Algoritam određivanja semantičke sličnosti između korisničkog profila i pitanja

Bojan Furlan, Jovana Stamenković, Boško Nikolić i Marko Mišić

1

***Apstrakt*—U radu je opisan softverski sistem koji vrši određivanje semantičke sličnosti između korisničkog profila i pitanja. Opisan je predloženi algoritam koji objedinjuje metode za određivanje semantičke sličnosti, kao i računanje sličnosti na nivou niza karaktera. Ovaj algoritam uvodi težinu dodeljenu svakoj ključnoj reči na osnovu identifikovane važnosti u tekstu koji opisuje pitanje ili profil korisnika. Na kraju su prikazani rezultati izvršene evaluacije predloženog rešenja i jednog od postojećih rešenja i data je njihova uporedna analiza.**

***Ključne reči*—question routing algorithm, similarity of user profiles and questions, corpus-based measures.**

I. Uvod

TRENUTNI napori u domenu Informacionih tehnologija uključuju i razvoj Semantičkog Veba, čiji je primarni cilj lakše pronalaženje i kombinovanje informacija. Jedan od sistema koji, doduše u manjoj meri, pokušava da ostvari ovu ideju je aplikacija "aLive" [1], koja predstavlja sistem za inteligentno prosleđivanje pitanja. Ključna funkcionalnost sistema je mogućnost da korisnik uputi pitanje sistemu koji će obezbediti kvalitetan odgovor pretragom dostupnih korisnika. Odgovor se obezbeđuje tako što sistem bira određeni broj korisnika koji su kompetentni (eksperti) za dato pitanje i prosleđuje im pitanje kao instant poruku. Odabrani korisnici mogu dati odgovor na pitanje, koji se zatim vraća korisniku koji ga je postavio.

S obzirom na prirodu problema, jedno od rešenja za određivanje kompetentnih korisnika za dato pitanje je upotreba nekog od algoritama za računanje sličnosti između dva kratka teksta (ili dve rečenice). U radu je opisan predloženi algoritam nazvan Profile-To-Question Similarity (P2QSim), koji predstavlja sinergiju postojećih algoritama [2] [3]. Novo rešenje uvodi težinu dodeljenu svakoj reči na osnovu njene identifikovane važnosti u tekstu koji opisuje pitanje ili profil korisnika. Fokus istraživanja opisanog u radu stavljen je na određivanje semantičke sličnosti između korisničkog profila i pitanja, dok ekstrakcija ključnih reči iz teksta i određivanje njihove težine predstavljaju zaseban istraživački problem. Jedno od rešenja može se naći u [4].

U narednim odeljcima data je analiza postojećih rešenja i prikazana je upotreba dostupnih alata u cilju izgradnje realizovanog softverskog sistema. Opisane su faze izvršavanja algoritma, uz naznačene modifikacije u odnosu na STS algoritam [2], prikazani rezultati izvršene evaluacije ova dva algoritma i data njihova uporedna analiza.

Bojan Furlan, Jovana Stamenković i Boško Nikolić - Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar Kralja Aleksandra 73, 11020 Beograd, Srbija (e-mail: bojan.furlan@etf.bg.ac.rs, jovana.stamenkovic@gmail.com, bosko.nikolic@etf.bg.ac.rs)

II. Postojeća i upotrebljena rešenja

U ovom poglavlju data je analiza postojećih rešenja, a zatim su navedeni upotrebljeni algoritmi i alati.

Na osnovu analize postojećih pristupa [5] zaključeno je da se nijedno od dostupnih rešenja ne može direktno primeniti na problem određivanja sličnosti između korisničkog profila i pitanja. Stoga je realizovan algoritam zasnovan na sinergiji pristupa [2] (Islam & Inkpen) i [3] (Mihalcea et al.) iz tri sledeća razloga.

Prvo, pristupi ne moraju koristiti bilo kakvu spoljnu bazu znanja (npr. WordNet), ručno kreirana pravila zaključivanja ili specifične jezičke alate, što bi predstavljalo prepreku u radu sa jezicima koji nemaju ovakave resurse.

Osim toga, pristup [2] ne koristi samo meru semantičke sličnosti, već uključuje meru sličnosti na nivou niza karaktera i daje bolje rezultate kod različitih oblika retkih imenica, što je jedna od glavnih slabosti metoda zasnovanih na rečniku [6]. Iako ova prednost nije toliko izražena kod engleskog jezika, može biti značajna kod jezika sa većim brojem fleksija, kakav je srpski. Shodno tome, rezultati evaluacije nad korpusom na engleskom mogu se znatno razlikovati od onih dobijenih za druge jezike.

Najzad, glavna razlika između pristupa [2] u odnosu na pristup [3] odnosi se na sledeće: pristup [2] pronalazi par najsličnijih reči iz rečenica Ra i Rb, uparuje ih i odstranjuje iz daljeg razmatranja. S druge strane, [3] dozvoljava da više reči iz rečenice Ra budu najsličnije sa jednom istom reči iz rečenice Rb, što je implicitno nemoguće u pristupu [2], jer se pronađen par reči odstranjuje iz daljeg razmatranja. Ova činjenica dovodi do nekih pogrešnih procena sličnosti. Kao primer uzmimo dve sintagme: „botanička bašta“ i „kraljev vrt“. Pristup [2] bi najpre uparo reči „bašta“ i „vrt“ kao najsličniji par reči i odstranio ih iz razmatranja dodelivši im visoku ocenu sličnosti. Zatim bi poredio jedine dve preostale reči – „botanička“ i „kraljev“, zaključio bi da one nemaju veliki stepen sličnosti i dodelio nisku ocenu. Ovakav pristup uravnotežuje ocenu sličnosti, tj. prepoznaje postojanje sličnosti među iskazima, kao i da ona nije naročito visoka. S druge strane, pristup [3] bi najpre analizirao prvu sintagmu i zaključio da je reč „bašta“ najsličnija reči „vrt“ iz druge celine, ali bi takođe zaključio da je i reči „botanička“ najsličnija reč „vrt“, tj. „vrt“ bi dva puta figurirao u oceni sličnosti. Zatim bi algoritam analizirao drugu sintagmu, uparo opet reč „vrt” sa „baštom“, a "kraljev" sa bilo kojom od dve

reči iz prve sintagme. Ovaj pristup, dakle, ima tendenciju precenjivanja sličnosti, jer dozvoljava da se više reči iz jedne rečenice upari sa jednom istom reči u drugoj rečenici. Problem je donekle prevaziđen u pristupu [3], pošto se porede samo iste vrste reči.

Poslednja stavka predstavlja značajan problem prilikom određivanja semantičke sličnosti dva kratka teksta, imajući u vidu da je bitno odrediti koji su parovi rečenica slični, ali takođe i koji parovi predstavljaju semantički različite rečenice [7]. To nije slučaj kod pronalaženja kompetentnih korisnika, tj. računanja sličnosti para (pitanje, korisnički profil), jer je neophodno odrediti samo koji je profil najsličniji datom pitanju, pa je samim tim potrebno odrediti najveću (maksimalnu) sličnost ovih parova, dok nije potrebno odrediti i one koji su različiti. Zbog toga se predloženo rešenje zasniva na pristupu [3], ali kombinuje i meru sličnosti na nivou niza karaktera iz [2], kako bi poboljšalo dobijene rezulatate.

Radi realizacije P2QSim sistema, u nastavku su razmotreni dostupni alati i algoritmi za uklanjanje završetka reči (*stemming*) i određivanje leksičke i semantičke sličnosti. Na kraju je utvrđena njihova primenljivost na dati problem i izabrano najbolje moguće rešenje.

Uklanjanje završetka reči (*stemming*) predstavlja transformaciju uklanjanja sufiksa, pri čemu se ne gubi osnovni semantički sadržaj. U ovom radu upotrebljen je standardni postupak za engleski jezik pod nazivom Porter stemmer. Leksička sličnost se zasniva na analizi poklapanja reči na nivou karaktera, odnosno njihovih delova [2]. Za semantičko poklapanje među rečima koristi se postojeća baza podataka [8] sa vektorskim reprezentacijama semantičke sličnosti, čijim se kosinusnim upoređivanjem dobija konačna semantička sličnost dveju reči. Sličnost između reči određuje se kao linearna kombinacija dobijene leksičke i semantičke sličnosti.

III. PREDLOŽENO REŠENJE

U nastavku je opisan predloženi algoritam, uz naznačene modifikacije u odnosu na STS algoritam [2]. P2QSim algoritam izvršava se u 11 koraka:

1. Ulazni podaci predstavljaju dva niza (1)(2) koje čine parovi tokena – ključnih reči označenih sa p i r i njima dodeljenih ponderisanih vrednosti – težina označenih sa q i u. P sadrži m parova, dok R sadrži n parova, gde je m<=n. Glavna razlika ovog koraka u odnosu na STS je što se rečima dodeljuje i njihova važnost u određenom tekstu, tj. težina te reči.

$$P = \{(p_1, q_1), (p_2, q_2), \dots (p_m, q_m)\} \quad (1)$$

$$R = \{(r_1, u_1), (r_2, u_2), \dots (r_n, u_n)\} \quad (2)$$

2. Nad tokenima iz P i R vrši se uklanjanje nastavaka reči. Težine reči kao i brojevi elemenata ovih nizova ostaju isti.
3. Konstruiše se matrica dimenzija $m \times n$ odnosno *matrica leksičkog poklapanja* $M1_p$: Računa se vrednost leksičkog poklapanja α za svaki token iz para (token, težina) iz nizova P i R. Množi se $\alpha_{ij} \times q_i$ i ta vrednost se smešta u vrstu *i* i kolonu *j* matrice *leksičke sličnosti P* (3).

$$M1_p = \begin{pmatrix} \alpha_{11} \times q_1 & \cdots & \alpha_{1n} \times q_1 \\ \vdots & \ddots & \vdots \\ \alpha_{m1} \times q_m & \cdots & \alpha_{mn} \times q_m \end{pmatrix} \quad (3)$$

4. Formira se matrica dimenzija $m \times n$, *matrica semantičke sličnosti* $M2_p$ (4), ali sada njeni elementi predstavljaju meru semantičke sličnosti. Za svaki token (iz parova) iz nizova P i R izračunava se semantička sličnost, gde se dobija vrednost β_{ij}, zatim se vrednost β_{ij} množi odgovarajućom težinom reči iz niza P - q_i i ta se vrednost smešta u vrstu *i* i kolonu *j*. Množenje težinama je jedina razlika u odnosu na STS algoritam u ovom koraku.

$$M2_p = \begin{pmatrix} \beta_{11} \times q_1 & \cdots & \beta_{1n} \times q_1 \\ \vdots & \ddots & \vdots \\ \beta_{m1} \times q_m & \cdots & \beta_{mn} \times q_m \end{pmatrix} \quad (4)$$

5. Konstruiše se udružena matrica (6) (*joint matrix*), koja je istih dimenzija kao i $M1_p$, $M2_p$. Formira se po (5):

$$M_p = \psi M1_p + \varphi M2_p, \psi + \varphi = 1 \quad (5)$$

ψ i φ predstavljaju težine (važnost) leksičkog i semantičkog poklapanja, respektivno. Važi: $\varphi = \psi = 0.5$.

$$M_p = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mn)} \end{pmatrix} \quad (6)$$

6. Vrši se sabiranje svih maksimalnih elemenata po kolonama matrice M_p i dobija se vrednost $maxSim_p$
7. Zbog uvedenih težina potrebno je računanje i matrica sličnosti R sa P. Dalji koraci ne postoje u originalnom algoritmu, pošto su zbog nedostatka težina matrice sličnosti P sa R i R sa P iste. Dakle, kostruiše se matrica dimenzija $n \times m$, odnosno matrica *leksičke sličnosti R* (7) na isti način kao u koraku 3.

$$M1_r = \begin{pmatrix} \alpha_{11} \times u_1 & \cdots & \alpha_{1m} \times u_1 \\ \vdots & \ddots & \vdots \\ \alpha_{n1} \times u_n & \cdots & \alpha_{nm} \times u_n \end{pmatrix} \quad (7)$$

8. Formira se matrica dimenzija $n \times m$, *matrica semantičke sličnosti R* (8) analogno koraku 4.

$$M2_r = \begin{pmatrix} \beta_{11} \times u_1 & \cdots & \beta_{nm} \times u_1 \\ \vdots & \ddots & \vdots \\ \beta_{n1} \times u_n & \cdots & \beta_{nm} \times u_n \end{pmatrix} \quad (8)$$

9. Konstruiše se udružena matrica (10) (*joint matrix*) istih dimenzija kao i $M1_r, M2_r$. Formira se po (9), analogno koraku 5.

$$M_r = \psi M1_r + \varphi M2_r, \psi + \varphi = 1, \varphi = \psi = 0.5 \quad (9)$$

$$M_r = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nm)} \end{pmatrix} \quad (10)$$

10. Vrši se sabiranje svih maksimalnih elemenata po kolonama matrice M_r i dobija se vrednost $maxSim_r$
11. Konačna sličnost S(P,R) dobija se (11).

$$S = \frac{1}{2} \times \frac{maxSim_p}{\sum_{i=1}^{m} q_i} + \frac{1}{2} \times \frac{maxSim_r}{\sum_{i=1}^{n} u_i} \quad (11)$$

IV. Formiranje skupa podataka

Iz korpusa Vikipedije slobodnim odabirom preuzeto je 20 članaka iz oblasti koje se međusobno ne preklapaju, tj. dovoljno su semantički različite (npr. biologija, fotografija, vajarstvo, arheologija, menadžment ljudskih resursa). Iz svakog članka odabrane su ključne reči i njima je dodeljena težina [0,1] – u odnosu na važnost ključne reči za tu oblast. Za članak o biologiji (https://en.wikipedia.org/wiki/Biology) dat je primer u Tabeli I sa ključnim rečima i dodeljenim težinama.

TABELA I
PRIMER KLJUČNIH REČI I NJIHOVIH TEŽINA ZA OBLAST BIOLOGIJA.

Ključna reč	**Težina**
biology	1
Cell	0.5
Ecology	0.5
gene	0.6
Heredity	0.2
living organism	0.7
molecular biology	0.7
organ	0.2
organ system	0.3
tissue	0.5

Ključne reči dodeljene su fiktivnom korisniku, koga nazivamo *biology* i koji će predstavljati eksperta iz oblasti biologija. Postupak je ponovljen za svaku od 20 odabranih oblast. Takođe, za svaku oblast generisan je određeni broj pitanja za koja su ručno određene ključne reči i njihove težine u pitanju. Ukupno je postavljeno 62 pitanja. Primer jednog pitanja sa parovima (ključna reč, težina) dat je u Tabeli II.

TABELA II
PRIMER PITANJA: "*HOW ARE LIVING ORGANISMS CONSUMING AND TRANSFORMING ENERGY?*" SA SKUPOM KLJUČNIH REČI I TEŽINA.

Ključna reč	**Težina**
Cosuming	0.4
Energy	0.8
living organisms	0.9
Transforming	0.4

Kompetentan korisnik koji može dati odgovor na pitanje prikazano u tabeli II biće korisnik *biology*. Kako bi se obezbedilo da više korisnika može biti kompetentno za istu oblast, što je realnost, dodato je još fiktivnih korisnika na sledeći način: za svaku od izabranih 20 oblasti određene su još 4 slične oblasti ili podoblasti. Za primer oblasti biologija uzete su sledeće podoblasti: cell biology (ćelijska biologija), genetics (genetika), biochemistry (biohemija) i microbiology (mikrobiologija). Iz teksta članaka za ove podoblasti određene su ključne reči i njihove težine tako da je za svaku oblast obezbeđeno po 5 fiktivnih korisnika, kompetentnih da daju odgovor na zadato pitanje. Na kraju, prilikom određivanja kompetentnosti korisnika u odnosu na pitanje, uzeto je da je jednako tačan odabir bilo kog od 5 korisnika dodeljenih oblasti iz koje je pitanje.

V. Evaluacija

Za evaluaciju opisanih algoritama upotrebljene su dve mere preciznosti P@N i MAP. Za svaki scenario postavljano je 62 pitanja iz 20 oblasti.

P@N predstavlja meru preciznosti kao odnos tačnih odgovora među svim postavljenim pitanjima. Ova vrednost se računa kao procenat tačno određenih kompetentnih korisnika za grupu pitanja. N predstavlja broj korisnika koji je uzet u razmatranje za jedno pitanje. U postupku evaluacije razmatrana je preciznost za N = 1, 3 i 5. U slučaju da je N>1 (više od jednog korisnika je uzeto u razmatranje), neophodno je bilo odrediti procenat tačno određenih kompetentnih korisnika, tj. onih koji će umeti tačno da odgovore na postavljeno pitanje. Određivanje preciznosti i rangiranje kompetentnih korisnika izvršeno je na sledeći način: u idealnom slučaju, sva tri korisnika bi predstavljala kompetentne korisnike koji bi mogli da odgovore na pitanje, pri čemu bi najkompetentniji bio rangiran kao prvi, drugi po kompetentnosti drugi, itd. Za svaku poziciju, odnosno rang korisnika, dodeljena je određena težina tako da se u zavisnosti od toga da li je izabrani korisnik kompetentan ili ne, kao i na kojoj poziciji se nalazi unutrar rangirane liste (dobijene kao rezultat izvršavanja algoritma) određuje vrednost kojom doprinosi u konačnom rezultatu. Određivanje preciznosti za korisnika U_i dobija se po (12).

$$p(U_i) = \frac{\sum_{i=1}^{N} R(i) * q(U_i)}{\sum_{i=1}^{N} R(i)} \quad (12)$$

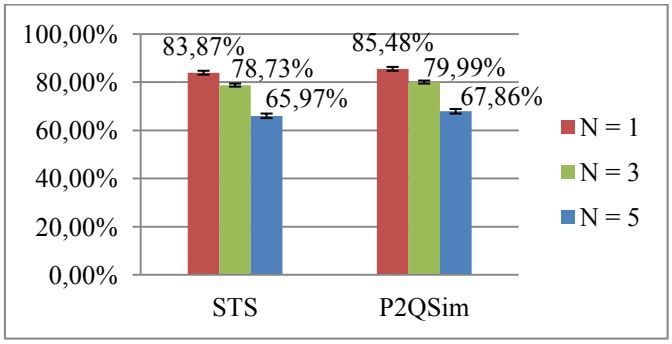
U (12) N = 1, 3 ili 5, a $q(U_i) = 1$ ukoliko je korisnik U_i kompetentan i nalazi se među prvih N razmatranih korisnika, u suprotnom $q(U_i) = 0$. Kompletna lista vrednosti R(i) za dati rang i data je u Tabeli III.

TABELA III
LISTA VREDNOSTI U ZAVISNOSTI OD RANGIRANE POZICIJE.

Rang - i	**Težina - R(i)**
1.	1
2.	0.9
3.	0.8
4.	0.7
5.	0.6

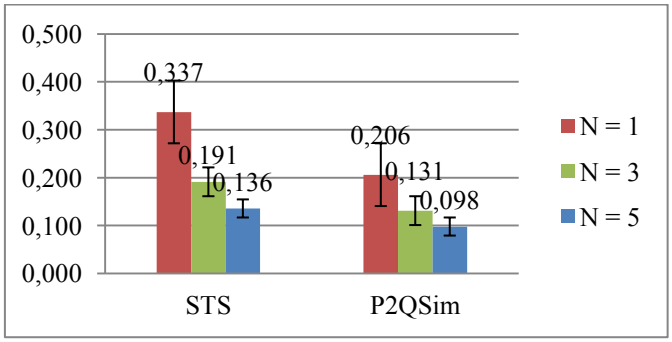
Na kraju, za svako pitanje se dodeljene vrednosti sumiraju i dele konstantom koja predstavlja maksimalan zbir vrednosti za dato N, npr. $\sum_{i=1}^{N=3} R(i) = 1 + 0{,}9 + 0{,}8 = 2{,}7$. Na ovaj način se za svako pitanje računa procenat tačno određenih korisnika u okviru razmatranog prozora N, nakon čega se dobijene vrednosti usrednjavaju za 62 pitanja iz 20 oblasti.

Rezultate dobijene preciznosti P@N za STS i P2QSim u postupku evaluacije nad prethodno opisanim skupom podataka pokazuje sl. 1. Za oba upotrebljena algoritma dobijena preciznost prevazilazi slučajnu vrednost od 25% (verovatnoća slučajno dobijenog korisnika za N=5 je 5:20). Algoritam P2QSim za sve tri vrednosti parmetra N daje bolje rezultate u odnosu na STS, pri čemu razlika između dobijene preciznosti ova dva algoritma ne opada ispod 1%.



Sl. 1. P@N - Rezultati dobijeni za oba algoritma u zavisnosti od parametra N

Drugi analizirani parametar je MAP. On predstavlja meru preciznosti koja se računa kao srednja prosečna preciznost za grupu pitanja. Prosečna preciznost (P) za svako pitanje računa se kao prosek vrednosti preciznosti svih tačno dobijenih kompetentnih korisnika, gde p predstavlja vrednost preciznosti za jednog tačno izabranog korisnika. Kao i za P@N, parametar N uzima vrednosti 1, 3 i 5. Postupak računanja parametra MAP se sastojao u izračunavanju prosečnog p u okviru jednog pitanja za sve tačno određene korisnike. Na osnovu dobijenih prosečnih vrednosti za određeno pitanje sračunata je srednja vrednost preciznosti P za svih 62 pitanja iz 20 oblasti (Sl. 2).



Sl. 2. MAP - Rezultati dobijeni za oba algoritma u zavisnosti od parametra N

Prilikom određivanja parametra MAP, STS daje bolje rezultate, pri čemu sa povećanjem N razlika između ova dva algoritma opada, gde za N=5 razlika iznosi samo 0.038.

VI. Zaključak

U radu je opisan predloženi algoritam P2QSim koji objedinjuje metode za određivanje semantičke sličnosti kao i računanje sličnosti na nivou niza karaktera. Rezultati dobijeni evaluacijom preciznosti P@N pokazuju da P2QSim za sve tri vrednosti parametra N daje bolje rezultate od STS. Razlika između dobijene preciznosti ovih algoritama ne opada ispod 1%. Mera preciznosti MAP pokazuje da STS daje bolje rezultate, ali sa povećanjem N razlika između ova dva algoritma opada, gde za N=5 iznosi 0.038.

Zahvalnica

Ovaj rad je delimično finansiran od strane Ministarstva nauke i prosvete Republike Srbije (projekti III44009, 44006, 32047).

Literatura

[1] Nikolić S., Furlan B., Josipović P., "aLive! – Intelligent Question Forwarding System," In YUINFO, Kopaonik, Serbia, 2008.

[2] A. Islam, D. Inkpen, Semantic Text Similarity Using Corpus-based Word Similarity and String Similarity, ACM Transactions on Knowledge Discovery from Data 2 (2) (2008) 1-25.

[3] R. Mihalcea, C. Corley, C. Strapparava, Corpus-based and Knowledge-based Measures of Text Semantic Similarity, Proceedings of the National Conference on Artificial Intelligence 21 (1) (2006) 775-780.

[4] Varga E., Furlan B., and Milutinovic V., "Document Filter Based on Extracted Concepts," Transactions on Internet Research, vol. 6, no. 1, pp. 5-9, January 2010.

[5] Furlan B., Nikolić B., Milutinović V., "A Survey and Evaluation of State-of-the-Art Intelligent Question Routing Systems," International Journal of Intelligent Systems, ISSN 1098-111X, 2013., DOI 10.1002/int.21597

[6] B. Furlan, V. Sivački, D. Jovanović, B. Nikolić, Comparable Evaluation of Contemporary Corpus-Based and Knowledge-Based Semantic Similarity Measures of Short Texts, Journal of Information Technology and Applications 1 (1) (2011) 65-71

[7] Furlan B., Batanović V., Nikolić B., "Semantic Similarity of Short Texts in Languages with a Deficient Natural Language Processing Support", Decision Support Systems, ISSN 0167-9236, 2013. DOI 10.1016/j.dss.2013.02.002

[8] Jovanović D., Furlan B., Nikolić B., "Softverski sistem za automatsko određivanje semantičke sličnosti kratkog teksta," ETRAN, Banja Vrućica (Teslić), R. Srpska, BIH, 6-9. Juna, 2011.

Abstract

This paper describes a software system for determining semantic similarity between a user profile and a question. The proposed algorithm combines methods for semantic similarity determination, as well as string similarity. Also, the weight assigned to each keyword, which determined by identified importance in text that describes question or user profile is introduced in the algorithm. The evaluation results of the proposed solution and one of the existing one are presented, and comparative analysis is given.

An Algorithm for Measuring the Semantic Similarity between a User Profile and a Question

Bojan Furlan, Jovana Stamenković, Bosko Nikolić, and Marko Mišić