsaslanır, sadə, etibarlı və parametrsiz üsuldur. Bu üsulun əsas üstünlüyü onun az parametrli olmasıdır, lakin qovşaqlar arasında oxşarlıqların (similarity) hesablanması çox vaxt tələb edir;
- Bayes klassifikatoru ehtimal nəzəriyyəsinə əsaslanan ən sadə, mətn təsnifatı üçün ən məşhur öyrətmə üsuludur, çünki bu sürətli, implementasiyası asan və yaxşı nəticə göstərən üsuldur;
- Dayaq vektorlar üsulu təsnifat məsələlərinin həlli üçün çox dəqiq və məşhur metoddur, sənədlər fəzasını optimal hiperxətt vasitəsilə kateqoriyalara ayırmağa əsaslanır. Bu üsulda əlamət fəzasının ölçüsü müstəqildir və çoxparametrlilik problemi yoxdur, lakin sənədlərin sayı artdıqca onları ayırmaq üçün optimal xəttin tapılması mürəkkəbləşir;
- Assosiasiyaya əsaslanan təsnifat qaydalara əsaslanaraq təsnifatı həyata keçirir, mətn sənədlərini mövzu ierarxiyalarına uyğun təsnifatlandırır. Assosiativ qaydalara əsaslanan təsnifat üsulunda dəqiqlik yüksəkdir, lakin burada da verilənlərin sayı artdıqca üsulun effektivliyi azalır;
- Terminlərin qraf modeli vasitəsilə təsnifat mətnlərdə terminlərin birgə istifadə sayına əsaslanır. Sənədlərin qrafla təsviri standart söz yığımı təsvirindən daha effektivdir, buna baxmayaraq mətn klassifikasiyasının qrafla təsvirinin mürəkkəbliyi bu üsulun ən başlıca çatışmazlığıdır;
- Süni neyron şəbəkələr klassifikasiya məsələlərində çox geniş istifadə olunur. Neyron şəbəkələr vasitəsilə təsnifatlaşdırma üsulu qeyri-xətti modeldir və kompleks əlaqələrin modelləşməsi üçün əlverişlidir, eyni zamanda özünü təkmilləşdirmə bacarığına malikdir, lakin girişlərin və gizli qovşaqların sayı artdıqca verilənlərin çox parametrlilik problemi yaranır.

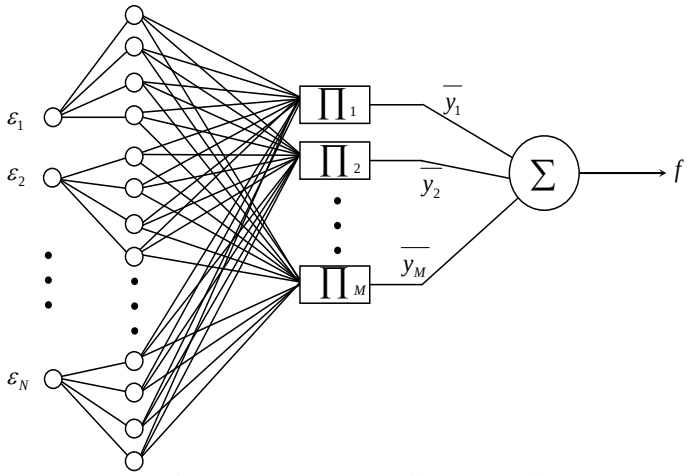
Bütün bu üsulların mənfi və müsbət cəhətlərini nəzərə alaraq, mətnlərin təsnifatı üçün qeyri-səlis neyron şəbəkələr (ANFİS) əsasında tərtib olunmuş alqoritmi tətbiq etməyi təklif edirik [3]. Sözlər və onların mənaları arasında qeyri-səlis münasibətlər mövcuddur, buna görə də neyron şəbəkələrin giriş elementləri kimi açar sözlərin linqvistik dəyişənlər olaraq

xarakterizə olunması qeyri-səlis neyron şəbəkələri baxılan məsələnin həlli üçün əlverişli edir.

III. QEYRİ-SƏLİS NEYRON ŞƏBƏKƏLƏR ƏSASINDA MƏTNLƏRİN TƏSNİFATI

Qeyri-səlis çoxluqlar nəzəriyyəsinin imkanlarından müxtəlif istiqamətlərdə mətnlərin təsnif olunması üçün istifadə olunur [4,5]. E-kitabxanalarda qeyri-səlis neyron şəbəkələr vasitəsilə elektron resursların avtomatik təsnif olunması aşağıdakı mərhələlər üzrə həyata keçirilir:

1. Biliklər bazasının seçilməsi - Biliklər bazası olaraq E-kitabxana Mərkəzində mövcud olan rəqəmsal resurslar və yaxud elmi əsərlər bazasından, həmçinin qlobal şəbəkədə mövcud olan elektron bazaların resurslarından istifadə edə bilərik;
2. Açar sözlərin avtomatik təyin olunması - Bəzi mətnlərdə açar sözlər müəlliflər tərəfindən qeyd olunur, ancaq bir çox mətnlərdə açar sözlər yoxdur. Bu səbəbdən təsnifat məsələsində əsas problemlərdən biri də açar sözlərin avtomatik təyin olunmasıdır. Bu məqsədlə mətnlərdə sözlərin işlənmə tezliyinə, onların bir-biri ilə əlaqəsinə, linqvistik əlamətlərinə əsaslanan müxtəlif üsullar tətbiq olunur [6,7];
3. Açar söz və açar sözlərin qruplaşmasının müxtəlif siniflərə mənsubiyyət dərəcələri təyin olunur, qaydalar bazası tərtib olunur, neyro-fazi şəbəkə qurulur. Şəbəkənin giriş verilənləri mətnin açar sözləri, çıxışı isə bu mətnin aid olduğu sinfin nömrəsidir (şəkil 1), birinci layda mənsubiyyət dərəcələri ikinci layda qaydalar bazasının qiymətləri hesablanır və sonda defazifikasiya vasitəsilə mətnin sinfi təyin olunur.



Şəkil 1. Açar sözlər vasitəsilə mətnlərin təsnifatı üçün qeyri-səlis neyron şəbəkə

NƏTİCƏ

Günümüzdə formalaşan informasiya bolluğu şəraitində, big data probleminin olduğu bir dövrdə mətnlərin avtomatik təsnif olunması təkcə e-kitabxanalarda deyil, müxtəlif sahələrdə uğurla tətbiq oluna bilər. Bu baxımdan mətnlərin təsnifatı üçün müxtəlif üsulların təklif olunması, mövcud üsulların effektivliyinin artırılması çox vacibdir.

ƏDƏBİYYAT

[1] M.E. Nidhi, V. Gupta, I.N. Sneddon “Recent Trends in Text Classification Techniques,” International Journal of Computer Applications, 2001, vol. 35, no.6, pp. 45-51.

[2] M. İkomakis, S. Kotsiantis, V. Tampakas “Text classification using Machine Learning Techniques,” Wseas Transactions on Computers, 2005, vol. 4(8), pp. 966-974.

[3] K. Aida-zade, S. Rustamov, E. Mustafayev, N. Aliyeva “Human-computer dialogue understanding hybrid system,” International Symposium on Innovations in Intelligent Systems and Applications, 2012, pp. (1-5).

[4] T.M. Nogueira, H.A. Camaigo, S.O. Rezende “Fuzzy Rules for Document Clasification to Improve Information Retrieval,” International Journal of Computer Information Systems and Industrial Management Applications, 2011, vol 3, pp. 210-217.

[5] G. Tsekouras, C. Anagnostopoulos, D. Gavalas, E. Dafni “Clasification of Web Documents using Fuzzy Logic Categorical Data Clustering,” IFIP The International Federation for Information Processing, 2007, vol 247, pp. 93-100.

[6] F. Liu, D. Pennell, Y.Liu “Unsupervised approaches for automatic keyword extraction using meeting transcripts”, The 2009 Annual Conference of the North american Chapter of the Association for Computational Linguistics, 2009, pp.620-628.

[7] X. Hu, B. Wu “Automatic keyword extraction using linguistic features”, Proc. of the Sixth IEEE International Conference on Data Mining Workshops, 2006, pp.19-23.