УНИВЕРЗИТЕТ У БЕОГРАДУ

ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

Бојан Д. Фурлан

МЕТОДОЛОГИЈА ПРОЈЕКТОВАЊА СИСТЕМА ЗА ИНТЕЛИГЕНТНО ПРОСЛЕЂИВАЊЕ ПИТАЊА НАПИСАНИХ НА ПРИРОДНОМ ЈЕЗИКУ

докторска дисертација

Београд, 2013

UNIVERSITY OF BELGRADE

SCHOOL OF ELECTIRCAL ENGINEERING

Bojan D. Furlan

A METHODOLOGY FOR THE INTELLIGENT QUESTION ROUTING SYSTEMS DESIGN

Doctoral Dissertation

Belgrade, 2013

Комисија за преглед и оцену докторске дисертације

МЕНТОР

др Бошко Николић, ванредни професор,
Универзитет у Београду – Електротехнички факултет

ЧЛАНОВИ КОМИСИЈЕ

др Вељко Милутиновић, редовни професор,
Универзитет у Београду – Електротехнички факултет

др Ненад Митић, ванредни професор,
Универзитет у Београду – Математички факултет

др Драган Милићев, ванредни професор,
Универзитет у Београду – Електротехнички факултет

др Милош Цветановић, доцент,
Универзитет у Београду – Електротехнички факултет

Датум одбране: _____

Захвалност

Желео бих прво да се захвалим мом ментору, проф. др Бошку Николићу, на подршци током израде ове дисертације као и током целог мог стручног рада. Такође, сматрам веома битним то што ми је с једне стране пружио велику слободу у избору теме и начину израде дисертације, а са друге стране што је умео својим саветима и упутствима да ме усмери и мотивише ка конкретним резултатима. Желео бих да се захвалим и професорима др Вељку Милутиновићу и др Ненаду Митићу с којима сам радио на темама блиским дисертацији, као и свим колегама са катедре за Рачунарску технику и информатику са којима сам сарађивао.

Овај рад не би био успешно реализован до краја да није било драгоцене помоћи сарадника који су учествовали на изради различитих верзија прототипских алата за одређивање семантичке сличности кратких текстова, као и алата за семантичку анализу и екстракцију информација из текста. Захваљујем се колеги Вуку Батановићу који је у оквиру свог мастер рада уложио велики труд приликом састављања и оцењивања корпуса парафраза на српском језику. Такође, он је реализовао моју идеју да је за одређивање семантичке сличности два кратка текста потребно узети у обзир и специфичности речи које ови текстови садрже. Ово ми је касније послужило као добра основа за реализацију алгоритма за одређивање семантичке сличности између питања и корисничких профила. Желео бих да се захвалим и Славку Житнику са Факултета за рачунарство и информатику из Љубљане, који ми је помагао приликом реализације корпуса питања и одговора, као и при евалуацији реализованог прототипа над овим корпусом.

Почетак мојих интересовања за ову тему датира још из периода основних студија на Електротехничком факултету када сам заједно са колегама Станком Николићем и Павлетом Јосиповићем у финалу такмичења Imagine Cup 2007 представио идеју реализације једног оваквог система. Са колегом Емилом Варгом радио сам на екстракцији информација из текста и њиховој визуелизацији, а са

Владисавом Јелисавчићем на моделовању тема. У експериментима са раним верзијама алата за поређење кратких текстова радио сам са колегама Давором Јовановићем и Владимиром Сивачким. Такође, са колегиницом Јованом Стаменковић радио сам на раној верзији алгоритма за поређење профила и питања и његовој евалуацији. Сарадња и рад са свима њима за мене је био веома инспиришући и пријатан, на чему сам им дубоко захвалан.

На крају, тешко ми је да пронађем речи којима бих се захвалио мојим родитељима и мојој девојци Драгани на неизмерној подршци, љубави, стрпљењу и разумевању које су ми пружили током година рада на овој теми. Уједно њима и посвећујем овај рад.

У Београду, *Бојан Фурлан*
децембра 2013. године

Наслов дисертације:

МЕТОДОЛОГИЈА ПРОЈЕКТОВАЊА СИСТЕМА ЗА ИНТЕЛИГЕНТНО ПРОСЛЕЂИВАЊЕ ПИТАЊА НАПИСАНИХ НА ПРИРОДНОМ ЈЕЗИКУ

Резиме:

Упркос великом развоју вештачке интелигенције, људски мозак је још увек супериорнији узимајући у обзир разумевање и манипулацију делимично познатим чињеницама. Једна од области у којој је ово нарочито истакнуто односи се на проблем одговарања на питања. Када је потребно дати одговор, посебно онај који се експлицитно не налази у тексту корпуса који се претражује, предности експерта – човека представљају различите способности као што је способност објашњавања, комбиновања сложених одговора и апстрактног резоновања.

Системи за Интелигентно Прослеђивање Питања (СИПП) имају за сврху размену знања на нивоу произвољне области експертизе и од значаја су за велики број апликација у којима се захтева интензивна комуникација између корисника. Корист од примене оваквих система укључује: (а) смањење непотребног оптерећења експерата који представљају вредан ресурс и (б) повећање квалитета услуга институције (универзитет, влада, предузеће), имајући у виду задовољство корисника с обзиром да су њихова питања прослеђена релевантним особама.

У овом раду представљена је методологија пројектовања система за интелигентно прослеђивање питања написаних на природном језику – СИПП. На почетку је дат детаљан преглед ове области где је посебан акценат стављен на реализацију фаза СИПП процеса. Такође, у овом поглављу представљена је оригинална презентациона парадигма која генерализује суштину свих расположивих СИПП решења из отворене литературе. Презентациона парадигма садржи три основне фазе извршавања које се односе на три главна проблема приликом реализације система: анализу питања, прослеђивање питања и профилисање корисничког знања. На основу ове парадигме извршена је детаљна

анализа и евалуација оваквих система, а као закључак наведен је предлог решења уочених проблема.

У наставку дисертације описана је реализација предложених решења у виду прототипа СИПП система. У оквиру модула за обраду питања реализован је приступ који омогућава визуализацију питања, што обезбеђује интуитивну представу специфичних односа између концепата, као и њиховог значаја у питању. Такође, овај приступ комбинује потпуно аутоматску обраду текста и ручну корекцију резултата, пружајући кориснику могућност повећања тачности излаза. Истовремено, реализовани модул за обраду текста употребљен је и за анализу одговора.

Након тога анализирани су и дискутовани постојећи приступи за одређивање семантичке сличности два кратка текста, погодни за језике са врло ограниченим електронским лингвистичким ресурсима, где је посебан акценат стављен на српски језик. На основу донетих закључака предложен је нови алгоритам, назван LInSTSS, који приликом одређивања семантичке сличности два кратка текста узима у обзир и специфичности речи које ови текстови садрже. Такође, реализован је корпус парафраза за српски језик над којим је извршена евалуацију. Резултати добијени над овим корпусом показали су да предложени алгоритам пружа боље резултате у односу на постојећа решења. Коначно, на основу евалуације над корпусима парафраза за српски и енглески извршено је фино подешавање параметара, а такође стечена искуства употребљена су за реализацију модула за одређивање семантичке сличности између питања и корисничког профила.

У оквиру реализације фазе прослеђивања питања, дискутоване су специфичности проблема поређења питања и корисничких профила, и предложен је нови алгоритам назван P2Q. Овај алгоритам одређује највећу (максималну) сличност између концепата идентификованих у питању и оних који се налазе у корисничком профилу. Коначно, анализирани су доступни веб портали и одабран је један чији подаци се употребљени за формирање корпуса питања и одговора. У

корпусу су издвојене три различите врсте корисника, које моделују: (1) интересовање, (2) знање и (3) истовремено и знање и интересовање. Формирани корпус је затим употребљен за евалуацију целокупног система и тестирање полазних хипотеза. Добијени резултати су показали да P2Q приступ пружа знатно боље резултате у односу на остале евалуиране приступе. Такође, уочено је да употреба семантичке екстракције информација из текста може побољшати резултате.

Допринос изложене докторске дисертације је у домену анализе и синтезе једног оваквог софтверског система, који треба да омогући интелигентно прослеђивање питања написаних на природном језику. Такође, на основу резултата евалуације закључено је да правилно додељене тежине могу побољшати перформансе целокупног система, али такође у случају да нису правилно постављене могу их знатно погоршати. Коначно, утврђено је да при профилисању компетентности корисника да пружи одговор на постављено питање, није важно само размотрити његове најбоље одговоре, односно профилисати његово знање, већ је такође потребно узети у обзир и питања која је поставио, с обзиром да она могу изразити интересовање.

Кључне речи:

Интелигентно прослеђивање питања, социјална претрага, сличност питања и корисничког профила, профилисање корисничког знања, семантичка сличност кратких текстова, екстракција информација из текста, креирање корпуса питања и одговора, креирање корпуса парафраза

Научна област:

Електротехника и рачунарство

Ужа научна област:

Рачунарска техника и информатика

УДК број:

621.3

Dissertation title:

A METHODOLOGY FOR THE
INTELLIGENT QUESTION ROUTING SYSTEMS DESIGN

Abstract:

In spite of great developments in artificial intelligence, human brain is still more powerful, concerning the comprehension and manipulation with partially known facts. One domain where this is prominent is related to a problem of question answering. When it comes to giving the answers, especially those that do not explicitly exist in the text corpus, the advantages of a human expert are abilities like explaining, combining complex answers, and abstract reasoning.

Intelligent Question Routing Systems (IQRS) serve as a knowledge exchange medium in an arbitrary field of expertise, where intensive communication between users is required. The benefit coming from deployment of such systems includes: (a) reducing unnecessary "pinging" of experts, which are a valuable resource and (b) increasing the system owners' (e.g. enterprise, government, university) quality of service, since users are more satisfied with answers, because their questions are answered by the right persons.

This dissertation represents a methodology for IQRS systems design. It starts with a survey of the existing research in this domain, where the emphasis was put on the implementation of phases of the IQRS process. The survey also introduces an original presentation paradigm that generalizes the essence of approaches found in the open literature. The presentation paradigm includes three basic processing stages related to the three major problems of system implementation: question analysis, question forwarding, and users' knowledge profiling. The outcome of this analysis is a proposal for a new approach that tackles identified problems.

The rest of the dissertation describes an IQRS prototype which implements the proposed ideas. The question analysis module implements an approach that enables question visualization, thus it provides an intuitive representation of specific relations

between concepts and their importance in the question. Also, this approach combines a fully automatic text processing and manual correction of the results, giving a users ability to increase the accuracy of the output. Simultaneously, the implemented text processing module is used for the analysis of answers.

Next, an analysis and discussion of existing approaches for determining the semantic similarity of two short texts is given, where particularly the focus was put on those suitable for languages with very limited electronic linguistic resources, like the Serbian language. Based on the conclusions a new algorithm is proposed, named LInSTSS, which for calculating the semantic similarity between two short texts includes the specificity of words that these texts contain. Additionally, a Serbian paraphrase corpus is constructed and the results obtained using this corpus showed that the proposed algorithm provides better results when compared to existing solutions. Finally, evaluation on paraphrase corpora both for English and Serbian is used to fine-tune algorithm parameters. Also, the lessons learned are applied to design the module for calculating the semantic similarity between a question and a user profile.

In the question forwarding phase issues related to the problem of comparing questions and user profiles are discussed and a new algorithm called P2Q is proposed. This algorithm determines the highest (maximum) similarity between the concepts identified in question and those from the user profile. Finally, the analysis of available web portals is carried in order to find the one suitable for the creation of a questions and answers corpus. In the created corpus three different types of users are extracted, which model: (1) interests, (2) knowledge, and (3) both knowledge and interests. Corpus is then used to test the initial hypothesis and to evaluate the overall system performances. The results showed that P2Q approach provides significantly better results than other evaluated approaches. It was also noted that the use of semantic information extraction from text can improve results.

Scientific contribution of the dissertation is in the field of analysis and synthesis of a software system, which should enable an intelligent questions routing. Also, based on the evaluation results it was found that properly assigned weights can

improve the overall performances of the system, but also if not assigned correctly performances can be significantly decreased. Finally, it was concluded that for profiling the user competence to give an answer for the provided question it is important not only to consider answers which the user best answered, i.e. to profile knowledge, but also it is important to consider questions which can express interests.

Keywords:

Intelligent question routing, social search, question-to-profile similarity, user knowledge profiling, semantic similarity of short texts, information extraction from text, questions and answers corpora construction, paraphrase corpora construction

Scientific field:

Electrical engineering and computer science

Specialized scientific field:

Computer engineering and information theory

UDK number:

621.3

Садржај

I Увод

Упркос великом развоју вештачке интелигенције, људски мозак је још увек супериорнији узимајући у обзир разумевање и манипулацију делимично познатим чињеницама. Једна од области у којој је ово нарочито истакнуто односи се на проблем одговарања на питања. Када је потребно дати одговор, посебно онај који се експлицитно не налази у тексту корпуса који се претражује, предности експерта – човека представљају различите способности као што је способност објашњавања, комбиновања сложених одговора и апстрактног резоновања.

Основна идеја Система за Интелигентно Прослеђивање Питања – СИПП (*Intelligent Routing Systems* – IQRS) написаних на природном језику може се неформално описати на следећи начин: не тражити од рачунара да разуме људе (што и није способан на тренутном нивоу технолошког развоја), већ тражити да пронађе особу која може пружити тачан одговор на постављено питање. Дакле, ако постоји потреба за неким саветом или упутством, уместо претраживања огромних количина информација које Интернет свакако нуди, може се питати особа која је довољно компетентна да на задато питање пружи кратак и људима разумљив одговор. Као резултат, интелектуални посао је и даље остављен човеку, али терет проналажења праве особе за постављено питање делегиран је ка рачунару.

Сматрајући да су мултидисциплинарност, сарадња и комуникација једни од главних покретача иновативности, СИПП имају за сврху размену знања на нивоу произвољне области експертизе. Из тог разлога развој једног оваквог система је од значаја за велики број апликација у којима се захтева интензивна комуникација између корисника, као што су е-управа, техничка подршка, информациони системи великих предузећа, здравствени систем, војска, итд. Друге примене укључују подршку у образовном и научном процесу, где СИПП омогућава ефикасну и ефективну размену знања између научника, истраживача, универзитетских кадрова и студената. Корист од примене оваквих система укључује: (а) смањење непотребног оптерећења експерата који представљају вредан ресурс и (б) повећање квалитета услуга институције (универзитет, влада,

предузеће), имајући у виду задовољство корисника с обзиром да су њихова питања прослеђена релевантним особама.

Структура и садржај рада

Докторска дисертација садржи шест поглавља, скуп неопходних прилога и преглед коришћене литературе. Прво поглавље представља увод у дисертацију. Друго поглавље даје опис проблема који се решава, као и осврт на повезаност истраживања из овог домена са осталим блиским областима вештачке интелигенције. На крају овог поглавља дат је оквирни опис фаза СИПП процеса. Треће поглавље садржи детаљан преглед области пројектовања система за интелигентно прослеђивање питања где је посебан акценат стављен на реализацију фаза СИПП процеса. Такође, у овом поглављу представљена је оригинална презентациона парадигма која генерализује суштину свих расположивих СИПП решења из отворене литературе и на основу које је извршена њихова детаљна анализа и евалуација. Као закључак наведен је предлог решења уочених проблема. Четврто поглавље даје приказ дизајна софтверског система који треба да омогући генерализован приступ профилисању корисничког знања, као и аналитички модел који треба да опише карактеристике предложеног решења. У оквиру реализације система приказана је и реализација алгоритама за одређивање семантичке сличности два кратка текста са посебним нагласком на српски језик. Такође, ово поглавље обухвата детаљан приказ реализације софтверског система са становишта имплементације, али и са становишта коришћења, као и преглед најзначајнијих проблема и начина на који су ти проблеми решени. Пето поглавље приказује евалуацију предложеног решења. У оквиру овог поглавља описан је корпус парафраза употребљен за евалуацију алгоритма за одређивање семантичке сличности два кратка текста написана на српском језику, као и опис корпуса питања и профила корисника на основу кога је извршена евалуација целокупног система. У шестом поглављу изложен је закључак. На крају су дати неопходни прилози и преглед литературе.

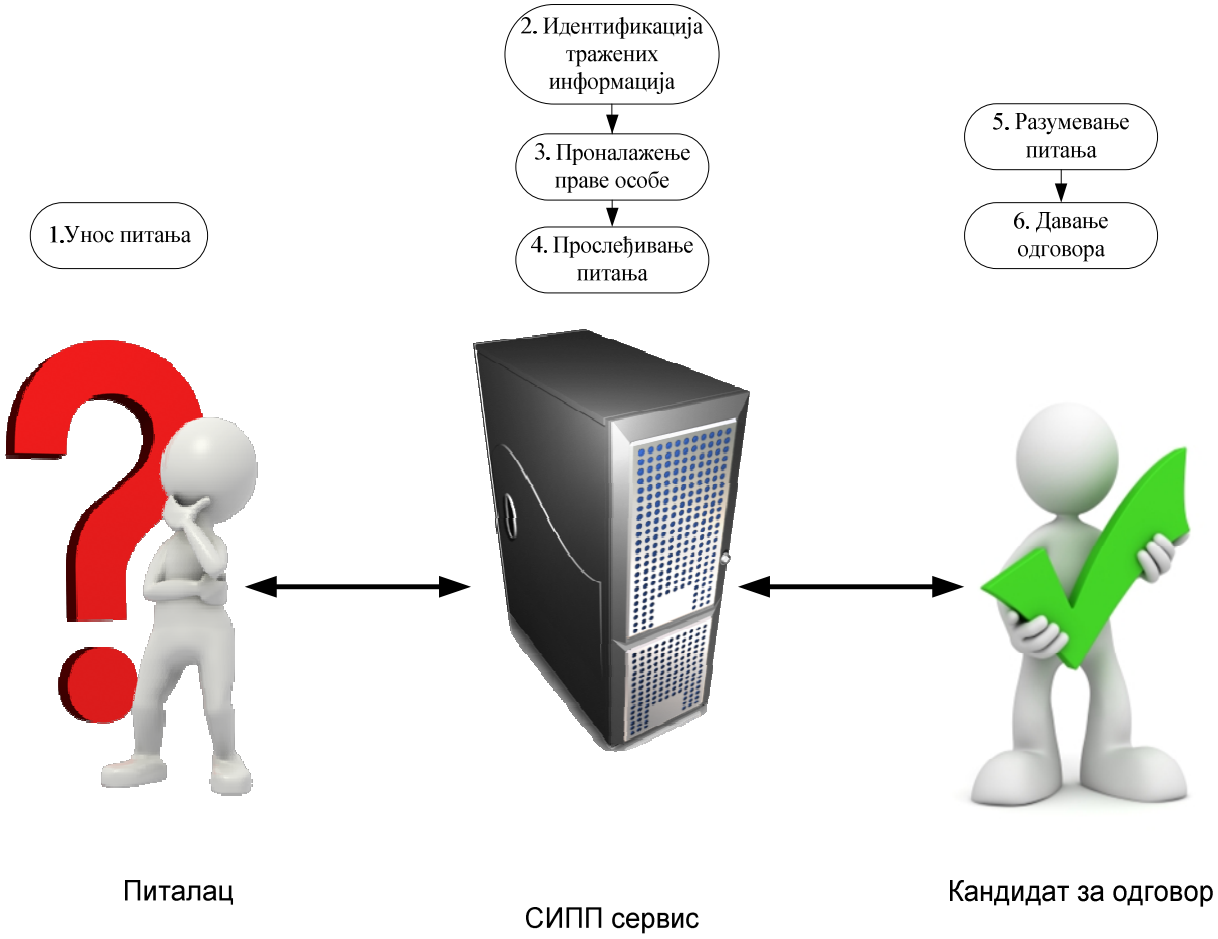
II Дефиниција проблема

Системи за одговарање на питања (*Question Answering* - Q/A) [1] имају задатак да аутоматски дају одговоре на питања постављена у природном језику. Основна сврха ових система је издвајање релевантних и концизних одговора како на општа, тако и на специфична питања. При том циљ је да се фокус измести са проблема проналажења докумената на проблем проналажења информација. За проналажење одговора Q/A системи користе различите изворе података од структуираних база података до великих колекција докумената писаних у природном језику – текстуалних корпуса. Ови текстуални корпуси могу се састојати од формалних докумената како што су обрађени новински чланци [2] или пак оних не стриктно форматираних као што су колекције веблог (*blog*) докумената [3], који могу садржати и различите неправилности или „шум“. Међутим, некада је неопходно усвојити основна знања и велику количину информација из различитих области. С друге стране, много експерта са потребним знањем постоји у оквиру неке институције, предузећа или универзитета. За младог истраживача, студента или некога заинтересованог за нову тему, било би веома корисно да се директно обрати особи која је компетентна у одређеној области и замоли је за савет или упутство. Стога, ефикасност у проналажењу праве особе може се добити помоћу софтверског система за интелигентно прослеђивање питање.

Такође, СИПП домен истраживања је повезан, али се разликују од домена проналажења експерата и анализе релација. Алгоритми засновани на релацијама као што су Рангирање страница (*PageRank*) и HITS (*Hyperlink-Induced Topic Search*) су успешно употребљени у друштвеним мрежама приликом проналажења експерата [4]. Међутим, наведени приступи не узимају у обзир проблем прослеђивања питања, имајући у виду да је фокус само на анализи стручности самог корисника, али не и на специфичности постављеног питања како би се утврдило да ли корисник може на њега да одговори. Коначно, постоје опсежна истраживања спроведена у областима семантичког рутирања упита (*Semantic Query Routing* - SQR) у локалним P2P (*Peer-to-Peer*) мрежама. Један од примера је техника изградње прекривајуће семантичке мреже [5] на основу које се проналазе чворови који су релевантни у односу на дати упит. Упити се усмеравају кроз

супер-чвор коме сваки чвор експлицитно оглашава свој садржај. Други пример је имплицитна идентификација садржаја на основу социјалних метафора [6]. Међутим, домен истраживања СИПП разматра питања у виду текста у слободној форми, насупрот структурираних или полу-структурираних упита. Стога, са напретком алата за обраду текста и недавном експанзијом друштвених мрежа, синергија између Q/A, SQR и система за проналажење експерата је постала могућа, тако да прослеђивање питање између корисника (СИПП) представља отворен научни проблем. Сходно томе, остатак ове дисертације описује истраживање фокусирано искључиво на СИПП и проблеме од значаја за парадигму уведену у овом раду.

Слика 1 илуструје типични СИПП сценарио који се састоји од следећих 6 фаза: (1) унос питања, (2) идентификација тражених информација, (3) проналажење праве особе, (4) прослеђивање питања, (5) разумевање питања и (6) давање одговора. У првој фази, корисник са питањем – питалац, уноси питање и шаље га СИПП сервису. Затим, сервис покушава да идентификује тражене информације из питања на основу којих проналази праву особу (потенцијалног кандидата за одговор) и затим то питање прослеђује овом кориснику. Коначно, остатак посла остаје на човеку: када потенцијални кандидат прими питање, након његовог читања и разумевања, он има могућност да одговори. Фазе (2), (3) и (4) се обрађују у оквиру СИПП сервиса, стога ће оне представљати главни фокус овог истраживања.



Слика 1. Илустрација СИПП процеса: Типичан сценарио

Претпоставке и ограничења

Полазна претпоставка ове студије је да су мултидисциплинарност, сарадња и комуникација једни од главних покретача иновативности, стога СИПП имају за сврху размену знања на нивоу произвољне области експертизе. Имајући ово у виду проблем прослеђивања питања представља сложен процес на који утичу различити статички и динамички параметри, тако да резултати овог истраживања могу имати ширу примену.

Детаљном анализом и евалуацијом расположивих софтверских решења уочено је да не постоји систем који на одговарајући начин може у потпуности да подржи целокупан СИПП процес. Такође, ни једно од анализираних решења није у потпуности погодно за употребу код језика са врло ограниченим електронским лингвистичким ресурсима, као што је српски језик.

Коначно, у овој студији фокус је стављен првенствено на проблем како моделовати компетентност корисника да пружи одговор на дато питање, али не и на вероватноћу да ће на крају одговорити, нпр. услед недоступности. Стога, у предложеном приступу неће бити моделовани атрибути корисника као што су повезаност ка другим корисницима, доступност, спам (непримерена или нежељена питања и одговори), или брзина одзива односно давања одговора. Остали наведени атрибути могу се третирати као независне димензије у проблему вишекритеријумског рангирања (*Multi-Criteria Rating problem*) [7].

Дефиниције појмова

- Кандидат за одговор (*answerer*) – особа односно корисник коме се може проследити постављено питање.

- Питалац (*asker*) – особа односно корисник који поставља питање.

- Шум (*noise*) – текстуалним шумом се може сматрати свака разлика између површинског облика кодиране репрезентације текста и жељеног, исправног, или оригиналног текста. Ово може бити услед нпр. типографске грешке или фразе које су увек присутне у природном језику, што обично смањује квалитет података на начин који чини текст мање доступним за аутоматизовану обраду, као што је обрада природних језика.

- P2P (*Peer-to-Peer*) је модел комуникације путем интернета, насупрот клиент/сервер модела и најчешће се користи за дељење фајлова. P2P је скраћеница од (енгл. *peer* to *peer*) што би се могло превести као вршњак - вршњаку.

- Дата мајнинг (*Data mining* – DM), или откривање законитости у подацима је интердисциплинарно поље информатике које се бави откривањем нових образаца у великим скуповима података. Она користи методе који су у пресеку вештачке интелигенције, машинског учења, статистике, и система база података. Свеукупни циљ овог приступа је екстракција новог знања из постојећих података и трансформација у облик подесан за даљу употребу.

- Машинско учење (*Machine learning* – ML) је подобласт вештачке интелигенције чији је циљ конструисање алгоритама који су способни да се адаптирају на аналогне нове ситуације, као и да уче на бази претходно прикупљеног искуства.

- Кластеровање (*clustering*) или груписање спада у групу алгоритама ненадзираног учења (*unsupervised learning*) и представља класификацију сличних предмета у различите групе – кластере (*clusters*), или прецизније, дељење скупа података у подскупове (групе) тако да подаци у сваком подскупу (идеално) деле неко заједничко обележје – често приближност према некој унапред дефинисаној величини удаљености.

- Стемовање (*stemming*), тј. уклањање завршетака речи представља трансформацију уклањања суфикса речи при чему се не губи основни семантички садржај.

- Веблог (*web log, blog*) чини низ хронолошки организованих уноса текста, који се приказују на веб-страницама (углавном су уноси сортирани од најновијих ка старијим), путем аутоматизованог софтвера који омогућује једноставно креирање и вођење.

- Прецизност (*precision*), осетљивост (*recall*) и Ф-мера (*F-measure*): Прецизност је вероватноћа да је неки, случајно одабран, документ, из пронађених докумената, релевантан. Осетљивост је вероватноћа да је неки, случајно одабран, релевантан документ пронађен уз помоћ датог алгоритма. Ф-мера се рачуна као хармонијска средина прецизности и осетљивости.

- 5W1H тип упита представља основну класификацију питања чији се одговори разматрају у фази прикупљања информација - ко, шта, када, где, зашто, како (*Who, What, When, Where, Why, How*).

- Уклањање стоп речи (*stop words filtering*) представља уклањање речи са ниским информационим садржајем. Такве речи су најчешће чланови, заменице, бројеви и остале помоћне речи без директне информативне вредности у контексту у коме се текст користи (нпр. екстракција концепата или идентификација области).

- Компетентност (*competence*) означава ниво знања везан за неку специфичну области, чија вредност може ићи од површног, на нивоу интересовања, до темељног познавања.

- Стручност или експертиза (*expertise*) представља низ компетентности, тј. познавања скупа области блиско повезаних унутар неког домена.

- Нит (*thread*) представља структуру која садржи питање са повезаним одговорима.

- STSS (*Short Text Semantic Similarity*) је техника одређивања семантичке сличности кратких текстова.

- Спам (*spam*) је порука која се електронском поштом шаље на више адреса, без сагласности прималаца. Обично садржи промоцију неког производа или услуге. Примаоце та порука најчешће иритира, јер им непотребно оптерећује "поштанско сандуче". Међу овакве поруке сврставају се: ланчана писма, игре на срећу, рекламе производа, услуга или порнографских страна на Интернету, понуда пиратског софтвера итд.

III Преглед постојећих и предлог новог решења

Гледиште овог истраживања најбоље је представљено појмовима приказаним на слици 2. Ова слика садржи оригиналну презентациону парадигму која генерализује суштину свих расположивих СИПП решења пронађених у отвореној литературу. Такође, ова парадигма ће бити употребљена за међусобно поређење анализираних система.

Пошто се процес питања и одговарања састоји из два дела "постављање питање" и "давање одговора", представљена структура је такође подељена на два дела који истовремено обрађују: (а) нова питања - процесирање питања (*Question Processing*) и (б) нове или постојеће кориснике - профилисање корисника (*User Profiling*). Оба ова дела се састоје од фаза представљених на слици 2. Свака фаза садржи један или више модула који се реализују коришћењем алгоритма из релативно великог скупа алгоритама. Фаза под називом Анализа Питања (*QA Stage*) садржи модул за *обраду питања* (*Question Analysis*), фаза Прослеђивање Питање (*QF Stage*) садржи модул за *поређење и рангирање* (*Matching&Ranking*) и модул за *достављање питања* (*Question Forwarding*) и на крају фаза за Профилисање Корисника (*UP Stage*) садржи следеће модуле: *профилисање корисничког знања на основу екстерних извора информација* (*User Knowledge Profiling: External Sources*), *профилисање корисничког знања на основу интерних извора информација* (*User Knowledge Profiling: Internal Sources*), *додатне информације* (*Additional Info*), као и *складиште профила корисника* (*User Profiles*).

Свака од ових фаза односи се на један од три основна проблема дефинисана помоћу три питања која су размотрена у наставку текста:

Питање # 1: Како идентификовати тражене информације из питања?

Анализа питања у СИПП систему односи се на проблем како да систем у довољној мери разуме питање како би га проследио компетентном кориснику за одговор. То је знатно једноставнији задатак у односу на изазов са којим је суочен идеални Q/A систем, који мора утврдити шта је тачна информација коју корисник тражи (нпр. да преведе потребне информације у кључне речи за претрагу), да процени да ли пронађени садржај укључује те информације и на крају да

пронађене информације прикаже у људски разумљивом формату. Насупрот томе, у СИПП процесу, човек је тај који даје одговор на питања. Стога, одговорност утврђивања релевантности датог одговора лежи на човеку, што је функција за коју је људска интелигенција врло погодна за обављање.

Процес анализе и екстракције информација из питања је приказан у оквиру модула за *обраду питања*. Овај модул обично издваја релевантне термине, тј. појмове из питања или класификује питање у једну или више унапред утврђених тема. Стога, излазни резултати овог модула су у форми идентификованих термина или тема који се даље прослеђује модулу за *поређење и рангирање*.

Критеријуми од интереса за Питање # 1:

1) Тип интеракције система са корисником:

    i. Са означавањем питања, у виду анотације (*tags*) или предефинисаних категорија.

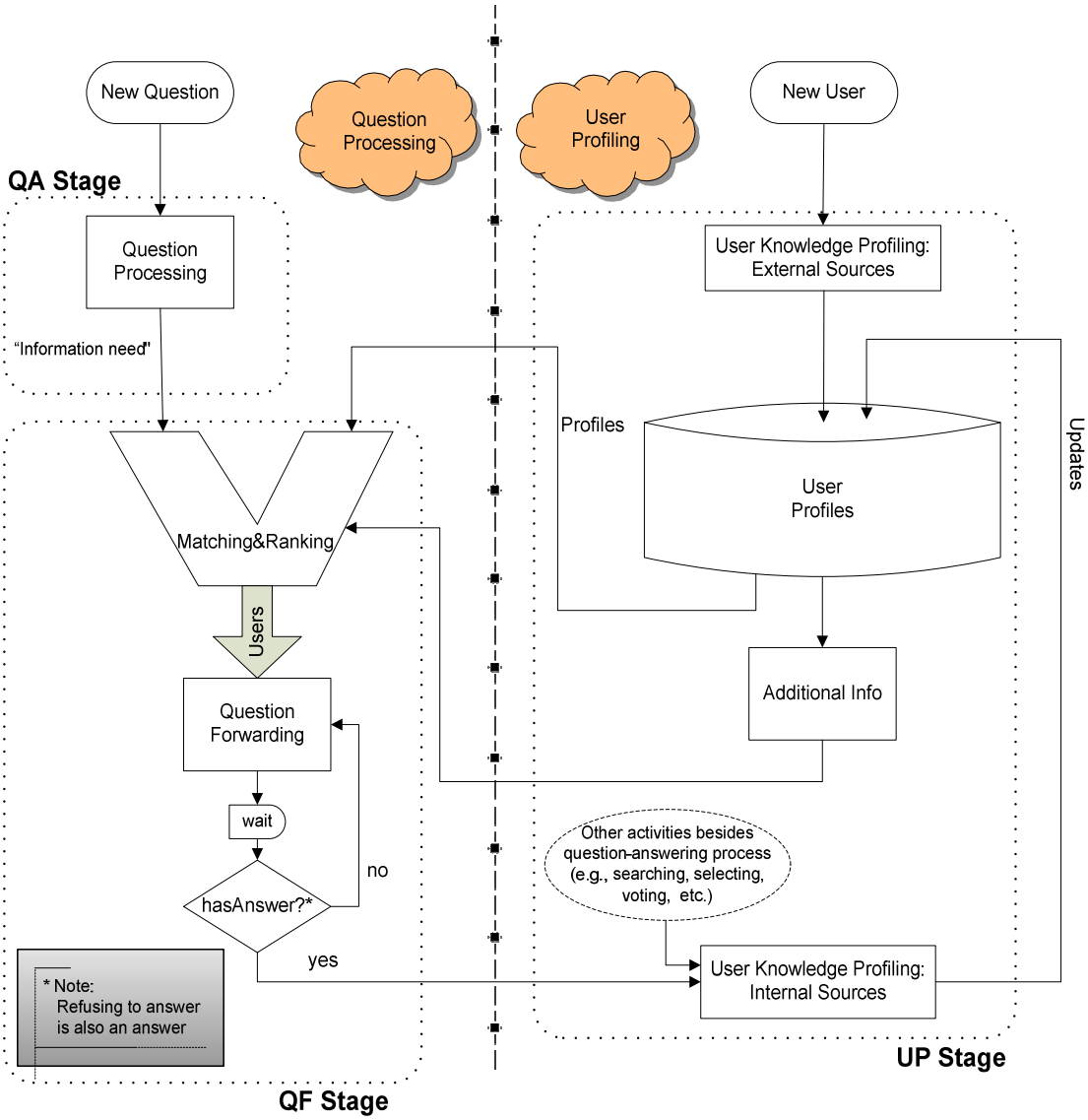
    ii. Без означавања питања.

2) Алгоритам екстракције информација:

    i. Технике обраде природног језика (*Natural Language Processing* – NLP), као што су уклањање завршетака речи (*stemming*), обрада и филтрирање врста речи (*Part-Of-Speech processing and filtering*), проналажење синонима, итд.

    ii. Технике машинског учења (*Machine Learning* – ML или *Data Mining* – DM) као што су моделовање тема (*topic modeling*) или тренирани класификатори тема (*trained topic classifiers*).

Могући правци за побољшање у вези Питања # 1: Визуелизација питања.



Слика 2. Оригинална презентациона парадигма: Генерализација приступа из отворене литературе

Питање # 2: Како пронаћи компетентне кориснике за дато питање?

Имајући у виду да су информације о питању добијене из модула за *обраду питања*, задатак проналажења компетентних корисника врши се поређењем добијених информација из питања и расположивих корисничких профила, чији је резултат рангирана листа корисника (или "кандидата за одговор"), које би требало

контактирати за одговор на дато питање. Ово поређење може бити реализовано егзактним поређењем или помоћу израчунавања семантичке сличности. Такође, организација модела може бити централизована (поређење и рангирање је реализовано у оквиру једног централног чворишта), или дистрибуирана (сви чворови су укључени у овај процес).

Као што је приказано на слици 2, улази модула за *Поређење и Рангирање* су: (а) тражене информације из питања, (б) доступни профили знања из складишта корисничких профила, и (в) додатне информације попут доступности, одзива, или ранга популарности. Излаз је рангирана листа корисника који се прослеђују модулу за *достављање питања*.

Критеријуми од интереса за Питање #2:

1) Модел организације
   i. Централизован
   ii. Дистрибуиран
2) Рачунање сличности
   i. Са егзактним поређењем
   ii. Са семантичким поређењем

Могући правци за побољшања у вези Питања #2: Проширено семантичко поређење.

Питање #3: Како прецизно проценити компетентност корисника на основу информација добијених из различитих извора?

Знање се може широко сврстати у две категорије: експлицитно и имплицитно (*tacit*) [8], [9]. Експлицитно знање се састоји од чињеница, правила, релација и смерница које се могу верно представити у папирном или електронском облику. Пошто је ово знање експлицитно изражено, може се

размењивати без потребе за дискусијом. Насупрот томе, имплицитно знање (или интуиција) захтева интеракцију. Ова врста знања наглашава личне вештине, и углавном је под утицајем погледа, вредности и убеђења. Његова размена захтева контакт лицем у лице или чак постепено учење. С обзиром да је индивидуално знање научено, тј. интернализовано унутар људског мозга, потребно је применити психолошки приступ посматрања карактеристика субјекта на основу уоченог понашања. У овом случају, уочено понашање представља садржај који корисник генерише, па у извесној мери тај садржај се може пресликати на поменуту поделу знања: експлицитно и имплицитно, где је експлицитно знање углавном изражено у виду објављених докумената, као што су научни радови, књиге, чланци или колекције веблогова, док комуникација путем електронске поште и садржај генерисан током процеса питања и одговарања може идентификовати имплицитно знање. Као резултат, обе врсте ових информација су драгоцене за профилисање корисничког знања. При том, компетентност означава ниво знања везан за неку специфичну области, чија вредност може ићи од површног, на нивоу интересовања, до темељног познавања. С друге стране, стручност представља низ компетентности, тј. познавања скупа области блиско повезаних унутар неког домена.

СИПП одржава корисничке профиле у складишту које се стално ажурира. Поред компетентности, што је основна информација која се чува о кориснику, кориснички профил може садржати и поменуте додатне информације (модул *Додатне Информације*). Ове информације нису директно повезане са знањем, али могу побољшати квалитет услуге система, нпр. брзину добијања одговора ако се питања усмеравају корисницима са високом стопом учестаности давања одговора. С друге стране, део корисничког профила који се односи на стручност може се креирати из различитих извора информација. Ови извори се могу сврстати у две категорије: (i) унутрашњи и (ii) спољни.

i. Унутрашњи извори сакупљају информације о активностима корисника унутар система, неке директно повезане са СИПП процесом као што су постављање питања и одговарање, или имплицитне попут

претраживања постојећих одговора, одабира одговора, гласања за најбољи одговор, итд. Даље, у зависности од употребљене технике профилисања ово може бити подељено у следеће две категорије: текст (нпр. за информације добијене из садржаја структуре питање-одговори) и друго (нпр. информације добијене из веза типа питање-одговори између корисника приликом бодовања одговора).

ii. Спољни извори се обично користе за креирање почетног профила знања корисника како би се избегао проблем новог корисника, тј. хладног старта, као и за пратеће измене у профилу. Ови извори сакупљају информације из других система, који нису саставни део СИПП, попут друштвених мрежа, веблогова, складишта електронске поште, итд. Такође, ови извори се даље могу поделити у исте две категорије: текст (нпр. поруке на друштвеним мрежама или веблоговима) и друго (нпр. демографски подаци о кориснику или социјалне везе).

Као што је представљено на слици 2, за новог корисника почетни профил се креира у модулу *профилисање корисничког знања на основу екстерних извора информација*. Након тога, током процеса одговарања на питања врши се ажурирање ових информација из модула *профилисање корисничког знања на основу интерних извора информација* (нпр. на основу добијених позитивних односно негативних оцена за одговоре). Такође, даље ажурирање информација о кориснику је могуће и из спољних извора помоћу модула за *профилисање корисничког знања на основу екстерних извора информација* (нпр. ручно мењање сопственог профила).

Критеријуми од интереса за Питање # 3, на основу методологије профилисања компетентности корисника:

1) Текст:

   i. NLP технике: нпр. *ad-hoc* екстракција именованих ентитета (*named entity extraction*) или уклањање завршетака речи (*stemming*),

   ii. Технике машинског учења: DM (нпр. класификација или груписање) или ML (моделовање тема),

   iii. Модели препоручивања (*Recommender System model* - RS).

2) Друго:

   i. *ad-hoc* (АН) модели,

   ii. Модели препоручивања (RS).

   iii. DM: нпр. Рангирање страница или HITS

Могући правци за побољшања у вези Питања # 3: Интеграција профила.

Преглед анализираних приступа

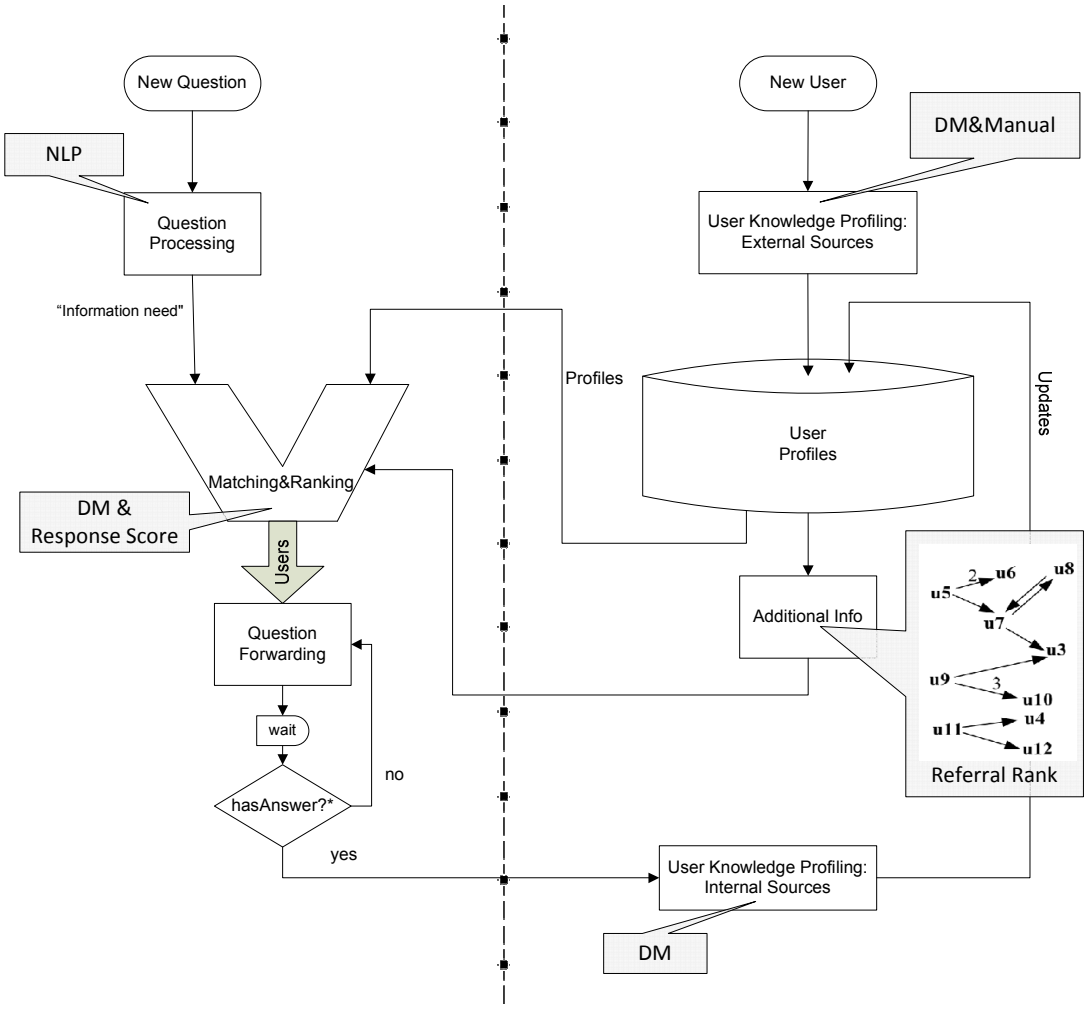
Приступи представљени у наставку односе се на три главна питања која дефинишу СИПП на карактеристичан начин. Ради лакшег поређења, сви приступи су такође приказани табеларно у прилогу А. Свака колона у табели одговара одређеном елементу СИПП процеса са слике 2. Сваки улаз у табели садржи назив и кратак опис (у оквиру критеријума од интереса). Такође, сви анализирани приступи су описани на сличан начин, укључујући информације према следећем шаблону: суштина, структура, релевантни детаљи, примене и предности и мане.

I. iLink

Davitz J. и сарадници [10] 2007. године су предложили модел за социјалну претрагу и управљање порукама под називом iLink. Главни фокус њиховог рада се односио на то како моделовати друштвене мреже и како те мреже решавају проблеме употребом P2P начина сарадње. iLink модел је примењен за развој система за генерисање најчешће постављаних питања (*Frequently Asked Questions* - FAQ) у оквиру друштвених мрежа – под називом FAQtory. Предложени систем омогућава генерисање складишта нити, тј. низова структура типа питања/одговори (*question/answer threads*), тако да када корисник пошаље питање систему, њему су представљене листе сродних парова питање/одговор, списак стручњака о темама које се налазе у питању, и на крају, као последња могућност, резултати претраге са Веба. Друштвена мрежа је представљена као граф са чворовима и везама. Параметри као што су стручност, резултат при одговарању (*response score*) и ранг повезаности (*referral rank*) се одржавају за сваки чвор, тј. корисника. За *обраду питања* користе се NLP технике, конкретно уклањање стоп речи (*stop words filtering*) и завршетака речи (*stemming*), као и проналажење синонима. Корисници такође могу анотирати питања произвољним речима (*tagging*) у циљу побољшања перформанси система. DM техника груписања (*clustering*) се користи за стварање профила корисничког знања из текстуалних

извора. Други параметар који се чува је резултат при одговарању који је у функцији учестаности давања одговора, тачности одговора, итд. Модул за *поређење и рангирање* рачуна семантичку сличности између термина. Ранг повезаности одржава се као *додатна информација* о кориснику, који одговара популарности датог корисника у односу на друге кориснике. Технике груписања се користе и за креирање почетног профила знања како би се идентификовале информације из спољних извора, као што су доступне друштвене мреже, складишта електронске поште или личне Веб презентације корисника. Такође, корисници могу начинити ручне измене у профилу. ILink модел је централно организован (као суперчвор у друштвеној мрежи), али се може користи и на децентрализован начин. Његова структура је приказана на слици 3 користећи образац уведен у презентационој парадигми.

Занимљива идеја коју iLink уводи је могућност инкременталног одговарања. На сваком кораку пропагације питања кроз мрежу, кориснички чворови могу допринети том питању са неком информацијом, чак и ако се та информација не квалификује као одговор. Ова информација може бити о самом питању или једноставно може бити доказ о томе где може постојати специфично знање у мрежи (нпр. ко зна нешто, ко зна некога). С друге стране, iLink не користи опште рачунање семантичке сличности, јер захтева претходно познавање релевантних тема у домену примене. Такође, недостаје модел отпорности на шум и неки детаљнији подстицајни модел за кориснике.

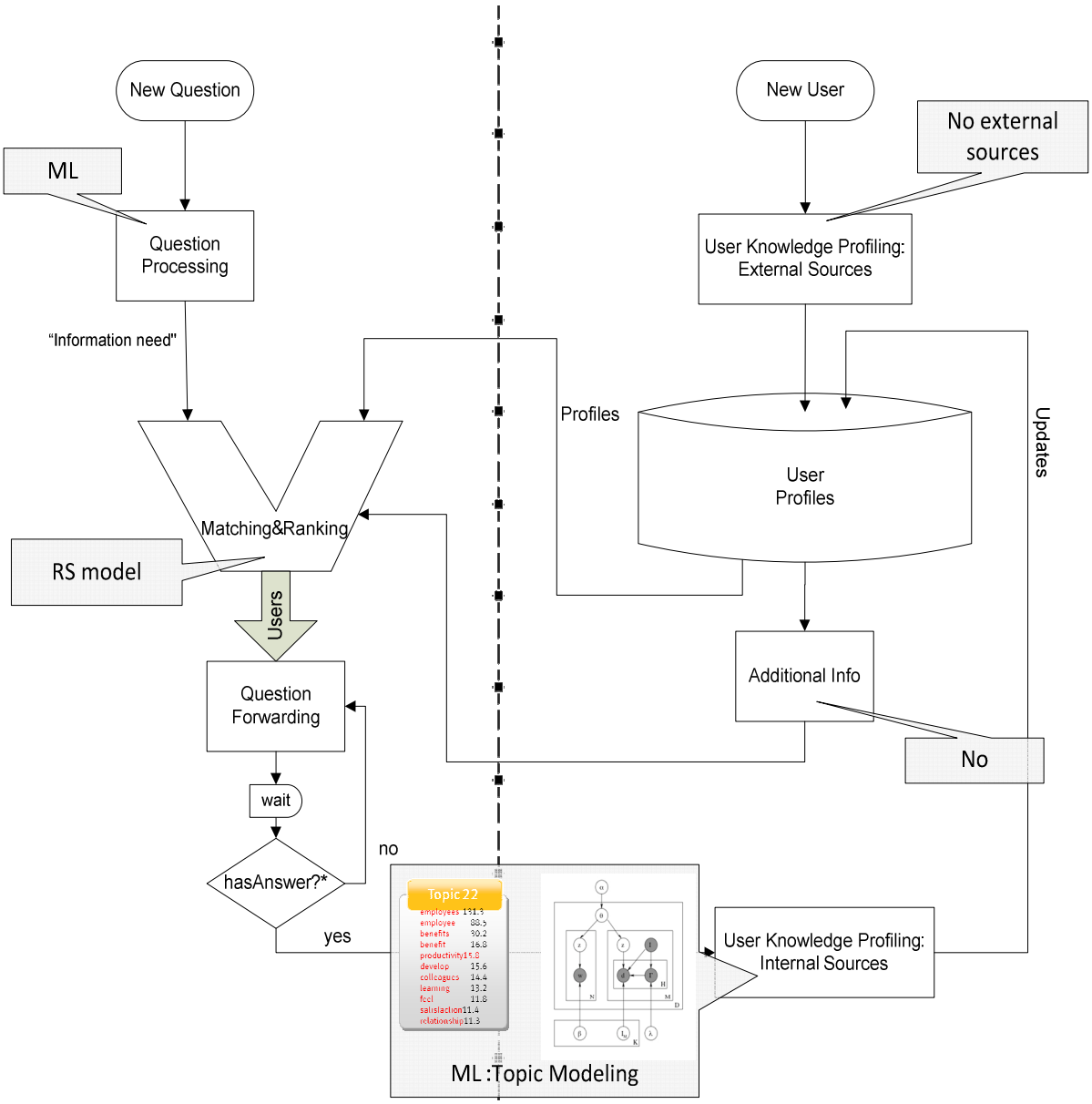


Слика 3. iLink: Структура

II. Пробабилистичка Латентна Семантичка Анализа у Порталима за Одговарање на Питања (PLSA in CQA)

Qu M. и сарадници [11] 2009. године су предложили модел за препоручивање питања употребом пробабилистичке латентне семантичке анализе (*Probabilistic Latent Semantic Analysis* - PLSA) који помаже корисницима да лоцирају занимљива питања у оквиру портала за одговарање на питања (*Community Question Answering Portals* - CQA), као што је Yahoo! Answers портал. Техника моделовања области заснована на PLSA се користи за *профилисање корисничког знања* из текстуалних извора. Такође, иста техника машинског учења (PLSA) се користи за *обраду питања*. *Поређење и рангирање* је централизовано и не користи семантичку сличност између екстрахованих појмова. Структура овог модела је приказан на слици 4.

Овај рад је укључен у преглед имајући у виду да уводи иновативан приступ у профилисању знања заснован на техници моделовања тема. Такође, предложена је нова метрика за евалуацију СИПП система која пореди ранг препорученог корисника са рангом корисника који је дао најбољи одговор на основу скупа података доступних са Yahoo! Answers портала. С друге стране, овом приступу недостаје много анализираних атрибута. Не постоји могућност анотације питања нити се чувају *додатне информације* о кориснику. Такође, не постоје други параметри укључени у профил знања.



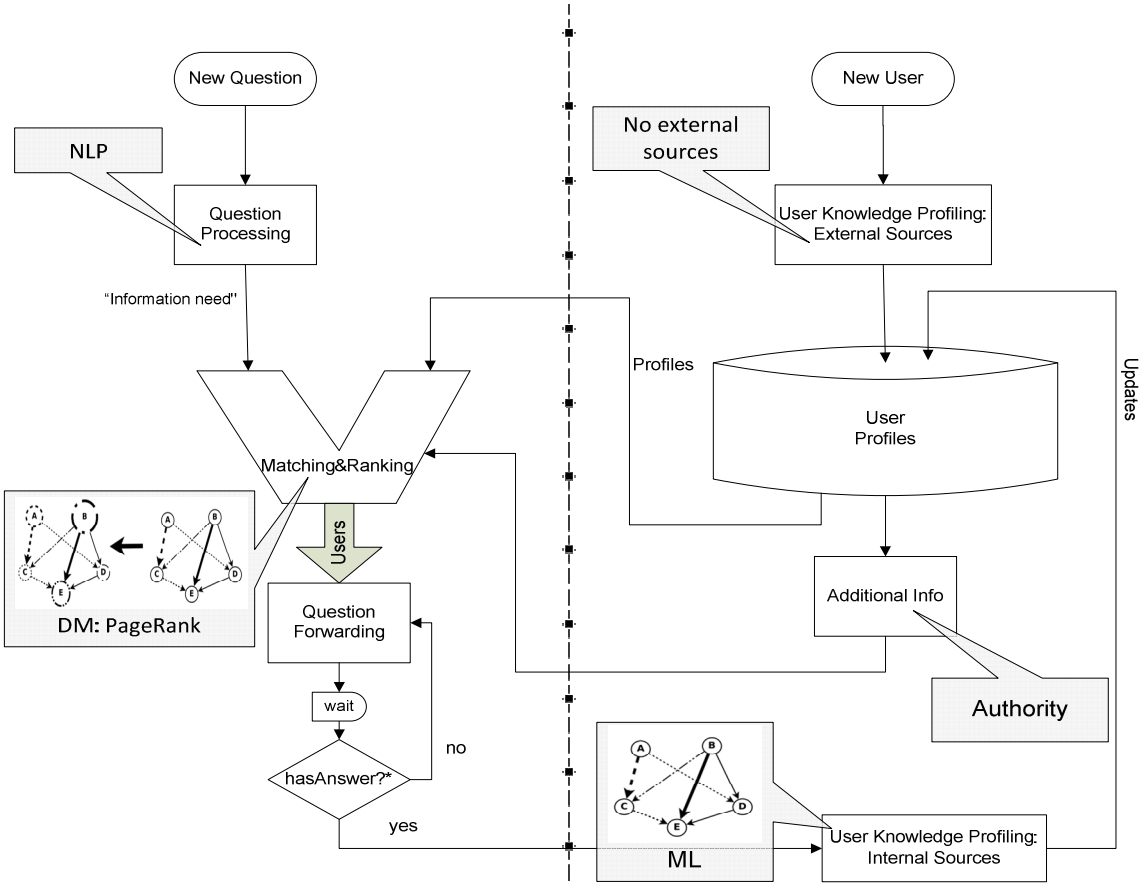
Слика 4. PLSA in CQA: Структура

III. Прослеђивање Питања Унутар Форума

Zhou Y. и сарадници [12] 2009. године су предложили комбинацију језичког модела и структуре форума (у облику питање-одговори) ради прослеђивања питања одговарајућим корисницима (*Routing within Forums*). Технике машинског учења засноване на вероватноћи појављивања термина су употребљене како би се креирали профили учесника Веб форума (*профилисање корисничког знања*). Тачније, три различита приступа су употребљена како би се интегрисала три различита језичка модела: модел заснован на профилу (*profile-based*), модел заснован на нитима (*thread-based*) и модел заснован на кластерима (*cluster-based*). Модел заснован на профилу креира профил за сваког корисника на основу одговора које је тај корисник дао и такође на основу одговарајућих питања на која је одговорио. У моделу заснованом на нитима, где нит представља питање са повезаним одговорима, свака нит служи као потенцијална област, тако да вероватноћа да је корисник стручњак за ново питање одређује се на основу асоцијације између нити и релевантног корисника. Дакле, сваки профил заснован на нити доприноси укупном рангу корисника на основу резултата асоцијације нити и тог корисника. Код модела заснованог на кластерима, групе нити са сличним садржајем се групишу у кластере и гради се група сличних нити за сваког корисника. Сваки кластер представља кохерентну тему и има везу према кориснику чија тежина указује на значај између тог кластера и корисника. За постављено ново питање ранг се израчунава за сваког корисника збрајањем свих кластера. Како би се избегла нулта вероватноћа, за непознате речи сва три језичка модела се поравнавају помоћу језичког модела у позадини, чиме је семантичко поклапање имплицитно употребљено. Питања као и одговори у оквиру сваке теме су претходно обрађена, што обухвата издвајање токена, филтрирање стоп речи, као и уклањање завршетака речи. Паралелно са језичким моделима, одвија се глобално рангирање корисника помоћу вредности ауторитета корисника. Ове вредности се рачунају користећи граф питања и одговора помоћу посебног алгоритама заснованог на алгоритму за рангирање страница (*PageRank-based algorithm*). Кориснички профили затим се комбинују са овим вредностима како

би се добио финални ранг експерата. Дакле, коначни ранг за сваког корисника се израчунава као вероватноћа која интегрише резултате ова два корака: вредности добијене од језичког модела и вредности ауторитета корисника. Структура овог приступа приказана је на слици 5.

Евалуација је извршена на основу података прикупљених са tripadvisor.com форума. Резултати су показали да употребљени алгоритми остварују значајан напредак у прецизности (*precision*) и осетљивости (*recall*). С друге стране, остаје неразрешен проблем ажурирања имајући у виду да се нове теме постављају свакодневно на форумима, што ствара потребу за ажурирањем инвертованих индекса. Овај поступак није тривијалан за моделе засноване на профилима и кластерима, док код модела заснованог на нитима није тако изражен због могућности инкременталног ажурирања. Такође, неке технике кластеровања могу се применити за генерисање ситнијих (*fine grained*) кластера уместо коришћења кластера добијених на основу подфорума. Коначно, проблем новог корисника постоји с обзиром да се не користе спољни извори информација за креирање иницијалног профила корисника.

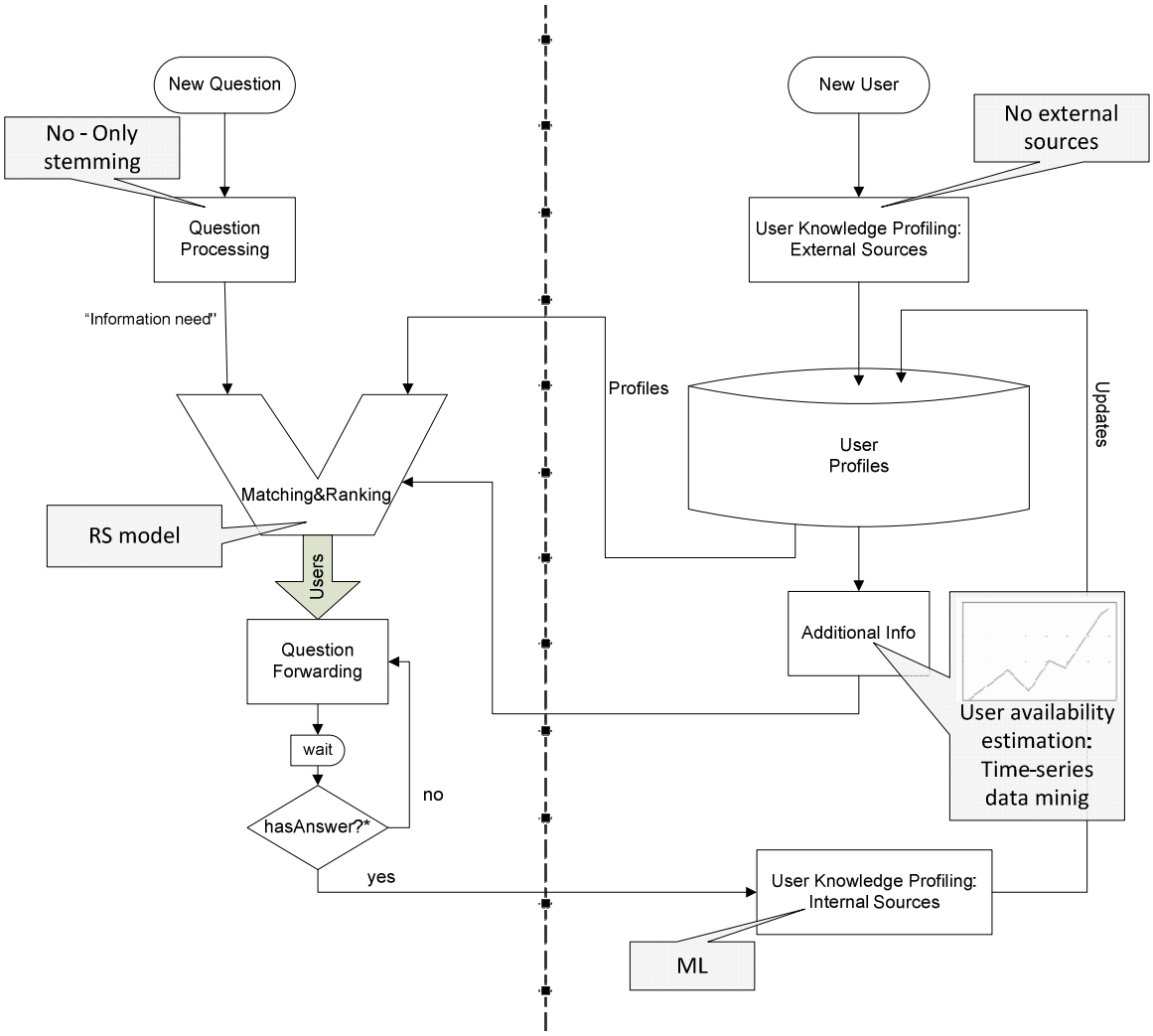


Слика 5. Routing within Forums: Структура

IV. Систем за Прослеђивање Питања

Li B. и King I. [13] 2010. године су предложили систем назван Систем за прослеђивање питања (*Question Routing Framework* - QR), који рангира кандидате за одговор унутар CQA система. *Профилисање корисничког знања* из текста је урађено двоструко: са и без разматрања квалитета одговора. Први модел процењује потенцијални квалитет одговора на основу квалитета раније одговорених питања од стране корисника. Други користи само фреквенције појмова за израчунавање сличности између датог питања и свих претходно одговорених питања. Такође, доступност се процењује као *додатна информација*. Претпоставља се да је корисник на располагању да пружи одговоре за прослеђена питања када је пријављен у систему, тако да се процена врши на основу модела анализе трендова DM техником временских серија (*time-series data mining*). *Поређење и рангирање* је централизовано и за сваког потенцијалног кандидата за одговарање QR систем израчунава резултат као линеарну комбинацију процењене експертизе и доступности корисника. Структура система је приказан на слици 6.

QR систем разматра како стручност корисника тако и његову доступност за пружање одговора у одређеном временском опсегу. Спроведени експерименти са Yahoo! Answers скупом података показали су потенцијал предложеног решења. Ипак, проблем новог корисника постоји пошто се не користе екстерни извори информација за креирање иницијалних профила. Такође, укључивање других атрибута у кориснички профил није разматрано, као ни семантичко поређење између екстрахованих појмова. Коначно, није подржана анализа питања нити њихова анотација.



Слика 6. The Question Routing Framework: Структура

V. G-Finder

Li W. и сарадници [14] у 2010. години представили су свој дизајн и имплементацију система G-Finder, алгоритма и алата који омогућава интелигентно прослеђивање питања унутар програмерских форума (нпр. посвећених Јава програмском језику). Рад је мотивисан емпиријским истраживањем које је спроведено над три популарна програмерска форума и које је показало да корисници форума наилазе на дуг период чекања на одговор, као и то да је мали број стручњака често затрпан питањима. Стога, њихов циљ је био да се искористе информације из изворног кода софтверског система коме је форум посвећен, како би се открили скривене везе између корисника форума. *Профилисање корисничког знања* врши се помоћу два алгоритма који користе изворни код програма са једне, и податке са форума са друге стране. На основу изворног кода креирају се мреже концепата, а на основу података са форума мреже корисника, које се затим интегришу у јединствен пробабилистички модел. За софтверски систем коме је форум посвећен динамички се гради мрежа концепата која се конструише тако што се издвајају везе између концепата из изворног кода (нпр. на основу хијерархије класа или графа позива метода) или пак преведеног кода (*bytecode*). За пример програмског језика Јава ова мрежа може представљати систем типова Јава класа. Истовремено, мреже корисника се конструишу као усмерени графови на основу релација добијених између питања и повезаних одговора. Постављено питање се онда користи како би се интегрисале ове две мреже и рангирали потенцијални стручњаци (*поређење и рангирање*). При том, израчунавање се врши по следећој формули:

$$\mathrm{P}(\text{корисник}|\text{питање}) = \sum_{\text{концепт}} \mathrm{P}(\text{корисник}|\text{концепт}) \times \mathrm{P}(\text{концепт}|\text{питање})$$

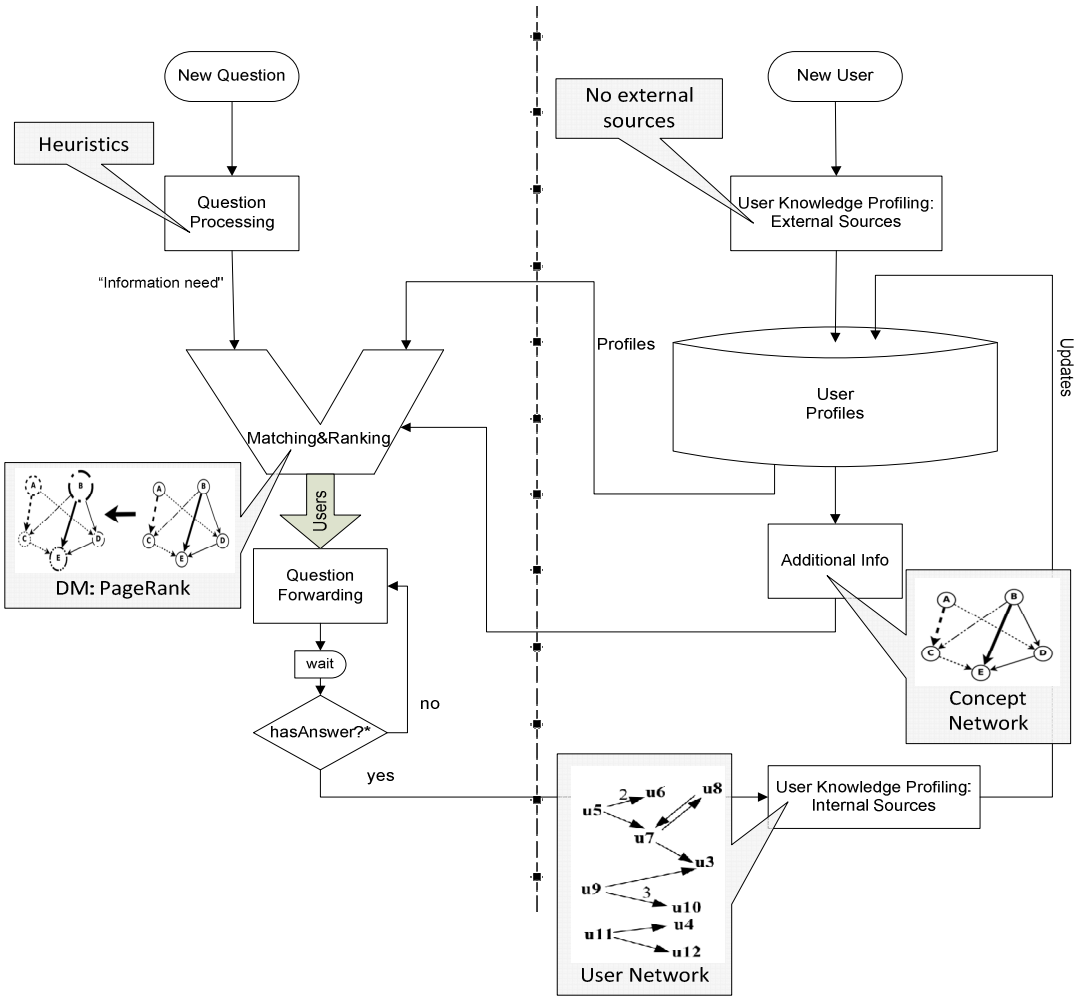
Вероватноћа P(корисник|концепт) рачуна се из два корака на следећи начин:

1) Изгради мрежу корисника за сваки концепт и примени прилагођени алгоритам Рангирања страница. У овом кораку, односи између концепата нису укључени.

2) Узми у обзир односе између концепата, које представља мрежа концепата, и за корисника израчунај вероватноћу P(корисник|концепт) засновану на семантичком груписању концепата.

Коначно, P(концепт|питање) односи на *обраду питања* која се обавља помоћу различитих хеуристика, као што су учестаност појављивања концепта у наслову или TF-IDF (*term frequency–inverse document frequency*) метрике концепта у оквиру изворног кода тела поруке. Структура G-Finder система представљена је на слици 7.

Евалуација спроведена над подацима из три велика програмерска форума (Java Forum, Java DevShed Forum и GEF Forum) показала је да G-Finder значајно побољшава прецизност предвиђања стручњака који могу пружити одговоре на програмерска питања. Главно ограничење је метод заснован на хеуристичком мапирању концепата, који у неким случајевима не успева да одреди концепте из питања или одговора. Такође, неки латентни обрасци и односи које модел не запажа су проблем новог корисника, тј. стручњака, проблем општег стручњака и проблем случајног стручњака. Проблем новог корисника односи на кориснике који никада нису одговарали на било шта раније па се њихова стручност не може одредити. Манифестација проблема општег стручњака се огледа у томе да питања одговорена од стране оваквог корисника немају нити експлицитне нити имплицитне семантичке везе међусобно. Проблем случајног стручњака односи на кориснике који активно учествују у темама са коментарима и сугестијама, али ретко са одговорима. Дакле, алгоритам даје исправно предвиђање да ће они вероватно дати одговор, али не може направити разлику између правих одговора и текстова коментара.



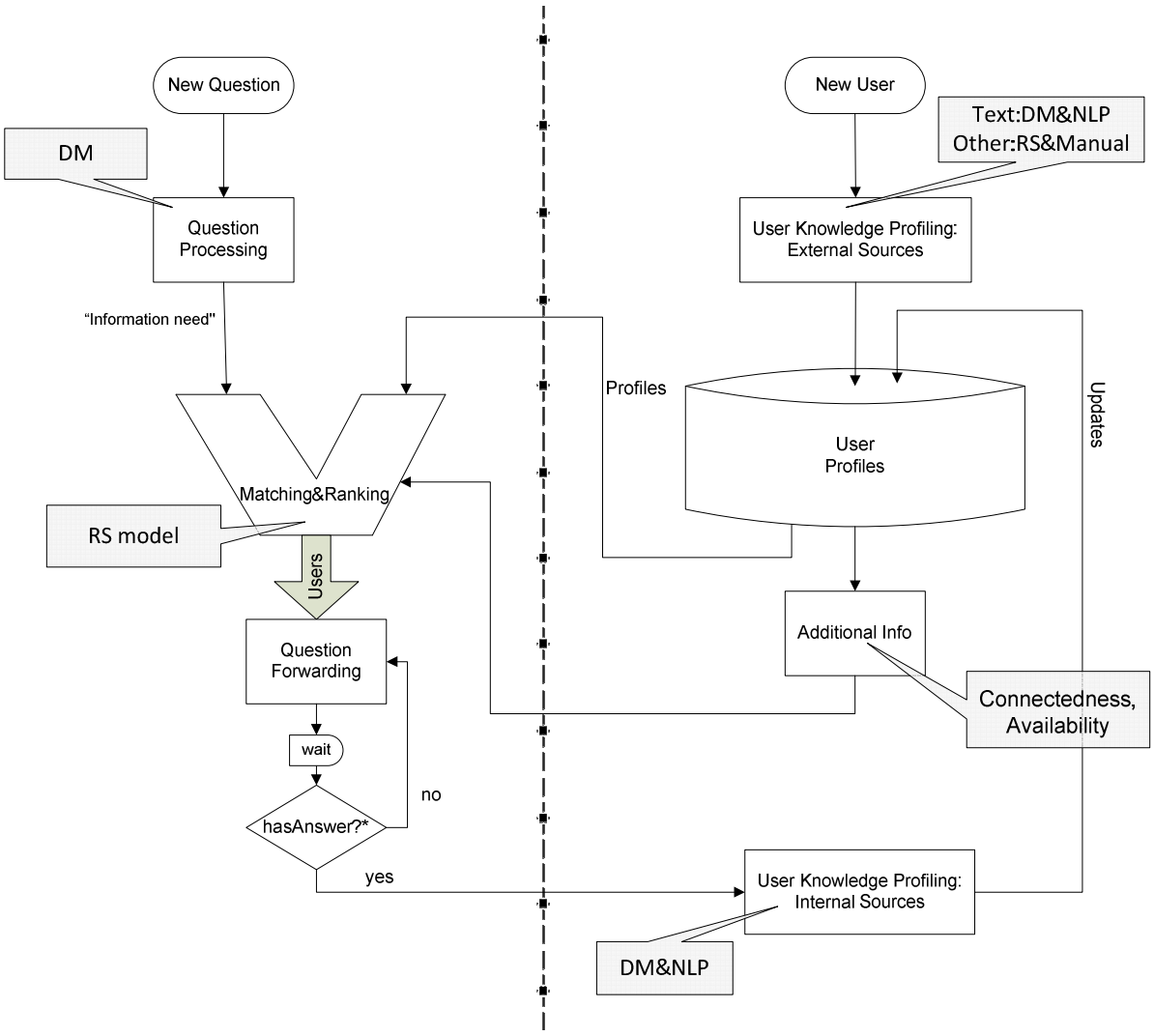
Слика 7. G-Finder: Структура

VI. Aardvark

Horowitz D. и Kamvar S. D. [15] у 2010. години представили су комерцијални систем под називом Aardvark. Овај систем представља сервис социјалне претраге (*social search engine*) где корисници могу поставити питање, било путем инстант поруке, електронске поште, веб улаза или гласа. Питања се анализирају DM техником (комбинацијом тренираних класификатора тема), а корисник их може додатно анотирати (*tagging*). Да би пронашао некога ко је највероватније у стању да одговори на питање, Aardvark прослеђује ово питање особама у проширеној друштвеној мрежи корисника. Стога, кориснички профил садржи проширену социјалну мрежу која индексира припадност и информације о пријатељству за сваког корисника као и за њихове пријатеље, представљајући друштвени граф пријатељства (*Friends-of-Friends social graph*). Корисник има могућност да увезе ове информације из постојећих друштвених мрежа као што су Facebook, LinkedIn или контакти из електронске поште, или пак ручно позивајући пријатеље да се придруже. Истовремено, за *профилисање корисничког знања* из текста, Aardvark одржава списак тема о којима корисник има одређени ниво интересовања. Ове теме су идентификоване из неколико извора: ручно наведене од стране корисника или његовог пријатеља који га је позвао, анализом странице профила или налога на коме редовно ажурира статус (нпр. Twitter или Facebook) и на крају, посматрањем понашања корисника приликом одговарања (или избора да не одговори) на питања о одређеним темама. Комбинација DM & NLP се користи за издвајање тема, конкретно SVM (*Support Vector Machine*) и *ad hoc* екстракција именованих ентитета. Такође, атрибути попут повезаности и доступност се воде за сваког корисника као *додатне информације*. Модул за *поређење и рангирање* користи заједно проширену социјална мрежу и теме добијене из текста како би рангирао потенцијалне кандидате за одговор. Сличност између корисника се рачуна употребом модела за препоручивање (*Recommender System - RS Model*) над атрибутима добијеним из друштвених мрежа, као што су демографска сличности, сличности профила, друштвене везе, итд. Сличност између издвојених тема из питања и тема из профила корисника се рачуна коришћењем семантичке

сличности засноване на текстуалном корпусу, као што је Википедија. Организација Aardvark система је централизована и његова структура је приказана на слици 8.

Aardvark алгоритам претраге ставља акценат на интимност, где је поверење корисника у добијени одговор првенствено засновано на познавању те особе (директно или индиректно из социјалне мреже). Дакле, питања су пре свега усмерена ка проширеној социјалној мрежи корисника. Како је наведено, то даје резултате за питања која су у контексту социјалне или демографске блискости корисника (нпр. давање мишљења о ресторану у близини или савет о изласку). Међутим, постоји још једна димензија поверења у добијени одговор која је заснована на репутацији кандидата за одговор. Ово се посебно односи на питања која су веома стручно оријентисана, где информација коју корисник тражи често не може бити пронађена у његовој проширеној друштвеној мрежи. У том контексту, потребан је други модел *профилисања корисничког знања* како би се питања проследила одговарајућем кориснику.



Слика 8. Aardvark: Структура

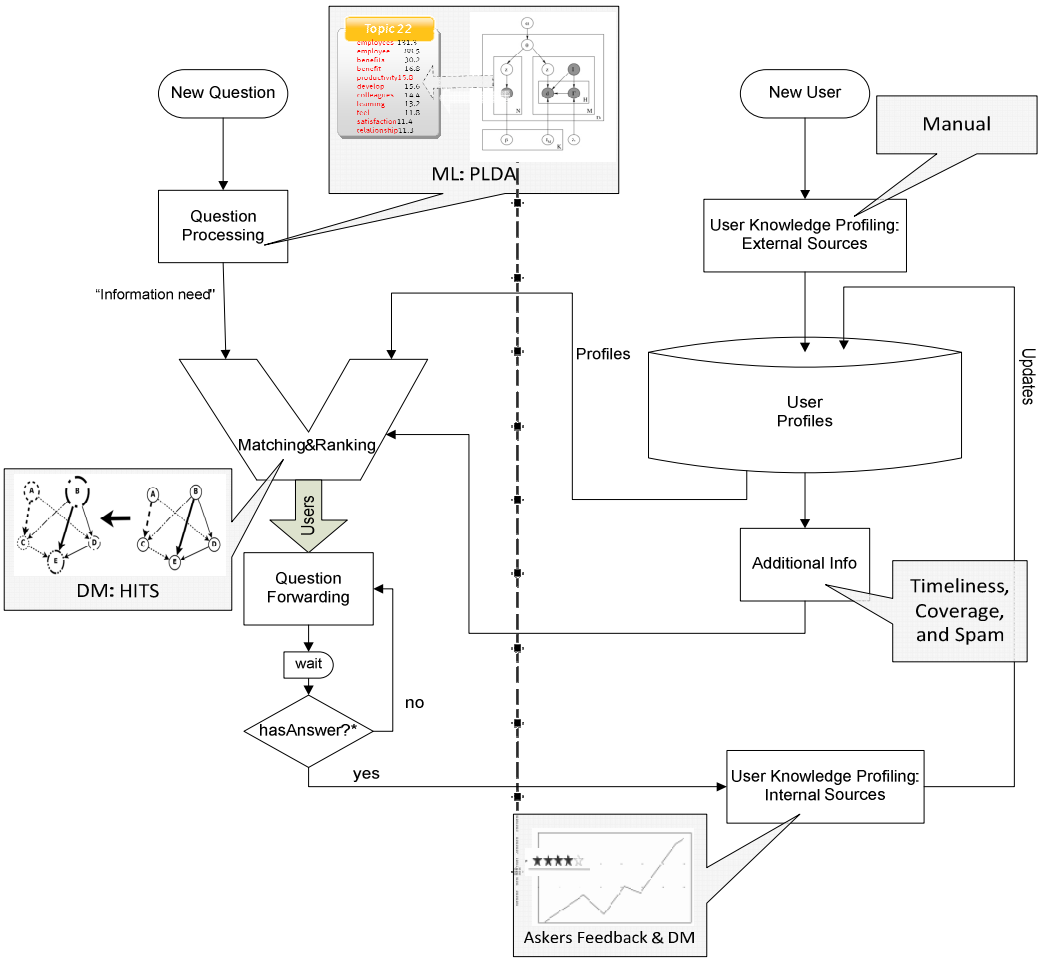
VII. Конфучије

Si X. и сарадници [16] у 2010. години представили су Конфучије (*Confucius*), Google сервис за питања и одговоре (*Questions&Answers* - Q&A) који користи шест DM процедура како би искористио синергију између веб претраге и друштвених мрежа. Када унети термини за претрагу указују на 5W1H тип упита (када, где, зашто, итд.) или када претраживач не може вратити довољно релевантан резултат (нпр. преклапање садржаја упита и потенцијално најбољих резултата је ниско), препоручује се Q&A сесија. На тај начин корисник који тражи неку информацију подстиче се да постави питање унутар Конфучије система. Паралелна имплементација латентне Дирихлеове алокације (*Parallel Latent Dirichlet Allocation* - PLDA) се користи за *обраду питања*. Када се унесе питање, систем ће предложити скуп наслова категорија за избор. Такође, нова категорија се може ручно додати питању. Док корисник уноси питање, а пре његовог слања човеку - кандидату за одговор, подсистем тражи слична раније постављена питања, као и њихове већ расположиве одговоре. Овај корак смањује време које је потребно да корисник добије задовољавајући одговор у случају да слично питање постоји. Поред тога, након слања питања, користе се сложене и временски захтевне NLP технике за аутоматско генерисање одговора. Овај корак не угрожава општи квалитет рада система, јер се одвија док се чека на људски одговор и такође у случају да је ниво поузданости произведеног одговора нижи од унапред одређеног прага, овај модул неће понудити никакав одговор.

Модул за *профилисање корисничког знања* сједињава неколико различитих атрибута као што су граф корисника, квалитет одговора и теме интеракција. За сваки одговор, поред прикупљених повратних информација добијених од питалаца (у виду селекције најбољег одговора или гласова прихватања односно неслагања са одговором), користи се и аутоматска рутина за процену квалитета одговора. Ова рутина је реализована као тренирани бинарни класификатор и заснива се на неколико фактора који укључују релевантност одговора као и његову оригиналност. Оцене квалитета се затим агрегирају за сваког корисника помоћу рутине за *поређење и рангирање*. Ова рутина

квантификује доприносе корисника и рангира их у зависности од теме, користећи тежински HITS алгоритам осетљив на теме како би се генерисао граф активности корисника који се добија претварањем Q&A структура у тематски пондерисане интеракције између корисника. Резултат извршавања алгоритма обухвата два аспекта репутације: (а) могућност корисника да досегне друге кориснике и (б) способност корисника да привуче пажњу других корисника. Коначан резултат се затим израчунава као линеарна комбинација ова два аспекта. Такође, семантичко поређење је имплицитно реализовано кроз PLDA модел. Конфучије систем је централно организован и његова структура је приказана на слици 9.

Добијени статистички резултати евалуације система Конфучије показали су да синергија између веб претраге и Q&A заједнице побољшава квалитет услуге. Такође, како би се избегао сувишан и непотребан посао, а самим тим и кашњење приликом слања питања директно људским корисницима, употребљене су различите технике, као што је аутоматско генерисање одговора или препоручивање сличних и већ одговорених питања. Ипак, проблем питања за мишљење је један од преосталих нерешених изазова. Наиме, тренирање овог модела се ослања на најбоље одговоре као позитивне узораке и не-најбоље одговоре као негативне узораке. Међутим, код питања за мишљење разлика између најбољих и не-најбољих одговора је субјективна. Да би се ово адекватно решило потребно је да модел такође евидентира субјективност питалаца (на пример као што је то учињено у систему Aardvark) што није предвиђено тренутним скупом атрибута.

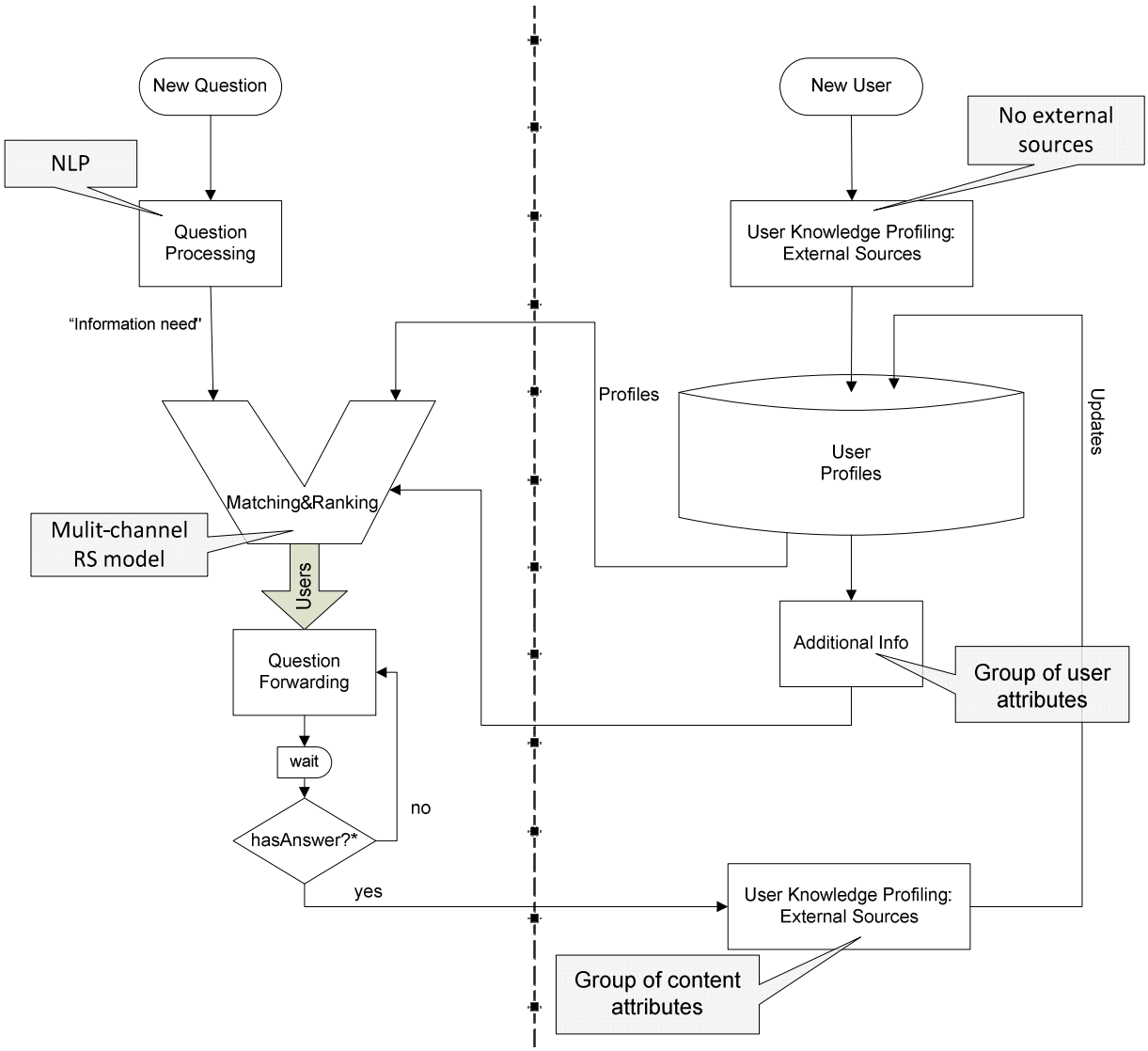


Слика 9. Конфучије: Структура

VIII. Yahoo! Answers Систем Препорука

Dror G. и сарадници [17] у 2011. години нагласили су потребу за механизмом препоручивања унутар CQA портала, конкретно Yahoo! Answers портала, који би могао да приближи питања корисницима који имају блиска интересовања и који евентуално могу одговорити. Овај систем је назван Yahoo! Answers системом препорука (*Yahoo! Answers Recommender System*). *Обрада питања* је реализована коришћењем NLP техника, као што су обрада и филтрирање различитих врста речи, затим уклањање стоп речи, као и уклањање завршетака речи. Такође, приликом постављања питања систему, корисник га мора анотирати додељујући једну или више категорија. Архитектура система је централизована, а *поређење и рангирање* се заснива на мулти-каналној технологији система препоручивања (*multi-channel recommender system technology*). Како би сјединио и генерализовао информације које с једне стране представљају различите сигнале генерисаног садржаја, а са друге стране сигнале социјалних веза корисника, конструисано је симетрично окружење које укључује и организује ове сигнале у више канала. Сигнали садржаја користе се за *профилисање корисничког знања* из текста и они се углавном односе на текстуалне атрибуте и категорије питања и повезаних одговора. Остали атрибути су такође укључени у облику социјалних сигнала, који евидентирају различите интеракције корисника са питањем, као што су постављање питања, одговарање, гласање, итд. Структура система је приказан на слици 10.

Основни циљ предложеног приступа је да се захтев корисника задовољи за различита питања, нека чињенична, али у већини субјективна где је појам експертизе ирелевантан. Ово се разликује од задатка претраживања експерата који покушава да идентификује ауторитативан одговор који би задовољио већину. Такође, у контексту Yahoo! Answers портала екстерни извори друштвених веза између корисника нису доступни, па је главни фокус био на томе како направити разлику између различитих интеракција корисника са питањем. Ово имплицира проблем новог корисника.



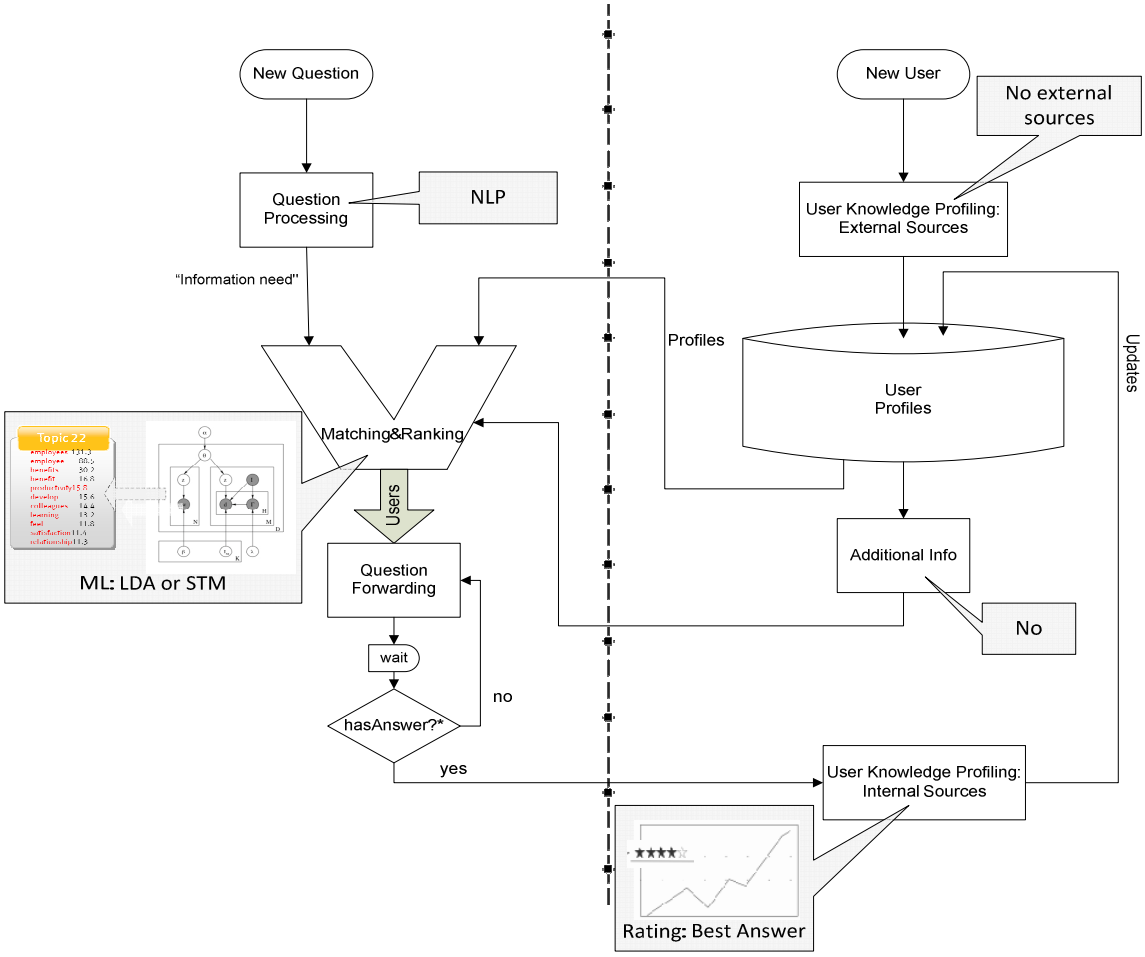
Слика 10. Yahoo! Answers Recommender System: Структура

IX. STM in CQA

Riahi F. и сарадници [18] у 2012. години објавили су своје истраживање о погодностима употребе две статистичке технике моделовања тема у СИПП системима и упоредили их насупрот традиционалним приступима у претраживању информација. Основни циљ њиховог истраживања био је да се израчуна вероватноћа која представља стручност сваког корисника за постављено питање. У предложеном моделу, свако питање има три дела: анотација питања – скуп додељених ознака (*tags*) од стране корисника који је поставио питање; наслов питања - кратак опис; и тело питања - детаљан опис. Интересовања корисника се моделују праћењем његове историје одговарања у CQA заједници. За сваког корисника, профил је креиран комбиновањем питања за која је тај корисник понудио одговор који је изабран за најбољи. На основу прикупљених корисничких профила, сличност између кандидата за одговор и новог питања (*поређење и рангирање*) мери се употребом неколико различитих метода: TF-IDF, језички модел са Дирихлеовом нормализацијом, латентна Дирихлеова алокација (*Latent Dirichlet Allocation – LDA*) и сегментирано моделовање тема (*Segmented Topic Model – STM*). Технике уклањања завршетака речи, затим уклањање стоп речи, као и не тако честих речи употребљене су за обраду питања и одговора. Генерална организација модела је централизована и његова структура је приказан на слици 11.

За евалуацију различитих предложених метода аутори су конструисали скуп података сакупљених са веб портала stackoverflow.com. Резултати добијени над овим скупом података су показали да употребљене технике моделовања тема надмашују остале методе. Даље, аутори су закључили да STM доследно даје боље резултате у односу на LDA. Међутим, предности екстерних извора информација за моделовање интересовања и експертизе корисника нису разматране, као ни други интерни атрибути међу којима су оцене одговора (*answers score*), број омиљених (*favorite count*) као и последњи датум уређивања. Такође, као што су аутори напоменули, корисничке информације често могу садржати метаподатке, у

виду бецева или репутације, који нису тренутно разматрани као додатни атрибути (*додатне информације*) за профилисање корисника.



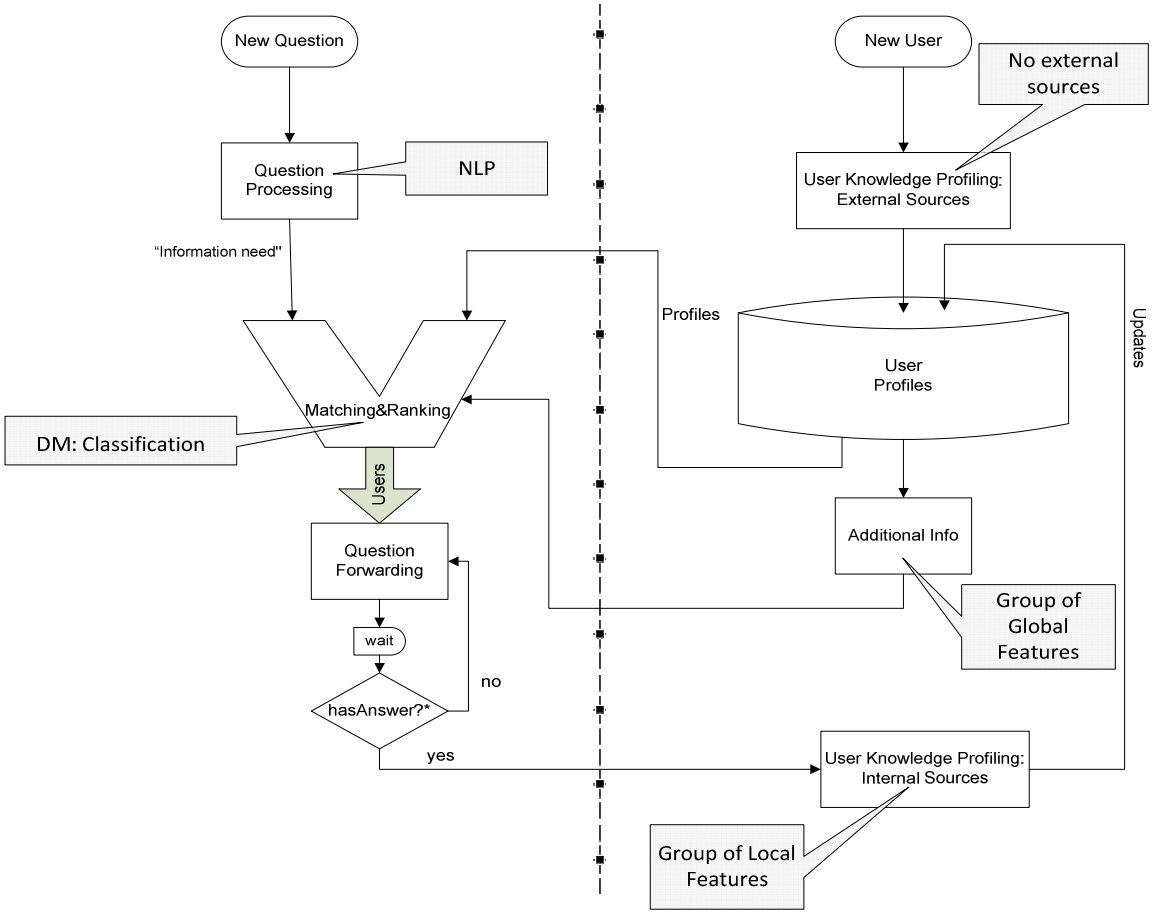
Слика 11. STM in CQA: Структура

X. Прослеђивање Питања Унутар CQA засновано на Класификацији

Zhou T.C. са сарадницима [19] у 2012. години приказали су метод назван Прослеђивање питања унутар CQA засновано на класификацији (*Classification-based Routing in CQA*). У свом раду разматрали су СИПП процес као проблем класификације, на основу кога су развили низ локалних и глобалних атрибута који осликавају различите аспекте питања, корисника и њихових међусобних релација. Локалне карактеристике укључују (а) атрибуте питања, нпр. дужина наслова, дужина тела, 5W1H тип питања; (б) атрибуте историје коришћења, нпр. дужина чланства, укупно сакупљених бодова, број најбоље пружених одговора, број постављених питања; и (в) атрибути односа корисник-питање, нпр. да ли је корисник на врху листе на основу његовог доприноса у категорији у којој питање припада. Поред тога, усвојени су неки атрибути који описују семантичку сличност језичког модела питања и језичког модела корисника, нпр. КЛ-дивергенцију између наслова актуелног питања и његових детаља, и свих наслова питања и њихових детаља за које је корисник обезбедио одговоре. Глобални атрибути узимају у обзир глобалне информацију о кориснику добијене из CQA сервиса и они се такође могу поделити у три категорије: (а) атрибути питања, нпр. просечна дужина наслова и просечна дужина детаља; (б) атрибути историје коришћења који приказују јединственост корисника, нпр. КЛ-дивергенција питања и одговора корисника и питања и одговора свих осталих корисника; и (в) атрибути односа корисник-питање који се заснивају на претпоставци да што су сличнији језички модели одговорених питања корисника и свих питања у категорији, вероватније је да тај корисник може дати одговор на питања из ове категорије, нпр. КЛ-дивергенција између наслова питања на која је корисник одговорио и његових детаља, и дивергенције наслова и детаља свих питања у категорији којој припада дато питање. За обраду питања и одговора употребљене су технике уклањања завршетака речи као и уклањање стоп-речи (осим 5W1H речи). *Поређење и рангирање* је разматрано као проблем класификације са две класе, при чему је фокус био на позитивној класи, што значи да за дати пар

питање-корисник се одређује који корисник би највероватније одговорио на питање. Као метод класификације употребљен је SVM. Модел организације је централизован и његова структура је илустрован на слици 12.

Експериментални резултати су добијени евалуацијом над скупом података из Yahoo!Answers система. Уколико је корисник одговорио на питање, пар корисник-питање се сматра позитивним примером, а ако је корисник поставио питање, овај пар се сматра негативним примером. Разлог за последње је да уколико је корисник поставио питање, то може значити да он не поседује знање о том питању, што показује да корисник није могао да одговори на њега. Међутим, као што аутори рада признају, поступак избора негативних примера је споран. Такође, поређење је извршено за различите врсте атрибута и одређено је колико они доприносе коначним резултатима. Ово је показало да однос атрибута питање-корисник игра кључну улогу у побољшању укупних перформанси. Ипак, проблем новог корисника постоји, као и проблем питања о мишљењу, у којима је разлика између најбољих и не-најбољих одговора субјективна.

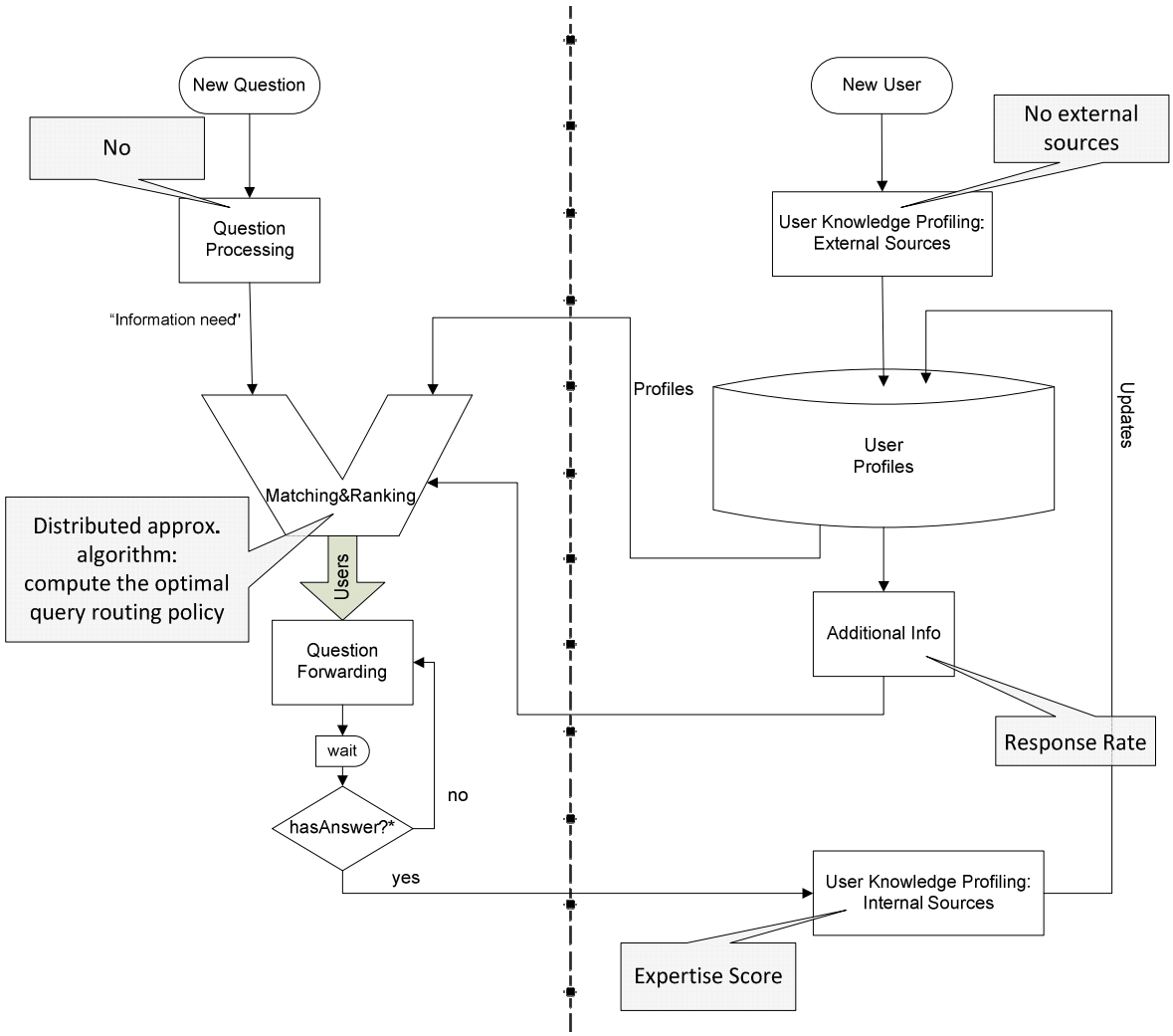


Слика 12. Classification-based Routing in CQA: Структура

XI. SQM

Banerjee A. и Basu S. [20] у 2008. години предложили су социјални модел упита (*Social Query Model* - SQM) за децентрализовану претрагу који представља алгоритам Рангирања страница и одређени Марковљев процес одлучивања као посебне случајеве. Друштвена мрежа је представљена као граф са чворовима и везама. Модел не разматра *обраду питања* и не подржава њихову анотацију. Профил знања корисника обухвата само оцену стручности, а као *додатни податак* уврштена је брзина одзива. Организација је децентрализована и *поређење и рангирање* се заснивају на дистрибуираном апроксимативном алгоритму који рачуна оптималну полису рутирања упита. Стога, у контексту модела ова полиса је истовремено оптимална за све чворове, у смислу да не постоји подскуп чворова који ће имати подстицаја да заједнички користе другачију локалну полису рутирања. Илустрација структуре SQM модела је представљена на слици 13.

У извесној мери сви претходно представљени приступи су комплементарни SQM моделу, јер њихов фокус није био на рутирању упита унутар чворова друштвене мреже, већ на идентификацији потенцијала корисника да пружи тачан одговор и упоређивању тог потенцијала за дато питање. Дакле, овај потенцијал се може окарактерисати различитим факторима, као што су стручност и брзина одзива, који представљају улазне параметре у оквиру SQM модела.



Слика 13. SQM: Структура

Евалуација анализираних приступа и предлог нове методологије

На основу анализе претходно приказаних решења, идентификовани су следећи проблеми, тако да остатак овог поглавља садржи идеје од интереса за могућа побољшања у вези са тим проблемима.

I. Визуализација питања

Питања се обично састоје од текста који није сувише дугачак, па једно решење за имплементацију модула за *обраду питања* је употребом NLP или DM/ML техника. Међутим, алати за аутоматску екстракцију информација у принципу могу бити недовољно прецизни и изоставити неке вредне информације. Такође, кратка питања често могу бити двосмислена. Имајући то у виду, међу свим анализираним, најбоље решење је предложено у систему Конфучије, јер за унето питање кориснику се предлаже скуп категорија за избор. Такође, нове категорије се могу ручно додати. Сходно томе, најефикасније решење је интерактивни кориснички интерфејс који омогућава комуникацију између модула за *обраду питања* и корисника који је то питање поставио. Овај приступ комбинује потпуно аутоматску обраду текста и ручну корекцију резултата, пружајући кориснику могућност повећања тачности излаза. С друге стране, аутоматска обрада може произвести више резултата који би обично били заборављени.

Имајући у виду све претходно наведено, као начин за представу добијених резултата могуће је применити приступ у коме се откривени концепти могу визуелно представити у форми облака концепата (*Tag Cloud* визуелизација). Једна од предности овог приступа је то да "што је концепт значајнији, има већу

додељену величину фонта", што обезбеђује интуитивнију представу специфичних односа између концепата, као и њиховог значаја у питању. Слика 14. представља пример генерисаног облака концепата.

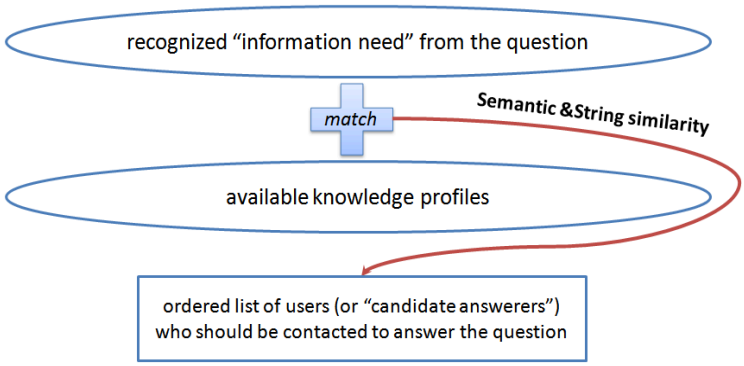
Слика 14. Визуелизација питања: Пример генерисаног облака концепата

II. Проширено семантичко поређење

Мера сличности између питања и профила корисника може се реализовати израчунавањем тачног поклапања између препознатих тема или појмова, или пак израчунавањем њихове семантичке сличности. Као што је поменуто, постављена питања су обично кратка, па одређено питање може бити врло семантички слично профилу корисника, али и даље лексички веома различито. Стога, бољи резултати могу се постићи употребом семантичке сличности. Неколико решења као што су Aardvark, Конфучије и G-Finder користе семантичку сличност у оквиру фазе *поређења и рангирања.* Међутим, питање или профил могу садржати различите облике ретких именица, као и грешке при куцању текста, што даље може умањити прецизност система.

Једно од решења проблема је употреба приступа заснованог на скупу речи (*bag of words*), који као меру сличности речи користи или ону базирану на текстуалном корпусу (*corpus-based*) или базирану на речнику (*knowledge-based*) [21]. За сваку реч у профилу, метод треба да идентификује највеће поклапање са речима из питања, а затим да их сједини у укупну меру семантичке сличности. Islam и Inkpen [22] предложили су побољшање која укључује алгоритам поређења на нивоу низа карактера заједно са мером семантичке сличности речи заснованом

на корпусу. Стога, побољшање се може наћи у овом правцу као што је илустровано на слици 15. Ова метода, поред семантичке сличности речи, укључује и меру лексичке сличности, па се може боље показати код грешка при куцању, затим код нових, популарних и не широко познатих речи или у случајевима различитих облика ретких именица. Ово последње је нарочито изражено код језика са великим бројем флексија речи као што је српски језик [23].



Слика 15. Коришћење семантичког поређења заједно са поређењем на нивоу низа карактера: Илустрација процеса упоређивања

III. Интеграција профила

Проблеми везани за креирање профила корисничког знања су: (1) коришћење Бајесове вероватноће, (2) проблем новог корисника, и (3) проблем интеграције. (1) Бајесова вероватноћа има чврсту теоријску основу и тренутно је у широкој употреби приликом реализација система који се базирају на поверењу. Међутим, овај приступ нема адекватан ниво изражајност и захтева неке вештачке конструкције. На пример, корисник А је одговорио на 100 питања о некој теми Ц и квалитет одговора оцењених од стране других корисника је 0,5. Затим, размотримо други случај у коме корисник А није одговорио ни на једно питање које се односи на тему Ц (немамо никаквих информација о нивоу знања корисника А о теми Ц). У оба случаја, коришћењем Бајесовог приступа, процењени степен поверења у знање корисника А о теми Ц је p(поверење) = 0,5 и p(неповерење) = 0,5. Дакле, Бајесовски приступ нема способност да разликује ова два случаја, између неповерења и незнања [24]. (2) Проблем новог корисника се

може наћи у домену система за препоручивање [25], јер је врло тешко направити профил за тек регистрованог корисника. Системи као што су Aardvark и iLink користе спољне изворе информација (нпр. информације са друштвених мрежа, блогова или ручног уноса) како би успоставили иницијални профил компетентности корисника. Такође, ове изворе затим користе и за његово повремено ажурирање. (3) Овакав приступ уводи трећи проблем, проблем интеграције, који се односи на то како на једноставан начин интегрисати информације о кориснику добијене из различитих извора. Стога, потребан је нови приступ за профилисање корисничког знања који ће омогућити униформно интегрисање различитих извора информација у облику софтверских агената.

Један могући правац ка решавању овог проблема може се наћи у Демпстер-Шафер (DST) теорији, математичкој теорије доказа која представља генерализацију Бајесове вероватноће. Она природно интегрише неодређеност и омогућава комбиновање доказа из различитих извора. Како би се дошло до одређеног степена поверења (представљеног функцијом поверења) DST узима у обзир све расположиве доказе. Поред тога, за интеграцију профила у СИПП могуће је употребити модел поверења заснован на Дезерт-Шмарандаш теорији (DSmT) [24], која представља генерализацију DST, стога има већу изражајност.

Предлог решења

Као резултат анализе решења пронађених у отвореној литератури и идентификованих проблема, предложене су три идеје за могућа побољшања. Ове три идеје односе се на реализацију све три фазе СИПП процеса, и то: (а) анализу питања и одговора, тј. идентификацију релевантних информација из питања или одговора, (б) прослеђивање питања, тј. проналажење компетентног корисника за постављено питање и (в) профилисање компетентности корисника из угла интересовања на основу различитих извора информација. Структура ових фаза и однос њихових компоненти представљени су у уведеној презентационој парадигми. На основу донетих закључака реализован је прототип софтверског система у оквиру кога је примењена предложена парадигма, а поменуте фазе су

реализоване применом наведених принципа визуелизације, проширеног семантичког поклапања и интеграције профила. Такође, једна од основних смерница у реду је била и то да се предложени приступ подједнако може употребити за енглески, српски или неки други језик са врло ограниченим електронским лингвистичким ресурсима.

IV Опис и реализација система

Анализа питања и одговора

Постоје бројни приступи за екстракцију информација (*Information Extraction*) из текста написаног на природном језику, неки се базирају на *text minig* техникама и статистичком приступу или други, који су засновани на принципима рачунарске лингвистике и обраде природног језика – NLP. Предности статистичког приступа су првенствено већа прецизност са повећањем корпуса за обраду, једноставнија реализација и дужа историја развоја и употребе. С друге стране, уколико је потребно обрадити краћи текст (питање или одговор), што је случај код СИПП система, овакав приступ неће дати најбоље резултате с обзиром на мали број речи које се могу анализирати. Из тог разлога за краће текстове може се употребити нека од техника рачунарске лингвистике. Међутим, алати за аутоматску екстракцију информација у принципу могу бити недовољно прецизни и изоставити информације вредне за даљу анализу и извршавање. Такође, кратка питања често могу бити двосмислена. Стога, најефикасније решење може се наћи у виду интерактивног корисничког интерфејса који омогућава комуникацију између модула за *обраду питања* и корисника који је то питање поставио. Овај приступ комбинује потпуно аутоматску обраду текста и ручну корекцију резултата, пружајући кориснику могућност повећања тачности излаза. С друге стране, аутоматска обрада може произвести више резултата који би обично били заборављени.

Имајући у виду претходно наведене проблеме за представу добијених резултата одабран је приступ у коме су откривени концепти визуелно представљени у форми облака концепата (*Tag Cloud* визуелизација) [26], [27]. Овај начин представе се природно уклапа у формат излазних резултата алата за аутоматску екстракцију информација из текста, с обзиром да њихов излаз чине концепти који се састоје од пара (кључна реч, тежина) – нпр. (биологија, 0,3). Такође, још једна од предности овог приступа је та да "што је концепт значајнији, има додељену већу величину фонта", што обезбеђује интуитивну представу

специфичних односа између концепата, као и њиховог значаја у питању. Такође, генерисани облак концепата представља скуп информација које описују постављено питање. Концепти у овом скупу чине специфичан контекст у ширем смислу који представља отисак (*fingerprint*) обрађеног текста. Овај отисак је специфичан за свако питање и сличан је код питања са истом тематиком и истим значењем. Коначно, уз помоћ овог отиска може се открити специфична релација између питања, као и однос између питања и тема којих се то питање дотиче.

С друге стране, за сваког корисника СИПП систем такође води евиденцију о његовом профилу који је представљен на исти начин и описује корисникову компетентност евидентирану на основу постављених питања и датих одговора. Стога, сваки идентификовани појам из питања или корисничког профила јединствено је описан паром (кључна реч, тежина) који се даље назива *концептом*.

Преглед употребљених алата

Antelope (*Advanced Natural Language Object-oriented Processing Environment*) [28] је програмски алат за обраду природног језика под .NET окружењем и садржи велики број различитих компонената и библиотека. Овај алат користи проширену верзију WordNet [29] лексикона која надограђује основни лексикон бољом концептуализацијом и интегрише формалне онтологије вишег нивоа. Основни скуп компонената пружа могућности синтаксичке и семантичке анализе текста, препознавање именских ентитета (*named entity recognition*), детекцију контекста, временског периода и локације, извлачење кореференци, као и могућност разлучивања смисла речи (*word sense disambiguation*). Такође, овај алат омогућава приступ другим библиотекама за анализу текста као што су Stanford Parser, WordNet и VerbNet.

ConceptNet [30] је семантичка мрежа која описује уопштено људско знање. Ова мрежа се састоји од речи и фраза међусобно повезаних релацијама. Релације су засноване на уобичајеним људским сазнањима, којих има преко

двадесет врста, као што су „служи за“, „је направљено од“, итд. Овде је битно указати на важан елемент ове семантичке мреже, који је посебно значајан приликом поређења концепата. Наиме, нова сазнања су унета у базу од стране различитх корисника, махом широм света, без постојања било каквог ограничења у погледу уноса. Стога, прикупљени подаци чине мрежу полуструктурираних фрагмената природног језика (речи или фраза и њихових релација) који представљају концепте. Такође, за разлику од лексикона као што је WordNet, ConceptNet у структури своје базе знања садржи поједине двосмислености и непрецизности. Овакве логичке некоректности су неминовне с обзиром на природу самог језика. Са друге стране, на овај начин ConceptNet је оптимизован за проналажење сродних концепата док за разлучивање двосмислених речи користи помоћ лексикона. ConceptNet нуди алате за проналажење концепата, повезивање односно проналажење веза између концепата, проналажење сродних концепата и проналажење теме докумената, као и лексичко, синтаксно и семантичко парсирање текста употребом библиотеке за обраду текста MontyLingua [31]. Сама база је вишејезична, док за сврхе овог истраживања, употребљен је њен део на енглеском језику који, између осталог, има и далеко највише унетих концепата и релација.

SemNet (*Semantic Network of Terms*) [32] је велика семантичка мрежа техничких термина која омогућава извршавање упита за неки термин дохватајући рангирану листу свих њему семантички сродних термина. Ова мрежа је изграђена аутоматски на основу именичких термина добијених из English Google Books Ngram Dataset употребом анализе кореференци. Она се састоји од 2,8 милиона различитих термина, који се састоје од једне или више речи, и 37,5 милиона тежинских релација између њих. Коначно, SemNet садржи велики број истих концепата и релација као и њему сличне семантичке базе знања, нпр. WordNet и ConceptNet.

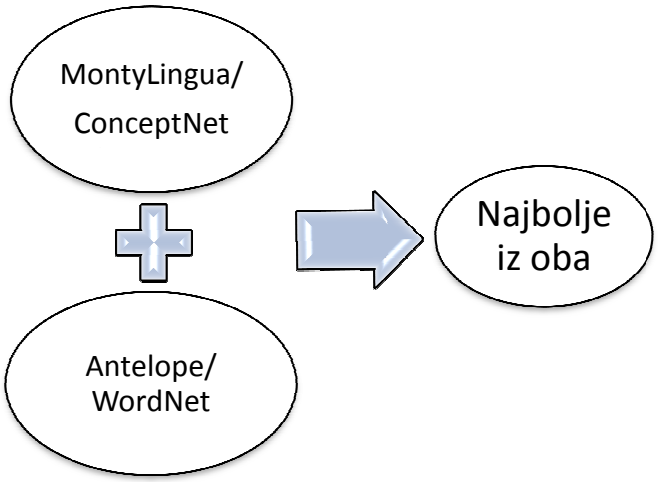
С обзиром на специфичност и комплексност реализације, наведени алати постоје само за веће светске језике као што је енглески, и евентуално француски и шпански, али не и за остале, нпр. српски језик. Из тог разлога размотрена је

употреба статискичких мера TF_{norm} и TF-IDF које су засноване на фреквенцији појављивања термина и које су независне од језика над којим се примењују. Њихов опис и употреба дати су у наредним секцијама.

На крају, цео подситем је реализован помоћу Microsoft .NET платформе при чему за израду корисничког интерфејса је употребљена Windows Presentation Fundation (WPF) технологија, а за потребе комуникације између клијентске и серверске стране је употребљена Windows Communication Fundation (WCF) технологија.

Реализација подситема за анализу питања

С обзиром да се Antelope заснива на знатно мањем, али прецизнијем WordNet лексикону, он је предвиђен за препознавање именских ентитета (нпр. имена људи, организација, држава, градова), као и мањег броја концепата. Такође, у експерименталној фази подржава и препознавање контекста што може допринети бољем препознавању концепата. С друге стране, ConceptNet садржи знатно богатију семантичку мрежу што даје боље резултате приликом идентификовања постојећих као и повезаних концепата, али нема могућности проналажења именованих ентитета нити препознавања контекста. Коначно, у другим студијама показано је да приликом семантичке експанзије упита (*query expansion*) комбиновање алата заснованих на ова два језичка ресурса (WordNet и ConceptNet) је дало боље резултате у односу на њихову појединачну употребу [33]. Стога, модул за анализу питања је реализован интеграцијом ова два алата као што је илустровано на слици 16. Овај приступ је назван Екстракција концепата (*Concept Extraction* – CE).



Слика 16. Интеграција алата Antelope и ConceptNet: илустрација приступа

С обзиром да оба алата, поред идентификованих концепата, одређују и њихову тежину као вредност у опсегу (0,1], приликом интеграције употребљен је следећи приступ: уколико је само један од алата пронашао неки концепт, тај концепт улази са својом тежином у коначни скуп, док уколико су оба алата идентификовала исти концепт резултујућа тежина се рачуна као њихов збир употебом пробабилистичке Т-конорме:

$$U(a,b) = a + b - a \cdot b \quad (1),$$

Ова формула има за нијансу бржу конвергенцију ка 1 у односу на Ајнштајнову Т-конорму:

$$U(a,b) = (a + b) / (1 + a \cdot b) \ (2).$$

Такође, како би се добили бољи резултати и смањио број погрешно идентификованих концепата, предузете су следеће мере:

1) Емпиријски је утврђена линеарна зависност између минималне вредности тежине идентификованог концепта (*threshold*) и дужине питања, тј. броја речи у питању:

$$minimalna\ težina\ za\ koncepte = 8{,}75\% + broj\ reči\ u\ tekstu \cdot 0{,}125\%$$

Ова вредност представља границу испод које се концепт избацује из листе уколико је његова додељена вредност нижа, чиме се смањује број погрешно идентификованих концепата. Код идентификације контекста није пронађена оваква зависност, па је одређена константна минимална гранична вредност.

2) Под претпоставком да је из опширног текста теже доћи до закључка којом се тематиком бави, него из краћег текста састављеног само од најбитнијих појмова, уведена је следећа измена приликом идентификације контекста: по проналаску битних концепата из ConceptNet-а и свих информација добијених од Antelope алата, листа добијених информација се поново прослеђује Antelope алату како би пронашао контекст, овог пута са већом прецизношћу.

3) Коначно, из скупа идентификованих концепата одстрањени су они који представљају стоп речи (*stop words filtering*) – речи са ниским информационим садржајем. У случају да ипак неке од ових речи имају информативни значај корисник их може накнадно унети у облак концепата.

Поред представљеног CE приступа, реализован је и посебан модул који користи SemNet и статистичке мере TF_{norm} и TF-IDF ради поређења добијених резултата. SemNet омогућава претрагу рангиране листе свих семантички сличних термина. Стога, након претпоцесирања улазног текста (поступак је детаљно описан у наредном поглављу под ставком *Претпроцесирање корпуса*) за сваку идентификовану реч дохвата се 3 семантички најсличнија термина. На пример, енглеска реч “car” (аутомобил) описана је са три следећа SemNet концепта: (“front”, 0.038), (“side”, 0.024) и (“truck”, 0.024).

Две доминантне статистичке мере засноване на фреквенцији појављивања термина су TF_{norm} (*Normalized Term Frequency*) и TF-IDF (*Term Frequency–Inverse Document Frequency*). Фреквенција појављивања неког термина одређује се простим пребројавањем понављања те речи у неком корпусу. У овом приступу

употребљни су исти корпуси над којима су креирани семантички простори за српски и енглески језик (описано у наредном поглављу под секцијом *Реализација одређивања семантичке сличности*). На тај начин у фази постпроцесирања корпуса срачунате су фреквенције појављивања свих речи које се налазе унутар корпуса.

За одређивање нормализоване фреквенције термина TF_{norm} употребљена је следећа формула:

$$TF_{norm} = \frac{TF_{\log}}{\max(TF_{\log})} \quad (3),$$

где је TF_{log} логаритамска вредност фреквенције појављивања дате речи у корпусу, а $\max(TF_{log})$ је логаритамска вредност фреквенције појављивања речи која се најчешће појављује у корпусу. TF_{log} се рачуна на следећи начин:

$$TF_{\log} = -\log(\frac{TF_{count}}{n}) \quad (4),$$

где TF_{count} означава број појављивања те речи у корпусу, а n је укупан број речи које се налазе у корпусу.

TF-IDF [34] је нумерички статистички податак који указује на то колико је важна одређена реч која се налази унутар документа у односу на колекцију докумената или цео корпус. Често се користи као тежински фактор за проналажење информација и *text mining*. Основна мотивација код овог приступа је да се уопштеним (и честим) терминима додели мања вредност у односу на термине који носе већу информациону вредност (нпр. именица „Београд“ у односу на везник „и“). Стога, поред фреквенције термина (*Term Frequency* – TF) уводи се и фактор инверзне документ фреквенције (*Inverse Document Frequency* – IDF) који умањује тежину термина који се у скупу докумената јављају врло често, а повећава тежину термина који се јављају ретко. Ово уједно представља и предност TF-IDF у односу на TF_{norm}, па се ова метрика доминантно користи,

изузев у случајевима када је објективно није могуће применити (нпр. када је цео корпус сачињен од једног великог документа).

TF-IDF се рачуна као производ два статитичка податка, фреквенције термина и инверзне документ фреквенције. Њена вредност расте сразмерно броју појављивања речи у документу, али је компензована фреквенцијом појављивања те речи у читавом корпусу, што има за последицу контролу ове вредности с обзиром да се неке речи генерално чешће појављују од других. TF-IDF се рачуна на следећи начин:

$$TF\text{-}IDF = TF(t,d) \cdot IDF(t,D) \quad (5),$$

где је D скуп свих докумената, тј. корпус, а t и d означавају термин односно документ. При том, документ може представљати питање, одговор или целолкупну нит (*thread*) која садржи питање са повезаним одговорима.

$$TF(t,d) = \frac{f(t,d)}{\max\{(f(w,d) : w \in d\}} \quad (6)$$

На сличан начин као код TF_{norm}, $TF(t,d)$ представља фреквенцију појављивања датог термина t у документу d, а како би се спречила пристрасност према дужини докумената, учестаност појављивања термина у документу је нормализована дељењем са највећом учестаности појављивања термина у том документу $\max\{(f(w,d) : w \in d\}$.

$$IDF(t,D) = \log \frac{|D|}{|1 + \{d \in D : t \in d\}|} \quad (7)$$

Инверзна документ фреквенција $IDF(t,D)$ је мера која показује да ли је термин уопштен или се ретко појављује. $|D|$ представља кардиналност скупа D, или укупан број докумената у корпусу, а $|1 + \{d \in D : t \in d\}|$ број докумената у којима се термин t појављује плус 1 како би се избегло дељење са 0.

На крају, извршени су различити експерименти комбиновања метрика SemNet и TF-IDF, при чему је слично као и код CE за агрегацију добијених вредности употребљена пробабилистичка T-конорма (1).

Изглед корисничког интерфејса

Након провере идентитета корисника приказује се основни кориснички интерфејс. Његов изглед је дат на слици 17. Горњи део корисничког интерфејса (2) односи се на евиденцију већ постављених питања, док се доњи (1) односи на визуализацију питања. Након уноса питања и притиска на дугме *GenerateContext!* појавиће се облак идентификованих концепата као на слици 18. Генерисани облак концепата корисник може изменити на два начина:

1) Додавањем или избацивањем неког концепта:

    a. Помоћу дугмета *Add* и *Remove*: Уколико је потребно, корисник може унети још значајних појмова у поље за унос новог концепта и додати га притиском на дугме *Add*. Ово поље има могућност аутоматског завршавања речи (*auto complete*) на основу претраге већ доступних кључних речи које су претходно евидентиране у систему. Такође, одабиром неког концепта из облака, а затим притиском на дугме *Remove* могуће је избацити га из листе.

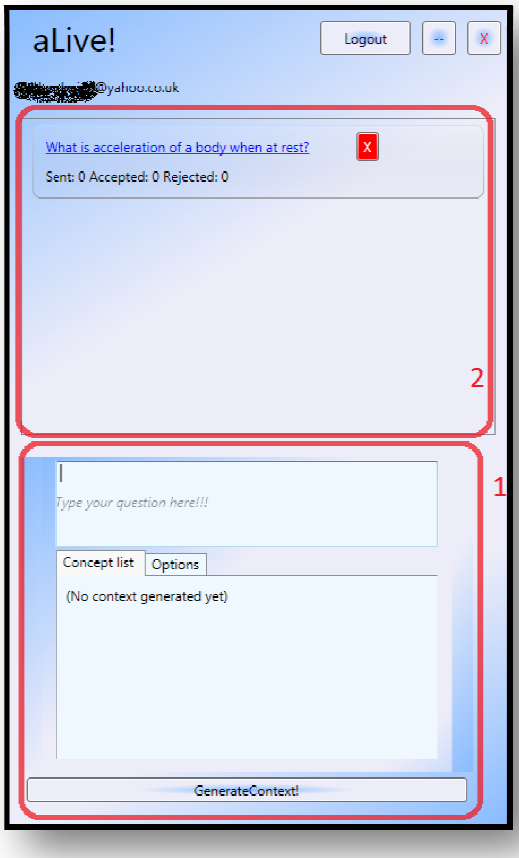
    b. Превлачењем у или из облака концепата: Операцијом превлачења (*drag and drop*) могуће је додати, односно избацити неки концепт из облака.

2) Променом величине неког концепта, односно његове тежине:

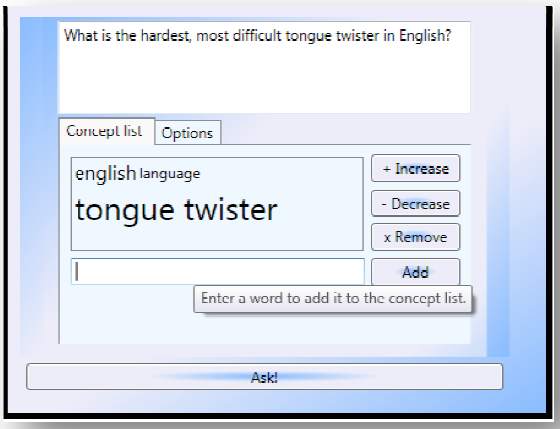
    a. Помоћу дугмета *Increase* и *Decrese* могуће је повећати односно смањити тежину одбраног концепта што ће визуелно бити приказано повећањем односно смањењем величине фонта.

b. Употребом думета за листње (*scrolling*) могуће је такође повећати односно смањити тежину одбраног концепта.

Такође, преласком курсора миша преко облака концепата јавља се помоћни балончић (*tool tip*) са различитим саветима. На крају, пристиском на дугме *Ask!* корисник има могућност слања унетог питања систему.



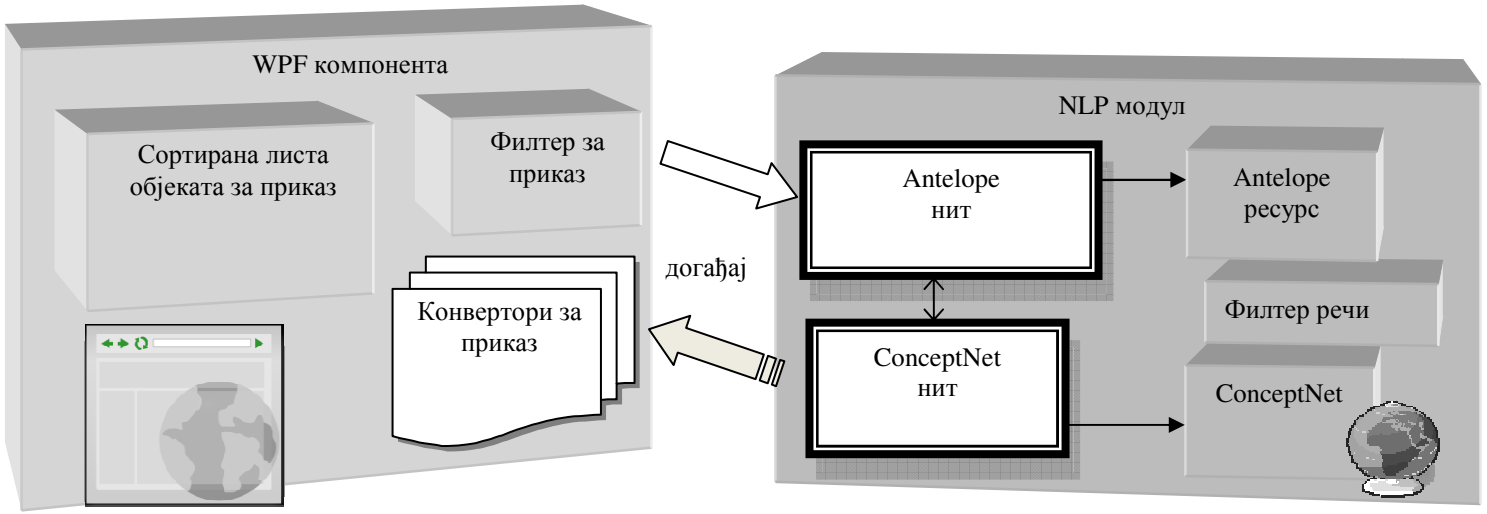
Слика 17. Изглед корисничког интерфејса: (1) део за визуализација питања, (2) део за евиденцију већ постављених питања



Слика 18. Изглед корисничког интерфејса за измену: Пример генерисаног облака идентификованих концепата

Архитектура подсистема за анализу питања

Архитектура реализованог решења је приказана на слици 19. Подсистем се састоји од презентационе WPF компоненте за приказ и интеракцију са облаком концепата и NLP модула за обраду и проналажење битних информација из текста на природном језику. NLP модул і WPF презентациона компонента комуницирају асинхроно путем механизма догађаја. На овај начин NLP модулу се путем асинхроне методе доставља текст, који затим он независно обрађује и на крају обавештава путем догађаја WPF компоненту да су подаци доступни. NLP модул се састоји од две интерне нити које конкурентно анализирају прослеђени текст помоћу Antelope и ConceptNet алата. Такође, NLP модул садржи филтер речи који одстрањује стоп речи. По завршетку обраде подаци се шаљу у облику сортиране листе WPF компоненти која их конвертује у потребни формат и приказује кориснику.



Слика 19. Архитектура подсистема за анализу питања

Анализа одговора и профилисање компетентности корисника

Врло је тешко, напорно, а чак некада и немогуће за корисника да ручно унесе комплетну листу свих области које су у његовом домену компетентности (за разлику од стручности која означава владање скупом блиско повезаних области, компетентност овде представља стварни ниво познавања неке специфичне области, чија вредност може ићи од површног, на нивоу интересовања, до темељног познавања). Такође, ова листа може бити субјективна, с обзиром да свака особа има различит субјективан осећај о нивоу сопствене компетентности (нпр. скромност или с друге стране претеривање), а чак за исту особу овај осећај се може разликовати од области до области. Стога, проблем који је потребно решити је како за сваког корисника аутоматски одредити ниво компетентности на основу информација прикупљених из СИПП система.

У предложеном приступу за сваког корисника креира се профил у коме се евидентира његова компетентност на основу (i) датих одговора и (ii) постављених питања.

i. Дати одговор од стране корисника свакако показује његову компетентност за скуп области на које се питање, а самим тим и одговор, односи. Ово поготово важи у случају да је тај одговор изабран од стране осталих корисника за најбољи међу свим понуђеним одговорима. Adamic са сарадницима [35] анализирала је питања из различитих категорија постављена на Yahoo! Answers порталу и донела закључак да одговори изабрани као најбољи одговори (*best answers*) углавном заиста и јесу најбољи одговори на дато питање.

ii. Поред одговора на питања других корисника, такође и сама постављена питања могу показати у најмању руку ниво заинтересованости корисника за неку област. Стога циљ је био

евидентирати сваку корисничку акцију унутар СИПП система, како би профил био што потпунији.

За анализу одговора употребљене су исте технике као и приликом анализе питања CE, SemNet и TF-IDF. Стога кориснички профил је састављен од скупа концепата описаних паром (кључна реч, тежина). Ово има за последицу да су и информације из питања и кориснички профили представљени на јединствен начин, па се поређење ова два скупа своди на израчунавање њихове сличности. Такође, овде је потребно нагласити да концепти представљају фино гранулиране (*fine grained*) области у односу на свеобухватне и апстрактне теме. На пример, није нужно тачно да неко ко је заинтересован за „математику“ (тема) може коректно одговорити на питање које се односи на „косинусну сличност“ (концепт). На крају, сваки евидентирани концепт (било из питања или одговора) складишти се унутар јединственог профила корисника. При том, у случају да профил већ садржи концепт са истом кључном речи, њихове тежине ће бити агрегиране помоћу пробабилистичке T-конорме (1). Овај приступ има следеће две особине:

i. Резултат агрегације не зависи од редоследа евалуације тј. има особину комутативности и асоцијативности.

ii. Када две оцене имају исту тенденцију, резултат агрегације мора да задржи исту тенденцију и појача је.

Нпр. за два концепта (информатика, 0,2) и (информатика, 0,8), агрегирана вредност, тј. резултујући концепт ће бити (информатика, 0,84). На тај начин вишеструко евидентирање концепата са истом кључном речи ће појачати тежину резултујућег концепта у профилу.

Прослеђивање Питања

Задатак проналажења компетентних корисника врши се поређењем препознатих информација из питања (добијених из модула за *обраду питања*) и расположивих корисничких профила (добијених из *складишта профила корисника*), чији је резултат рангирана листа корисника или "кандидата за одговарање". На основу ове рангиране листе могуће је изабрати једног или више корисника које би затим требало контактирати за одговор на питање.

С обзиром да су добијене информације из питања, као и расположиви кориснички профили, представљени на јединствен начин, у виду концепата, тј. листе парова (кључна реч, тежина), поређење ова два скупа се своди на израчунавање њихове сличности. Одређивање сличности може бити реализовано егзактним поређењем, тј. одређивањем тачног поклапања између речи, или помоћу израчунавања семантичке сличности. С обзиром да одређено питање може бити врло семантички слично неком профилу корисника, али и даље лексички веома различито, бољи резултати се могу постићи употребом семантичке сличности. Као пример који илуструју ову тврдњу може се узети појава синонима, речи која је по значењу идентична или врло слична некој другој речи, али се од ње разликује по свом облику (нпр. речи ученик и ђак). Стога, у наставку дисертације посебан фокус је стављен на одређивање семантичке сличност. Такође, као што је претходно наведено, и овде је једна од основних смерница била то да се предложени приступ подједнако може употребити за енглески, српски или неки други језик са врло ограниченим електронским лингвистичким ресурсима.

Одређивање сематничке сличности две речи

Семантичка сличност представља концепт додељивања метрике скуповима израза или докумената засноване на сличности њиховог значења. Овај концепт један је од кључних за разумевање природних језика, јер омогућава

прављење смислених поређења и закључивања. Због тога одређивање семантичке сличности игра важну улогу у аутоматској категоризацији и сумаризацији текста, машинском превођењу, проналажењу информација и другим областима вештачке интелигенције. Проблем семантичког поређења кратких текстова (*Short Text Semantic Similarity* – STSS) има посебан значај, јер су кратки текстови у широкој употреби на Интернету, у форми натписа и описа производа, анотација слика и веб страница, кратких новинских наслова и вести, итд. Такође, овај проблем игра важну улогу у питањима везаним за образовање и учење, као што су аутоматско тестирање и оцењивање задатака [36].

У отвореној литератури постоји релативно велики број предложених мера за рачунање сличности између две речи, полазећи од оних које се заснивају на одређивању удаљеност између речи унутар семантичке мреже речи, до оних које се заснивају на статистичким моделима дистрибутивне сличности срачунате над великим текстуалним корпусима [21]. Мере засноване на одређивању удаљености се још називају и мерама базираним на речнику (*knowledge-based*) с обзиром да за израчунавање удаљености између речи користе лексикон (речник) који представља семантичку мрежу речи. WordNet [29] је пример једне овакве семантичке мреже где речи представљају чворове графа над којим је могуће израчунати удаљеност. Мере засноване на дистрибутивној сличности се називају и мерама базираним на текстуалном корпусу (*corpus-based*) с обзиром да користе велике текстуалне корпусе над којима рачунају дистрибутивну сличност. Основни поступак код овог приступа је изградња семантичког простора користећи дистрибуцију речи унутар текстуалног корпуса. У таквом простору свака реч има свој контекстни вектор, а семантичка сличност две речи представљена је односом њихових вектора. Овај закључак је последица дистрибутивне хипотезе која тврди да речи са сличним значењем имају тенденцију да се појављују у сличним контекстима. Хипотеза не имплицира да се речи морају појављивати једна поред друге, већ да би требало да се појаве унутар истог скупа речи заједно са осталима.

Велика предност другог наведеног приступа је та што не захтева постојање лексикона нити било каквих других ресурса или алата за обраду

природних језика. Ово представља значајну предност посебно уколико се узме у обзир то да стварање ових ресурса изискује доста времена и труда. Због тога, за разлику од енглеског језика, овакви ресурси још увек нису доступни за многе језике, чинећи многа постојећа STSS решења за енглески језик неприменљива на друге језике. Овај недостатак је посебно евидентан у мањим језицима или језицима са комплексним граматичким правилима као што је српски језик. Стога за потребе реализације употребљен је приступ заснован на текстуалном корпусу погодан за употребу код језика са врло ограниченим електронским лингвистичким ресурсима.

Анализа доступних технологија и алата

У отвореној литератури постоји велики број квалитетних STSS решења за енглески језик, која као меру сличности између речи користе било оне засноване на речнику или текстуалном корпусу. Ипак, имајући у виду да се ова решења често ослањају на напредне и језички специфичне технике обраде текста само неколицина се може применити над језицима са ограниченим електронским лингвистичким ресурсима.

Mihalcea са сарадницима [21] је предложила метод за одређивање семантичке сличности два кратка текста (реченице или пасуса) користећи заједно мере сличности речи засноване на речнику и текстуалном корпусу. За сваку реч у тексту, овај метод идентификује најбоље поклапање са другом речи у супротном тексту и потом тај резултат придодаје укупној мери семантичке сличности. Овај приступ доприноси високом резултату Ф-мере (*F-measure score*), али је рачунски захтеван и захтева употребу семантичког модела речи. Такође, овај приступ има тенденцију прецењивања сличности два текста, што је објашњено у наставку. Islam и Inkpen [22] предложили су побољшање овог приступа који укључује алгоритам поређења на нивоу низа карактера заједно са мером семантичке сличности речи засноване на текстуалном корпусу. Семантичка сличност такође игра важну улогу у задатку препознавања текстуалне дедукције (*recognition of textual entailment* – RTE) и дели многе особине са њом, нпр. [37]. Међутим, као

што је објашњено у [38], RTE представља асиметричан задатак, док одређивање семантичка сличности не, па самим тим изискује и другачији приступ.

Li и сарадници [39] су предложили метод за одређивање сличности реченица који користи плитко парсирање текста. Именичке фразе, глаголске фразе као и фразе предлога издвојене су из прослеђених реченица и коначна сличност се рачуна као комбинација сличности ове три врсте израза. Oliva и сарадници [38] у њиховом раду су такође комбиновали семантичке и синтаксичке информације. Анализа синтаксе се врши кроз процес дубоког парсирања како би се издвојиле фразе у свакој реченици. На крају, сличност између свих појмова који играју исту синтаксичку улогу се израчунава помоћу лексичке базе. Такође, извршени су експерименти са коришћењем психолошке прихватљивости додељивањем другачијих тежина различитим синтаксичким улогама, чиме су потврђена претходна сазнања о томе да људи различито вреднују различите синтаксичке улоге приликом одређивања семантичке сличност. Аутор је дошао до сличних сазнања која су објављена исте године [23].

Нажалост, за велики број језика, међу које спада и српски језик, алати за дубоко и плитко парсирање нису доступни што наведене приступе [38], [39] чини практично неприменљивим. Имајући ово у виду утврђено је да ни једно од постојећих решења се не може директно применити на проблем одређивања STSS тако да се може подједнако употребити за енглески, српски или неки други језик са врло ограниченим електронским лингвистичким ресурсима. Стога, закључено је да је потребно створити сопствени приступ, заснован на модификацији алгоритма [22] из следећих разлога:

1. Овај приступ не користи неку спољну базу знања као што је нпр. *WordNet*, ручно креирана правила закључивања, нити алате за анализу или парсирање текста, који би били препрека у раду са језицима који немају овакве ресурсе.

2. Тачност (*accuracy*) је проценат исправно начињених идентификација од стране система. С обзиром да узима у обзир и лажно позитивне и лажно

негативне ситуације (*false positives and false negatives*), овај параметар представља један од доминантних приликом поређења квалитета различитих мера семантичке сличности. Поред овог постоје и други параметри, као што су прецизност, осетљивост и Ф-мера. Свака од разматраних метода евалуирана је над MSRPC корпусом (*Microsoft Research Paraphrase Corpus*), највећим корпусом парафраза за енглески језик који се састоји од 5801 пара реченица [40]. Методе које користе напредне технике парсирања текста [38], [39] показале су високе резултате код параметара осетљивости и Ф-мере, али нису достигле ниво тачности знатно већи од [22], приликом евалуације над овим корпусом.

3. Овај метод не користи само меру семантичке сличности речи, већ укључује и алгоритам поређења на нивоу низа карактера, што може дати боље резултате у случају грешка при куцању, као и код нових, популарних и не широко познатих речи (*hot words*). Такође, за разлику од енглеског језика, може дати боље резултате у случају поређења различитих облика ретких именица што је нарочито изражено у језицима са великим бројем флексија речи као што је српски језик [23], па резултати евалуације над корпусом на енглеском се могу значајно разликовати од резултата добијених за друге језике.

Како би се конструисао овакав STSS систем било је неопходно анализирати и постојеће алате за уклањање завршетака речи, као и алгоритме за одређивање семантичке и лексичке сличности. Затим је утврђена њихова применљивост на дати проблем и изабрано је најбоље могуће решење.

Уклањање завршетака речи (*stemming*) представља трансформацију код које може доћи до уклањања суфикса речи при чему се не губи основни семантички садржај. Овај поступак се може схватити и као процес нормализације у којем се неколико морфолошких варијанти мапира у исти облик. Овај поступак тако смањује број различитих речи јер се све речи са истом основом мапирају у

исти облик. Нпр, речи шума, шумски и шумовит се све преводе у облик „шум“. Треба још напоменути да је ово једини језички зависан део система који се користи у предложеном приступу.

За уклањање завршетака речи у енглеском језику развијен је већи број различитих решења. Једно од најпознатијих је Портеров стемер [41] које је и овде употребљено. Међутим, за многе језике, укључујући и српски, овакви алати нису јавно доступни или барем не бесплатно. Кешељ и Шипка [42] су предложили општи суфиксни метод за конструисање стемера за језике са богатом флексијом и оскудним ресурсима. Овај приступ има експериментално утврђену тачност од 81,83% за српски језик и примењен је и у овом раду.

Лексичка сличност се заснива на анализи лексичког поклапања речи односно делова речи. Показано је да се бољи резултати могу постићи комбиновањем семантичке и лексичке сличности [22]. Стога у овом раду за одређивање лексичке сличности користе се три варијанте алгоритма најдуже заједничке подсеквенце (*Longest Common Subsequence* – LCS) уз одговарајуће нормализације, а затим се узима просек њихових оцена. То су:

- **NLCS** – *Normalized Longest Common Subsequence*
- **$MCLCS_1$** – *Maximal Consecutive Longest Common Subsequence starting at character **1***
- **$MCLCS_N$** – *Maximal Consecutive Longest Common Subsequence starting at character **N***

Све три варијанте користе поступак нормализације у коме се добијена сличности дели са дужином низа карактера који се пореди. Ово представља предност у поређењу са осталим методима који не узимају у обзир дужину краћег низа, који у неким ситуацијама може имати приметан утицај на коначну оцену. Осим тога, комбинација узастопних и не узастопних мера подсеквенце служе балансирању резултата.

За семантичко поређење речи искоришћени су готови алгоритми за процесирање текстуалних корпуса из S-Space пакета [43], који такође укључује разне алате за његово претпроцесирање и постпроцесирање. Алгоритми за процесирање корпуса се заснивају на коришћењу матрице заједничког појављивања речи (*co-occurrence matrix*). У њој свака врста представља јединствену реч, а свака колона репрезентује контекст. Свака ћелија, тј. елемент матрице, садржи број појављивања речи у датом контексту. Контекст може бити документ или регион неке друге речи, у зависности од алгоритма. У случају документа, димензије вектора одговарају укупном броју докумената, док у случају речи димензије вектора могу, у најгорем, одговарати укупном броју различитих речи које се могу наћи у корпусу. У предложеном приступу употребљени су следећи алгоритми:

- **COALS** – *Correlated Occurrence Analogue to Lexical Semantic* [44]
- **RI** – *Random Indexing* [45]

COALS алгоритам је одабран с обзиром да постиже већу прецизност од старијих алгоритама као што је HAL (*Hyperspace Analogue to Language*) [44]. Осим тога, за разлику од LSA (*Latent Semantic Analysis*) [46] који као улаз очекује скуп докумената, COALS употребљава неиздиференцирани корпус текстова над којим користи покретни прозор за одређивање груписања речи. На овај начин матрица заједничког појављивања речи COALS алгоритма је скоро фиксна, за разлику од LSA матрице чије димензије су пропорционалне броју докумената. Стога, COALS алгоритам је далеко скалабилнији и лакши за примену у ситуацијама које захтевају коришћење великог текстуалног корпуса. Међутим, ова скалабилност се постиже коришћењем рачунски захтевних техника за смањење димензија простора речи под називом декомпозиција сингуларних вредности (*Singular Value Decomposition* – SVD), алгебарске операције која користи факторизацију и декомпозицију матрице. Овај приступ је скуп у смислу потрошње меморије, а чак може бити и неприменљив, посебно за велике корпусе где почетна величина простора речи може бити огромна. Из тог разлога размотрен

је и RI алгоритам који због свог специфичног инкременталног рада не захтева посебну фазу смањења димензија и нарочито је погодан за обраду великих текстуалних корпуса.

На крају, за одређивање сличности пара речи одређује се његова лексичка и семантичка сличности чије се вредности комбинују у коначни резултат. Ови резултати добијени за сваки пар речи се затим користе за израчунавање укупне сличност два фрагмента текста. При том, за сваки пар сличност се одређује узимајући у обзир и тежину речи које га чине. Ово је последица претходно наведеног принципа да "што је концепт значајнији, има већу додељену тежину", који је употребљен и приликом обраде питања и приликом профилисања корисника.

Реализација подсистема за прослеђивање питања

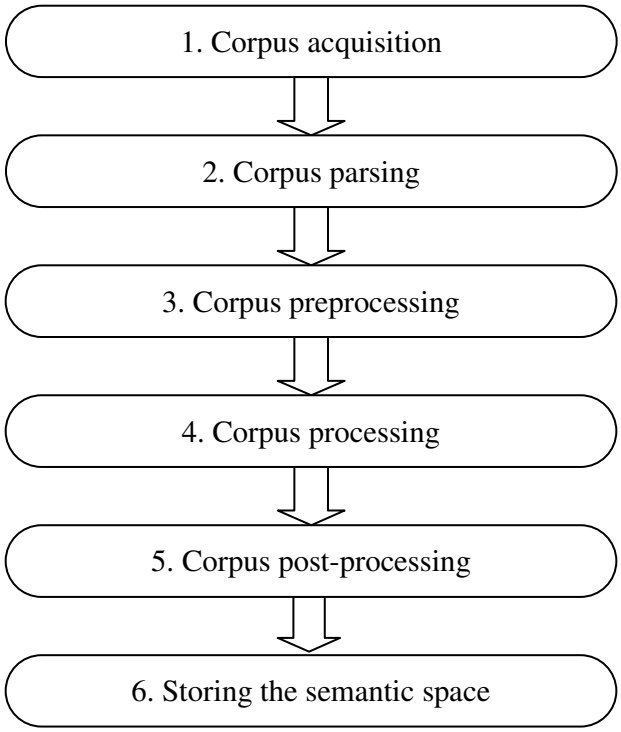
У наставку је прво описана реализација одређивања семантичке сличности по фазама, а затим је наведен предложени алгоритам и дат пример његовог извршавања.

Реализација одређивања семантичке сличности

Ток процеса реализације система за одређивање семантичке сличности је приказан на слици 20. која описује фазе стварања семантичког простора.

1. *Прибављање корпуса* (*Corpus acquisition*) подразумева проналажење довољно великог, јавно доступног и бесплатног корпуса текстова који би могао да послужи као основа за креирање семантичког простора. Као добро решење показали су се корпуси апстраката чланака са Википедије на српском и енглеском језику који су доступни у виду XML фајла. У време прибављања, корпус на енглеском чинило је преко милион и по страница укупне величине око 1,4 гигабајта, а корпус на српском скоро пола милиона страница у укупном износу од преко 186 мегабајта.

2. *Парсирање корпуса* (*Corpus parsing*) је неопходно како би се уклониле све сувишне информације из даљег разматрања и издвојио текст од интереса. У употребљеним корпусима текст сваког апстракта налази се између специфичних XML ознака, док остатак фајла чине информације небитне за креирање семантичког простора. Стога је било потребно извршити екстракцију жељеног текста.
3. *Претпроцесирање корпуса* (*Corpus preprocessing*) служи како би се смањио укупан број различитих речи у корпусу, чиме се смањују димензије контекстних вектора речи, а тиме и оптерећење рачунарских ресурса. У предложеном приступу, претпроцесирање се врши у три корака:
    i. *Чишћење текста* – подразумева уклањање карактера који спадају у друга писма, уклањање бројева и речи које садрже бројеве, уклањање интерпункције, и изједначавање малих и великих слова.



Слика 20. Ток процеса реализације система: Фазе стварања семантичког простора.

ii. *Уклањање стоп речи* – стоп речи су помоћне речи као што су предлози, заменице и везници, који носе занемарљив семантички садржај, али које се често појављују због њихове језичке функције. Уклањањем ових речи, смањује се укупан број различитих речи у корпусу што као резултат има смањење семантичког простора и повећање тачност семантичких алгоритама, јер везе између семантички важних речи постају више наглашене. За енглески језик употребљена је стандардна листа стоп речи, док за српски језик фромирана је стоп-листа речи настала прикупљањем најчешћих речи из текст корпуса [47]. С обзиром на то да употребљени корпус садржи опште, енциклопедијско знање, за очекивати је да ће поједине фреквенције речи у њему релативно тачно одражавати и опште фреквенције речи у самом језику. Информације о фреквенцијама речи у корпусу добијене у овом кораку сачуване су за потребе каснијег рачунања разних фреквенција за сваку реч.

iii. *Стемовање* – за корпус на енглеском језику употребљен је претходно наведени Портеров стемер. Корпус чланака са Википедије за српски језик написан је делимично на ћириличном, а делимично на латиничном писму и кодиран је у UTF-8 формату. Ово је представљало проблем зато што коришћени стемер за српски језик прихвата као улаз само речи написане у специјалном дуал1 кодирању код кога се сваки дијакритик кодује комбинацијом два недијакритичка слова. Из тог разлога, било је потребно најпре извршити конверзију која ће текст корпуса превести са ћирилице и латинице у ово специјално кодирање, а затим извршити стемовање.

4. *Процесирање корпуса* (*Corpus processing*) подразумева избор жељеног алгоритма за креирање семантичког простора и задавање улазног фајла који садржи претпроцесирани текст корпуса.
5. *Постпроцесирање* (*Corpus post-processing*) остварује редукцију димензија контекстних вектора, односно редукцију димензија матрице заједничког појављивања речи, чиме се смањује комплексност израчунавања приликом

одређивања сличности реченица. Сваки алгоритам одвојено спроводи постпроцесирање, при чему је овај поступак енкапсулиран у самим алгоритмима који су део S-Space пакета. Као засебан део постпроцесирања срачунате су фреквенције појављивања сваке речи у корпусу чије вредности су касније употребљене за рачунање TF-IDF и TF_{norm} описаних у предходном поглављу.

6. *Чување семантичког простора* на хард диску (*Storing the semantic space*) је неопходно како би се избегло његово поновно креирање при сваком покретању програма. Чување у виду фајла није практично због лоших перформанси насумичног приступа једном делу огромног фајла. Зато се семантички простор смешта у базу података која користи индексну структуру, што као резултат има прихватљиву брзину приликом претраживања и приступа подацима. База садржи две табеле – једну намењену семантичком простору који је добијен коришћењем COALS алгоритма и другу намењену семантичком простору добијеном коришћењем RI алгоритма. Обе табеле имају идентичну структуру и садрже колоне за кључ, реч и одговарајући контекстни вектор који се чува као дугачак низ карактера. Додатна табела базе података посвећена је вредностима TF-IDF и TF_{norm} срачунатих за сваку реч која се појављује у корпусу.

Алгоритам одређивања сличности између питања и корисничког профила

За одређивање сличност између питања и корисничког профила употребљен је језички модел на принципу „вреће речи“ (*bag-of-words*), у коме се разматрају само речи које сачињавају скуп, али не и њихове међусобне зависности у реченици. Предложени приступ заснован је на модификацији I&I-STSS алгоритма [22], у ком за сваку реч у краћем тексту налазимо најсличнију реч у дужем тексту. Модификација је инспирисана методом предложеним у [21], названим SemSim, који се ослања на сличан приступ, али узима у обзир такође и специфичност речи како би израчунао њихову узајмну сличност. У наставку су

прво дискутоване предности и мане ова два приступа, а затим је дат предложени алгоритам.

Главна разлика између ова два приступа је у томе што I&I-STSS, за разлику од SemSim, проналази пар најсличнијих речи који одстрањује из даљег разматрања. С друге стране, SemSim дозвољава да више речи из једне реченице буду најсличније са истом речи из друге реченице. Ова чињеница доводи до неких погрешних процена сличности. Као пример узмимо две синтагме: „ботаничка башта“ и „краљев врт“. SemSim би најпре анализирао прву синтагму и закључио да је реч „башта“ најсличнија речи „врт“ из друге целине, али би такође закључио да је и речи „ботаничка“ најсличнија реч „врт“, тј. „врт“ би два пута фигурирао у оцени сличности. Затим би алгоритам анализирао другу синтагму, упарио опет реч „врт” са „баштом“, а "краљев" са било којом од две речи из прве синтагме. Овај приступ, дакле, има тенденцију прецењивања сличности, јер дозвољава да се више речи из једне реченице упари са једном истом речи у другој реченици, без обзира на то да ли је реч већ упарена. Због ове праксе, неки парови различитих реченица ће погрешно бити протумачени сличним. С друге стране, ово је имплицитно немогуће код I&I-STSS приступа, јер се пронађени пар речи одстрањује из даљег разматрања, па би у наведеном примеру овај приступ најпре упарио речи „башта“ и „врт“ као најсличнији пар речи и одстранио их из разматрања доделивши им високу оцену сличности. Затим би поредио једине две преостале речи – „ботаничка“ и „краљев“, закључио да оне немају велики степен сличности и доделио им ниску оцену. Стога, овај приступ уравнотежује оцену сличности међу реченицама, тј. даје реалистичнију оцену укупне сличности.

Представљени проблем SemSim донекле превазилази тако што пореди само речи које припадају истој врсти (*part-of-speech*). Међутим, као што је раније примећено, напредни алати за анализу текста нису доступни за многе језике, па овај приступ није примењљив у том случају. Ипак, SemSim побољшава резултате узимајући у обзир специфичности речи приликом одређивања њихове сличности. Из тог разлога ова два приступа [21], [22] су комбинована употребом нормализоване тежине, где је циљ био да се с једне стране превазиђе наведени

проблем, а с друге стране да се на природан начин укључи и тежина концепта коме одређена реч припада (добијена анализом питања или профилисањем корисника).

Такође, претходно наведени проблем је значајан код одређивања семантичке сличности два кратка текста, имајући у виду да је битно одредити који парови реченица су слични, али такође и који су различити. Међутим, један од циљева истраживања је био и испитати да ли је ово случај приликом проналажења компетентних корисника, тј. рачунања сличности пара (питање, кориснички профил), јер за дато питање је неопходно одредити само који је профил најсличнији, али не и који су профили различити у односу на ово питање. Стога, у извесној мери наведени проблем представља проблем класификације једне класе (*one-class-classification*) с обзиром да постоје само позитивни примери, нрп. познат је корисник који је најбоље одговорио на питање, док су негативни примери непознати, тј. није познато који корисници нису у стању да дају најбољи одговор. Из тог разлога уведен је специфичан, модификован алгоритам, назван P2Q (*Profile-to-Question*), који тежи одређивању највеће (максималне) сличност између концепата из питања ка концепатима из профила, али не и обратно.

У наставку је прво дат опис основног приступа, назван LInSTSS (*Language Indepented STSS*) који представља унапређење постојећих STSS алгоритама, с обзиром да се приликом одређивања семантичке сличности узима у обзир и тежине додељене концептима који се пореде. Након тога описан је специфичан приступ – P2Q, који је прилагођен проблему одређивања сличности између питања и корисничког профила.

Основни приступ – LInSTSS

У предложеном основном приступу сличност $S(q_i, u_j)$ између два концепта q_i и u_j рачуна се тако што се сличност добијена између парова речи које ова два концепта садрже $\gamma_{\mathrm{i,j}}$ множи нормализованом тежином $w(q_{\mathrm{i}}, u_{\mathrm{j}})$:

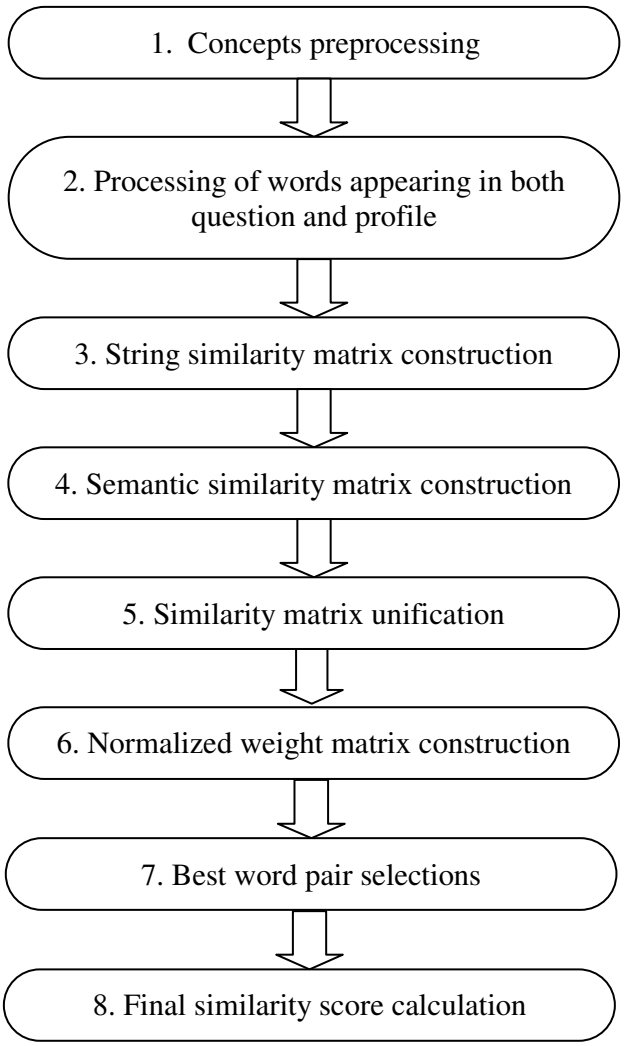
$$S(q_i, u_j) = \gamma_{ij} \cdot w(q_i, u_j) \quad (4)$$

Нормализована тежина $w(q_\text{i}, u_\text{i})$ се рачуна на следећи начин:

$$w(q_i, u_j) = 2^{w(q_i) \cdot w(u_j) - 1}, \text{ где } w(q_\text{i}), w(u_\text{j}) \in (0,1] \quad (5)$$

При том $w(q_\text{i})$ представља тежину концепата q_i из питања, а $w(u_\text{j})$ тежину концепта u_j из профила. Користећи ову технику нормализације, могу се добити нормализоване вредности за сваки пар речи у опсегу (0.5, 1]. Пар концепата који имају додељену ниску важност, па самим тим и тежину чија вредност је блиска 0, имаће нормализовану тежину блиску 0,5, док код оних са великом тежином вредност нормализоване тежине ће тежити или бити једнака 1. Заправо, када се користи нормализована тежина концепата за пондерисање резултата сличности, код пара сачињеног од концепата са високом тежином резултат сличности ће задржати потпуну или скоро потпуну оцену сличности, док код парова који садрже не тако битне концепте сличност ће бити умањена за највише 50%.

Процес извршавања предложеног алгоритма приказан је на слици 21. која описује фазе одређивања семантичке сличности између концепата из питања и корисничког профила. Опис извршавања алгоритма по фазама је дат у наставку.



Слика 21. Процес извршавања алгоритма: Фазе одређивања сличности између питања и корисничког профила.

1. Као и приликом фазе претпроцесирања корпуса описане у претходној секцији, фаза претпроцесирања листе концепата (Concepts preprocessing) започиње поступком чишћења текста, уклањањања стоп речи и стемовања, чиме се обрађују кључне речи садржане у сваком улазном концепту. Након ове обраде, уколико постоје концепти код којих се кључна реч састоји од више речи, за сваку реч се креира нови концепт са додељеном истом тежином као и оригинални, нпр. за концепт (ботаничка башта, 0.5) биће креирана два нова концепта (ботаничка, 0.5) и (башта, 0.5). Након тога, у случају да постоји више концепата који садрже исту кључну реч, сви такви

концепти се спајају у један израчунавањем резултујуће тежине помоћу пробабилистичке Т-конорме (1).

Након претпроцесирања улазних података питање је представљено скупом Q датим у (6), а профил скупом P датим у (7). Q садржи парове $\left(q_{\mathrm{i}}, \mathrm{w}(q_{\mathrm{i}})\right)$ који представљају концепте, где q_{i} означава кључну реч која одређује концепт, а $\mathrm{w}(q_{\mathrm{i}})$ додељену тежину. Слично, P садржи парове $\left(u_{\mathrm{j}}, \mathrm{w}\left(u_{\mathrm{j}}\right)\right)$. Кардиналност скупа Q је m, а P је n.

$$Q = \left\{(q_1, \mathrm{w}(q_1)), \left(q_2, \mathrm{w}(q_2)\right), \ldots (q_m, w(q_{m)})\right\} \quad (6)$$

$$U = \left\{(u_1, \mathrm{w}(u_1)), \left(u_2, \mathrm{w}(u_2)\right), \ldots (u_n, w(u_{n)})\right\} \quad (7)$$

2. Процесирање речи које се налазе и у питању и профилу (*Processing of words appearing in both question and profile*) започиње идентификацијом ових речи које се затим уклањају из даље обраде заједно са концептима којима припадају. С обзиром да се у овом случају ради о идентичним речима, њихова сличност износи 1 па је крајња вредност сличости између концепата одређена нормализованом тежином из (5). На крају, све вредности се сумирају чинећи суму сличности S_{same} а затим се ови концепти избацују из даљег разматрања.
3. Приликом креирања матрице лексичке сличности (*String similarity matrix construction*) добија се матрица M_1 димензија $m \times n$ где свака ћелија заузима нумеричку вредност α која се налази између 0 и 1 и представља сличност на нивоу низа карактера између речи-колоне и речи-реда. Редови матрице се користе за преостале речи из питања, а колоне представљају преостале речи из профила. Вредност нула означава потпуно различите садржаје ова два низа, а вредност један указује на идентичне. Начин одређивања лексичке сличности помоћу алгоритма најдуже заједничке подсеквенце описан је раније у оквиру овог поглавља.

$$M_1 = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{m1} & \cdots & \alpha_{mn} \end{pmatrix} \qquad (8)$$

4. Одређивањем матрице семантичке сличности (*Semantic similarity matrix construction*) добија се матрица M_2 димензија $m \times n$ где свака ћелија заузима нумеричку вредност β која се налази између 0 и 1 и представља семантичку сличност између речи-колоне и речи-реда. Редови матрице се користе за речи из питања, а колоне представљају речи из профила. Слично као и код матрице лексичке сличности, вредност нула означава потпуно различите семантичке садржаје, а вредност један указује на идентичне. Семантичка сличност пара речи одређује се израчунавањем косинусне сличности њихових контекстних вектора добијених из предходно креиране базе података семантичке сличности речи.

$$M_2 = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1n} \\ \vdots & \ddots & \vdots \\ \beta_{m1} & \cdots & \beta_{mn} \end{pmatrix} \qquad (9)$$

5. Унификација матрица сличности (*Similarity matrix unification*) комбинује матрице лексичке и семантичке сличности у једну множећи их одређеним тежинским фактором и сумирајући их као у (10) чиме се добија унификована матрица сличности M_3. Овде су употребљене емпријиски утврђене вредности $\psi = 0{,}45$ и $\varphi = 0{,}55$.

$$M_3 = \psi M_1 + \varphi M_2,\ \psi + \varphi = 1 \qquad (10)$$

$$M_3 = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mn} \end{pmatrix} \qquad (11)$$

6. Нормализована тежинска матрица (*Normalized weight matrix construction*) добија се множењем сваке ћелије $\gamma_{\mathrm{i,j}}$ матрице M_3 одговарајућом нормализованом тежином $w(q_{\mathrm{i}}, u_j)$. Изглед ове тежинске матрице M_4 приказан је у (12) при чему тежине $w(q_{\mathrm{i}}, u_j)$ су добијене помоћу (5).

$$M_4 = \begin{pmatrix} w(q_1, u_1) \cdot \gamma_{11} & \cdots & w(q_1, u_n) \cdot \gamma_{1n} \\ \vdots & \ddots & \vdots \\ w(q_{\mathrm{m}}, u_1) \cdot \gamma_{m1} & \cdots & w(q_{\mathrm{m}}, u_n) \cdot \gamma_{mn} \end{pmatrix} \qquad (12)$$

7. Одабир најбољих парова речи (*Best word pair selections*) се врши над финалном матрицом сличности. Циљ је упарити речи из оба скупа на основу вредности њиховог међусобног поклапања, тј. одабрати парове речи који имају највећу додељену вредност унутар коначне мартице сличности. Када се пронађе овакав пар, вредност дате ћелије придодаје се суми $S_{different}$, а затим се уклања ред и колона матрице којој припада изабрана ћелија. На овај начин се одбацују сви други парови речи у којима се појављују речи из одабраног пара. Поступак се понавља све док не нестане редова и/или колона у матрици.
8. Израчунавање коначног резултата сличности (*Final similarity score calculation*) врши се помоћу следеће формуле:

$$S(Q,U) = \frac{(S_{same} + S_{different}) \times (m+n)}{2mn} \quad (13)$$

Другим речима, коначни резултат сличност $S(Q,U)$ се добија сабирањем резулатата сличност речи (концепата) које се појављују у оба текста (S_{same}) и резултата формираних од парова речи јединствених у једном од текстова ($S_{different}$). На крају се овај износ множи реципрочном хармонијском средином дужине оба текста, како би се постигао коначан резултат сличности између 0 и 1.

Приликом реализације алгоритма, с обзиром да израчунавање сличности између одређеног питања и сваког корисника понаособ представља независтан поступак, тј. кориснички профил је независтан за сваког корисника, употребљена је техника конкурентног извршавања како би се повећале перформансе система. Употребљени приступ заснован је на принципу „вреће задатака“ (*bag-of-tasks*) где се више нити-радника (*workers*) извршава конкурентно на следећи начин:

```
while (true) {
 // dohvati jedan zadatak, tj. korisnika iz skupa
 if (nema preostalih zadataka) break;
 //izvrši zadatak, tj. izračunaj sličnost S(korisnik, pitanje)
}
```

Овај приступ, поред тога што је једноставан за реализацију, уводи могућност балансирања оптерећења с обзиром да за дати случај број нити-радника је знатно мањи у односу на број корисничких профила који се пореде са постављеним питањем, а такође број концепата који профил може садржати се разликује од корисника до корисника, па самим тим и време извршавања одређеног задатка.

Специфичан приступ – P2Q

Код овог приступа кораци 1, 3 и 4 се извршавају на идентичан начин као и у основном приступу. Корак 2. (Процесирање речи које се налазе и у питању и профилу) је изостављен, с обзиром да је намера одредити највећу (максималну) сличност између ова два скупа – питања Q и профила P. Стога се ниједна од речи не избацује из даљег разматрања, што оставља могућност да се више речи из P упари са истом речи у Q.

Након корака 4. извршава се крајњи корак 5. - израчунавање коначног резултата сличности (*Final similarity score calculation*). Код P2Q приступа овај корак се извршава на другачији начин, тако што се за сваки ред из унифициране матрице сличности M_3, проналази максимална γ_{ij} вредност. С обзиром да се сваки ред у овој матрици односи на речи из питања, тј. садржи вектор сличности ове речи ка свим речима из профила са којим се врши поређење, на овај начин се проналази максимална вредност између свих речи из питања ка речима из профила. Ова максимална вредност i-тог реда означена је са $max_i(\gamma_{ij})$, где $i \in 0,..m$. На крају, за сваки ред ова вредност се множи одговарајућом нормализованом тежином $w(q_i, u_j)$. Коначан резултат сличности се добија као аритметичка средина ових вредности:

$$S(Q,U) = \frac{1}{m}\sum_{i=0}^{m} w(q_i,u_j)\cdot \max_i(\gamma_{ij}) \qquad (14)$$

Пример извршавања алгоритма

Предложени алгоритам (основни приступ) могуће је употребити и за одређивање семантичке сличности над паром (питање, профил), као и над паром реченица, с обзиром да је текст сваке реченице могуће представити као скуп концепата. Такође, овај приступ је применљив за енглески језик, али је погодан за употребу и код језика са врло ограниченим електронским лингвистичким ресурсима, као што је српски језик. У наставку је демонстрирано извршавање алгоритма за пар реченица на српском језику које су приказане на слици 22. Овај пар узет је из корпуса парафраза српског језика, чији је опис дат у поглављу *V Евалуација*.

1. Након фазе претпроцесирања, пар реченица из примера има облик приказан на слици 23. Број концепата, односно идентификованих токена у првом тексту је 12, а у другом 17. Тежине концепата одређене су помоћу претходно описане TF_{norm} метрике.
2. У фази идентификације истих речи биће издвојене речи: podizn, spusxtajucy, platfor, osob, ju, pescacxk, prolaz, terazij. За речи које се не налазе у корпусу, па самим тим ни у бази семантичке сличности, TF_{norm} тежина ће бити аутоматски постављена на 1. Табела 1. приказује TF_{norm} тежину и коначну сличност свих ових речи. За дати примеру вредност суме S_{same} ће бити 6,319.
3. Табела 2. приказује матрицу лексичке сличности која одговара пару реченица из примера.
4. Табела 3. приказује матрицу семантичке сличности добијене коришћењем COALS алгоритма за пар реченица из примера.
5. Табела 4. приказује јединствену матрицу сличности која одговара пару реченица.
6. Табела 5. приказује норамлизовану мартицу тежина за дати пар реченица, а табела 6. финалну матрицу сличности.

Rečenica 1 (R1):
Podizno-spuštajuća platforma za osobe sa posebnim potrebama je juče puštena u rad u pešačkim prolazima na Terazijama.

Rečenica 2 (R2):
Osobe sa invaliditetom i svi koji imaju poteškoće sa kretanjem od juče mogu da koriste podizno-spuštajuću platformu u podzemnom pešačkom prolazu na Terazijama u Beogradu.

Слика 22. Пример пара реченица узет из корпуса парафраза српског језика.

Rečenica 1 (R1):
podizn spusxtajucy platfor osob posebn potre ju pusx rad pesxacxk prolaz terazij

Rečenica 2 (R2):
osob invaliditet svi ima potesxkocy kretany ju mo kor podizn spusxtajucy platfor podzemn pesxacxk prolaz terazij beograd

Слика 23. Пример пар реченица након фазе претпроцесирања текста.

ТАБЕЛА 1

TF_{NORM} WEIGHTS AND SIMILARITY SCORES OF WORDS APPEARING IN BOTH TEXTS

Word	podizn	spusxtajucy	platfor	osob	ju	pescacxk	prolaz	terazij
TF_{norm} weight	1.000	1.000	0.733	0.548	0.561	0.849	0.643	0.889
Similarity score	1.000	1.000	0.726	0.616	0.622	0.824	0.666	0.865

ТАБЕЛА 2

STRING SIMILARITY MATRIX

Word	invaliditet	svi	ima	potesxkocy	kretany	mo	kor	podzemn	beograd
posebn	0.010	0.037	0.000	0.094	0.039	0.055	0.037	0.189	0.016
potre	0.030	0.000	0.000	0.224	0.075	0.066	0.110	0.160	0.047
pusx	0.000	0.055	0.000	0.116	0.000	0.000	0.000	0.035	0.000
rad	0.050	0.000	0.073	0.000	0.079	0.000	0.073	0.031	0.283

ТАБЕЛА 3

SEMANTIC SIMILARITY MATRIX

Word	invaliditet	svi	ima	potesxkocy	kretany	mo	kor	podzemn	beograd
posebn	0.000	0.127	0.211	0.000	0.132	0.181	0.195	0.083	0.093
potre	0.000	0.105	0.152	0.000	0.097	0.139	0.106	0.062	0.047
pusx	0.000	0.029	0.054	0.000	0.027	0.034	0.048	0.020	0.043
rad	0.000	0.131	0.191	0.000	0.150	0.200	0.204	0.085	0.198

ТАБЕЛА 4

UNIFIED SIMILARITY MATRIX

Word	invaliditet	svi	ima	potesxkocy	kretany	mo	kor	podzemn	beograd
posebn	0.005	0.086	0.116	0.042	0.090	0.124	0.124	0.131	0.058
potre	0.014	0.058	0.084	0.101	0.087	0.106	0.108	0.106	0.047
pusx	0.000	0.040	0.030	0.052	0.015	0.019	0.026	0.027	0.024
rad	0.023	0.072	0.138	0.000	0.118	0.110	0.145	0.061	0.236

ТАБЕЛА 5

NORMALIZED WEIGHT MATRIX

Word	invaliditet	svi	ima	potesxkocy	kretany	mo	kor	podzemn	beograd
posebn	0.711	0.640	0.593	0.707	0.646	0.580	0.620	0.659	0.598
potre	0.756	0.669	0.611	0.751	0.676	0.595	0.644	0.691	0.617
pusx	0.810	0.702	0.632	0.804	0.711	0.613	0.672	0.730	0.640
rad	0.679	0.620	0.580	0.676	0.625	0.569	0.603	0.636	0.585

ТАБЕЛА 6

FINAL SIMILARITY MATRIX

Word	invaliditet	svi	ima	potesxkocy	kretany	mo	kor	podzemn	beograd
posebn	0.003	0.055	0.069	0.030	0.058	0.072	0.077	0.086	0.035
potre	0.010	0.038	0.051	0.076	0.059	0.063	0.069	0.073	0.029
pusx	0.000	0.028	0.019	0.042	0.010	0.011	0.018	0.020	0.015
rad	0.015	0.045	0.080	0.000	0.074	0.063	0.088	0.039	**0.138**

ТАБЕЛА 7

BEST WORD PAIR SELECTION – STEP 2

Word	invaliditet	svi	ima	potesxkocy	kretany	Mo	kor	podzemn
posebn	0.003	0.055	0.069	0.030	0.058	0.072	0.077	**0.086**
potre	0.010	0.038	0.051	0.076	0.059	0.063	0.069	0.073
pusx	0.000	0.028	0.019	0.042	0.010	0.011	0.018	0.020

ТАБЕЛА 8

BEST WORD PAIR SELECTION – STEP 3

Word	invaliditet	svi	ima	potesxkocy	kretany	mo	kor
potre	0.010	0.038	0.051	**0.076**	0.059	0.063	0.069
pusx	0.000	0.028	0.019	0.042	0.010	0.011	0.018

ТАБЕЛА 9

BEST WORD PAIR SELECTION – STEP 4

Word	invaliditet	svi	ima	kretany	mo	kor
pusx	0.000	**0.028**	0.019	0.010	0.011	0.018

7. Одабир најбољих парова речи је приказан по корасцима у табелама 6-11, где је највећа вредност ћелије означена подебљаним писмом:

   Табела 6 – Највећи резултат у финалној матрици сличности има пар речи (rad, beograd). Након овог првог корака сума $S_{different}$ има вредност 0,138.

   Табела 7 – Највећи резултат у кораку 2 у финалној матрици сличности има пар речи (posebn, podzemn). Након овог корака сума $S_{different}$ има вредност 0,224.

Табела 8 – Највећи резултат у кораку 3 у финалној матрици сличности има пар речи (potre, potesxkocy). Након овог корака сума $S_{different}$ има вредност 0,3.

Табела 9 – Највећи резултат у кораку 4 у финалној матрици сличности има пар речи (pusx, svi). Након овог корака сума $S_{different}$ има вредност 0,328.

8. За дати пар реченица коначан резултат сличности има следећу вредност:

$$S(R1, R2) = \frac{(6{,}319 + 0{,}328) \cdot (12 + 17)}{2 \cdot 12 \cdot 17} = 0{,}472$$

Као што ће бити приказано у поглављу *V Евалуација*, оптимална вредност прага када се користи COALS алгоритам је 0,407. Стога, систем ће исправно идентификовати пар реченица из примера као реченице које су веома семантички сличне.

V Евалуација

У овом поглављу ће бити представљена евалуација предложеног решења. На почетку ће бити дат опис два текстуална корпуса парафраза написана на српском и енглеском језику. Ови корпуси су употребљени за евалуацију алгоритма за одређивање семантичке сличности између два кратка текста. Добијени резултати евалуације су даље употребљени како би се извршило фино подешавање параметара и такође стечена искуства применила при реализацији алгоритма за одређивања семантичке сличности између питања и профила. У наставку је дат опис корпуса питања и одговора на основу кога је извршена евалуација целокупног система, затим су представљени резултати евалуације, као и њихова дискусија. При том, коначан циљ евалуације целокупног система је био испитати следеће хипотезе:

h1: Употреба семантичке анализе питања и одговора побољшава резултате у односу на случај када се она не користи.

h2: Тежине додељене концептима играју улогу, тј. правилно додељене тежине могу побољшати резултате целокупног система.

h3: Специфичан приступ – P2Q, прилагођен одређивању семантичке сличности између питања и корисничког профила, даје боље резултате него општи приступ – LInSTSS, који је намењен одређивању семантичке сличности између два кратка текста.

h4: Приликом профилисања компетентности корисника на основу његове активности унутар СИПП система, потребно је узети у обзир, поред датих одговора, и питања која је он поставио.

Евалуација система за одређивање семантичке сличности

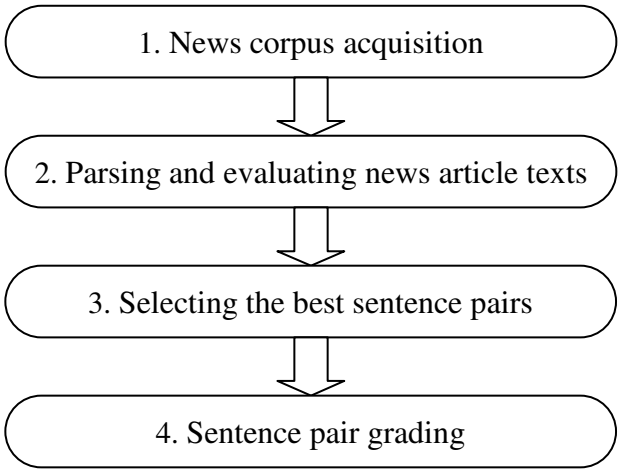
Приликом евалуације STSS неопходно је одредити вредности параметара рада система за које он постиже максималну прецизност, при чему се под прецизношћу подразумева степен поклапања оцена система са оценама сличности које би дао човек. У ту сврху неопходно је одредити оптималан праг за оцене сличности које враћа систем, тј. одредити оптималну вредност између 0 и 1 која би представљала границу, тако да се све оцене изнад ње могу третирати као процена семантичке сличности, а све оцене испод ње као процена семантичке различитости. Да би се овај праг одредио, потребан је довољно велики скуп парова реченица. Реченице које сачињавају пар потребно је да се преклапају у свом семантичком садржају, тако да неке од њих представљају стварне парафразе, док друге треба само у одређеној мери да буду семантички сличне (али нису парафразе). У сваком случају, колекција оваквих парова реченица се обично назива корпусом парафраза. Након што се овакав корпус парова реченица креира, портебно је ручно оценити их бинарним оценама како би се помоћу статистичке анализе, поређењем ручно додељених и машински одређених оцена, одредила вредност прага за коју систем постиже оптималну прецизност.

Корпус парафраза

За евалуацију предложеног приступа за енглески језик употребљен је MSRPC корпус (*Microsoft Research Paraphrase Corpus*), највећи корпус парафраза за енглески језик који се састоји од 5801 пара реченица [40]. Сваки пар реченица је оцењен, тј. дата је оцена семантичке сличности од стране двоје судија. При том, додељиване оцене су бинарне, где оцена један указује да су реченице сличне, а нула обратно. Случајеве у којима је дошло до неслагања додељених оцена решавао

је трећи судија. Од укупно 5801 пара реченица, њих 3900 (67%), је проглашено семантички сличним, док је остатак представљао семантички различите парове. Коначно, корелација у оценама између троје судија износи 83%, што представља и горњу границу прецизности коју систем може постићи над овим корпусом.

Као полазна тачка за креирање корпуса парафраза на српском језику сагледана су искуства и резултати истраживања приликом реализације MSRPC корпуса. Основни приступ за изградњу корпуса заснива се на проналажењу више новинских извештаја који се баве истом вешћу. По новинарској конвенцији, прва реченица извештаја или прве две реченице извештаја обично представљају сумаризацију садржаја вести, па су ове реченице, добијене из различитих извештаја о истој вести, добри кандидати за постојање парафразног односа. Ток процеса стварања корпуса парафраза на српском језику је приказан на слици 24. У наставку је описан начин на који је овај корпус састављен и обрађен пре него што је стављен у употребу.



Слика 24. Ток процеса реализације корпуса парафраза на српском језику.

1. Прибављање корпуса вести (*News corpus acquisition*) - Главни захтев при изградњи корпуса вести односио се на постојање бесплатно доступне и сређене веб архиве вести, која омогућава лак приступ важним вестима за жељени датум. Као једно од најбљих решења показао се сајт www.vesti.rs. Овај сајт представља агрегатор вести који обједињује вести свих већих медијских кућа у Србији, како телевизијских станица и штампаних медија, тако и многих Интернет магазина и портала. Укупно се користи преко 210 различитих извора вести. Одлучено је да се за скупљање парафраза користе само најважније вести за сваки датум, јер је за њих највећа вероватноћа постојања извештаја из више извора. Сем тога, на тај начин је могуће прибавити парове реченица који се тичу разних области живота и разних места и актера дешавања, чиме се спречава опасност од фокусираности корпуса парафраза на неку специфичну тему. Да би се обезбедило довољно материјала за изградњу корпуса, обрађене су најважније вести из целе 2010. године и из првих седам месеци 2011. године.
2. Обрада и евалуација текстова чланака (*Parsing and evaluating news article texts*) - Након што се добије текст свих извештаја једне вести, потребно је тај текст пречистити, поделити на реченице и проценити квалитет тих реченица. Парсирање и чишћење текстова вести је врло проблематичан задатак због потпуно слободног формата у коме се текстови вести појављују, употребе тачке на местима која не представљају крај реченице и разних варијанти непотребних информација које треба уклонити. Такве су, на пример, информације о месту дешавања, времену дешавања, извору вести, и сл. Сваком пару реченица који испуњава одређене минималне критеријуме у погледу дужине реченица и броја семантички битних речи у њима, придружују се атрибути који описују тај пар и који се користе приликом одређивања најбољег тј. најквалитетнијег пара реченица за сваки чланак. Ти атрибути су: број дугачких речи у краћој реченици, број дугачких речи у дужој реченици и број дугачких речи које се јављају у обе реченице, без узимања у обзир понављања речи. Под дугачким речима подразумевају се речи од бар шест слова, тј. оне речи за које је извесно да

су семантички релевантне. Минимална дужина за коју се реч сматра семантички релевантном зависи, наравно, од језика који је у питању. На пример, у српском језику постоји много предлога и присвојних заменица који су дужине пет слова. Стога је закључено да у овом случају треба поставити минималну дужину на шест слова.

3. Одабир најбољих парова реченица (*Selecting the best sentence pairs*) - Од свих парова реченица придружених једном чланку, потребно је одредити један за који је највероватније да је најквалитетнији. Овај процес селекције захтева израчунавање нумеричке оцене квалитета за сваки пар на основу придружених атрибута. За најбољи или најквалитетнији пар може се сматрати онај за који је вероватно да његове реченице садрже исту семантичку информацију, и то речену на заиста другачији начин. Парови реченица који се могу сматрати лошим кандидатима су: (а) оне које су очигледно семантички различите реченице, или (б) оне које јесу семантички исте, али само због велике лексичке сличности међу њима. Пример квалитетног пара реченица је већ приказан приликом илустрације извршавања алгоритма на слици 22. Пример два пара реченица лошег квалитета за случајеве (а) и (б) дата су у овом редоследу на слици 25. Дакле, главни циљ при изградњи корпуса парафраза је постићи довољно велики проценат заступљености заиста семантички истих парова реченица, избегавајући при томе колико је год могуће просте примере сличности који произлазе из лексичког поклапања.

2.1 Građani Južnog Sudana masovno glasali tokom prvog dana jednonedeljnog referenduma o nezavisnosti te oblasti od vlasti u Kartumu.

2.2 Dan uoči referenduma o nezavisnosti južnog Sudana, njegov lider Salva Kir rekao je da ne postoji alternativa mirnoj koegzistenciji između severa i juga.

3.1 Grčka je otkazala naručenih 12,3 miliona doza vakcine protiv novog gripa i traži vraćanje uplaćenog avansa.

3.2 Grčke vlasti su otkazale naručenih 12,3 miliona doza vakcine protiv novog gripa A (H1N1) i zatražile vraćanje uplaćenog avansa.

Слика 25. Пример два пара реченица лошег квалитета: случај (а) представља пар 2.1 и 2.2, а (б) 3.1 и 3.2

Уочено је да је најбоље фаворизовати реченице сличне дужине које имају око 50% истих речи. Наиме, код парова реченица доста различитих дужина смањује се вероватноћа да су заиста у питању парафразе. Показује се да када је проценат сличних речи висок, тада су највероватније у питању две исте реченице од којих је једна проширена неком семантички небитном информацијом. С друге стране, када је проценат истих речи низак, тада је највероватније да је у питању неки семантички различит део исте вести. Поред тога, додатна тежина се даје кратким паровима реченица, јер код њих свака реч носи пропорционално већу тежину у креирању коначне оцене.

4. Оцењивање парова реченица (*Sentence pair grading*) је извршено ручно. За неке парове реченица непосредно је било јасно коју оцену треба доделити, али такође постојали су и парови чије семантичке информације су донекле биле сличне, али не у потпуности. С обзиром да је било пуно оваквих појава, било је неопходно утврдити одређене смернице за оцењивање, који ће обезбедити највећу могућу уједначеност критеријума оцењивања. Ове смернице су приказане у облику псеудокода на слици 7. Такође, приликом ручног оцењивања спровођено је и исправљање словних и других грешака које су последица несавршености изворног текста. Оцењивање је извршено од стране једног човека - судије, након чега је други судија оценио део насумично одабраних парова реченица, који је износио 30% парова целокупног корпуса. Ова двострука провера је учињена у циљу процене слагања додељених оцена између двојце судија, која је важан параметар, јер диктира горњу границу тачности система. За креирани корпус проценат слагања додељених оцена између двојце судија је био 78,27%.

Коначно, користећи претходно наведену процедуру добијен је корпус од 1194 пара реченица. Од тога, 553 пара су оцењена као семантички подударна, а 641 пар као семантички различит. Процентуално, семантички подударних парова има 46,31%, а семантички различитих 53,69%. Овај корпус је назван SRPC (*Serbian Paraphrase Corpus*).

```
if semantic contents of sentences are completely different then
   assign grade 0;
else begin
   remove from consideration differences arising from the use of pronominal and noun
   phrase anaphora;
   if it is unclear whether sentences refer to the same event then
      assign grade 0;
   else if the sentences have the same subject matter but employ different rhetorical
   structures then
      assign grade 0;
   else if the sentences have the same subject matter but emphasize different aspects
   of it then
      assign grade 0;
   else begin
      if one sentence represents a semantic subset of the other then
      begin
         extract the information present only in the semantically richer sentence;
         if that information is not particularly important then
            assign grade 1;
         else
            assign grade 0;
      end;
      else begin
         compare the subjects, predicates and other important semantic features of both
         sentences;
         if there is a semantic discrepancy then
            assign grade 0;
         else
            assign grade 1;
      end;
   end;
end;
```

Слика 26. Смернице за оцењивање пара реченица: Ове смернице су неопходне како би се осигурала општа униформност критеријума оцењивања.

Евалуација

Одређивање вредности прага врши се прво на већем скупу података тј. парова реченица, који се назива скупом података за тренирање (*training data set*). Оптимална вредност прага је она за коју систем достиже максималну могућу тачност. Тако добијена вредност прага се проверава на независном скупу података за тестирање (*test data set*) како би се одредиле коначне перформансе система. Скуп података за тренирање чини 70% укупног корпуса, што за корпус парафраза на српском – SRPC чини 835 парова реченица, док за MSRPC износи 4076. Тест скуп износи преосталих 30% укупног корпуса, што је 359 парова за SRPC и 1725 за MSRPC. Код предложеног LInSTSS приступа заа одређивање тежине сваког

идентификованог концепта из текста уоптребљене су вредности добијене помоћу TF_{norm} (3) метрике приликом стварања семантичких простора за српски и енглески (фаза *постпроцесирање корпуса* из предходног поглавља).

Током евалуације над SRPC корпусом, употребом COALS и RI алгоритама и низа вредности за праг, добијени су различити резултати тачности система који су приказани у табелама 10 и 11. Највећа вредност означена је подебљаним писмом. За одређивање оптималног прага посматран је параметар тачност (*Accuracy*) који представља однос броја правилно оцењених парова и укупног броја парова реченица у корпусу над којим се врши евалуација. Гранична вредност је повећавана у корацима од 0,001, како би се постигла максимална могућа тачност.

Израз „стварно позитивни“ (*True Positives* – TP) односи се на оне парове реченица који представљају парафразе и правилно су означени као такви од стране алгоритма. „Стварно негативни“ (*True Negatives* – TN) представљају несличне парове реченица које је алгоритам правилно препознао. „Лажно позитивни“ (*False Positives* – FP) чине парови семантички различитих реченичнца, који су погрешно означени као парафразе. Коначно „лажно негативни“ (*False Negatives* – FN) су парови реченица који јесу парафразе, али су погрешно оцењени као семантички различити. На основу ових вредности могуће је одредити прецизност (*precision* – P), осетљивост (*recall* – R) и Ф-меру (F). Ове мере интензивно се користе у теорији претраживања информација и рачунају се на следећи начин:

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F = \frac{2PR}{P + R}$$

У контексту STSS, прецизност се може схватити као однос броја правилно идентификованих парова парафраза и укупног броја парова означених парафразама од стране алгоритма. Осетљивост представља однос између тачно идентификованих парова парафраза и стваног броја парова парафраза у корпусу. Ф-мера се рачуна као хармонијска средина прецизности и осетљивости.

ТАБЕЛА 10

AN OVERVIEW OF SENTENCE PAIR SCORES GAINED BY USING THE COALS ALGORITHM ON SRPC

Threshold	Sentences correctly identified by the COALS algorithm					
	Semantically equivalent $\frac{TP}{TP+FN}$		Semantically diverse $\frac{TN}{TN+FP}$		Overall $\frac{TP+TN}{TP+FP+TN+FN}$	
	Training set	Test set	Training set	Test set	Training set	Test set
0.1	100%	100%	0%	0%	46.23%	46.52%
0.2	99.48%	100%	2.67%	2.6%	47.43%	47.91%
0.3	96.11%	97.01%	33.63%	38.02%	62.51%	65.46%
0.4	72.02%	71.86%	75.28%	77.6%	73.77%	74.93%
0.407	70.73%	71.26%	77.28%	81.25%	**74.25%**	**76.6%**
0.5	36.79%	35.93%	96.66%	96.88%	68.38%	68.52%
0.6	8.03%	8.98%	99.78%	100%	57.37%	57.66%
0.7	0.52%	0%	100%	100%	54.01%	53.48%
0.8	0%	0%	100%	100%	53.77%	53.48%
0.9	0%	0%	100%	100%	53.77%	53.48%

ТАБЕЛА 11

AN OVERVIEW OF SENTENCE PAIR SCORES GAINED BY USING THE RI ALGORITHM ON SRPC

Threshold	Sentences correctly identified by the RI algorithm					
	Semantically equivalent $\frac{TP}{TP+FN}$		Semantically diverse $\frac{TN}{TN+FP}$		Overall $\frac{TP+TN}{TP+FP+TN+FN}$	
	Training set	Test set	Training set	Test set	Training set	Test set
0.1	100%	100%	0%	0%	46.23%	46.52%
0.2	99.48%	100%	1.56%	1.56%	46.83%	47.35%
0.3	97.41%	97.6%	30.29%	33.85%	61.32%	63.51%
0.4	73.58%	73.65%	72.38%	75.52%	72.93%	74.65%
0.417	69.95%	69.46%	77.95%	82.81%	**74.25%**	**76.6%**
0.5	38.08%	37.72%	95.99%	96.35%	69.22%	69.08%
0.6	8.55%	9.58%	99.78%	100%	57.6%	57.94%
0.7	0.52%	0%	100%	100%	54.01%	53.48%
0.8	0%	0%	100%	100%	53.77%	53.48%
0.9	0%	0%	100%	100%	53.77%	53.48%

Највећи проценат исправно идентификованих парова употребом COALS алгоритма над SRPC корпусом се добија за вредност прага од 0,407, што доводи до тачности система од 76,6% као што је приказано у Табели 10. С друге стране, резултати постигнути током оцењивања SRPC корпуса употребом RI алгоритма и низа вредности за праг приказани су у Табели 11. Оптимална вредност прага је 0,417, што доводи до исте тачности система од 76,6%.

COALS и RI су достигли сличан ниво прецизности над SRPC. Међутим, мерење времена извршавања приликом креирања семантичког простора за српски

је показало да је COALS доследно бржи. Ово је вероватно последица релативно мале величине текстуалног корпуса коришћеног за креирање семантичког простора, што је онемогућило RI алгоритам првенствено осмишљен за велике корпусе, да покаже своје предности у том погледу, као што је то био случај приликом креирања семантичког простора за енглески.

На исти начин извршено је одређивање прага над MSRPC корпусом. Како би се извршило поређење, како над SRPC тако и над MSRPC, реализовани су кораци предложеног I&I-STSS алгоритма из [22], при чему за одређивање семантичке сличности пара речи употребљени су COALS и RI. Најбољи резултати над оба корпуса, за ову импелментацију I&I-STSS алгоритма, добијени су употребом резултата COALS алгоритма, стога овај приступ је назван полазним (*baseline*), а добијени резултати су узети као полазиште за поређење са предложеним приступом.

Табела 12 приказује поређење основних карактеристика неких претходно описаних метода, као и предложеног приступа. Прва три реда (1-3) приказују резултате преузете из радова Mihalcea са сарадницима [21], Islam и Inkpen [22] и Li са сарадницима [39]. Ове вредности добијене су над MSRPC корпусом. Такође, наредна четири реда (4-7) приказују резултате добијене над MSRPC корпусом, за полазни приступ и предложени LInSTSS. Ред 4 (*Baseline (MSRPC)*) садржи резултате у којима су изостављене стоп речи из разматрања, као што је то наведено у раду [22], док ред 6 (*Baseline' (MSRPC)*) садржи резултате у којима су и стоп речи узете у разматрање. На исти начин је евалуиран и предложени LInSTSS приступ, где ред 5 (*LInSTSS (MSRPC)*) садржи резултате без стоп речи, а ред 7 (*LInSTSS' (MSRPC)*) резултате у којима су укључене и стоп речи.

На основу поређења вредности редова 2 и 4 види се да су добијени резултати полазног алгоритма (имплементације I&I-STSS алгоритма, где за стварање семантичког простора је употребљен COALS алгоритам над корпусом апстракта чланака Википедије) се добијају нешто лошији резултати него они наведени у раду [22]. Могући разлог за добијање различитих резултата може се

пронаћи у употреби различитих алгоритама и корпуса за креирање семантичког простора (референцирани приступ је евалуиран употребом знатно богатијег корпуса - British National Corpus (BNC) који садржи 100 милиона речи, величине 5,2 GB у односу на корпус апстракта чланака Википедије од 1,4 GB)

У оба случаја, са и без узимања у обзир стоп речи, полазни приступ даје боље резултате у односу на предложени. Разлог за ово могу бити додељене тежине које су добијене помоћу TF_{norm} (3) метрике. Ова метрика узима у обзир само фреквенцију појављивања термина, али не и његову заступљеност у документима у односу на целу колекцију документа (инверзну документ фреквенцију), па неким терминима који су често јављају, али у свега пар докумената, може бити погрешно додељена ниска тежина. С обзиром да употребљени корпус садржи апстракте чланака Википедије који су за енглески језик знатно дужи и богатији речима у односу на онај употребљен за српски, овај проблем може бити изражен.

На крају, на основу поређења добијених резултата у којима су стоп речи узете у разматрање и оних где нису, може се закључити да иако стоп речи представљају речи са малим информационим садржајем, њихов садржај није занемарљив приликом одређивања семантичке сличности, па избацивање може погоршати резултате.

ТАБЕЛА 12

A COMPARISON OF THE CHARACTERISTICS OF VARIOUS STSS METHODS

	Method	Optimal threshold	Accuracy	Precision	Recall	F-Measure
1.	Mihalcea et al. (MSRPC)	0.5	70.3%	69.6%	97.7%	81.3%
2.	Islam and Inkpen (MSRPC)	0.6	**72.64%**	**74.65%**	89.13%	81.25%
3.	Li et al. (MSRPC)	0.4	70.8%	70.3%	**97.4%**	**81.6%**
4.	Baseline (MSRPC)	0.626	70.32%	72.92%	88.05%	79.77%
5.	LInSTSS - COALS (MSRPC)	0.337	69.45%	71.95%	88.58%	79.41%
6.	Baseline' (MSRPC)	0.619	**72.69%**	**73.97%**	90.93%	**81.58%**
7.	LInSTSS' - COALS (MSRPC)	0.337	71.53%	72.37%	**92.5%**	81.21%
8.	Baseline (SRPC)	0.599	76.04%	**82.93%**	61.08%	70.34%
9.	LInSTSS - COALS (SRPC)	0.407	**76.6%**	76.77%	**71.26%**	**73.91%**
10.	LInSTSS - RI (SRPC)	0.417	76.6%	77.85%	69.46%	73.42%

Последња три реда (8-10) у Табели 12. приказују резултате евалуације над SRPC. Резултати за полазни алгоритам су приказани у реду 8, а редови 9 и 10 садрже резултате предложеног LInSTSS приступа, у обе варијанте COALS и RI. Оптимална вредност прага над овим корпусом за LInSTSS уптребом COALS и RI гравитирала је ка 0,4. Такође, имплементација полазног I&I-STSS алгоритма, има скоро исту оптималну вредност прага као и код оригиналног приступа одређеног над MSRPC.

Над SRPC корпусом предложени LInSTSS приступ доводи до највеће тачности, која је незнатно већа од полазне импелментације I&I-STSS алгоритма (*baseline*), и која је само неколико процената мања од максималне могуће тачност система, имајући у виду проценат слагања додељених оцена између судија (78,27%). Овај приступ такође има знатно већу осетљивост, по цену ниже прецизности. Разлика између ове две мере је смањена у односу на полазни метод (*baseline*), па побољшава вредности укупне Ф-мере. Такође, пошто предложени приступ узима у обзир специфичност речи како би се другачије одмерила сличност парова речи које се пореде, предложени приступ донекле строжије додељује оцену семантичке сличности, што се може приметити поређењем вредности, у близини оптималног прага, у колонама семантички еквивалентни (*Semantically equivalent*) и семантички различити (*Semantically diverse*) у табелама 10. и 11.

Као закључак, употреба тежина додељених речима може побољшати резултате. Стога, на основу евалуације над корпусима парафраза извршено је фино подешавање параметара, а такође стечена искуства употребљена су за реализацију модула за одређивање семантичке сличности између питања и корисничког профила.

Евалуација целокупног система

У наставку је дат преглед и анализа доступних веб портала чији подаци се могу употребити за формирање корпуса питања и одговора. На основу анализе одабран је један веб портал помоћу кога је формиран корпус, над којим је затим извршена евалуација целокупног система.

Корпус питања и одговора

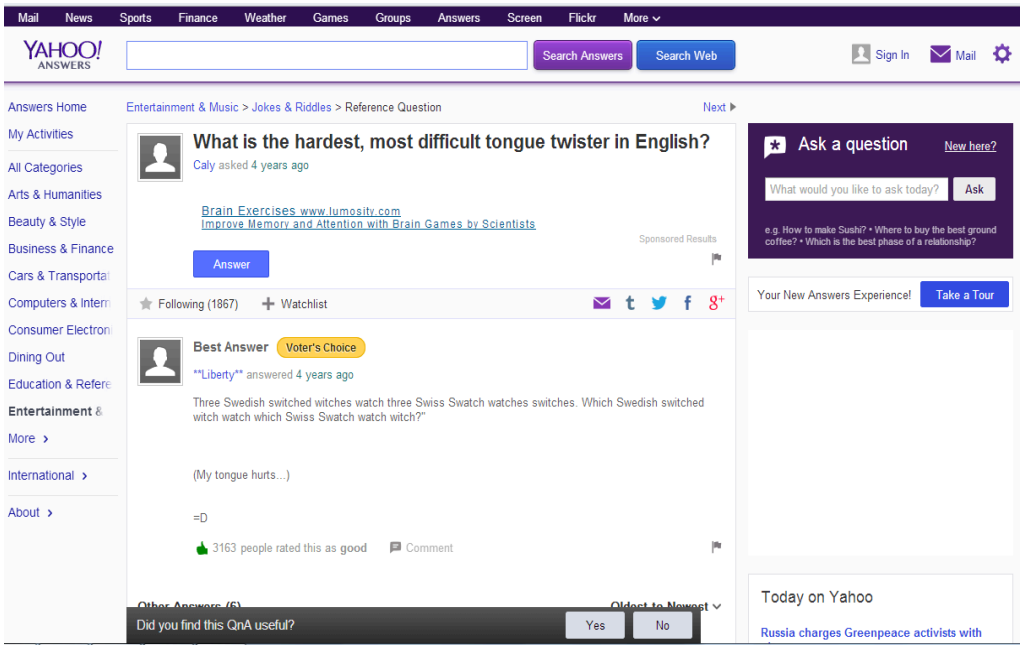
Постоје бројни веб сајтови, форуми и портали за питања и одговоре (*Question Answering* – QA), који се још називају и друштвеним заједницама за питања и одговоре (*Community Question Answering* – CQA). Међу сајтовима овог типа најпознатији су:

1) Yahoo! Answers, http://answers.yahoo.com,
2) StackOverflow, http://stackoverflow.com,
3) AskVille, http://askville.amazon.com,
4) Mahalo, http://mahalo.com,
5) Quora, http://quora.com,
6) AllExperts, http://allexperts.com,
7) Ask.com, http://ask.com,
8) Answers.com, http://answers.com.

Последњих шест наведених сајтова (3-8) делимично или у потпуности не дозвољавају јавни приступ подацима или неким функцијама, стога даље нису разматрани. Сајт StackOverflow карактерише велики број корисника, као и јавни

приступ подацима, али за дате податке (питања и одговоре) нису доступне категорије, и оно што је још важније, целокупан сајт је оријентисан ка једном домену (програмерским проблемима). Такође, и на овом порталу неке функције нису јавно доступне, и то углавном оне које се односе на оцене постављених питања или датих одговора. Стога је одлучено да се за евалуацију одабере Yahoo!Answers портал, с обзиром да садржи широк спектар различитих категорија питања и да у свакој категорији постоји довољан број активних корисника.

Yahoo! Answers представља QA портал на коме људи могу поставити питања и дати одговоре који су јавно доступни сваком веб кориснику. На слици 27. дат је пример једног питања и придруженог одговора са овог портала. За изградњу корпуса за евалуацију, употребљен је конкретно скуп података Yahoo! Answers Webscope L6 dataset [48] скраћено назван L6. Овај скуп података прикупљен је са Yahoo! Answers портала током 2007. године и обухвата сва питања (њих 4.483.032) и њихове придружене одговоре. Поред ових података укључени су и неки анонимни метаподаци (нпр. идентификатор корисника), па је могуће повезати одређено питање или одговор са корисником који га је поставио, тј. његовим власником.



Слика 27. Пример питања и одговора доступних на сајту Yahoo! Answers

Питања и одговори представљају инстанце типа поруке (*post*), тј. порука може представљати питање или одговор у зависности од придруженог типа. С друге стране, једно питање са свим расположивим одговорима формира нит. Свака инстанца поруке састоји се од текстуалног тела поруке, избране главне категорије, основне категорије и поткатегорије (Cat1,2,3), као и идентификатора власника, тј. корисника који је написао поруку. Питања додатно садрже наслов и идентификатор одговора који је одабран као најбољи. Код порука типа одговор, власник је познат само у случају да је овај одговор изабран као најбољи. Ово последње намеће ограничење да само за корисника који је пружио најбољи одговор се могу одредити издвојени концепти. Другим речима, код одговора, власник је познат само у случају да је тај одговор изабран за најбољи, док за остале одговоре њихов власник није познат. Стога, на овај начин могуће је креирати профил корисника на основу постављених питања и датих одговора који су проглашени за најбољи одговор.

Приликом профилисања корисника, поред концепата добијених употребом евалуираних приступа (CE, SemNet, TF-IDF), придодати су и концепти који представљају додељене категорије (Cat1,2,3). С обзиром да ове категорије директно поставља корисник, што представља најпрецизнију информацију, овим концептима је додељена највећа тежина 1. Такође, у зависности од типа поруке из које су екстраховани концепти, разликују се три врсте извора информација, па самим тим и три врсте добијених концепата: питање (*question*), одговор (*answer*) и нит (*thread*). Коначно, у L6 скупу података кориснички налози су потпуно анонимни, па су моделовани помоћу доступног јединственог идентификационог броја.

Из целокупног L6 скупа података издвојена су три типа базе података:

- Тип 1: Овај скуп података моделује интересовање, јер садржи кориснике који су поставили најмање десет питања и свако од ових питања мора имати најмање пет одговора.

- Тип 2: Како би се моделовало знање, издвојени су корисници који су најбоље одговорили најмање на десет питања. Опет, свако питање треба да има најмање пет одговора.

- Тип 3: Како би заједно било заступљено и знање и интересовање, издвојени су корисници који су поставили најмање пет питања и најбоље одговорили на најмање пет питања. Такође, свако од ових питања, за које је корисник повезан, било директно за питање или за најбољи одговор, треба да има најмање пет одговора.

Услов да се разматрају питања која имају најмање пет одговора обезбеђује одређени ниво квалитета питања (нпр. питање није тривијално и привлачи пажњу осталих корисника). Сваки тип базе података садржи 100 корисника који су јединствени над целим L6 скупом података и не налазе се у преостала два типа. Пошто постоји много различитих категорија у оригиналном скупу података, одабрано је пет репрезентативних из којих је издвојено по 100 корисника за сваки тип. Расподела корисника по категоријама приказана је у Табели 13. Такође, за сваког корисника издвојено је још једно додатно питање за које је тај корисник дао одговор који је изабран за најбољи одговор на ово питање. Ово питање је затим употребљено за евалуацију предложеног система, као што је објашњено у наредној секцији.

ТАБЕЛА 13

DISTRIBUTION OF POST CATEGORIES IN EACH DATABASE TYPE

Category	Number of selected users
Society & Culture	35
Food & Drink	35
Computers & Internet	15
Travel	10
Cars & Transportation	5
Total	100

Евалуација

Adamic са сарадницима [35] у својој студији анализирала је питања из различитих категорија постављена на Yahoo! Answers порталу и донела закључак да одговори изабрани као најбољи одговори (*best answers*) углавном заиста и јесу најбољи одговори на дато питање. Стога информација да је корисник пружио одговор на дато питање који је затим одабран за најбољи може се узети за позитиван пример, тј. може послужити као основна истина (*ground truth*) за евалуацију система. Другим речима, приликом евалуације предложеног система потребно је проценити у којој мери ће систем за дато питање препознати, тј. високо рангирати овог корисника. С друге стране, имајући у виду претходно наведено ограничење расположивог корпуса, да је власник одговора познат само ако је тај одговор изабран за најбољи, у извесној мери своди проблем евалуације предложеног система на проблем евалуације класификације једне класе (*one-class-classification*). Другим речима, ограничење представља постојање само позитивних примера, тј. познат је корисник који је најбоље одговорио на питање, док су негативни примери непознати, тј. није познато који корисници нису у стању да дају најбољи одговор. Стога, ово ограничење је уједно одредило и метрике које су употребљене за евалуацију система.

Евалуација система је извршена на следећи начин:

i. Прво је за сваког корисника, у зависности од типа базе података којој припада, начињен његов профил компетентности на основу постављених питања и датих одговора помоћу једног од предложених приступа. Затим је за сваког од ових корисника издвојено још једно, додатно питање, за које је корисник дао одговор, при чему је тај одговор изабран за најбољи одговор на ово питање. Наравно, додатно питање и придружени одговори нису узети у обзир приликом креирања профила за датог корисника.

ii. Додатно питање је затим анализирано, тј. екстраховани су сви концепти из питања помоћу једног од предложених приступа. Затим је за ово питање, помоћу једног од предложених алгоритама, одређена сличност ка свим

корисничким профилима из базе података датог типа. На крају је извршено рангирање свих корисника на основу добијене сличности за свако додатно питање (којих укупно има 100 с обзиром да има и 100 корисника унутар једне базе).

iii. Коначно, с обзиром да је потребно проценити у којој мери ће систем за дато додатно питање препознати, тј. високо рангирати корисника који је пружио најбољи одговор, за евалуацију перформанси система употребљене су две широко примењиване метрике, просечни реципрочни ранг (*Mean Reciprocal Rank* – MRR) и прецизност на *N* (*Precision@N* – P@N).

Ове метрике су дефинисане на следећи начин:

MRR је метрика преузета из домена претраживање информација (*Information Retrieval*) у којој се листа могућих одговора на упит (питање) сортира на основу вероватноће њихове исправности. Формула по којој се рачуна MRR (14) дефинисана је као аритметичка средња вредност реципрочног ранга за скуп питања Q, где за свако питање q из скупа података можемо одредити сличност $S(q,p)$ ка свим расположивим профилима ($p \in P$) и сходно томе рангирати кориснике. Затим на основу тога је могуће израчунати ранг корисник који је дао најбољи одговор ($rank_i$).

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \qquad (14)$$

Прецизност такође потиче из домена претраживања информација и представља проценат преузетих докумената (у овом случају корисника) који су релевантни за упит (у овом случају питање). Овде се за релевантног корисника узима онај корисник који је пружио најбољи одговор на дато питање. P@N је стога прецизност која се процењује за дату граничну вредност ранга *N*, тј. након рангирања свих корисника на основу сличности $S(q,p)$, у обзир се узима само првих *N* резултата, међу којима се тражи релевантан корисник. Другим речима, за скуп питања Q, мери се проценат исправно одговорених питања међу првих *N*

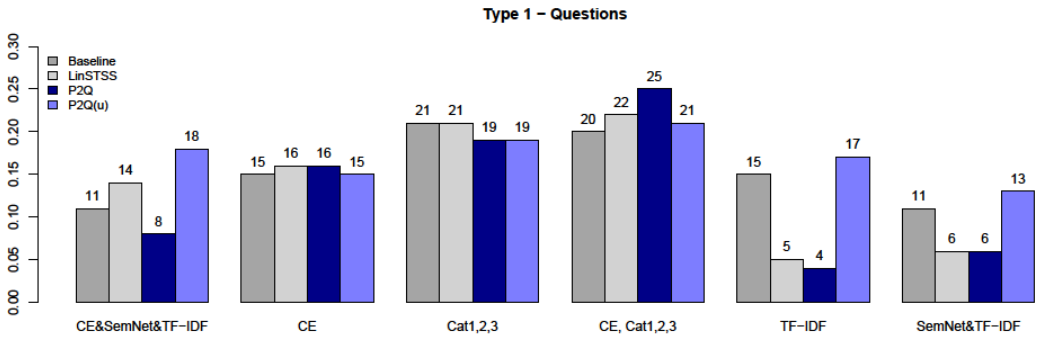
најбоље рангираних корисника или интуитивно колика је вероватноћа да ће се за дато питање добити тачан одговор уколико се пошаље ка првих *N* најбоље рангираних корисника.

Пошто база података типа 1 садржи само питања постављена од стране корисника који се користе за евалуацију, из ове базе података издвојени су само концепти типа питање. За типове 2 и 3 издвојени су концепти типа одговора (који се односе на најбољи одговор) и концепти типа нити (који се односе на питање са свим одговорима). Резултати евалуације за сва три типа базе података дати су у Прилогу Б. У табелама 15. и 16. приказани су добијени MRR резултати, а у табелама 17.-21. резултати P@N.

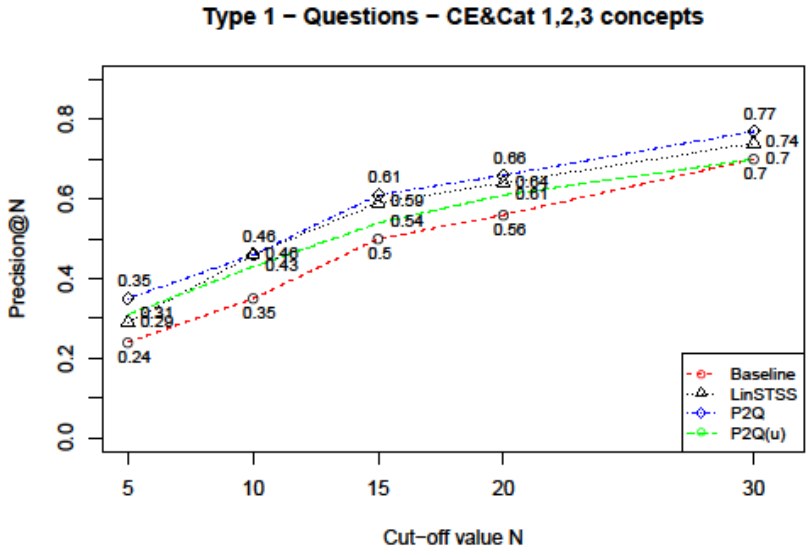
Графички приказ MRR резултата за сва три типа базе података дати су на сликама 28-33, изузев слике 29. Baseline означава полазни приступ, односно имплементацију I&I-STSS приступа, а затим следе предложени приступи, LinSTSS и P2Q, и на крају су приказани резултати P2Q-unweighted, означени са P2Q(u), који представља верзију P2Q алгоритма у којој се тежине концепата не узимају у обзир. Треба такође напоменути да за све концепте који представљају категорије (Cat1,2,3) додељена тежина има исту вредност 1, с обзиром да су ти концепти непосредно добијени од корисника, па према њиховом значају додељена им је највећа тежина. Стога, за ове концепте тежине немају значаја, па полазни и LinSTSS дају исте резултате, као и P2Q и P2Q(u).

За тип 1 најбољи MRR резултати добијени су коришћењем P2Q приступа над комбинованим концептима из CE и категорија (Cat1,2,3). Ово показује да комбиновањем концепата идентификованих од стране корисника и оних добијених семантичком анализом, тј. употребом семантичке експанзије упита, се могу побољшати резултати. TF-IDF и SemNet дају слабије резултате, као и приликом комбиновања са другим изворима. Такође, интересантно је да приступи који не узимају у обзир тежине (Baseline и P2Q(u)), евалуирани над овим концептима су показали знатно боље резултате од њихових верзија које узимају у обзир тежине (LinSTSS и P2Q). Разлог зашто увођење тежина помоћу TF-IDF и

SemNet погоршава резултате може се наћи у чињеници да неке идентификоване речи које су ретке, па самим тим имају и већу тежину, не морају бити од велике важности за то питање. Такође, за семантичку експанзију упита, SemNet узима у обзир само једну реч, а не цео текст поруке као CE, чиме се уводи већи степен грешке. P@N резултати за ову базу података за концепте добијене комбиновањем категорија и CE приказани су на слици 29.



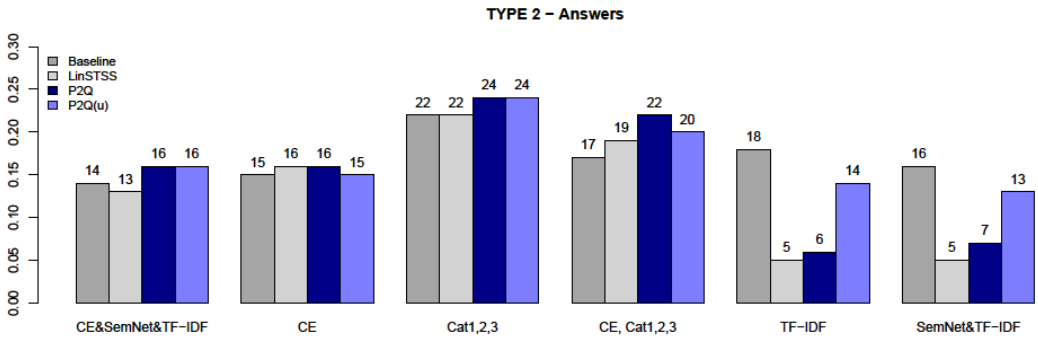
Слика 28. Графички приказ MRR резултата за тип 1 базе података.



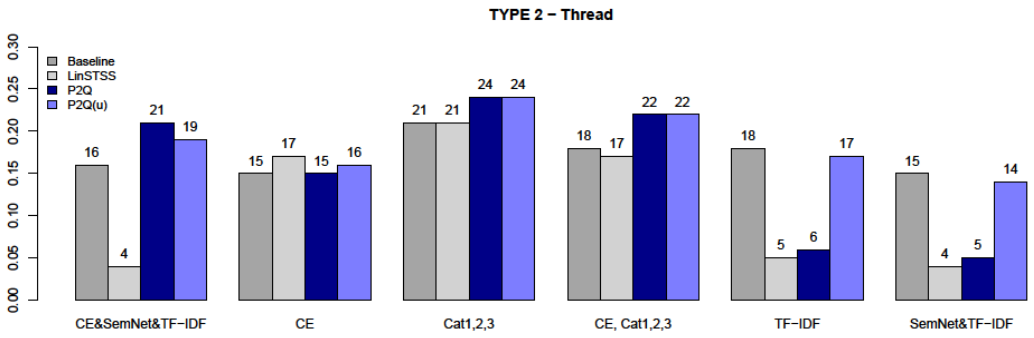
Слика 29. Графички приказ P@N резултата за тип базе података 1 за концепте добијене комбиновањем категорија и CE

Графички приказ MRR резултата за концепте типа одговор и нити, над базом типа 2 дат је на сликама 30 и 31 респективно. За овај тип базе података,

подједнако и за концепте типа одговор и за оне типа нити, P2Q приступ даје најбоље резултате над концептима из Cat1,2,3. CE сада пружа нешто слабије резултате, чак и у комбинацији са кориснички додељеним категоријама, па су најбољи резултати добијени само над концептима из Cat1,2,3. Разлог за ово може бити то што, у односу на Тип 1, из ове базе су екстраховани концепти типа одговора и нити, што представља знатно дужи текст, па повећава вероватноћу грешке. Такође, само текст одговора може сугерисати другу тему у односу на тему питања, а и семантичка анализа помоћу CE у неким случајевима не може тачно издвојити концепте из текста у одговору, јер овај текст представља само наставак питања.



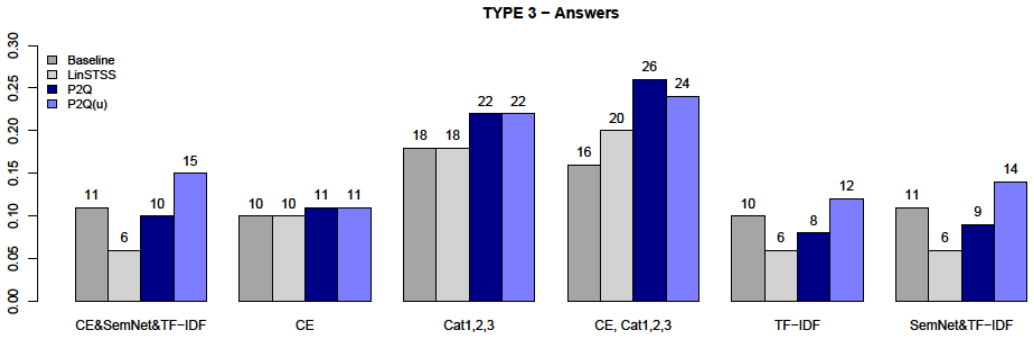
Слика 30. Графички приказ MRR резултата за тип 2 базе података и концепте типа одговора.



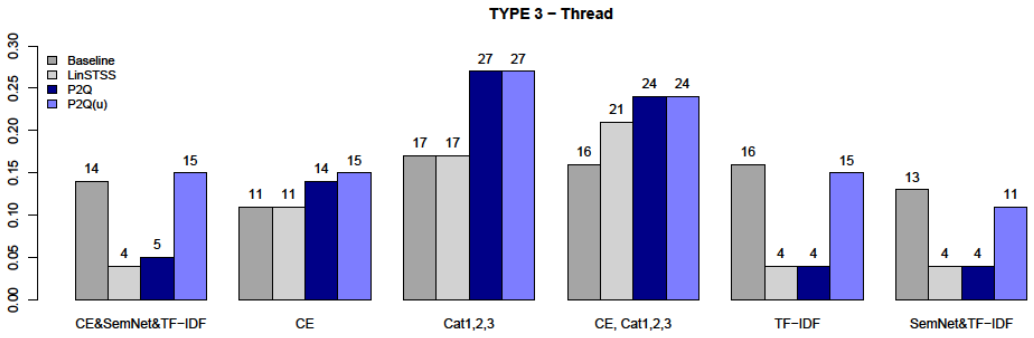
Слика 31. Графички приказ MRR резултата за тип 2 базе података и концепте типа нити.

Графички приказ MRR резултата за концепте типа одговор и за оне типа нити, над типом 3 базе података дат је на сликама 32 и 33 респективно. Код ове базе, за концепте типа одговор P2Q приступ је дао најбоље MRR резултате над

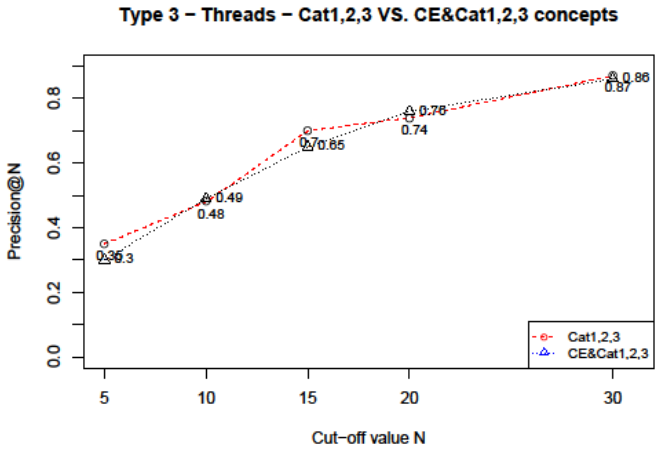
комбинованим концептима добијеним из CE и Cat1,2,3. За концепте типа нити P2Q приступ је опет пружио најбоље резултате, али сада само над концептима из Cat1,2,3. Уједно ови резултати представљају најбоље постигнуте MRR резултате. Такође, за P2Q приступ занимљиво је издвојити P@N резултате за случај да су концепти идентификовани помоћу комбинације CE и Cat1,2,3 и само Cat1,2,3. Ово је приказано на слици 34. са које се може видети да оба начина имају сличан раст и дају врло приближне вредности, што значи да ова разлика настаје поређењем рангирања на нижим позицијама (након 30 места).



Слика 32. Графички приказ MRR резултата за тип 3 базе података и концепте типа одговора.



Слика 33. Графички приказ MRR резултата за тип 3 базе података и концепте типа нити.



Слика 34. Графички приказ P@N резултата за тип базе података 3 и концепте типа нити добијене комбиновањем категорија и CE у односу на само категорије

Упоредном анализом најбољих резултата може се уочити да P2Q константно пружа најбоље резултате чиме је доказана хипотеза *h3*. Стога, даља дискусија резултата ће се односити само на овај алгоритам, осим у случају када је другачије наглашено. Што се тиче тврдње *h1* она важи само у случају да нису доступне кориснички додељене категорије (Cat1,2,3). У супротном, комбинација ових категорија и концепата добијених помоћу CE пружа најбоље резултате само код анализе питања и у случају типа 3 базе података и концепата типа одговора, док код осталих анализа одговора и целих нити најбољи резултати се добијају употребом само Cat1,2,3 концепата. Стога, хипотеза *h1* није потврђена, а разлог за ово вероватно лежи у начину комбиновања тежина додељених концептима, односно у томе што подједнаку важност имају и оне добијене помоћу CE и оне које представљају кориснички дефинисане категорије.

На основу упоредне анализе алгоритама који узимају у обзир тежине (LinSTSS и P2Q) и оних који не узимају (полазни и P2Q(u)), може се закључити да тежине, уколико су правилно додељене, могу побољшати резултате целокупног система, што потврђује *h2*. У случају да нису правилно додељене (нпр. гледајући резултате само добијене помоћу TF-IDF или у комбинацији са SemNet) ове тежине могу знатно погоршати добијене резултате. Стога, за језике са врло ограниченим

електронским лингвистичким ресурсима код којих не постоје језички алати слични алатима Antelope и ConceptNet, као што је српски језик, најбоље је користити полазни приступ над свим идентификованим речима, с обзиром да тежине додељене помоћу TF-IDF нису од користи.

Такође, упоредном анализом најбољих резултата добијених над типом 3 са онима добијеним над типом 2, може се уочити да се бољи резултати добијају над типом 3. Ово потврђује хипотезу *h4*, с обзиром да тип 3 моделује и знање и интересовање у односу на тип 2 који моделује само показано знање, тј. код типа 3 издвојени су корисници који су поставили најмање пет питања и најбоље одговорили на најмање пет питања, док код типа 2 издвојени су корисници који су само најбоље одговорили на најмање десет питања. Стога, може се закључити да за профилисање компетентности корисника да пружи одговор на постављено питање, није важно само размотрити најбоље одговоре за овог корисника, односно профилисати његово знање, већ је такође потребно узети у обзир и питања која је поставио, с обзиром да она могу изразити интересовање.

На крају, у Табели 14. дат је однос између случаја насумично (*rand*) одабраних корисника и најбољих резултата добијених помоћу P2Q. Ова вредност варира од 7,2 пута за P@5 до 3 пута за P@30, где је за P@5 вероватноћа да је корисник који је дао најбољи одговор међу првих 5 рангираних корисника износи 36%, у односу на 5% у случају насумичног одабира, а за P@30 вероватноћа да је овај корисник међу првих 30 рангираних корисника износи 90%, у односу на 30% у случају насумичног одабира.

ТАБЕЛА 14

BEST ACHIEVED P@N SCORES COMPARED WITH RANDOM RESULTS

N	*5*	*10*	*15*	*20*	*30*
rand	0.05	0.10	0.15	0.20	0.30
P2Q	0.36	0.57	0.66	0.78	0.90
Ratio P2Q/rand	7.20	5.70	4.40	3.90	3.00

VI Закључак

Основни циљ ове студије био је изучавање система за интелигентно прослеђивање питања – названих СИПП системима, као и реализација прототипа једног оваквог система. Рад на дисертацији обухватао је следеће научне методе истраживања:

1. систематско проучавање домаће и иностране литературе из области дисертације;
2. развој аналитичког СИПП модела;
3. критичку анализу проблема интелигентног прослеђивања питања написаних на природном језику;
4. евалуацију софтверских система за интелигентно прослеђивање питања са становишта три основне фазе које укључује СИПП процес;
5. реализацију прототипа циљног софтверског система и верификацију полазних хипотеза.

Допринос изложене докторске дисертације је у домену анализе и синтезе једног оваквог софтверског система, који треба да омогући интелигентно прослеђивање питања написаних на природном језику. Као саставни делови дисертације садржани су следећи научни доприноси:

1. Идентификација домена истраживања СИПП система, као и генерални преглед области вештачке интелигенције везане за одговарање на питања написана на природном језику, проналажење експерата и семантичког рутирања упита.
2. Систематизација и класификација постојећих решења и генерализација њихових функционалности са становишта унапред дефинисаног СИПП процеса.
3. Увођење оригиналне презентационе парадигме која генерализује суштину свих расположивих СИПП решења пронађених у отвореној литератури и која омогућава упоредну анализу и евалуацију оваквих система.

4. Формирање методологије пројектовања СИПП система на основу класификације, анализе и евалуације решења која су била примењена, која се примењују или која се могу применити у оквиру ове дисциплине.

5. На основу изведених закључака и уочених проблема дат је предлог и имплементација новог софтверског система који треба да омогући интелигентно прослеђивање питања написаних на природном језику.

6. У оквиру предложеног приступа реализован је нови приступ за обраду питања, који омогућава њихову визуелизацију, што обезбеђује интуитивну представу специфичних односа између концепата, као и њиховог значаја у питању. Такође, овај приступ комбинује потпуно аутоматску обраду текста и ручну корекцију резултата, пружајући кориснику могућност повећања тачности излаза. Истовремено, реализовани модул за обраду текста употрбљен је и за анализу одговора.

7. Анализирани су и дискутовани постојећи приступи за одређивање семантичке сличности два кратка текста, погодини за језике са врло ограниченим електронским лингвистичким ресурсима, где је посебан акценат стављен на српски језик.

8. На основу донетих закључака предложен је нови алгоритам, назван LInSTSS, који приликом одређивања семантичке сличности два кратка текста узима у обзир и специфичности речи које ови текстови садрже. Такође, реализован је корпус парафраза за српски језик над којим је извршена евалуацију. Резултати добијени над овим корпусом показали су да предложени алгоритам пружа боље резултате у односу на постојећа решења.

9. У оквиру реализације фазе прослеђивања питања, дискутоване су специфичности проблема поређења питања и корисничких профила, и предложен је нови алгоритам назван P2Q. Добијени резултати показали су да овај приступ пружа зантно боље резултате у односу на остале евалуиране приступе.

10. Дат је преглед и анализа доступних веб портала чији подаци се могу употребити за формирање корпуса питања и одговора. На основу извршене анализе одабран је један помоћу кога је формиран корпус, а који је затим употребљен за евалуацију целокупног система и тестирање полазних хипотеза.

Резултати и објашњења овог истраживања су од интереса за све оне који желе да уђу у ово ново подручје истраживања, да схвате основне појмове и да употребе предложену методологију за реализацију једног оваквог софтверског система. Такође, аутор се нада да ће ова дисертација допринети покретању нових веб портала овога типа, као и унапређењу размене знања уопштено. Коначно, ауторова жеља је да дисертација буде од користи за будуће генерације студената докторских студија, инжењере, практичаре и истраживаче који имају додира са овом облашћу и који желе да дају свој допринос.

С обзиром да свако одговорено питање, ствара још пуно нових питања, односно сваки решен проблем отвара још пуно нових нерешених проблема, тако и ово истраживање као резултат има неколико нових питања и нерешених проблема. У наставку, наведене су смерница за даљи рад.

1. С обзиром да семантичка анализа текста може побољшати резултате, требало би испитати да ли предложени CE приступ се може употребити за постављање вредности тежина речи како би се одредила семантичка сличност кратких текстова. Такође, требало би испитати да ли разна лингвистичка својства текста се могу употребити и комбиновати у духу машинског учења како би се унапредили резултати.

2. Идентификовани проблеми везани за креирање профила корисничког знања су: (1) коришћење Бајесове вероватноће, (2) проблем новог корисника, и (3) проблем интеграције. Предлог идеје за решење ових проблема дат је у поглављу 3 под називом *Интеграција профила*. Међутим, у реализацији ове идеје наишло се на проблем недоступности потребних података како би се она тестирала. Наиме, сви јавно доступни подаци који су анализирани, у конкретном случају подаци доступни са потрала StackOverflow и Yahoo!

Answers, су анонимни. Стога, није било могуће извршити интеграцију са другим изворима информација, као што су социјалне мреже. Такође, у овим подацима недостају, делом или у потпуности, оцене питања и одговора. Код L6 скупа података ове оцене нису доступне у потпуности, док код StackOverflow потрала недостају само негативне оцене. Ово је и разумљиво с обзиром да јавност негативних оцена добијених од стране других корисника може резултирати осветом, односно негативно оцењен корисник може из освете давати негативне оцене онима који су њему дали овакву оцену. Као резултат овог недостатака, није било могуће извршити евалуацију модела поверења заснованог на Дезерт-Шмарандаш теорији (DSmT) [24]. Неки од начина како се ово може решити:

i. тражити податке од већ постојећих система,
ii. направити нови систем (*startup*).

Друго решење (самосталан систем) је уједно и најбоље, с обзиром да експеримент уживо, на великом скупу стварних корисника, представља реалније окружење за евалуацију у односу на податке добијене од других портала. С друге стране, ово уводи додатне проблеме као што су неопходни ресурси (нпр. сервери, одржавање, итд.), реализацију система у потпуности (предложено решење је само прототип) и последње, и оно најважније је како привући кориснике. Ови проблеми иако нису научни, представљају велики изазов, јер захтевају висок степен креативност и иновативност.

У закључку, пошто питања и сврсисходни одговори чине суштину СИПП система, смернице у овом раду могу се најбоље описати изреком: „*Prudens quaestio dimidium scientiae* - Пола науке је поставити право питање“ Аристотел (384 п.н.е. - 322 п.н.е.). Или прикладније: „Половина одговора је право питање“. Стога, основу ове дисертације представљала су три суштинска питања. Помоћу ова три питања генерализована су сва доступна решења у виду презентационе парадигме. Затим, употребом ове парадигме извршена је анализа и уочени су постојећи проблеми. Коначно, за уочене проблеме предложена су и реализована нова решења, што уједно представља и главни допринос овог рада.

Литература

[1] O. Kolomiyets and M.-F. Moens, “A survey on question answering technology from an information retrieval perspective,” *Inf. Sci. (Ny).*, vol. 181, no. 24, pp. 5412–5434, Dec. 2011.

[2] H. T. Dang, J. Lin, D. Kelly, and C. Hill, “Overview of the TREC 2006 Question Answering Track,” in *TREC*, 2006.

[3] I. Ounis, C. Macdonald, and I. Soboroff, “Overview of the TREC-2008 Blog Track,” in *TREC*, 2008.

[4] T. Lappas, K. Liu, and E. Terzi, “A Survey of Algorithms and Systems for Expert Location in Social Networks,” in in *Social Network Data Analytics*, 1st ed., C. Aggarwal, Ed. US: Springer, 2011, pp. 215–241.

[5] D. Faye, G. Nachouki, and P. Valduriez, “Semantic Query Routing in SenPeer , a P2P Data Management System,” in *Network-Based Information Systems*, 2007, pp. 365–374.

[6] C. Tempich, D.- Karlsruhe, S. Staab, and A. Wranik, “REMINDIN : Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphors,” in *WWW*, 2004.

[7] G. Adomavicius and Y. Kwon, “New Recommendation Techniques for Multicriteria Rating Systems,” *IEEE Intell. Syst.*, vol. 22, no. 3, pp. 48–55, May 2007.

[8] I. Rus and M. Lindvall, “Knowledge management in software engineering,” *IEEE Softw.*, vol. 19, no. 3, pp. 26–38, May 2002.

[9] J. Wyatt, “Management of explicit and tacit knowledge.,” *J. R. Soc. Med.*, vol. 94, pp. 6–9, 2001.

[10] J. Davitz, J. Yu, S. Basu, D. Gutelius, and A, “iLink: search and routing in social networks,” in *WWW*, 2007, pp. 931–940.

[11] M. Qu, G. Qiu, X. He, and C. Zhang, “Probabilistic question recommendation for question answering communities,” in *WWW*, 2009, pp. 1229–1230.

[12] Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao, “Routing Questions to the Right Users in Online Communities,” in *ICDE*, 2009, pp. 700–711.

[13] B. Li and I. King, “Routing questions to appropriate answerers in community question answering services,” in *CIKM*, 2010, pp. 1585–1588.

[14] W. Li, C. Zhang, and S. Hu, “G-Finder: Routing Programming Questions Closer to the Experts,” in *OOPSLA/SPLASH*, 2010, pp. 62–73.

[15] D. Horowitz and S. D. Kamvar, “The anatomy of a large-scale social search engine,” in *WWW*, 2010, pp. 431–441.

[16] X. Si, E. Chang, Z. Gyongyi, and M. Sun, “Confucius and its intelligent disciples: integrating social with search,” in *VLDB*, 2010, pp. 1505–1517.

[17] G. Dror, Y. Koren, Y. Maarek, and I. Szpektor, “I want to answer; who has a question?: Yahoo! answers recommender system,” in *KDD*, 2011, pp. 1109–1117.

[18] F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios, “Finding expert users in community question answering,” in *WWW – CQA Workshop*, 2012, pp. 791–798.

[19] T. C. Zhou, M. R. Lyu, and I. King, “A classification-based approach to question routing in community question answering,” in *WWW – CQA Workshop*, 2012, pp. 738–790.

[20] A. Banerjee and S. Basu, “A social query model for decentralized search,” in *SNA-KDD*, 2008.

[21] R. Mihalcea, C. Corley, and C. Strapparava, “Corpus-based and knowledge-based measures of text semantic similarity,” in *National Conference on Artificial Intelligence*, 2006, vol. 21, no. 1, pp. 775–780.

[22] A. Islam and D. Inkpen, “Semantic text similarity using corpus-based word similarity and string similarity,” *ACM Trans. Knowl. Discov. Data*, vol. 2, no. 2, pp. 1–25, Jul. 2008.

[23] B. Furlan, V. Sivački, D. Jovanović, and B. Nikolić, “Comparable Evaluation of Contemporary Corpus-Based and Knowledge-Based Semantic Similarity Measures of Short Texts,” *J. Inf. Technol. Appl.*, vol. 1, no. 1, pp. 65–71, 2011.

[24] J. Wang and H.-J. Sun, “A new evidential trust model for open communities,” *Comput. Stand. Interfaces*, vol. 31, no. 5, pp. 994–1001, Sep. 2009.

[25] G. Adomavicius and A. Tuzhilin, “Toward the Next Generation of Recommender Systems : A Survey of the State-of-the-Art and Possible Extensions,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[26] O. Kaser, S. John, and D. Lemire, “Tag-Cloud Drawing : Algorithms for Cloud Visualization,” in *WWW*, 2007.

[27] M. A. Hearst and D. Rosner, “Tag Clouds: Data Analysis Tool or Social Signaller?,” in *HICSS*, 2008.

[28] “Antelope: Proxem resources for Natural Language Processing,” 2008. [Online]. Available: www.proxem.com.

[29] G. Miller, “WordNet: a lexical database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[30] H. Liu and P. Singh, “ConceptNet — A Practical Commonsense Reasoning Tool-Kit,” *BT Technol. J.*, vol. 22, no. 4, pp. 211–226, Oct. 2004.

[31] H. Liu, “Montylingua: A Free, Commonsense Enriched Natural Language Understander for English,” *Technical Report of the MIT University*, 2004. [Online]. Available: http://web.media.mit.edu/~hugo/montylingua/.

[32] H. Agt and R.-D. Kutsche, “Automated Construction of a Large Semantic Network of Related Terms for Domain-Specific Modeling,” in *Advanced Information Systems Engineering*, 2013, pp. 610–625.

[33] M. Hsu, M. Tsai, and H. Chen, “Combining WordNet and ConceptNet for automatic query expansion: a learning approach,” *Inf. Retr. Technol.*, vol. 4993, pp. 213–224, 2008.

[34] C. D. Manning, P. Prabhakar, and H. Schütze, “Scoring, term weighting, and the vector space model,” in in *An Introduction to Information Retrieval*, 1st ed., C. D. Manning, Ed. Cambridge: Cambridge University Press, 2009, pp. 100–123.

[35] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, “Knowledge sharing and yahoo answers: everyone knows something,” in *WWW*, 2008, pp. 665–674.

[36] M. Mohler and R. Mihalcea, “Text-to-text semantic similarity for automatic short answer grading,” in *European Chapter of the Association for Computational Linguistics*, 2009, pp. 567–575.

[37] R. Wang and G. Neumann, “Recognizing textual entailment using sentence similarity based on dependency tree skeletons,” in *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007, no. June, pp. 36–41.

[38] J. Oliva, J. I. Serrano, M. D. del Castillo, and Á. Iglesias, “SyMSS: A syntax-based measure for short-text semantic similarity,” *Data Knowl. Eng.*, vol. 70, no. 4, pp. 390–405, Apr. 2011.

[39] L. Li, Y. Zhou, B. Yuan, J. Wang, and X. Hu, “Sentence similarity measurement based on shallow parsing,” in *Fuzzy Systems and Knowledge Discovery*, 2009, pp. 487–491.

[40] B. Dolan, C. Quirk, and C. Brockett, “Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources,” in *International Conference on Computational Linguistics*, 2004.

[41] M. Porter, “An algorithm for suffix stripping,” *Progr. Electron. Libr. Inf. Syst.*, vol. 14, no. 3, pp. 130–137, 1980.

[42] V. Kešelj and D. Šipka, “A suffix subsumption-based approach to building stemmers and lemmatizers for highly inflectional languages with sparse resource,” *INFOTHECA–Journal Informatics Librariansh.*, vol. IX, no. 1–2, p. 24a–33a, 2008.

[43] D. Jurgens and K. Stevens, “The S-Space package: an open source package for word space models,” in *ACL System Demonstrations*, 2010, pp. 30–34.

[44] D. L. T. Rohde, L. M. Gonnerman, and D. C. Plaut, “An improved method for deriving word meaning from lexical co-occurrence,” *Cogn. Psychol.*, vol. 7, pp. 573–605, 2004.

[45] M. Sahlgren, “An introduction to random indexing,” in *In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, 2005.

[46] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, “Indexing by latent semantic analysis,” *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[47] R. T. W. Lo, B. He, and I. Ounis, “Automatically building a stopword list for an information retrieval system,” *J. Digit. Inf. Manag.*, vol. 5, no. Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR), pp. 17–24, 2005.

[48] M. Surdeanu, M. Ciaramita, and H. Zaragoza, “Learning to Rank Answers on Large Online QA Collections.,” in *ACL:Human Language Technologies*, 2008, pp. 719–727.

Прилози

Прилог А: Поређење анализираних приступа

	1. Question Analysis		2. Matching & Ranking		3. User Knowledge Profiling				4. Additional Info.
	Annotation	*Analysis*	*Model Organization*	*Semantic Matching*	*Internal*		*External*		
					Text	*Other*	*Text*	*Other*	
A. iLink	Tagging	NLP	Centralized (or Distributed)	No	DM	Response Score	DM	Manual	Referral Rank
B. PLSA in CQA	No	ML	Centralized	No	RS & ML	No	No		No
C. Question Routing Framework	No	No	Centralized	No	RS & ML	No	No		Availability
D. Routing within Forums	No	NLP	Centralized	Yes	ML	DM	No		Authority
E. G-Finder	No	Heuristics	Centralized	Yes	DM	DM	No		Concept Network
F. Aardvark	Tagging	DM	Centralized	Yes	DM & NLP	RS	DM & NLP	RS & Manual	Connectedness Availability
G. Confucius	Categories	DM	Centralized	Yes	DM	DM	No		Timeliness, Coverage, and Spam
H. Yahoo! Answers Recommender System	Categories	NLP	Centralized	No	RS	RS	No		Group of User Attributes
I. STM in CQA	Tagging	NLP	Centralized	Yes	ML	No	No		No
J. Classification-based Routing in CQA	Categories	NLP	Centralized	Yes	DM	DM	No		Global Features
K. SQM	No	No	Distributed	No	No	Expertise Score	No		Response Rate

Прилог Б: Резултати евалуације

ТАБЕЛА 15

ACHIEVED MRR SCORES ON DB TYPES 1 AND 2

DB TYPE 1 - QUESTION CONCEPTS

	Baseline	*LInSTSS*	*P2Q*	*P2Q(u)*
CE, SemNet, TF-IDF	0.11	0.14	0.08	0.18
CE	0.15	0.16	0.16	0.15
Cat1,2,3	0.21	0.21	0.19	0.19
CE, Cat1,2,3	0.20	0.22	0.25	0.21
TF-IDF	0.15	0.05	0.04	0.17
SemNet, TF-IDF	0.11	0.06	0.06	0.13

DB TYPE 2 - ANSWER CONCEPTS

	Baseline	*LInSTSS*	*P2Q*	*P2Q(u)*
CE, SemNet, TF-IDF	0.14	0.13	0.16	0.16
CE	0.15	0.16	0.16	0.15
Cat1,2,3	0.22	0.22	0.24	0.24
CE, Cat1,2,3	0.17	0.19	0.22	0.20
TF-IDF	0.18	0.05	0.06	0.14
SemNet, TF-IDF	0.16	0.05	0.07	0.13

DB TYPE 2 - THREAD CONCEPTS

	Baseline	*LInSTSS*	*P2Q*	*P2Q(u)*
CE, SemNet, TF-IDF	0.16	0.04	0.21	0.19
CE	0.15	0.17	0.15	0.16
Cat1,2,3	0.21	0.21	0.24	0.24
CE, Cat1,2,3	0.18	0.17	0.22	0.22
TF-IDF	0.18	0.05	0.06	0.17
SemNet, TF-IDF	0.15	0.04	0.05	0.14

ТАБЕЛА 16

ACHIEVED MRR SCORES ON DB TYPE 3

DB TYPE 3 - ANSWER CONCEPTS

	Baseline	*LInSTSS*	*P2Q*	*P2Q(u)*
CE, SemNet, TF-IDF	0.11	0.06	0.10	0.15
CE	0.10	0.10	0.11	0.11
Cat1,2,3	0.18	0.18	0.22	0.22
CE, Cat1,2,3	0.16	0.20	0.26	0.24
TF-IDF	0.10	0.06	0.08	0.12
SemNet, TF-IDF	0.11	0.06	0.09	0.14

DB TYPE 3 - THREAD CONCEPTS

	Baseline	*LInSTSS*	*P2Q*	*P2Q(u)*
CE, SemNet, TF-IDF	0.14	0.04	0.05	0.15
CE	0.11	0.11	0.14	0.15
Cat1,2,3	0.17	0.17	0.27	0.27
CE, Cat1,2,3	0.16	0.21	0.24	0.24
TF-IDF	0.16	0.04	0.04	0.15
SemNet, TF-IDF	0.13	0.04	0.04	0.11

ТАБЕЛА 17

ACHIEVED P@N SCORES ON DB TYPE 1 - QUESTION CONCEPTS

	N	*5*	*10*	*15*	*20*	*30*
CE, SemNet, TF-IDF	Baseline	0.13	0.27	0.34	0.45	0.56
	LInSTSS	0.13	0.28	0.36	0.46	0.57
	P2Q	0.07	0.12	0.21	0.24	0.42
	P2Q(u)	0.27	0.36	0.47	0.54	0.65
CE	Baseline	0.16	0.25	0.39	0.46	0.63
	LInSTSS	0.19	0.27	0.38	0.45	0.62
	P2Q	0.24	0.33	0.40	0.48	0.59
	P2Q(u)	0.21	0.36	0.41	0.50	0.55
Cat1,2,3	Baseline	0.29	0.44	0.55	0.61	0.8
	LInSTSS	0.29	0.44	0.55	0.61	0.8
	P2Q	0.27	0.45	0.60	0.72	0.88
	P2Q(u)	0.27	0.45	0.60	0.72	0.88
CE, Cat1,2,3	Baseline	0.24	0.4	0.58	0.62	0.7
	LInSTSS	0.29	0.46	0.59	0.64	0.74
	P2Q	0.29	0.48	0.59	0.71	0.8
	P2Q(u)	0.26	0.48	0.56	0.64	0.77
TF-IDF	Baseline	0.21	0.27	0.40	0.51	0.65
	LInSTSS	0.06	0.09	0.15	0.21	0.33
	P2Q	0.05	0.09	0.14	0.21	0.35
	P2Q(u)	0.25	0.37	0.48	0.54	0.68
SemNet, TF-IDF	Baseline	0.13	0.24	0.32	0.41	0.55
	LInSTSS	0.06	0.09	0.16	0.2	0.37
	P2Q	0.07	0.12	0.17	0.23	0.34
	P2Q(u)	0.17	0.29	0.43	0.52	0.65

ТАБЕЛА 18

ACHIEVED MRR SCORES ON DB TYPE 2 - ANSWER CONCEPTS

	N	*5*	*10*	*15*	*20*	*30*
CE, SemNet, TF-IDF	Baseline	0.21	0.3	0.39	0.46	0.57
	LInSTSS	0.20	0.29	0.39	0.45	0.57
	P2Q	0.08	0.17	0.24	0.37	0.5
	P2Q(u)	0.24	0.36	0.47	0.54	0.68
CE	Baseline	0.21	0.28	0.32	0.36	0.48
	LInSTSS	0.20	0.28	0.31	0.35	0.47
	P2Q	0.22	0.32	0.37	0.39	0.50
	P2Q(u)	0.23	0.34	0.35	0.39	0.51
Cat1,2,3	Baseline	0.33	0.48	0.59	0.69	0.87
	LInSTSS	0.33	0.48	0.59	0.69	0.87
	P2Q	0.36	0.57	0.66	0.78	0.90
	P2Q(u)	0.36	0.57	0.66	0.78	0.90
CE, Cat1,2,3	Baseline	0.24	0.35	0.50	0.56	0.70
	LInSTSS	0.27	0.41	0.49	0.63	0.70
	P2Q	0.35	0.46	0.61	0.66	0.77
	P2Q(u)	0.31	0.43	0.54	0.61	0.70
TF-IDF	Baseline	0.25	0.32	0.43	0.48	0.62
	LInSTSS	0.04	0.09	0.14	0.16	0.31
	P2Q	0.04	0.10	0.17	0.27	0.40
	P2Q(u)	0.17	0.37	0.46	0.57	0.67
SemNet, TF-IDF	Baseline	0.23	0.31	0.41	0.52	0.63
	LInSTSS	0.04	0.09	0.14	0.20	0.32
	P2Q	0.06	0.13	0.20	0.28	0.44
	P2Q(u)	0.15	0.33	0.42	0.52	0.65

ТАБЕЛА 19

ACHIEVED MRR SCORES ON DB TYPE 2 - THREAD CONCEPTS

	N	*5*	*10*	*15*	*20*	*30*
CE, SemNet, TF-IDF	Baseline	0.19	0.29	0.46	0.55	0.71
	LInSTSS	0.02	0.06	0.10	0.14	0.19
	P2Q	0.27	0.41	0.53	0.58	0.64
	P2Q(u)	0.27	0.38	0.51	0.55	0.63
CE	Baseline	0.22	0.26	0.34	0.42	0.56
	LInSTSS	0.20	0.27	0.36	0.46	0.63
	P2Q	0.24	0.37	0.49	0.56	0.66
	P2Q(u)	0.23	0.35	0.45	0.57	0.65
Cat1,2,3	Baseline	0.31	0.48	0.60	0.70	0.87
	LInSTSS	0.31	0.48	0.60	0.70	0.87
	P2Q	0.36	0.57	0.66	0.78	0.90
	P2Q(u)	0.36	0.57	0.66	0.78	0.90
CE, Cat1,2,3	Baseline	0.26	0.40	0.53	0.62	0.79
	LInSTSS	0.28	0.42	0.56	0.69	0.80
	P2Q	0.31	0.50	0.69	0.78	0.85
	P2Q(u)	0.31	0.47	0.65	0.74	0.81
TF-IDF	Baseline	0.21	0.33	0.46	0.57	0.74
	LInSTSS	0.04	0.08	0.13	0.18	0.22
	P2Q	0.04	0.10	0.17	0.27	0.40
	P2Q(u)	0.25	0.37	0.54	0.62	0.70
SemNet, TF-IDF	Baseline	0.16	0.31	0.49	0.56	0.73
	LInSTSS	0.04	0.07	0.09	0.13	0.17
	P2Q	0.04	0.07	0.15	0.19	0.34
	P2Q(u)	0.23	0.29	0.44	0.52	0.65

ТАБЕЛА 20

ACHIEVED MRR SCORES ON DB TYPE 3 - ANSWER CONCEPTS

	N	*5*	*10*	*15*	*20*	*30*
CE, SemNet, TF-IDF	Baseline	0.14	0.23	0.27	0.32	0.49
	LInSTSS	0.05	0.14	0.18	0.34	0.45
	P2Q	0.10	0.18	0.28	0.37	0.53
	P2Q(u)	0.24	0.33	0.40	0.52	0.62
CE	Baseline	0.14	0.20	0.27	0.35	0.46
	LInSTSS	0.12	0.20	0.25	0.34	0.48
	P2Q	0.13	0.28	0.31	0.42	0.48
	P2Q(u)	0.15	0.26	0.35	0.42	0.50
Cat1,2,3	Baseline	0.22	0.41	0.54	0.66	0.81
	LInSTSS	0.22	0.41	0.54	0.66	0.81
	P2Q	0.31	0.51	0.68	0.77	0.87
	P2Q(u)	0.31	0.51	0.68	0.77	0.87
CE, Cat1,2,3	Baseline	0.23	0.39	0.50	0.55	0.64
	LInSTSS	0.33	0.41	0.51	0.59	0.65
	P2Q	0.36	0.53	0.58	0.63	0.74
	P2Q(u)	0.33	0.48	0.53	0.65	0.71
TF-IDF	Baseline	0.15	0.29	0.35	0.45	0.59
	LInSTSS	0.04	0.10	0.15	0.22	0.40
	P2Q	0.07	0.14	0.22	0.31	0.45
	P2Q(u)	0.15	0.29	0.38	0.53	0.63
SemNet, TF-IDF	Baseline	0.15	0.24	0.31	0.39	0.51
	LInSTSS	0.04	0.11	0.18	0.29	0.44
	P2Q	0.10	0.19	0.26	0.37	0.52
	P2Q(u)	0.19	0.30	0.42	0.49	0.60

ТАБЕЛА 21

ACHIEVED MRR SCORES ON DB TYPE 3 - THREAD CONCEPTS

	N	*5*	*10*	*15*	*20*	*30*
CE, SemNet, TF-IDF	Baseline	0.23	0.35	0.46	0.55	0.63
	LInSTSS	0.04	0.06	0.11	0.18	0.27
	P2Q	0.04	0.12	0.18	0.27	0.46
	P2Q(u)	0.25	0.39	0.51	0.55	0.69
CE	Baseline	0.12	0.26	0.33	0.42	0.57
	LInSTSS	0.14	0.26	0.37	0.43	0.55
	P2Q	0.24	0.35	0.45	0.53	0.57
	P2Q(u)	0.25	0.36	0.48	0.52	0.57
Cat1,2,3	Baseline	0.22	0.39	0.54	0.61	0.75
	LInSTSS	0.22	0.39	0.54	0.61	0.75
	P2Q	0.35	0.48	0.70	0.74	0.87
	P2Q(u)	0.35	0.48	0.70	0.74	0.87
CE, Cat1,2,3	Baseline	0.18	0.34	0.45	0.56	0.76
	LInSTSS	0.25	0.48	0.63	0.65	0.83
	P2Q	0.30	0.49	0.65	0.76	0.86
	P2Q(u)	0.36	0.49	0.62	0.69	0.83
TF-IDF	Baseline	0.21	0.30	0.41	0.49	0.68
	LInSTSS	0.04	0.08	0.12	0.18	0.28
	P2Q	0.04	0.09	0.15	0.22	0.35
	P2Q(u)	0.19	0.29	0.44	0.54	0.69
SemNet, TF-IDF	Baseline	0.18	0.32	0.41	0.52	0.65
	LInSTSS	0.04	0.06	0.11	0.18	0.22
	P2Q	0.04	0.09	0.16	0.21	0.40
	P2Q(u)	0.14	0.27	0.40	0.53	0.61

Прилог В: Списак радова везаних за дисертацију

1. **Радови у часописима са импакт фактором (SCI листа)**
   1.1. **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, Vol. 55, Issue 3, pp. 710–719, June 2013. (**IF 3.037**), doi: 10.1016/j.dss.2013.02.002
   1.2. **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, Vol. 28, Issue 7, pp. 686–708, July 2013., (**IF 1.579**) doi: 10.1002/int.21597
2. **Радови у иностраним научним часописима**
   2.1 **Furlan B.**, Sivački V., Jovanović D., Nikolić B. "Comparable Evaluation of Contemporary Corpus-Based and Knowledge-Based Semantic Similarity Measures of Short Texts," *JITA*, vol. 1, no. 1, ISSN 2233-0194 (online), pp. 65-71, June 2011. (**nema impakt faktor**)
   2.2 Varga E., **Furlan B.**, and Milutinovic V., "Document Filter Based on Extracted Concepts," *Transactions on Internet Research*, vol. 6, no. 1, ISSN 1820 – 4503 (online), pp. 5-9, January 2010. (**nema impakt faktor**)
3. **Радови у зборницима радова међународних конференција**
   3.1 **Furlan B.**, Žitnik S., Nikolić B., Bajec M., "The Role of Semantic Similarity for Intelligent Question Routing," in Informatics, Spišská Nová Ves, Slovakia, November 5th – 7th, 2013.
   3.2 Zitnik S., Subelj L., Jankovic M., **Furlan B.**, Draskovic D., Kojic N., Misic M., Bajec M., "Iterative End-to-end Information Extraction based on Linear Models," in ERK, Portorož, Slovenia, September 2013.
   3.3 **Furlan B.**, Nikolic B., Milutinovic V., "A Survey of Intelligent Question Routing Systems," in IEEE Intelligent Systems, Sofia, Bulgaria, September 2012.
   3.4 Jelisavčić V., **Furlan B.**, Protić J., Milutinović V., "Topic Models and Advanced Algorithms for Profiling of Knowledge in Scientific Papers," in MIPRO, Opatija, Croatia, May 2012.
4. **Саопштења на међународним конференцијама и скуповима објављена у изводу**
   4.1 **Furlan B.**, Nikolic B., Milutinovic V., "Intelligent Question Routing: An Overview of Some Recent Advances and Open Problems", VIPSI, Miločer, Crna Gora, 2011.
   4.2 **Furlan B.**, "An Intelligent Question Routing System," VIPSI, Pisa, Italy, 2008.

5. Радови у зборницима радова домаћих конференција

5.1 **Furlan B.**, Stamenković J., Nikolić B., Mišić M., "Algoritam određivanja semantičke sličnosti između korisničkog profila i pitanja," ETRAN, Zlatibor, Srbija, 3 - 6. Juna 2013.

5.2 Jelisavčić V., **Furlan B.**, Protić J., Milutinović V., "Knowledge Modeling and Classification of Scientific Papers Based on Topic Modeling," in YUINFO, Kopaonik, Serbia, March 2012. pp. 664-669

5.3 Batanović V., **Furlan B.**, Nikolić B., "Softverski sistem za određivanje semantičke sličnosti kratkih tekstova na srpskom jeziku," TELFOR, Beograd, Srbija, 22-24. Novembra, 2011.

5.4 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.

5.5 **Furlan B.**, Nikolić B., "Veb-Baziran Sistem za Efikasno Dobijanje Odgovora," ETRAN, Palić, Srbija, 2008.

5.6 Nikolić S., **Furlan B.**, Josipović P., "aLive! - Sistem za inteligentno prosleđivanje pitanja," YUINFO, Kopaonik, Srbija, 2008.

Биографија аутора

Бојан Фурлан је рођен 04.09.1982. године у Панчеву, од оца Дарка и мајке Иванке. Завршио је средњу електротехничку школу „Никола Тесла“ у Београду са одличним успехом. Током школовања био је учесник такмичења у знању, активно тренирао кошарку и бавио се спортом.

Електротехнички факултет Универзитета у Београду уписао је 2001. године, смер Рачунарска техника и информатика. Током студија, био је предавач у Едукационом центру Електротехничког факултета и учесник *Microsoft Student Partner* програма, у оквиру кога је држао курсеве и помагао студентима заинтересованим да овладају Microsoft технологијама.

Након дипломирања и завршетка студија, 2007. године, два пута краће борави на Институту за информатику, Техничког универзитета у Минхену, у оквиру DAAD програма Немачке владе где учествује на истраживачком пројекту SimLab. По повратку уписује докторске студије на Електротехничком факултету, где је у јануару 2008. године примљен у звање асистент у настави за ужу научну област Рачунарска техника и информатика. У периоду од септембра 2009. до јуна 2010. године служи војни рок на Војној академији, где учествује на развоју информационог система за размену докумената у процесу војног одлучивања.

На Електротехничком факултету у Београду од 2008. године до данас држи вежбе на основним и академским мастер студијама из више предмета. Такође, неколико година је учествовао у формирању, припреми и вођењу екипа које су освојиле бројне награде на такмичењу из програмирања на Електријади.

На докторским студијама положио је све испите са просечном оценом 10. Поред предмета предвиђених наставним планом и програмом похађао је и неколико летњих школа за докторанте из области блиских његовом стручном интересовању.

Аутор је четири рада у међународним часописима, међу којима су и два рада објављена у часописима са *impact* фактором (SCI листа), један из категорије М21 (врхунски међународни часопис) и један из категорије М22 (истакнути међународни часопис). Такође, аутор је једне скрипте са решеним задацима чији је издавач Електротехнички факултет и која се користи у настави на академским мастер студијама. Има више радова који су представљени на међународним и домаћим научним конференцијама, као и техничка решења развијена у оквиру домаћих и међународних пројеката.

Тренутно је рецензент у међународним часописима *Information Sciences* и *Decision Support Systems*, као и домаће конференције ТЕЛФОР. Добитник је награде за најбољи рад конференције IEEE Intelligent Systems IS'12, као и неколико стипендија за краће посете иностраним научно истраживачким институцијама. Говори енглески и шпански и служи се немачким и руским.

Изјава о ауторству

Потписани-а Бојан Фурлан

број уписа 2007/5006

Изјављујем

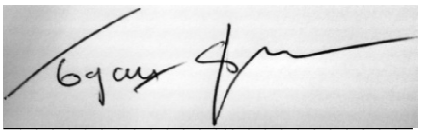
да је докторска дисертација под насловом

МЕТОДОЛОГИЈА ПРОЈЕКТОВАЊА СИСТЕМА ЗА ИНТЕЛИГЕНТНО ПРОСЛЕЂИВАЊЕ ПИТАЊА НАПИСАНИХ НА ПРИРОДНОМ ЈЕЗИКУ

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанта

У Београду, 29.11.2013.



Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Бојан Фурлан

Број уписа 2007/5006

Студијски програм Софтверско инжењерство

Наслов рада Методологија пројектовања система за интелигентно прослеђивање питања написаних на природном језику
Ментор др Бошко Николић, ванредни професор

Потписани Бојан Фурлан

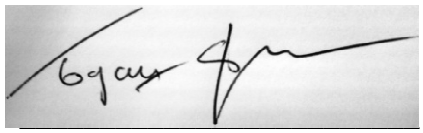
изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанта

У Београду, 29.11.2013.



Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

МЕТОДОЛОГИЈА ПРОЈЕКТОВАЊА СИСТЕМА ЗА ИНТЕЛИГЕНТНО ПРОСЛЕЂИВАЊЕ ПИТАЊА НАПИСАНИХ НА ПРИРОДНОМ ЈЕЗИКУ

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

④ Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

Потпис докторанта

У Београду, ___29.11.2013._____