CZU 025.4.036:004

INFORMATION RETRIEVAL SYSTEMS EVALUATION METHODOLOGIES

METODOLOGII DE EVALUARE A SISTEMELOR DE INVESTIGARE A INFORMAŢIEI

Viorica LUPU

„Dacă nu poţi măsura, nu poţi îmbunătăţi”
(“If you can not measure it, you can not improve it.”)
Lord Kelvin

Abstract: *Evaluation is an important step in the development of information retrieval systems, as it allows for the successful measurement of a system designed to help users meet their information goals. Several scholars have conducted a series of tests to calculate the efficiency of information retrieval systems. Among the developed methodologies, the Cranfield model remains the dominant approach in evaluating information systems. This paper presents the steps to be taken to test an information system, the indicators used to evaluate the effectiveness of information systems (recall and precision). Studying the evaluation and its role in the functioning of information retrieval tools can essentially contribute to improving the performance of these tools and the quality of their results.*
Key words: *information retrieval system, evaluation, efficiency, relevance, precision, recall*

Dezvoltarea rapidă a noilor tehnologii de informare şi comunicare, precum şi creşterea impresionantă a volumului de informaţii, date şi documente disponibile face ca identificarea informaţiei de care avem nevoie să devină o operaţiune extrem de importantă, cu implicaţii majore în activitatea cotidiană, dar mai ales în cea academică ori de cercetare. Astfel, este imperios să găsim în termene optime conţinuturi de o relevanţă maximă, care să răspundă cel mai bine nevoilor noastre cognitive. Anume în acest scop au fost elaborate instrumentele de investigare a informaţiilor (baze de date, sisteme de informare, cataloage etc.), pe care bibliotecile le implementează activ în activităţile lor informaţionale.

În cadrul proiectului Tempus „Servicii informaţionale moderne pentru îmbunătăţirea calităţii studiilor” (MISISQ), şapte biblioteci universitare din Republica Moldova şi-au unit eforturile şi au creat catalogul colectiv partajat al bibliotecilor universitare, un sistem menit să asigure accesul operativ şi de calitate al utilizatorilor la informaţie.

Una dintre cerinţele înaintate faţă de acest produs informaţional este descoperirea şi livrarea informaţiei de o plenitudine şi precizie maxime. Descoperirea informaţiilor presupune obţinerea, cu ajutorul sistemului de informare, a resurselor relevante dintr-o colecţie de resurse informaţionale. Alcătuit din cataloagele a şapte biblioteci, funcţionarea catalogului colectiv trezeşte mai multe întrebări şi suspiciuni, în special în ceea ce priveşte performanţa acestuia şi relevanţa rezultatelor pe care le oferă. Iată de ce este nevoie de a aprecia măsura în care sistemul răspunde necesităţilor de informare ale utili-

zatorilor. Fără o evaluare adecvată în acest sens nu putem stabili cât de bine funcţionează un sistem, nu putem compara în mod obiectiv nivelul acestuia de investigare a informaţiei cu cel al altor sisteme.

Scopul articolului de faţă nu este de a evalua eficienţa unui sistem concret de informare, ci de a identifica şi analiza unele metodologii de evaluare a instrumentelor de investigare a informaţiei, metodologii care, în perspectivă, ar putea fi aplicate la măsurarea eficienţei sistemelor de acest fel din Republica Moldova.

Evaluarea sistemelor de investigare a informaţiei este una dintre cele mai mari provocări pentru specialiştii în ştiinţele informării. Determinarea performanţei unui sistem este condiţionată de aprecierea gradului de relevanţă a documentelor furnizate de sistem în raport cu nevoile de informare ale utilizatorului. Evaluarea permite cuantificarea şi măsurarea succesului unui sistem de investigare a informaţiilor, fiind importantă pentru proiectarea, dezvoltarea şi menţinerea unor sisteme eficiente de informare.

Există două modalităţi de evaluare: evaluarea nemijlocită a sistemului şi evaluarea bazată pe utilizator. Prima modalitate presupune măsurarea capacităţii sistemului de a clasifica documentele în funcţie de relevanţă, în timp ce a doua se axează pe percepţia utilizatorului şi măsoară gradul de satisfacţie al acestuia în legătură cu funcţionarea sistemului. Abordarea centrată pe utilizator examinează sarcina de căutare a informaţiilor în contextul comportamentului uman, urmărind a înţelege complexitatea interacţiunii utilizatorului cu un sistem informatic. Se porneşte de la premisa că înţelegerea comportamentului utilizatorului facilitează proiectarea mai eficientă a sistemului şi stabileşte criteriile de utilizare în evaluarea relaţiilor dintre utilizator şi sistem.

Cu toate avantajele pe care le prezintă, evaluarea bazată pe utilizator este extrem de costisitoare şi dificil de realizat corect. Pentru o apreciere corespunzătoare este nevoie de un eşantion suficient de mare şi reprezentativ de utilizatori efectivi ai sistemului de informare (a cărui rutină zilnică va fi întreruptă de procesul de evaluare), fiecare dintre sistemele supuse evaluării trebuie să fie la fel de complete şi dezvoltate, cu o interfaţă optimă de căutare a informaţiei, prietenoasă cu utilizatorul, iar fiecare subiect/utilizator trebuie să fie instruit la fel de bine în toate sistemele (Voorhees, E. M. 2002). Astfel de considerente îi determină pe cercetători să utilizeze opţiunea mai puţin costisitoare, adică evaluarea centrată pe sistem.

Testele de laborator sunt mult mai puţin costisitoare decât evaluările bazate pe utilizator, oferind în special informaţii de diagnosticare a comportamentului sistemului.

Primele experimente riguroase vizând evaluarea sistemelor de informare au fost realizate în cadrul Bibliotecii Colegiului Aeronautic Cranfield din Marea Britanie, sub conducerea bibliotecarului şi omului de ştiinţă în domeniul informaticii Cyril Cleverdon. Acesta a iniţiat o serie de experimente, numite Cranfield, prin care au fost puse bazele cercetării privind evaluarea sistemelor de investigare a informaţiilor. Autorii şi-au propus îmbunătăţirea eficienţei sistemelor studiate prin dezvoltarea unor limbaje şi metode de indexare mai eficiente. Experimentele au fost efectuate în două etape principale. Prima etapă, numită în mod obişnuit Cranfield 1, a cuprins perioada anilor 1957–1961, iar cea de-a doua, Cranfield 2, s-a desfăşurat între 1963 şi 1966 (Cleverdon, Cyril 1967). Cleverdon şi colegii săi s-au confruntat cu provocări serioase în realizarea experimentelor. Menţionăm că experimentele nu au fost computerizate, ci efectuate manual, destul de laborios, cu indicii scrise manual, ca şi cataloagele de scoruri şi căutări efectuate prin încrucişarea informaţiilor despre aceşti indici de scor.

Experimentele Cranfield au stabilit metodologia experimentală standard în domeniu: evaluarea sistemului în baza unei colecţii fixe de teste – un set de documente, un set

de cereri de informare şi un set de hotărâri de relevanţă (evaluări ale documentelor).

La baza concepţiei lui Cleverdon stă noţiunea de relevanţă, clasificată în două tipuri: relevanţa semantică, ce presupune corespunderea conţinutului informaţiei livrate cu conţinutul cererii de informare a utilizatorului, şi relevanţa formală, axată pe corespunderea modelului de investigare a documentului cu modelul de investigare a cererii de informare.

Experimentele de la Cranfield au culminat cu instituirea indicatorilor de precizie şi plenitudine, care au devenit criterii clasice de evaluare a sistemelor de regăsire a informaţiei.

Precizia exprimă capacitatea sistemului de a livra documentele relevante pentru o cerere, adică informaţia ce coincide cu conţinutul cererii de informare. De exemplu, dacă precizia este de 50%, înseamnă că printre documentele găsite jumătate sunt relevante şi jumătate irelevante.

Pentru calcularea preciziei informaţiei livrate se propune următoarea formulă:

$$Pr = \frac{a}{a+b} \; x \; 100\%,$$

unde *a* este cantitatea de documente relevante livrate, iar *b* – cantitatea de documente livrate de către sistem.

Plenitudinea este raportul dintre numărul de documente relevante livrate de sistem şi numărul total de documente relevante nelivrate de sistem:

$$Pl = \frac{a}{a+c} \; x \; 100\%.$$

De exemplu, dacă plenitudinea este de 50%, înseamnă că jumătate din documentele relevante nu sunt găsite de sistem.

Între indicatorii de precizie şi plenitudine există o corelaţie strânsă. Odată cu creşterea excesivă a preciziei scade plenitudinea şi invers, dacă ridicăm plenitudinea peste 90% precizia scade brusc, de aceea este foarte importantă existenţa unui echilibru între aceşti doi indicatori. Pentru a stabili acest echilibru este necesar mai întâi de toate de a determina nivelul de profunzime necesară indexării. Stabilirea acestui nivel trebuie să se realizeze în funcţie de categoriile de utilizatori pentru care este destinat sistemul şi de tipul bibliotecii în care este exploatat sistemul. Creşterea plenitudinii este posibilă numai în baza scăderii preciziei şi invers.

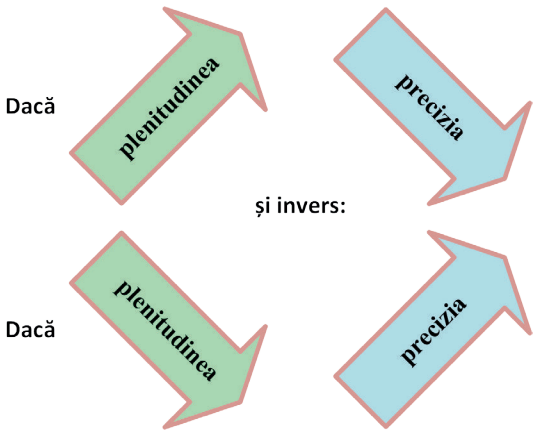

Fig. 1. Relaţia dintre precizie şi plenitudine

Urmărind logica testelor Cranfield, modalitatea desfăşurării lor şi indicatorii folosiţi, credem că această metodologie ar putea fi aplicată cu succes la evaluarea sistemelor de informare existente la noi.

Pentru a testa un sistem de informare şi a compara strategiile de căutare este nevoie de a parcurge mai multe etape, prima dintre care este *pregătirea unei colecţii de documente pentru a permite testarea semnificativă a rezultatelor.* Sistemele de informare indexează documente care sunt identificate ca răspuns la cererile utilizatorilor. O colecţie de teste trebuie să conţină un set static de documente care să reflecte tipurile de documente ce pot fi găsite în setările sau domeniile

operaţionale. Acest lucru ar putea implica colecţii de biblioteci digitale, cataloage sau seturi de pagini web. Noţiunea de colectare a documentelor statice este importantă deoarece asigură posibilitatea de a reproduce rezultatele la reutilizarea colecţiei de teste. Întrebările specifice care ar putea fi luate în considerare la colectarea documentelor includ:

- Câte documente ar trebui colectate? În cazul colecţiilor de biblioteci digitale mai mici, toate elementele ar putea fi incluse în colecţia de teste. Totuşi, în situaţii care implică colecţii mai mari sau colecţii în continuă schimbare, de exemplu pagini web, trebuie de colectat un eşantion de documente.
- Ce elemente ar trebui să fie eşantionate pentru a crea colecţia de documente? Elementele din colecţia de testare ar trebui să le reprezinte cât mai fidel posibil pe cele găsite într-un cadru operaţional. Dacă elementele trebuie să fie selectate pentru o colecţie de documente, atunci poate fi oportună selectarea unui eşantion aleatoriu sau a unui anumit subset de documente.

Modelarea unei aplicaţii reale de utilizator, cu nevoi realiste de informare, este o altă etapă de evaluare, în cadrul căreia se stabileşte un set de cereri de informare tipice ale utilizatorilor. În baza acestora se creează modelul de investigare a informaţiei (traducerea cererilor de informare într-un limbaj de indexare şi întocmirea unui şir de cuvinte-cheie în baza cărora se va efectua căutarea informaţiei). Fiecare cerere trebuie să reprezinte o necesitate reală de informare şi trebuie să fie exprimată corect şi fără ambiguitate. În practică, pentru obţinerea unor rezultate de evaluare fiabile ar trebui incluse cel puţin 50 de cuvinte-cheie sau vedete de subiect.

Următoarea etapă, *investigarea informaţiei şi evaluarea relevanţei,* presupune accesarea sistemului şi adresarea cererilor de informare. În rezultat, pentru fiecare cerere sistemul va livra o cantitate de documente ce trebuie evaluată din punctul de vedere al relevanţei.

Logica cercetării este următoarea: în orice masiv informaţional al unui sistem, faţă de o cerere de informare există o parte de documente relevante. O parte din aceste documente este livrată utilizatorului, iar altă parte, din cauza unor imperfecţiuni tehnice, rămâne în sistem.

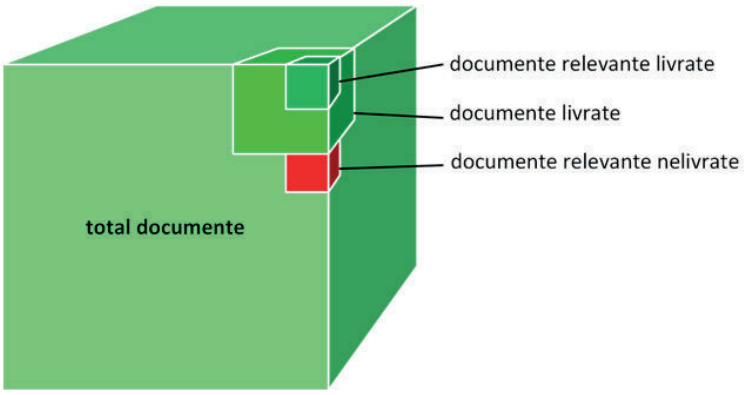

Fig. 2. Categoriile de documente în sistemul de informare
(figură adaptată după Hinkelmann, Knut)

Determinarea relevanţei este extrem de importantă pentru evaluarea de ansamblu a sistemelor de regăsire a informaţiei şi este efectuată în permanenţă de experţi în domeniu. Judecăţile emise în cadrul acestui proces sunt afectate de multe caracteristici ale documentelor şi utilizatorilor, precum şi de factori situaţionali (Harter, S.P. 1996).

Într-un studiu recent al procesului de evaluare a relevanţei, cercetătorii Cuadra şi Katter au identificat patru tipuri principale de factori care pot afecta rezultatul unei judecăţi de relevanţă: tipul de document care este evaluat/judecat, inclusiv subiectul, nivelul de dificultate, stilul etc.; condiţiile în care trebuie să fie pronunţate judecăţile, adică timpul disponibil, ordinea de prezentare şi mărimea setului de documente, tipul de specificaţie a sarcinii etc.; declaraţia care specifică cerinţa de informare ce determină relevanţa; evaluatorii folosiţi pentru a face judecăţile, adică experienţa, cunoştinţele, atitudinea lor etc. Aceste variabile sunt rezumate în tabelul din figura 3.

Trebui să admitem totuşi că aprecierea relevanţei presupune o mare doză de subiectivism, întrucât chiar şi aceeaşi persoană poate aprecia diferit în momente diferite, fiind influenţată de numărul de documente ce trebuie evaluate, de ordinea în care ele sunt prezentate, de cunoştinţele în cauză ale celui ce apreciază şi, eventual, chiar de dispoziţia lui.

Relevanţa din perspectiva umană este subiectivă (depinde de opinia unui anumit utilizator), situaţională (se referă la nevoile curente ale utilizatorului), cognitivă (depinde de percepţia umană) şi dinamică (modificări în timp) (Zuva, K., Zuva, T. 2012).

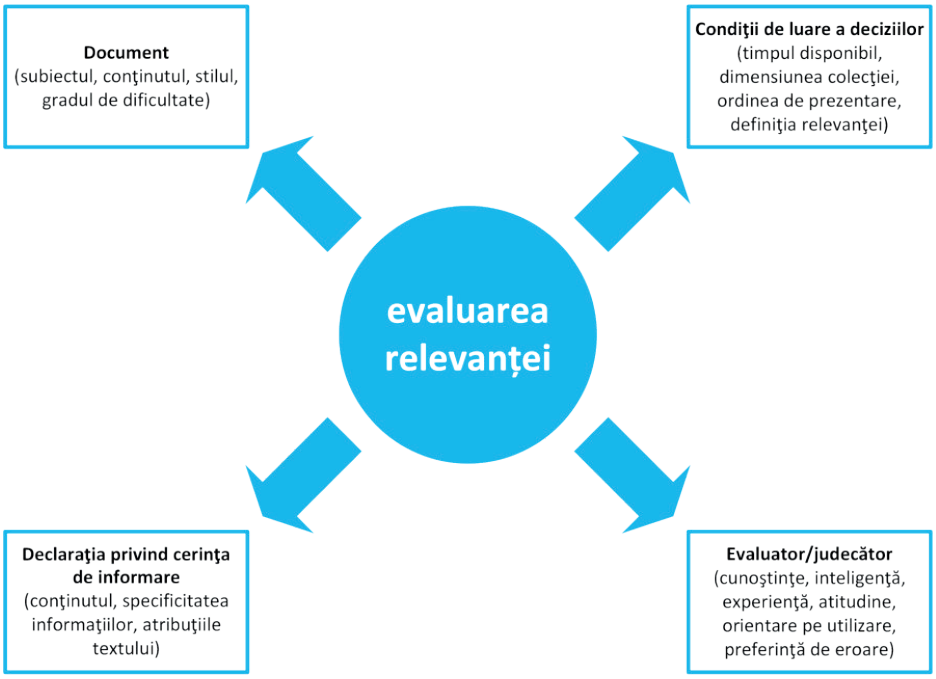

Fig. 3.Variabile privind evaluarea relevanţei *(Lesk, M.E., Salton, G. 1968)*

Evaluatorii de relevanţă vor emite judecăţi de relevanţă pentru fiecare document, acestea fiind ulterior utilizate pentru a calcula eficacitatea sistemului. Generarea unor asemenea evaluări este adesea extrem de consumatoare de timp şi de intensitate a forţei de muncă. Aşa cum am arătat mai sus, judecăţile de relevanţă tind să difere în funcţie de cine face evaluarea, iar valorile de plenitudine şi de precizie obţinute prin utilizarea acestor evaluări se pot dovedi apoi instabile.

Pentru a genera valori coerente de plenitudine şi de precizie se va efectua mai întâi examinarea conţinutului documentelor în vederea aprecierii gradului de coincidenţă a

informaţiei livrate de sistem cu conţinutul cererii de informare. Ulterior se va face raportarea documentelor livrate la cele nelivrate, precum şi a documentelor livrate relevante la cele nerelevante.

În vederea asocierii relevanţei cu livrarea informaţiei poate fi utilizat tabelul de mai jos.

	documente relevante	**documente nerelevante**	
documente livrate	a	b	a+b
documente nelivrate	c	d	c+d
	a+c	b+d	a+b+c+d

Utilizând acest tabel putem determina indicatorii de eficienţă a sistemului de informare (precizie şi plenitudine).

Să presupunem că pentru cererea de informare C1, sistemul a livrat un număr total de 100 de documente, dintre care 65 sunt relevante şi 35 nerelevante. În rezultatul căutării, în masivul de documente s-a depistat că referitor la cerinţa respectivă în colecţie mai există un număr de 20 de documente relevante, dar pe care sistemul din mai multe motive nu le-a livrat. Astfel, pentru cererea de informare C1 plenitudinea va constitui 76,4%, iar precizia – 65%.

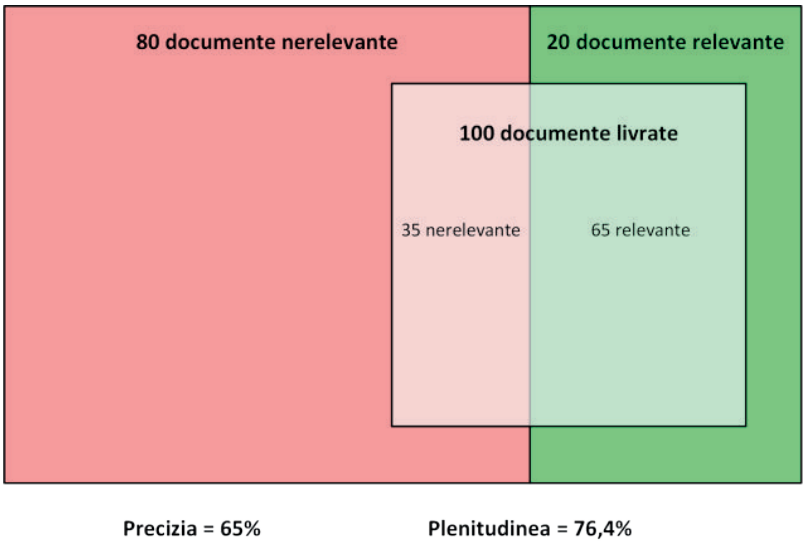

Fig. 4. Matricea precizie-plenitudine (adaptată după James D. McCaffrey 2016)

Testele Cranfield au introdus o nouă paradigmă în evaluarea sistemelor de informare şi continuă să fie un model experimental dominant utilizat în eforturile de evaluare ale mai multor grupuri internaţionale precum SMART (System for the Mechanical Analysis and Retrieval of Text), Text Retrieval Conference (TREC), Cross-Language Evaluation Forum (CLEF) şi NII-NACSIS Test Collection for IR Systems (NTCIR).

Proiectul SMART a adaptat această metodologie la mediul computerizat şi a promovat un model de cercetare şi publicare bazat pe un experiment automatizat. Acesta a fost iniţiat de Gerard Salton şi colaboratorii săi la începutul anilor 1960, mai întâi la Universitatea Harvard, apoi la Universitatea Cornell, şi a continuat într-o formă sau alta până la sfârşitul anilor 1990. Evaluarea SMART a folosit modelul dezvoltat la Cranfield, iar după ce a devenit disponibilă în formă uşor de citit, colecţia Cranfield a fost inclusă în suita de testare SMART. Noutatea pe care o aduce modelul SMART este importantă atât pentru metoda de regăsire, cât şi pentru metodologia de evaluare şi presupune că răspunsul la o căutare ar trebui să fie mai degrabă o listă de documente clasificate în funcţie de gradul (sau probabilitatea) de relevanţă decât un set neordonat de potriviri exacte.

Întregul potenţial pentru evaluarea la scară largă a procesului de regăsire a informaţiei nu a fost realizat însă până la producerea de colecţii de teste TREC. Proiectul TREC, iniţiat în 1992 la Institutul Naţional de Standarde şi Tehnologie (NIST, o agenţie guvernamentală a Statelor Unite), a avut drept motivaţie principală necesitatea de a crea colecţii de dimensiuni realiste şi de a consolida şi extinde cercetarea în tehnologia de regăsire a informaţiei prin utilizarea datelor experimentale comune şi de înaltă calitate, precum şi a tehnicilor standard de evaluare (Prange, 1996; Harman, 1992b). A fost un exerciţiu experimental de mare anvergură, colaborativ şi comparativ, cu participarea a peste 250 de grupuri internaţionale de cercetare (Voorhees şi Harman, 2005a), având rezultate inovatoare nu doar în ceea ce priveşte mărimea colecţiei, ci şi cu referire la metoda experimentală. Impactul TREC asupra cercetării de căutare a informaţiilor a fost la fel de semnificativ. Au fost construite colecţii mari de teste atât pentru regăsirea tradiţională ad-hoc, cât şi pentru noi sarcini, cum ar fi regăsirea în limbajul transversal, regăsirea în limbaje controlate (vocabulare). Proiectul TREC a standardizat metodologia de evaluare utilizată pentru a evalua calitatea rezultatelor de regăsire, atât prin dimensiunea colecţiei de date, cât şi prin eficacitatea metodologiei (Voorhees, Ellen M. 2006).

În lipsa unui consens la aplicarea metodelor alternative de evaluare a sistemelor informative, modelul Cranfield rămâne abordarea de bază în domeniul regăsirii informaţiei, fiind aplicată în majoritatea proiectelor la această sferă.

Referinţe bibliografice:

1. cLEVERDON, Cyril (1967). The Cranfield tests on index language devices. In: *Aslib Proceedings*, Vol. 19, Issue 6, pp. 173-194. doi:https://doi.org/10.1108/eb050097.
2. HINKELMANN, Knut. Information Retrieval and Knowledge Organisation [online]. Available at: http://www.hinkelmann.ch/knut/lectures/irko/IRKO-InformationRetrieval_eval.pdf.
3. HARTER, Stephen P. (1996). Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness. In: *Journal of the American Society for Information Science*, vol. 47(1), pp. 37-49. Available at: https://asis.org/Publications/JASIS/Best_Jasist/1997Harter.pdf
4. LESK, M. E., SALTON, G. (1968). Relevance assessments and retrieval system evaluation. In: *Information Storage and Retrieval* [online], vol. 4, issue 4, pp. 343-359. Available at: https://doi.org/10.1016/0020-0271(68)90029-6.
5. SALTON, Gerard (). The Smart environment for retrieval system evaluation-advantages and problem areas [online]. Available at: http://sigir.org/files/museum/Information_Retrieval_Experiment /pdfs/p316-salton.pdf

6. VOORHEES, Ellen M. (2002) The Philosophy of Information Retrieval Evaluation. In: PETERS C., BRASCHLER M., GONZALO J., KLUCK M. (eds) *Evaluation of Cross-Language Information Retrieval Systems*. CLEF 2001. Lecture Notes in Computer Science [online], vol 2406. Springer, Berlin, Heidelberg. Available at: https://link.springer.com/chapter/10.1007/3-540-45691-0_34#citeas
7. VOORHEES, Ellen M. (2006). TREC: Improving information access through evaluation. In: *Bulletin of the American Society for Information Science and Technology* [online], vol. 32, pp. 16-21. doi:10.1002/bult.2003.1720320105. Available at: https://onlinelibrary.wiley.com/doi/full/10.1002 /bult.2003.1720320105
8. ZUVA, K., ZUVA, T. (2012). EVALUATION OF INFORMATION RETRIEVAL SYSTEMS. In: *International Journal of Computer Science & Information Technology* (IJCSIT), Vol 4, No 3, pp. 35-43. DOI : 10.5121/ijcsit.2012.4304 35