УНИВЕРЗИТЕТ У БЕОГРАДУ
ГРАЂЕВИНСКИ ФАКУЛТЕТ

Ђорђе Љ. Недељковић

ИЗДВАЈАЊЕ И ВИЗУЕЛИЗАЦИЈА ЗНАЊА ИЗ ТЕКСТУАЛНИХ ИЗВОРА ЗА ПОТРЕБЕ УПРАВЉАЊА ИНВЕСТИЦИОНИМ ПРОЈЕКТИМА У ГРАЂЕВИНАРСТВУ

Докторска дисертација

Београд, 2018

UNIVERSITY OF BELGRADE
FACULTY OF CIVIL ENGINEERING

Đorđe Lj. Nedeljković

KNOWLEDGE EXTRACTION AND VISUALIZATION FROM TEXTUAL SOURCES INTENDED FOR CONSTRUCTION PROJECT MANAGEMENT

Doctoral dissertation

Belgrade, 2018

Подаци о ментору и о члановима комисије

Ментор:

др Милош Ковачевић, ванредни професор

Универзитет у Београду, Грађевински факултет

Чланови комисије:

др Милош Ковачевић, ванредни професор

Универзитет у Београду, Грађевински факултет

др Ненад Иванишевић, ванредни професор

Универзитет у Београду, Грађевински факултет

др Наташа Прашчевић, ванредни професор

Универзитет у Београду, Грађевински факултет

др Бранислав Ивковић, редовни професор у пензији

Универзитет у Београду, Грађевински факултет

др Вељко Милутиновић, редовни професор у пензији

Универзитет у Београду, Електротехнички факултет

Датум одбране:

Захвалница

Велику захвалност дугујем ментору, др Милошу Ковачевићу, на безрезервној подршци током рада на дисертацији. Уз његове вредне савете, израда дисертацији је била занимљиво путовање кроз свет научно-истраживачког рада.

Захваљујем се др Браниславу Ивковићу, који је у раним фазама препознао могућности и идеје предложеног решења и обезбедио неопходне податке са стварних пројеката, без којих не би било могуће спровести истраживање.

Захвалност дугујем и колегиницама и колегама са Катедре за управљање пројектима у грађевинарству, који су ми квалитетним сугестијама помогли да усмерим истраживање у правом смеру.

На крају, захваљујем се и својој породици, чија су ми љубав и подршка помогли да стигнем на циљ.

Сажетак:

Током животног циклуса инвестиционог пројекта ствара се велики корпус неструктуираних и полуструктуираних докумената. Традиционални приступи у складиштењу и организовању информација из неструктуираних података су оријентисани на рад са документима, што их чини неподесним за анализу и издвајање знања. У неструктуираним документима је отежано прикупљање, анализа и поновно коришћење релевантних информација у интегралном облику, што може изазвати проблеме на пројекту услед неблаговремених или неодговарајућих одлука.

У овој дисертацији је приказана репрезентација информација издвојених из неструктуираних текстуалних докумената у облику графа значајних фраза, који корисницима треба да омогући визуелизацију и анализу значајних чињеница на пројекту са минималном количином уложеног труда. Са циљем да се конструише доменски независна репрезентација са минималним трудом експерта за претходно конфигурисање, значајне фразе су детектоване у вишејезичном окружењу применом статистичких мера за одређивање корелисаности пара речи. Граф садржи аутоматски издвојене значајне фразе које су повезане на основу сличности семантичких контекста.

Репрезентација је имплементирана у графовској бази података што корисницима омогућава да детектују и визуелизују различите скривене обрасце у подацима. Неинформативне фразе су филтриране кроз поступке одређивања ентропије скупа контекста и динамичности суседства фразе кроз више графова који представљају тренутке у времену. Приказана је хеуристика за издвајање комплексних концепата, заснована на итеративној процедури за детекцију блиских фраза које припадају истом семантичком подграфу. Могућности примене предложене репрезентације су демонстриране на графу конструисаном за постојећи корпус докумената са међународног инвестиционог пројекта.

Кључне речи: неструктуирани подаци, издвајање значајних фраза, ентропија, семантичка мрежа, релација, граф значајних фраза, визуелизација, динамичност суседа, управљање пројектима

Научна област: Грађевинарство

Ужа научна област: Примена информационих технологија у грађевинарству и геодезији

УДК број: 624:005.8(043.3)

Abstract:

During a construction project lifecycle, an extensive corpus of unstructured or semi-structured text documents is generated. Traditional approaches for information storing and organizing are document-oriented, which is highly inconvenient for data analysis and knowledge extraction. The nature of unstructured sources impedes users' acquisition, analysis, and reuse of relevant information, leading to possible negative effects in the project management process.

This dissertation suggests a procedure for automatic extraction of relevant project concepts from unstructured text documents. Concepts are organized in the form of a key-phrase network, intended to provide users with the possibility to visualize and analyze valuable project facts with less effort. With the objective of constructing a domain-independent and language-independent key-phrase network, with minimal expert involvement for configuration, an approach to detect key phrases was examined by using measures of correlation for word pairs. A network contains key phrases automatically extracted from various types of unstructured documents, with relations based on the similarity of semantic contexts.

The representation was implemented as a graph database, enabling project participants to extract and visualize various patterns in data. The problem of noisy key phrases was reduced by introducing the entropy score for a set of co-occurring contexts and the measure of phrase neighborhood dynamics throughout construction project lifecycle. A heuristic for extraction of complex concepts is presented, based on the iterative procedure for detection of adjacent key phrases belonging to a same semantic subnetwork. Possible applications, such as concept tracking through time or determination of communication patterns between project participants, is demonstrated using a key-phrase network generated for the existing document corpus from an international construction project.

Keywords: unstructured data, key-phrase extraction, entropy, semantic network, relationship, key-phrase network, visualization, neighborhood dynamics, project management

Scientific field: Civil engineering

Scientific subfield: Application of information technology in civil engineering and geodesy

UDC number: 624:005.8(043.3)

Садржај

1 Уводна разматрања

Инвестициони пројекти у грађевинарству су *јединствени* и *непоновљиви*. Сваки објекат који је предмет пројекта се уговара, пројектује и гради под јединственим условима. Према (Ivković & Popović 2005), "инвестициони пројекат представља комплексан техничко-технолошки, организациони, финансијски и правни подухват, који се састоји од скупа координисаних и констролисаних активности са јасно дефинисаним почетком и крајем, чији је циљ изградња, реконструкција, модификације и/или опремање објекта или објеката који су потребни инвеститору". Различите међузависне активности, свака са својим особеностима и временским оквиром, треба да се изврше у одређеном поретку како би се пројекат успешно реализовао.

Комплексна природа пројекта ставља се под *контролу* применом *стандардизованих* процедура у различитим фазама његовог животног циклуса. Да би у сваком тренутку успешно управљали пројектом, експерти морају имати *тачне*, *јасне* и *правовремене* информације. Велики број учесника, који су у обавези да *поделе* статус својих и *прате* активности других учесника, диктирају да *размена информација* током управљања пројектом буде једна од кључних активности.

За успешно управљање инвестиционим пројектом неопходно је ефикасно праћење и контролисање протока информација (Russell et al. 2009). Стални пораст обим информација које се размењују на пројекту је последица подељености пословних процеса, што изискује значајан ниво међусобне координације, контроле и комуникације међу учесницима. Ако се проток информација на пројекту не контролише, експерти задужени за управљање су „затрпани" документима које морају да анализирају како би имали потребне елементе за доношење одлука. У (Songer et al. 2006) подвлачи се да су на

грађевинским пројектима заступљене ситуације описане као *богате подацима, а сиромашне информацијама*. Аутори наводе да пројекат обилује информацијама које су квантитативно описане у више димензија, а без значајног доприноса за додатно разумевање проблема.

Да презасићеност информацијама буде већа, доприноси и корпоративна култура управљања: руководиоци често сматрају да контрола и поседовање информација представљају ствар престижа и ауторитета, те захтевају и податке који нису неопходни за текуће задатке (Haksever 2000). Са друге стране, подређени имају интерес да истакну своју улогу, па надређенима прослеђују повећани обим информација које ексклузивно стварају, а које нису нужно релевантне (Pietroforte 1997).c

У техничком извештају[1] који описује пораст обима података кроз време, утврђено је да је типичан пројекат из 2004. године садржао око 100 гигабајта података. Сличан пројекат истог грађевинског предузећа је 2014. године садржао 6.6 терабајта података, што представља повећање од 66 пута! Посебно се указује на повећан обим порука електронске поште (100.000 порука у 2004. години, насупрот 288.000 за 2014. годину).

1.1 Основни појмови

Концепт „знања“ је изузетно широк и могуће га је, у зависности од дисциплине која га дефинише, различито интерпретирати. У овом истраживању, појам знања разматра се са становишта примењивости у реалном пословном окружењу. Према капацитету да се пренесе значење, разликујемо *податак*, *информацију* и *знање*, аналогно познатом концепту пирамиде знања (Ackoff 1989).

Податак представља „основну јединицу описа ствари, догађаја, активности или трансакције“ (Kenneth C. Laudon & Laudon 2012). Да би се податак

[1]http://www.johnsiskandson.com/sites/default/files/page/701/2424-526740-future-cons.pdf

превео у структуру која носи виши степен значења, потребно је да се *процесира*, *повеже* са другим подацима и стави у одговарајући *контекст*, након чега постаје део *информације*.

Информација представља податке преведене у форму која поседује значење и сврху за особу која је тумачи (Kenneth C. Laudon & Laudon 2012).

Знање настаје када се синтетисаним информацијама, у одређеном контексту, додају експертска искуства и правила (Pearlson & Saunders 2010). У пословном окружењу, знање представља променљив оквир искустава, правила и информација који омогућава процену и присвајање нових знања (Davenport & Prusak 1998).

На нивоу предузећа, знање је похрањено у *документима* и *базама података*, као и у *пословним праксама* и *процесима*. Међутим, ако је похрањено на неодговарајући начин, знање може да деградира у форму са мањом семантичком вредношћу (Davenport & Prusak 1998). Према томе како су подаци који чине делове знања структуирани и описани приликом њиховог похрањивања, разликују се три форме складиштења: структуирана, полуструктуирана и неструктуирана.

Структуирана форма се односи на податке са високим нивоом организације и формалним описом појединих делова који одређује њихово значење. Формални опис подразумева да се делови података именују, да им се одреди домен вредности и правила ажурирања. Подаци се не могу уносити а да не одговарају формалном опису. Најчешће се складиште у релационим базама података или еквивалентном окружењу (глава 8).

Полуструктуирана форма података поседује формални опис делова, али подаци *не морају* да га прате (изостављена су правила ажурирања). Опис садржи ознаке које раздвајају семантичке или хијерархијске целине. У пројектном окружењу, најзаступљенији полуструктуирани подаци односе се на различите *табеларне* формате (*xlsx*, *csv*). Показано је да 67% експерата који се баве

управљањем пројектима, за праћење и извештавање користе табеларне формате[2].

Неструктуирани подаци се карактеришу одсуством формалног описа и структуре. Значење појединих делова потребно је, због одсуства формалног описа, накнадно интерпретирати. Пример неструктуираних података је и текст ове тезе у којој се, без икаквог ограничења, може мењати њен садржај. Како рачунарски систем *нема* претходно дефинисана правила за тумачење таквог садржаја, отежана је његова претрага и анализа, у поређењу са полуструктуираним и структуираним подацима (Sint et al. 2009).

Из претходног излагања закључује се да су, за доношење одлука, најпогоднije информације похрањене у структуираном и полуструктуираном облику. Међутим, највећи део садржаја који се користи за доношење одлука на инвестиционом пројекту је у форми неструктуираних података *(Caldas et al. 2002)*. Према другом истраживању, неструктуирани подаци који се најчешће налазе у текстуалном облику, чине око 80% пословних информација у предузећу (Blumberg & Atre 2003).

Како би се савладао проблем презасићености информацијама на пројекту, неопходно је коришћење одговарајућих информационих система. Међутим, највећи део информационих система који се користе у грађевинској индустрији, заснива се на складиштењу које по природи одговара класичним папирним формама (Zhu et al. 2007).

У (Matthies 2015), наводе се резултати интервјуисања експерата задужених за управљање пројектима, по питању побољшања информационе подршке:

- Неопходна је *боља подршка* у претрази и анализи *неструктуираних* података из докумената;

[2] The 5thAnnual Construction Technology Report," JB Knowledge

- Неопходан је стандардизован поступак за чување релевантних искустава са пројекта. Наводи се да „*није потребно више информација, неопходне су праве информације*“;
- Значајна је *концептуализација* пројектног знања. Уместо читања дугачког документа, како би се издвојило неколико значајних информација, упутно је *означити информације* у документима према специфичним ситуацијама.

1.2 Циљеви истраживања

Имајући у виду значај информација садржаних у текстуалним изворима на пројекту, научни циљеви истраживања су да:

- класификује системе за рад са документима за потребе одлучивања на пројекту,
- опише репрезентације знања погодне за визуелизацију и одлучивање,
- дефинише и експериментално верификује модел за аутоматско препознавање значајних концепата из неструктуираног текста,
- дефинише и експериментално верификује модел за успостављање веза између концепата,
- имплементира систем за одговарајуће складиштење, претрагу и визуелизацију дефинисане репрезентације,
- илуструје корисне ефекте предложеног система за издвајање знања из текстуалних извора на примерима из праксе.

1.3 Организација дисертације

Анализа тренутне праксе у домену коришћења информација из неструктуираног текста, у грађевинском сектору Републике Србије, приказана је у глави 2. Истраживање је спроведено кроз две анкете у којима су испитани постојећи поступци у раду са документима. Истражено је како грађевинска предузећа обављају интерну и екстерну комуникацију, као и како се обрађују подаци значајни за доношење одлука.

У глави 3 разматрају се постојећи системи за рад са документима који настају током животног циклуса пројекта, са становишта могућности издвајања информација неопходних за управљање пројектом. Показане су специфичности поступка извођења новог знања у окружењу система датотека, система за управљање информацијама, као и система заснованог на онтологији. Утврђено је да су постојећи системи или захтевни за имплементацију на различитим пројектима, или не поседују одговарајуће алате за закључивање из неструктуираних података.

На основу анализе постојећих решења за издвајање значајних информација из неструктуираног текста, у глави 4 су формулисане основне претпоставке, описан концепт и наведена ограничења предложеног поступка за издвајање и визуелизацију знања из текстуалних извора.

У глави 5 је дефинисан поступак издвајања релевантних концепата из неструктуираних докумената на пројекту. Концепти су представљени значајним паровима речи које се издвајају применом језички независних статистичких мера за одређивање међузависности. Предложен је поступак за филтрирање неинформативних парова, заснован на ентропији контекста појављивања. На овом месту, извршена је експериментална провера свих метода предложеног поступка, као и резултати експеримента којим се утврђује семантички капацитет парова да опишу препознате концепте на пројекту.

Глава 6 разматра постојеће репрезентације знања које би представљале основу за организовање издвојених концепата у погодну структуру за извођење нових знања. Разматране су семантичке мреже, концептуални графови и оквири.

Доменски и језички независан поступак конструисања репрезентације издвојених концепата, као графа значајних фраза, описан је у глави 7. Дефинисан је критеријум за успостављање релација између значајних парова, заснован на заједничком контексту појављивања унутар докумената у корпусу. Добијени граф користи се за проналажење значајних фраза састављених од већег броја речи, као и за додатно рангирање фраза према динамичности суседа.

У глави 8 се разматрају погодна окружења за складиштење и обраду предложене репрезентације. Пореде се релациона и графовска база података, са становишта брзине извршавања и једноставности задавања упита. Посебно се разматра могућност визуелизације резултата упита, као неопходне опције која омогућава правилну корисничку интерпретацију приликом извођења нових знања.

Глава 9 описује интерактивни рад у окружењу графа значајних фраза. Дефинисани су поступци за одређивање блиских значајних концепата на пројекту, као и њихово праћење кроз време. Предложена је хеуристика за одређивање комплексних концепата који покривају одређене теме на пројекту. Описано је могуће семантичко проширивање предложене репрезентације, увођењем кориснички дефинисаних ентитета, попут датума, особе и акције. Проширивање се може обавити уз минимално претходно ангажовање експерта. Могућности проширене репрезентације илустроване су на примеру одређивања интеракције између учесника на пројекту, а на основу обраде записника са састанака.

Дисертацију закључује глава 10, у којој су дата закључна разматрања о предложеном решењу. Приказане су могућности примене графа значајних фраза у окружењу инвестиционог пројекта. На крају су наведени предлози могућих праваца за будуће истраживање.

2 Коришћење неструктуираних текстуалних информација у грађевинском сектору Републике Србије

Да би се испитале постојеће праксе примене и искоришћености неструктуираних података за грађевинска предузећа у Републици Србији, спроведене су две анкете: анкета *A2015* (стање у 2015. години) и анкета *A2016*. Испитаници су одабрани у складу са релевантношћу њихове позиције у предузећу према постављеним питањима. Анкетирана предузећа су категорисана по броју запослених, у складу са чланом 6. важећег Закона о рачуноводству[3] (табела 2.1):

Табела 2.1: Категорије пореских обвезника према броју запослених.

Категорија	Критеријум
Микро правно лице	просечан број запослених ≤ 10
Мало правно лице	10 < просечан број запослених ≤ 50
Средње правно лице	50 < просечан број запослених ≤ 250
Велико правно лице	250 < просечан број запослених

Највећи број питања у обе анкете је конципиран тако да описује одређене поступке из пословне праксе. Понуђени одговори су организовани у облику Ликартове скале за градирање значаја понуђених опција (Carifio & Perla 2007). Од испитаника се тражило да додели оцену на скали од 1 до 4, где оцене

[3] https://www.nbs.rs/internet/latinica/20/zakoni/rac_racunovodstvo.pdf

представљају степен значаја описаног поступка за предузеће, према шеми описаној у табели 2.2:

Табела 2.2: Значај описаног поступка за предузеће.

Оцена	Значај
1	није значајан
2	мало значајан
3	Значајан
4	веома значајан

У анкети *A2016* су испитани и потенцијални недостаци или препреке за примену одређених поступака, који су оцењени истом скалом.

Примери:

- Поступак: *Приступ и претрага ел. поште* – веома значајан (4);
- Поступак: *Коришћење програма за рад са табелама за обраду података значајних за доношење одлука* – значајан (3);
- Недостатак: *У комуникацији корисници не добијају информације на време* – мало значајан (2).

2.1 Структура анкета

A2015: Поред основних података о предузећу, од испитаника се тражило да оцене значај понуђених поступака у раду са документима. Питања су груписана према томе да ли се односе на основне или напредне поступке у раду са документима.

A2016: Поред основних података о предузећу, испитаници су добили групе питања којима се, на нивоу целог предузећа, оцењују поступци и недостаци у:

- интерној и екстерној комуникацији;
- доношењу пословних одлука;
- обради података од значаја за доношење одлука.

У табелама 2.3 и 2.4 су приказани резултати анкета *A2015* и *A2016*. Резултати су груписани по категоријама предузећа. За свако питање су приказани *модус* (оцена која се најчешће јављала; у случају једнаких фреквенција изабрана је оцена која изражава мањи значај критеријума који се вреднује), као и *консензус* унутар групе оцењивача. Овде се под консензусом подразумева степен слагања групног става по одређеном питању (Tastle & Wierman 2007). С обзиром да су понуђени одговори организовани тако да одговарају Ликартовој скали (Carifio & Perla 2007), у обзир је узета препорука да се резултати прикажу модусом (или медијаном), а не средњом вредношћу (Boone & Boone 2012). За свако питање консензус испитаника из исте групе рачуна се према (Tastle & Wierman 2007):

$$Kon = 1 + \sum_{i=1}^{n} p_i log_2 \left(1 - \frac{|X_i - \mu_x|}{d_x}\right) \qquad (2.1)$$

где је:

n – број оцена (овде 4)

N – број испитаника (зависи од групе предузећа)

d_x – распон оцена (овде 4 – 1 = 3)

X_i – i – та оцена

p_i – вероватноћа да је на питање одговорено i – том оценом,

$$p_i = \frac{broj\ ispitanika\ koji\ su\ dali\ ocenu\ X_i}{N}$$

μ_x – очекивана оцена за питање, $\mu_x = \sum_{i=1}^{n} p_i X_i$

У случају када је расподела оцена униформна, консензус испитаника једнак је нули па се из одговора не могу извући поуздани закључци, сем да постоји неслагање. Обрнут случај, када сви испитаници дају исту оцену једном питању, даје консензус један – сви испитаници су потпуно сагласни. У овом истраживању

се већи значај даје одговорима за које је утврђен већи степен слагања испитаника ($Kon > 0.5$).

Табела 2.3: Резултати анкете о поступцима у раду са документима (*A2015*). Сиво су означени одговори за које је консензус мањи од 0.5.

	Категорија предузећа			
Питање	**Микро**	**Мало**	**Средње**	**Велико**
Број предузећа	12	17	9	7
Просечан број пословних партнера са којима се комуницира током месеца:	29	31.18	78.33	93.57
Коришћење електронске поште	4 (0.55)	4 (0.94)	4 (0.7)	4 (1)
Архивирање постојеће документације	4 (0.71)	4 (0.89)	4 (0.77)	4 (1)
Поштовање предефинисаних радних процедура и контроле квалитета у раду са документима	3 (0.74)	4 (0.82)	4 (0.9)	4 (0.44)
Претрага података на интернету	4 (0.69)	4 (0.85)	4 (0.7)	3 (0.5)
Контрола приступа документима у предузећу	4 (0.77)	4 (0.71)	3 (0.74)	4 (0.63)
Увид у документацију на терену	4 (0.57)	4 (0.67)	4 (0.77)	3 (0.72)
Проналажења документа према садржају унутар фајла	3 (0.6)	4 (0.6)	3 (0.74)	4 (0.74)
Претрага спољних стручних база података	3 (0.61)	3 (0.73)	3 (0.82)	3 (0.42)
Увид у део документа са жељеном информацијом без његовог отварања	3 (0.6)	3 (0.72)	2 (0.77)	2 (0.74)
Аутоматско груписање сличних докумената	4 (0.43)	3 (0.74)	3 (0.66)	3 (0.78)
Аутоматска класификација докумената према типу	4 (0.44)	3 (0.58)	2 (0.6)	3 (0.74)
Аутоматско описивање документа најважнијим концептима у њему	3 (0.58)	3 (0.68)	2 (0.74)	4 (0.44)
Ручно лабелирање докумената мета-подацима	2 (0.65)	3 (0.73)	3 (0.87)	3 (0.78)
Аутоматско додавање ознака документима	4 (0.4)	3 (0.65)	2 (0.65)	2 (0.63)

Табела 2.4. Резултати анкете о постојећој пракси у поступцима комуникације и обради података при доношењу одлука (*A2016*). Сиво су означени одговори за које је консензус мањи од 0.5.

	Категорија предузећа			
Питање	Микро	Мало	Средње	Велико
Број предузећа	10	13	9	7
Коришћење докумената на више језика	2 (0.58)	4 (0.2)	4 (0.49)	4 (0.63)
Коришћење докумената за које не постоји одговарајућа верзија у електронском облику	2 (0.61)	2 (0.4)	3 (0.74)	4 (0.36)
За интерну комуникацију у писаном облику користе се:				
електронска пошта	4 (0.5)	4 (0.5)	4 (0.9)	4 (1)
преносиви уређаји за складиштење података (CD, USB, HDD, ...)	4 (0.48)	4 (0.5)	3 (0.67)	4 (0.87)
писма и факс	1 (0.58)	2 (0.39)	1 (0.38)	3 (0.67)
клауд сервиси	2 (0.58)	1 (0.32)	1 (0.38)	1 (0.52)
платформе за колаборацију	1 (0.54)	1 (0.51)	1 (0.43)	1 (0.78)
За екстерну комуникацију у писаном облику користе се:				
електронска пошта	4 (0.78)	4 (0.56)	4 (0.9)	4 (1)
преносиви уређаји за складиштење података (CD, USB, HDD, ...)	2 (0.45)	4 (0.47)	3 (0.61)	4 (0.74)
писма и факс	3 (0.48)	2 (0.43)	3 (0.49)	3 (0.74)
клауд сервиси	1 (0.54)	1 (0.35)	1 (0.27)	3 (0.5)
платформе за колаборацију	1 (0.54)	1 (0.47)	1 (0.49)	1 (1)
Документи који се користе у раду су:				
потпуно оригинални	4 (0.6)	4 (0.36)	3 (0.48)	4 (0.87)
типски са минималним изменама	2 (0.54)	2 (0.55)	2 (0.54)	2 (0.52)
За доношење одлука приликом управљања пројектом користе се:				
експертско знање и искуство руководиоца	4 (0.68)	4 (0.5)	4 (0.9)	4 (0.87)
савети и искуства из фирме	3 (0.64)	4 (0.43)	4 (0.74)	4 (0.87)
услуге консалтинга	3 (0.69)	3 (0.57)	3 (0.61)	2 (0.78)
структуирани подаци фирме	3 (0.53)	3 (0.34)	3 (0.67)	3 (0.78)
структуирани подаци ван фирме	2 (0.54)	3 (0.34)	3 (0.82)	2 (0.74)
неструктуирани подаци фирме	2 (0.73)	2 (0.64)	3 (0.6)	2 (0.53)
неструктуирани подаци ван фирме	3 (0.63)	2 (0.67)	3 (0.6)	1 (0.65)

	Категорија предузећа			
Питање	Микро	Мало	Средње	Велико
За обраду података значајних за доношење одлука користе се:				
оловка и папир	2 (0.52)	4 (0.28)	3 (0.41)	3 (0.72)
програми за рад са табелама	4 (0.51)	4 (0.38)	4 (0.32)	4 (0.78)
стручни софтверски алати	4 (0.43)	4 (0.28)	4 (0.48)	3 (0.72)
специјализовани алати за управљање пројектима	3 (0.45)	1 (0.27)	3 (0.69)	3 (0.67)
web апликације	2 (0.58)	2 (0.52)	1 (0.25)	2 (0.7)
Недостаци који се јављају у комуникацији:				
корисници добијају непотпуне информације	2 (0.54)	2 (0.64)	1 (0.37)	2 (0.67)
корисници добијају непотребне информације	2 (0.54)	2 (0.74)	1 (0.43)	2 (0.72)
корисници не добијају информације на време	2 (0.61)	2 (0.7)	1 (0.49)	2 (0.72)
постоје уска грла у токовима комуникације	2 (0.61)	2 (0.67)	1 (0.49)	2 (0.72)
не поштују се дефинисани токови комуникације	2 (0.61)	1 (0.69)	1 (0.49)	2 (0.74)
Препреке за прелазак на напреднији алате за обраду података у процесу доношења одлука:				
недостатак интерних ресурса	2 (0.52)	2 (0.59)	1 (0.4)	1 (0.44)
некомпатибилност са постојећим пословним процесима	2 (0.64)	3 (0.51)	1 (0.77)	2 (0.53)
недостатак функционалности	2 (0.83)	1 (0.59)	3 (0.61)	2 (0.83)
могуће грешке и проблеми са безбедношћу	2 (0.61)	1 (0.52)	1 (0.54)	2 (0.83)
Препреке за оптимално коришћење неструктуираних података у процесу доношења одлука:				
значајни подаци се налазе на различитим местима	3 (0.6)	1 (0.36)	3 (0.54)	2 (0.78)
значајне податке је тешко одвојити од осталих	2 (0.69)	1 (0.43)	3 (0.44)	2 (0.72)
подаци нису у формату за претрагу и анализу	2 (0.9)	1 (0.47)	2 (0.56)	2 (0.67)

2.2 Анализа анкета

Анализа резултата анкете *A2015* указује на следеће закључке:

1. Најзначајнији поступак у раду са документима је *коришћење електронске поште* (висок консензус по свим групама предузећа).
2. Сви испитаници су, уз висок консензус, истакли значај *архивирања* и *поштовања претходно дефинисаних процедура* у раду са документима.
3. *Претрага података на интернету*, је оцењена као веома значајна или значајна (за велика предузећа).
4. *Могућност увида у документацију на терену* је оцењена као веома значајна или значајна (за велика предузећа).
5. Микро (мањи консензус) и мала предузећа дају већи значај напредним опцијама за рад са документима (аутоматско додавање мета-података).

   Претпоставља се да је недостатак запослених у мањим предузећима, који би били ангажовани на ручном обележавању и груписању документа, главни мотив за повећану заинтересованост.
6. *Ручно лабелирање докумената* је оцењено као значајније од средњих и великих предузећа, уз виши консензус.

   Претпоставља се да заинтересованост за иновативне поступке опада са порастом броја запослених.

Уочени трендови су даље анализирани у анкети *A2016*, са фокусом на протоколе комуникације и обраде података при доношењу одлука. Анализа резултата анкете указује на следеће закључке:

1. Са порастом броја запослених *расте* удео докумената на *више* језика, као и докумената *без* одговарајуће верзије у електронском облику.
2. У свим предузећима комуникација се најчешће обавља преко *електронске поште* и *преносивих уређаја* за складиштење података.

3. Писма и факс су *значајно мање* заступљени у интерној у односу на екстерну комуникацију.
4. Највише се користе *потпуно оригинални* документи у којима запослени креирају целокупан садржај. Типски документи са предефинисаним шаблонима се користе у *значајно мањој* мери.
5. Приликом доношења одлука, све групе предузећа се највише ослањају на сопствена *искуства и знања* запослених, док се у значајано *мањој* мери користе неструктуирани подаци.
6. За обраду података од значаја за доношење одлука највише се користе програми за *рад са табелама*. У овој групи питања изражен је *мали* консензус, посебно код малих и средњих предузећа.
7. Код малих и микро предузећа, прелазак на напредније алате отежава *недостатак* сопствених ресурса и *некомпатибилност* са постојећим пословним процесима.
8. Корисне информације у оквиру неструктуираних података налазе се на *различитим местима*, што отежава њихово коришћење.

Анкете указују да грађевинска предузећа у Републици Србији највећи део пословних процеса везују за размену информација *електронском поштом*. За пословну комуникацију испитаници користе *различите технологије* и сматрају да се она обавља *без већих проблема*[4]. Према резултатима анкете, у предузећима се не користе у значајној мери *напреднији алати* за управљање пројектима. Простор за *побољшање* процеса доношења одлука је у *бољој искоришћености доступних података*. Истраживање показује да би *обједињавање* раздвојених неструктуираних података, који представљају значајну информацију, омогућило корисницима да их боље искористе.

[4] Ауторово субјективно мишљење је да испитаници не показују у значајној мери негативан став према предузећу, или постојећим пословним навикама.

3 Постојећи системи за рад са документима у оквиру којих се изводе информације за потребе управљања пројектом

У овом поглављу ће бити анализирани постојећи системи за рад са документима који настају током животног циклуса пројекта, са становишта могућности издвајања информација неопходних за управљање пројектом. Биће приказани уобичајени поступци за организовање издвојених информација у репрезентације погодне за анализу и извођење нових знања са пројекта.

У општем случају постојећа решења се могу категорисати према нивоу структуираности података са којим манипулишу:

- *системи датотека* – неструктуирани подаци;
- *системи за управљање информацијама на пројекту* – полуструктуирани подаци;
- *експертски системи засновани на онтологијама* - структуирани подаци.

Приликом издвајања и повезивања релевантних информација из различитих неструктуираних извора, неопходно је спровести одговарајуће активности у сваком од наведених система. Активности, које треба спровести у оквиру наведених система, биће илустроване на примеру добијања информација о *кашњењу* током извођења радова (*Кашњење*):

Кашњење: Приказати све позиције на пројекту на којима је било кашњења. Приказати све описе позиција, одговорне учеснике, фазу пројекта, тип радова, трајање кашњења и разлоге због којих је дошло до кашњења.

Овакви и слични скупови повезаних информација од значаја , под условом да се могу издвојити у току реализације или након завршетка пројекта, могу значајно умањити ризик од будућих кашњења или омогућити да се позицијама које касне боље управља. Делови информација које се односе на задати пример су у општем случају организоване на следећи начин:

- У документима типа *Извештај*[5] (недељни или месечни) дате су *шифре позиција* са изведеним радовима, у којима се наводи *одговорна компанија*, да ли је било *кашњења* и колико је њихово *трајање*.
- На основу познате *шифре*, из документа *Списак позиција* могуће је одредити *опис* и на који *тип радова* се позиција односи.
- *Фаза пројекта* није експлицитно наведена ни у једном документу.
- *Разлози за кашњење* су наведени имплицитно, кроз документе различитих типова (*Записник са састанака*, *Преписка*, итд.)

Претпоставка је да су подаци о позицијама у документима типа *Списак позиција* и *Извештај* организовани као табеле, а да су разлози за кашњење у форми отвореног текста, где се о једном истом догађају може говорити у различитим документима.

Потребно је истаћи да је организација докумената у одговарајуће логичке целине основни предуслов за коришћење било ког од наведених приступа за манипулацију неструктуираним подацима. У зависности од потреба компаније, потребно је дефинисати типове докумената, конвенције за именовање, унутрашњу структуру докумената и сл. Систем датотека, који ће бити приказан у наставку, је најопштији и најједноставнији приступ који корисницима омогућава да слободно организују документацију на пројекту.

[5] У наведеном примеру се под наглашеним речима које почињу великим словом подразумева тип документа (*Извештај*), а нагалшеним речима које почињу малим словом атрибут позиције (*компанија*).

3.1 Систем датотека

Без обзира да ли су у форми штампаних докумената или у електронском облику, документи са пројекта морају да се организују како би могли ефикасно да се користе. Најједноставнији, мада и најограниченији са становишта функционалности, приступ за организовање документа је коришћење одговарајуће директоријумске структуре на диску која моделира логичке целине на пројекту – *система датотека*.

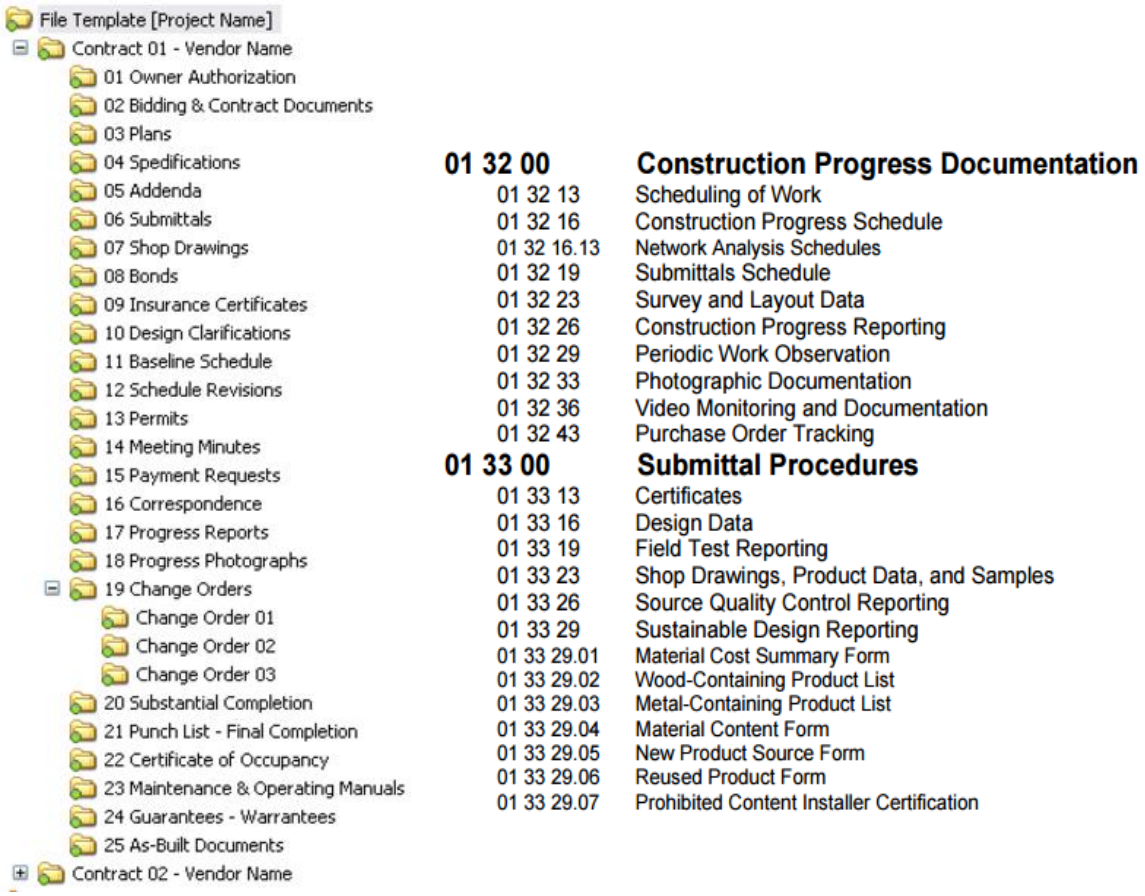
Корисници имају пуну слободу у одређивању категорија и поткатегорија које дефинишу директоријуме: на слици 3.1 лево је приказана директоријумска структура организована према препорукама из (Civitello 2000), заснована на уговорним обавезама учесника. Структуру је могуће другачије организовати према активностима на пројекту, где се могу користити претходно дефинисане спецификације, попут MasterFormat[6] стандарда, које дефинишу стандардизоване информације по типовима пројекта (слика 3.1 десно).

Поред дефинисања директоријумске структуре, у систему датотека пожељно је кодификовано именовање докумената. Корисници могу самостално да одреде одговарајућу конвенцију која у општем случају име документа дефинише као скуп информација од интереса за појединачни документ. Пример обрасца за име документа:

datum_tip_kompanija_ime_verzija, где је:

datum – датум креирања документа
tip – тип документа (нпр. *Записник са састанка*)
kompanija – компанија запосленог који је креирао документ
ime – назив документа који описује његов садржај
verzija – број ревизије документа

[6] Стандард за организацију техничких спецификација у градитељству - Construction Specifications Institute, http://www.masterformat.com/



Слика 3.1: Два приступа у организовању датотека - систем заснован на *уговорним обавезама* учесника (лево) и спецификација за организацију датотека према *активностима* на пројекту у зависности од његовог типа (десно).

Складиштење докумената према усвојеној структури и усвајање конвенције за њихово именовање, треба да омогуће корисницима да једноставно пронађу жељени документ коришћењем неког од програма за прегледање датотека. У случају када је потребно пронаћи неки садржај унутар документа, а корисник не зна унапред у ком се документу тражени садржај налази, примењују се програми за *текстуалну претрагу* (Lu et al. 2007).

Када корисници самостално дефинишу систем датотека, он је у општем случају централизован на једној локацији (локална радна станица или сервер), те

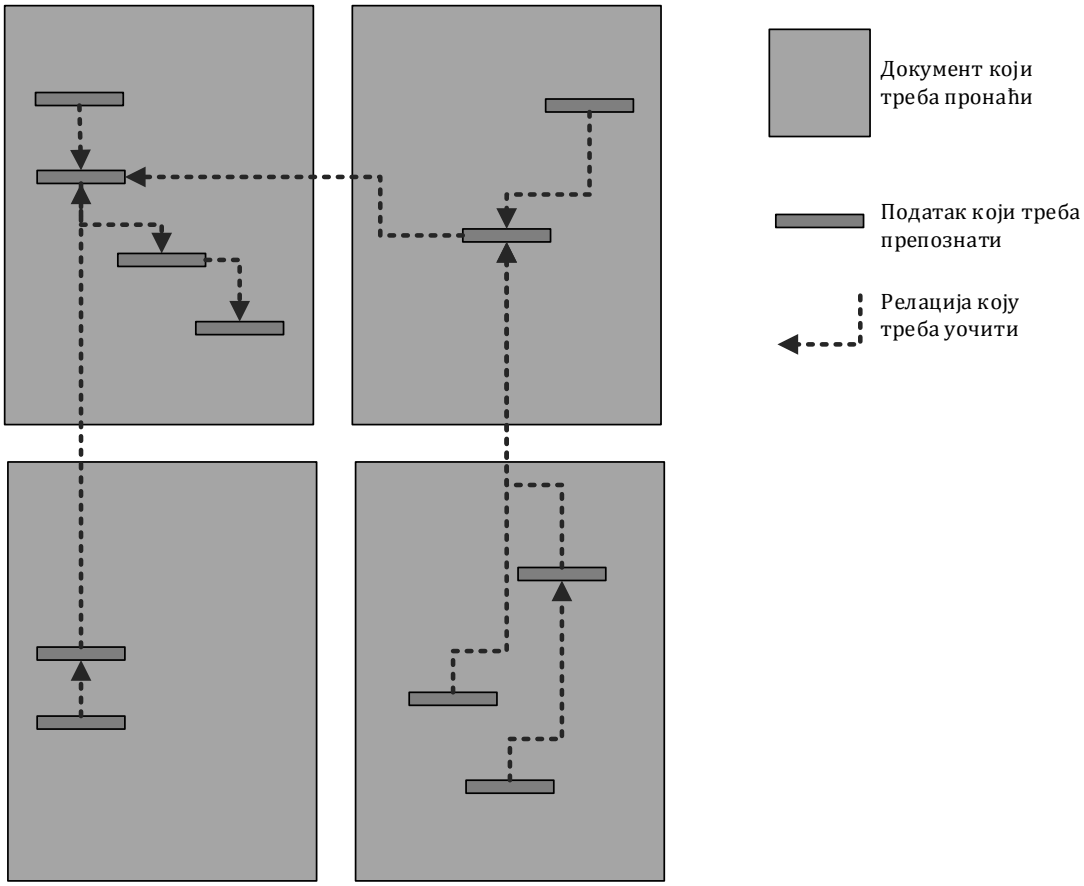
се могу користити решења специјализована за текстуалну претрагу датотека. Текстуална претрага је један од најраспрострањенијих концепата који се користи у рачунарским системима и овде се неће детаљно објашњавати. Потребно је напоменути да програми за претрагу текста могу имати широк спектар функционалности – од претраге по задатим кључним речима, до детекције ентитета попут броја, датума, адресе ел. поште и сл. Међутим, без обзира на комплексност и доступне функционалности, у општем случају резултат претраге је документ или део документа који корисник мора *самостално да интерпретира* – да утврди релевантност резултата за задати упит и да га потом стави у контекст са осталим прикупљеним информацијама.

У случају да је доступна само текстуална претрага над системом датотека, поступак за издвајање информација из примера *Кашњење* би био следећи (слика 3.2):

1. Корисник треба да зна како да дефинише и пронађе документе типа *Извештај* и *Списак позиција* (употребом одговарајућих кључних речи).
2. Свакој позицији из *Извештаја* која касни треба придодати одговарајуће податке (*одговорна компанија*, *трајање*)
3. Пронаћи *описе* и *типове радова* за позиције које касне из *Списка позиција.*
4. Имплицитно извести *фазу пројекта* кроз анализу осталих позиција из *Извештаја.*
5. Пронаћи документе који описују околности које доводе до кашњења и упарити их са одговарајућим позицијама. Корисник треба *самостално* да препозна концепт кашњења у документима.

Може се запазити да све ставке у одређеној мери захтевају да корисник буде *упознат* са процесима на пројекту (задатак је тешко решив за особу која није упозната са пројектом!). Ставка (5) је најкомпликованија јер захтева да се на њој ангажује учесник на пројекту који се директно бавио проблемима кашњења

(састављао дописе, дискутовао на састанцима и сл.). Чак и када су ти учесници доступни (што често није случај), морали би мануелно да обраде велику количину текста да би пронашли одговарајуће делове информација за придруживање позицијама. Могући поступак за решавање ставке (5) би био да се, за једну позицију, над целокупним корпусом текста на пројекту изврши текстуална претрага за најинформативније делове из њеног описа. Добијене резултате је потребно додатно филтрирати према томе да ли имају везе са кашњењем.



Слика 3.2: Поступак издвајања знања са пројекта када је доступна текстуална претрага у систему датотека. Тип докумената је имплицитно дефинисан његовим местом у директоријумској структури или његовим именом. Потребно је пронаћи одговарајуће документе, релевантне податке унутар њих, као и повезати их на одговарајући начин.

Може се закључити да у општем случају:

- Сваки резултат претраге мора да се интерпретира од стране корисника што успорава поступак и повећава могућност грешке;
- Није могуће аутоматизовати поступак издвајања информација, нити поново користити претходно дефинисане упите;
- Информације издвојене из различитих датотека корисник мора самостално да повеже.
- Полуструктуиране информације у документима (у форми табела) се не могу аутоматски издвојити.

Описани приступ је адекватан за мање комплексне пројекте где се јављају типске ситуације, а битни корисници су укључени у већину активности на пројекту. Са порастом комплексности пројекта, обима документације и усложњавањем знања које је потребно издвојити, систем датотека није ефикасан и препоручљиво је коришћење напреднијих решења.

3.2 Системи за управљање информацијама

У овој тези се под кровним термином *Системи за управљање информацијама* подразумевају решења за аутоматизацију поступака прикупљања, обраде и дистрибуције информација на пројекту из различитих извора[7].

[7] Преглед и рангирање софтвера за управљање информацијама: https://project-management.zone/

Према природи информација које обрађују, решења се могу поделити на софтвере за:

- Управљање електронским документима
  (Document management system – DMS)
  (Chassiakos & Sakellaropoulos 2008), (Björk 2002)
- Управљањем садржајем у предузећу
  (Enterprise content management – ECM) (Moses et al. 2008)
- Планирање ресурса предузећа
  (Enterprise resource planning – ERP) (Voordijk et al. 2003)

Како се ради о познатим концептима из савремене пословне праксе, они неће бити појединачно описивани. У овом истраживању наведена решења посматрају се кроз функционалност рада са неструктуираним подацима, иако се базирају на фундаментално различитим идејама (нпр. ERP решења се примарно баве пословним процесима, док су DMS решења специјализована за рад са документима).

Укључивање неког од наведених решења на комплексном пројекту обезбеђује да неструктуирани документи и подаци у њима буду боље описани *релевантним мета-подацима*. Алати који то обезбеђују су:

- мануелно додељивање ознака постојећим документима (приликом уноса у систем или након претраге),
- мануелно успостављање веза између постојећих докумената (на основу логике пословних процеса),
- коришћење претходно дефинисаних образаца за креирање типских докумената у оквиру система (омогућава касније аутоматско процесирање према предефинисаним правилима).

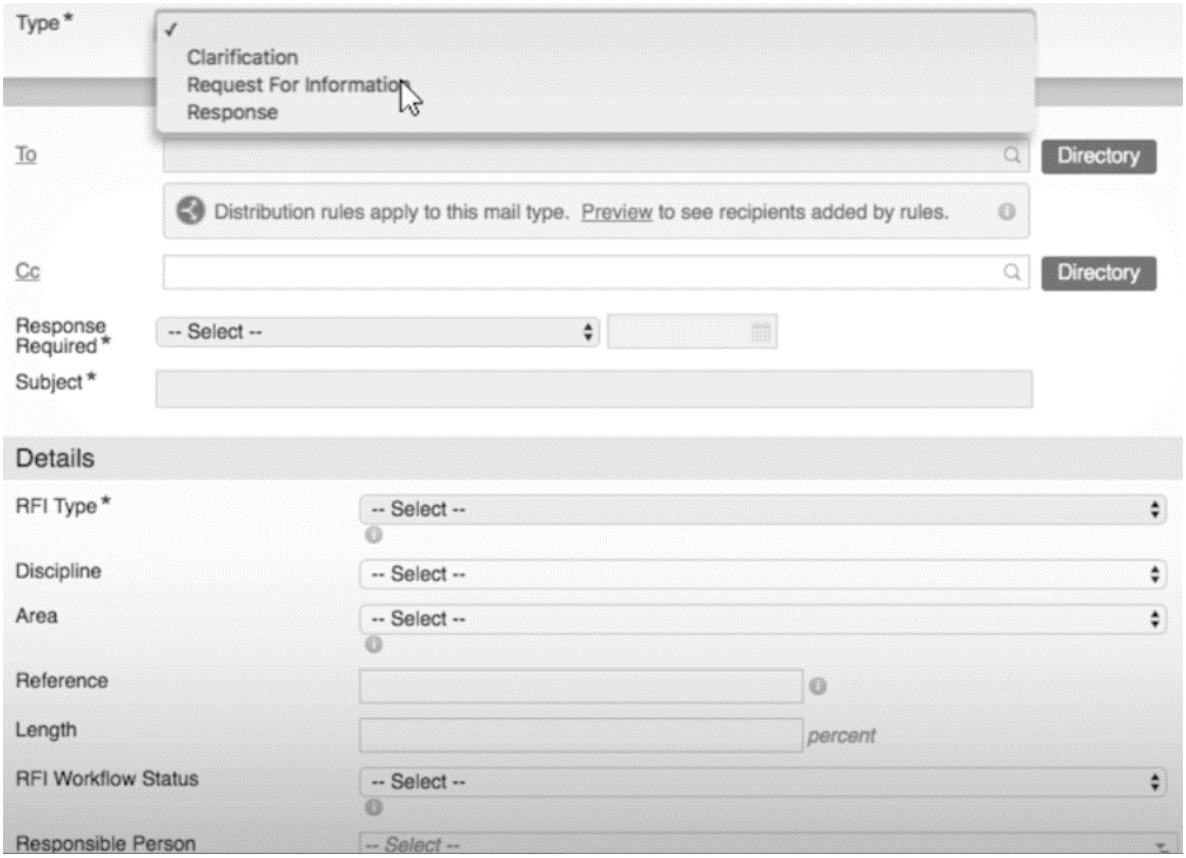
Захваљујући додатим мета-подацима омогућено је једноставније идентификовање релевантних информација и њихово повезивање, што значајно

олакшава поступак издвајања потребног знања са пројекта. На пример, могуће је издвојити податке о електронској пошти (наслов, датум, пошиљалац, прималац, прилози, итд.) и организовати их у облик погодан за претрагу и анализу (слика 3.3).

Subject	Date	From	From Organization	Recipients	Status	Type	Attachments
Re: Emi-6 - Reduced Flow to Sewer	2011-02-22	Lewis Miller	AJ Hydraulic Services	Tim Yeung	N/A	Consultants Advice Notice	
Re: Emi-6 - Reduced Flow to Sewer	2011-02-22	Lewis Miller	AJ Hydraulic Services	Tim Yeung	Responded	Variation Request	
Re: Conf docs	2016-09-21	Lewis Miller	AJ Hydraulic Services	James Wong	Outstanding	Consultants Advice Notice	pdf-sample.pdf
PLease register these attachments	2014-10-14	Jamal Abbudin	Arif Consulting	Patrick O'Leary	N/A	Credit Interpretation Request	OutlookEmail.msg rr.png Chrysanthemum.jp Desert.jpg Hydrangeas.jpg Jellyfish.jpg Koala.jpg Lighthouse.jpg New TabNew TabNe
Invoice - Logitech H34011	2014-10-09	Gretchen Pitcher	Ashton Design & Drafting	Patrick O'Leary	N/A	Contractors Advice	OutlookEmail.msg 20141007_15220 20141007_15244
Notification of licence request	2014-10-16	Gretchen Pitcher	Ashton Design & Drafting	Patrick O'Leary	N/A	Contractors Advice	OutlookEmail.msg
Fwd: testing for missing element	2012-04-16	Patrick O'Leary	Majestic Builders		Outstanding		
	2014-09-17	Patrick O'Leary	Majestic Builders		N/A	Internal Memorandum	
Re: Test_04	2014-10-15	Patrick O'Leary	Majestic Builders	Patrick O'Leary	N/A		
	2014-10-15	Patrick O'Leary	Majestic Builders		N/A	Defect	
	2015-05-20	James Wong	Majestic Builders	Abdul Hussein	Outstanding	Defect	
	2015-05-20	James Wong	Majestic Builders	Abdul Hussein	Outstanding	Defect	
	2015-05-20	James Wong	Majestic Builders	Abdul Hussein	N/A	Defect	
	2015-05-20	James Wong	Majestic Builders		N/A		

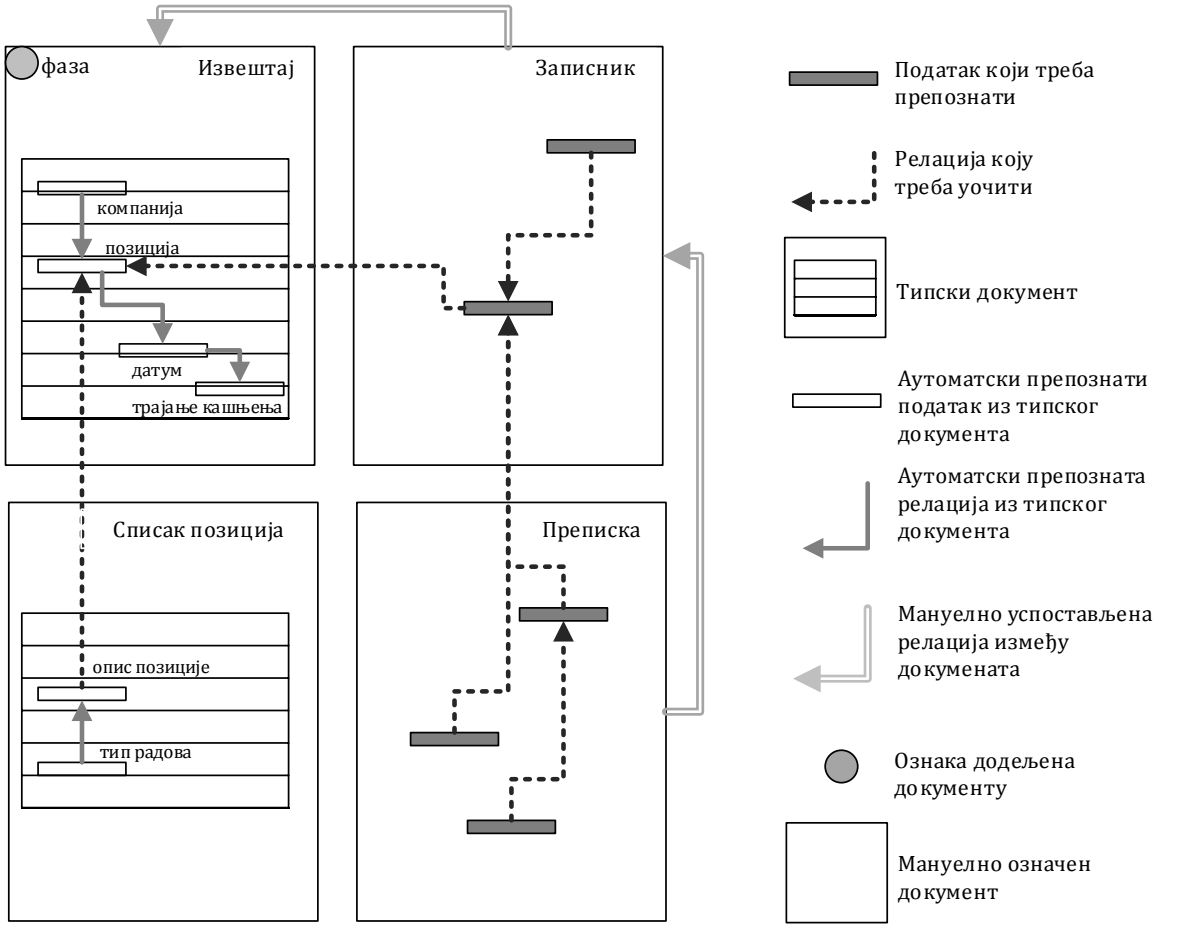
Слика 3.3: Пример аутоматски издвојених података из пословне кореспонденције. Тип поруке (претпоследња колона) се изводи аутоматски на основу претходно дефинисаних правила. Извор: *rampiva.com*

Ако су типски документи организовани по претходно дефинисаним обрасцима (слика 3.4), приликом креирања, подаци унети у њих се аутоматски издвајају и повезују (нпр., из типског документа *Извештај* могу се препознати подаци о позицији и њихов међусобни однос). Могуће је дефинисати правила по којима се документима аутоматски додају ознаке (*фаза пројекта* из *Извештаја*).



Слика 3.4: Пример претходно дефинисаног обрасца за документ *Захтев за информацијом*. Дефинисани су атрибути *тип*, *област*, *статус*, *одговорна особа*, итд. Извор: *aconex.com*

У оваквом систему би, приликом спровођења задатка из примера *Кашњење,* поред текстуалне претраге, на располагању биле и информације о документима попут оних приказаних на слици 3.5.



Слика 3.5: Поступак издвајања знања од стране корисника у окружењу неког од система за управљање информацијама. Документи имају препознате типове, уз могућност обележавања додатних категорија (*фаза пројекта*). За предефинисане обрасце могуће је аутоматско издвајање информација (нпр., ако *Извештај* садржи типску табеларну структуру у коју се уносе *компанија*, *позиција*, *датум* и *трајање кашњења*). Ако је за посматрани процес дефинисан радни ток, приликом уноса у систем између докумената се успостављају релације.

Поступак решавања задатка из примера *Кашњење* би, у окружењу система за управљање информацијама, био следећи:

1. Корисник може одмах да приступи *Списку позиција* и *Извештају* коришћењем филтера типа документа у текстуалној претрази.
2. Ако је *Извештај* типски документ, сви подаци о позицијама су издвојени у структуру коју је могуће анализирати и филтрирати, укључујући и *фазу* која је генерисана аутоматски.
3. Ако није позната веза између табела у различитим типским документима, треба *ручно* пронаћи описе и типове радова за позиције које касне из *Списка позиција.*
4. Документи који описују околности од интереса за позиције су, у општем случају, повезани са одговарајућим извештајима. Међутим, корисник и у овом случају треба да зна како да препозна концепт кашњења у повезаним документима.

Задатак из примера *Кашњење* је сада, у односу на пројекат на коме се користи систем датотека, једноставнији – значајан део података је додатно обележен и повезан, те је проналажење и повезивање информација у структуру којом се репрезентује знање једноставније. Рад са типским документима и припадајућим подацима може се делимично аутоматизовати (чување претходно дефинисаних упита, креирање скрипти са правилима из пословног процеса). Горе наведене околности омогућавају да се, за проналажење релевантних информација везаних за ставке 1-3, ангажују учесници са *мањим степеном знања* о конкретном пројекту.

Међутим, за ставку 4 и даље је неопходна интерпретација релевантних информација из неструктуираног текста које описују разлоге за кашњење. Иако је сада простор који треба ручно претражити мањи (јер су у општем случају познате везе између докумената), неопходан је учесник који је директно упознат

са проблемима кашњења, а коме на располагању стоји обична текстуална претрага.

Системи датотека и системи за управљање информацијама, у општем случају, омогућавају да се релативно једноставно издвоје потребне информације из полуструктуираних података. Тако издвојене информације, заједно са структуираним подацима (различите базе, регистри и сл.), експерти користе како би закључили *шта се дешава на пројекту*. У случају комплексних међународних инвестиционих пројеката, системи за управљање информацијама могу значајно олакшати процес закључивања у односу на приступ са системом датотека. Међутим, одговор на питање *зашто се нешто дешава* и даље захтева значајно ангажовање експерта како би се правилно интерпретирали *неструктуирани подаци* и пронашао одговор.

У наставку ће бити приказани различити експертски системи који преводе неструктуиране податке у одговарајуће репрезентације које омогућавају богатију анализу и закључивање.

3.3 Експертски системи засновани на онтологијама

Да би се неструктуирани подаци из текста превели у структуиране информације из којих се може изводити ново знање, неопходно је:

- дефинисати формалну спецификацију репрезентације у коју се подаци преводе,
- детектовање релевантних информација из докумената и њихово пресликавање у одговарајуће категорије из формалне спецификације.

У истраживањима везаним за област управљања пројектима у грађевинарству, најчешће коришћен приступ је дефинисање *онтологија* којима се формално описује посматрани проблем. У домену информационих технологија, онтолгија је дефинисана као спецификација концепата и релација између њих који могу формално да постоје за агента или групу агената -

спецификација концептуализације (Gruber et al. 1995). У претходној дефиницији, под агентом се подразумева аутономни систем који може да извлачи закључке. Нешто релаксиранија дефиниција представља онтологију као „опис појмова и начина на који су они повезани“ (Welty 2003).

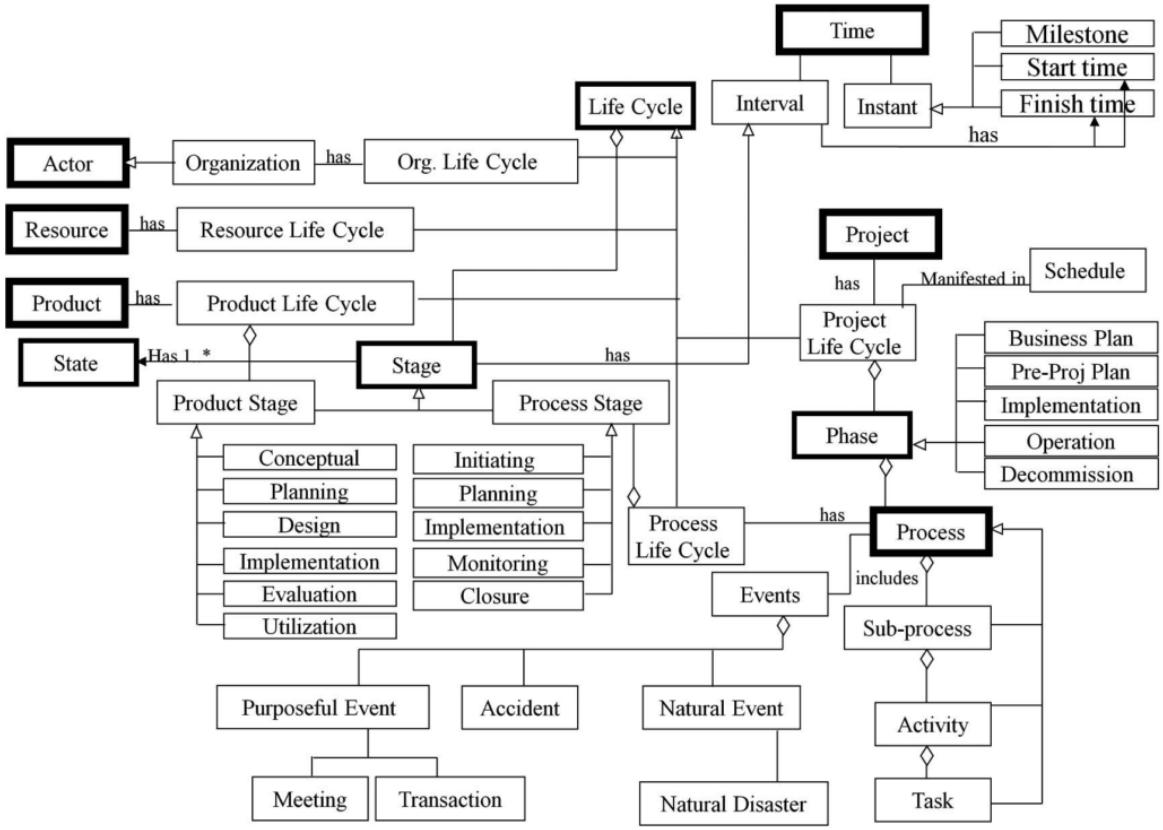
На слици 3.6 је приказан део онтологије *Пројекта* преузетог из (El-Diraby 2012), где су дати значајни концепти *Акције*, *Процеси*, *Учесници*, итд., заједно са појмовима који их додатно дефинишу и њиховим међусобним релацијама. Могу се уочити одређене сличности са шемом базе података, која ће детаљније бити приказана у поглављу 8.1. Ипак, онтологија се суштински разликује од базе података (Horrocks 2013):

- у базама података сваки ентитет има јединствено име, док у онтологијама ентитет може имати више имена;
- у онтологији се могу изводити имплицитни закључци;
- у базама података важи претпоставка затвореног света (исказ је тачан само ако се зна да је тачан). У онтологијама важи претпоставка отвореног света (исказ може бити тачан иако се не зна да је тачан).

Преглед истраживања у области грађевинарства где су примењиване онтологије је дат у (Issa et al. 2015). Потребно је истаћи да је поступак дефинисања онтологије комплексан, услед чега се проблем издвајања информација из неструктуираних извора фокусира на специфичне потпроблеме:

- Безбедност на раду и записници о проблемима на градилишту (H.-H. Wang et al. 2011);
- Процеси и учесници на пројекту (El-Diraby 2012), (El-Gohary & El-Diraby 2010);
- Процена трошкова (Ma et al. 2016), (Lee et al. 2014);
- Провера усаглашености прописа (Yurchyshyna & Zarli 2009);
- Ланац снабдевања (Pandit & Zhu 2007);

- Процена количина материјала за давање понуде (Quantity take off) (Liu et al. 2016);
- Управљање ризиком (Ding et al. 2016).



Слика 3.6: Приказ дела онтолошког модела за *Пројекат*, са фокусом на концепт *Процес*. Концепти се представљају као примарни ентитети у моделу (подебљани правоугаоници) или секундарни ентитети (обични правоугаоници). Релације са ◊ - *је_део*, релације са ▲ – *садржи*. Пример: *задатак* (правоугаоник *Task*) је део активности, *активност* је део потпроцеса, а *потпроцес* је део процеса, *процес* садржи животни циклус процеса (правоугаоник *Process Life Cycle*). Извор: (El-Diraby 2012)

Када је формирана одговарајућа онтологија, релевантне информације *препознају се* и *издвајају* из неструктуираних докумената, па потом *пресликавају* у одговарајуће онтолошке категорије. У ту сврху примењују се различите методе,

попут техника машинског учења, анализе текста, обраде природног језика и других метода вештачке интелигенције. Треба истаћи да су наведени поступци препознавања веома комплексни и да се, уколико је захтевана висока тачност пресликавања, спроводе тако што се дефинишу *типски документи* са *предефинисаним форматом*, који олакшавају препознавање одговарајућих ентитета.

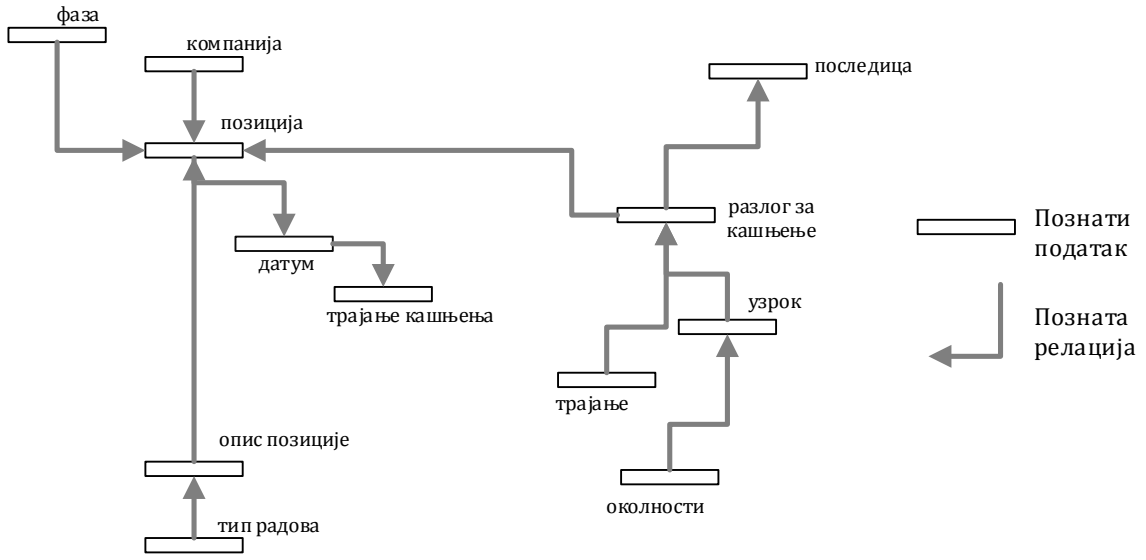
Крајњи резултат је база знања за уско дефинисани домен, која се може анализирати одговарајућим језиком за задавање упита[8] (слика 3.7).

```
PREFIX ontoCC: <http://our_ontology.owl#>
SELECT ?portique display xml
where { ?portique rdf:type ontoCC:PortiqueSecurite
OPTIONAL
{ ?portique ontoCC:overallWidth ?width
FILTER ( xsd:integer(?width) >= 80) }
FILTER (! BOUND( ?width) ) }
```

Слика 3.7: Упит у SPARQL језику којим се, у оквиру онтологије, испитује да ли ширина сигурносних врата може бити већа или једнака од 80цм. Извор: (Yurchyshyna & Zarli 2009)

Под претпоставком да постоји одговарајућа онтологија којом се моделирају позиције и кашњења, задатак из примера *Кашњење* би се решавао у окружењу које је потпуно независно од текстуалне претраге (слика 3.8). Корисник више не претражује неструктуиране податке него директно изводи релевантне чињенице из формиране и „напуњене“ онтолошке структуре. Могуће је директно добити одговор на сва питања од интереса, уз предуслов познавања модела и синтаксе језика којим се изводи ново знање.

[8] пример језика за упите у онтологијама:
SPARQL - Simple Protocol and RDF Resource Description Framework Query Language, https://www.w3.org/2009/sparql/wiki/Main_Page



Слика 3.8: Поступак издвајања знања у окружењу експертског система заснованог на онтологијама. Сви подаци од интереса су детектовани и представљени као појмови повезани одговарајућим релацијама. Концепт кашњења је у потпуности издвојен из различитих текстуалних извора и потом структуиран.

И поред заступљености у научној литератури и неоспорне адекватности примене онтологије за извођење знања из неструктуираног текста, треба имати на уму ограничења која спречавају ширу примену овог приступа. Сам процес конструисања и ажурирања онтологије је веома комплексан и захтева мултидисциплинарна знања и највиши ниво разумевање самог проблема који се моделира. Поред тога, неопходно је дефинисати процесе којима се, из природног језика у писаној форми, препознају ентитети и концепти које је потребно издвојити и додати онтологији. Услед недостатка комерцијалних решења заснованих на овом приступу, компаније би морале интерно да дефинишу и одржавају овакве системе – што је, и поред евидентних користи, за велику већину компанија из грађевинског сектора данас недоступно.

3.4 Технике за аутоматско издвајање информација из неструктуираног текста

Независно од приказаних поступака, информације из неструктуираног текста се могу издвајати применом различитих техника попут *регуларних израза*, *машинског учења* или *обраде природног језика*. О регуларним изразима више речи биће у поглављу 5.5, када се буде излагао приступ којим се могу препознати и издвојити датуми.

Машинско учење бави се алгоритмима који решавају проблем на бази учења из претходног искуства. За алгоритам се каже да учи ако му се, за решавање одређене класе задатака, перформансе побољшавају са повећањем искуства (Mitchell 1997). Алгоритми машинског учења се, према начину учења, могу раздвојити на алгоритме за *нагледано* и *ненадгледано* учење.

Концепт надгледаног учења подразумева да постоје обележени примери за учење на којима се креира модел. Неки од најчешће примењиваних алгоритама за надгледано учење у области издвајања информација из текста су:

- Метод потпорних вектора (Support Vector Machines) (Peshkin & Pfeffer 2003), (Mahfouz 2011);
- Бајесове мреже (Li et al. 2005).

Алгоритми за ненадгледано учење изводе модел којим се проналазе скривене структуре из необележених података. Најчешће коришћени приступи у обради неструктуираног текста су:

- Кластерисање (Larsen & Aone 1999);
- Латентна семантичка анализа (Landauer et al. 1997).

Обрада природног језика (ОПЈ) је дисциплина која проучава алгоритме за разумевање природног језика и говора. Задатак алгоритама из области ОПЈ је да, применом формалних поступака из области рачунарске лингвистике, као што је *обележавање дела текста* (Derose 1988), реши следеће проблеме:

- Одређивање именованих ентитета (*локације*, *личности*, *организације*, *догађаји*, ...) (Nadeau 2007);
- Издвајање релација (*живи у*, *ради за, ...*) (Ramakrishnan et al. 2006).

У општем случају се подразумева да је претходно дефинисана структура у коју се подаци из текста преводе, мада се неке технике могу користити и када то није случај. Наведене структуре могу бити једноставне (типови према којима треба класификовати документе), преко сложених регулаторних правила ручно издвојених из докумената, до онтологија које описују одређени домен.

Један од првих покушаја аутоматске класификације неструктуиране пројектне документације коришћењем техника машинског учења је приказан у (Caldas et al. 2002), где су аутори поредили различите класификаторе текста. У (Al Qady & Kandil 2010), аутори су користили плитко парсирање да издвоје семантичко знање из уговорне документације. Слично, у (Kim et al. 2010), парсирани су документи са статичким прорачуном како би се издвојиле семантичке структуре. (Lin et al. 2012) су коришћењем доменске онтологије издвајали секције од значаја као независне документе, да би добили информације из области земљотресног инжењерства. Показано је да се добијеним сегментима побољшавају резултати претраге за дугачке документе са великим бројем значајних појмова. Семантичке асоцијације у еволутивној онтологији коришћене су да би се добили резултати релевантни за задате кључне речи (Costa et al. 2013). У (Al Qady & Kandil 2014), да би превазишли недостатак података за тренирање класификатора по свим могућим типовима документа, аутори су користили ненадгледано груписање (clustering), да би организовали документе у међусобно дисјунктне класе. Аутори у (Zhang & El-Gohary 2015) су предложили приступ

заснован на правилима за ОПЈ којим се аутоматски издвајају правила из уговорних докумената. (Fan et al. 2015) су, коришћењем речника специфичног за пројекат и техника ОПЈ, побољшали резултате претраге неструктуираних докумената на пројекту. Аутори су посебно истакли да је издвајање семантичких концепата и релација између њих, као и разумевање контекста у којима се јављају, и даље значајан изазов за проблем управљања документацијом на грађевинском пројекту.

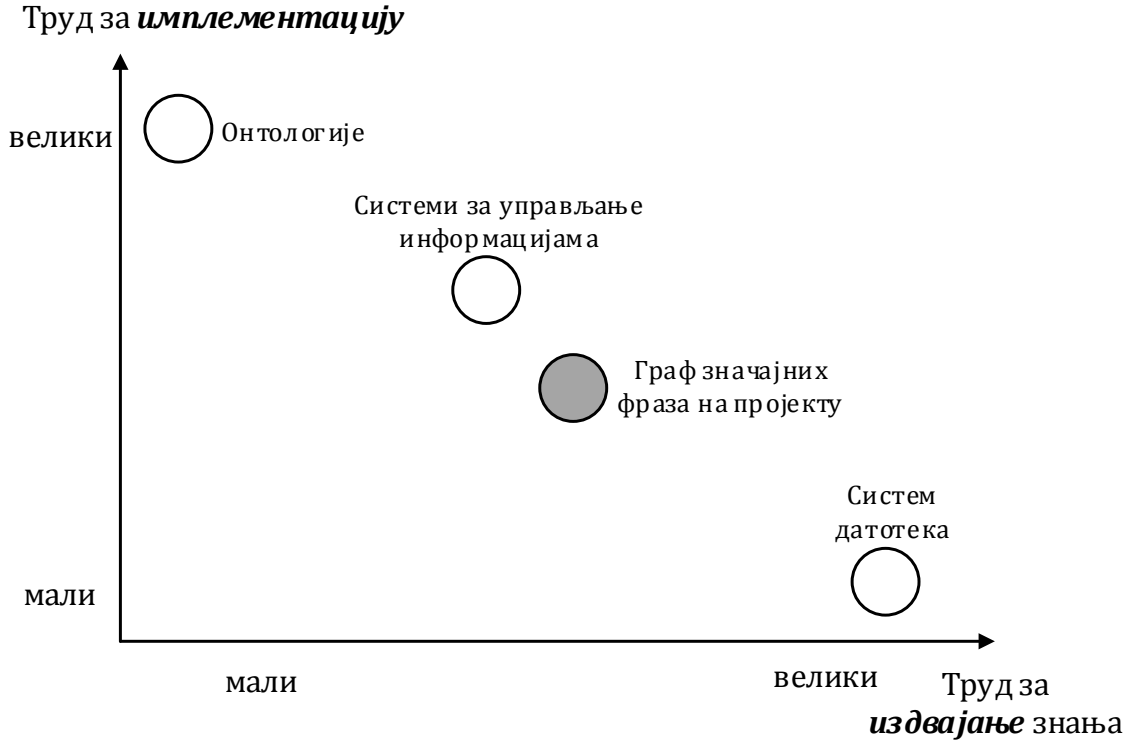
Већина приказаних поступака се у значајној мери ослања на *експерта који мора да дефинише правила* за издвајање значајних информација. Иако је показано да ангажовање експерта даје несумњиво боље резултате у односу на класичне технике претраге, *велики обим посла* за имплементацију ових приступа *значајно умањује* њихову применљивост у практичним проблемима.

3.5 Поређење постојећих система за рад са документима у погледу извођења знања из неструктуираног текста

У складу са претходно изнетим чињеницама, закључује се да постојећи системи за рад са документима, када је у питању издвајање знања из неструктуираних текстуалних података, не могу адекватно да одговоре на све специфичности окружења грађевинског пројекта. На слици 3.9 су приказани односи постојећих решења са становишта труда који је неопходно уложити за:

- почетно конфигурисање система;
- преносивост на различите домене,
- издвајање знања.

Систем датотека, заједно са текстуалном претрагом, је најједноставнији за имплементацију и може се одмах применити на било ком типу пројекта. Међутим, за проналажење и повезивање информација из различитих извора корисник мора да уложи значајан труд.



Слика 3.9: Релативни однос постојећих система и предложеног решења за издвајање и репрезентовање информација из неструктуираних података (које ће бити описано почевши од главе 5). Имплементација подразумева *почетно конфигурисање* система и *преносивост* на различите домене.

Системи за управљање информацијама представљају постојећи стандард у раду са неструктуираним подацима. Помоћу њих корисници могу, уз одговарајући уложен труд, да пронађу информације од значаја и изводе ново знање. Овај приступ се може, уз одређене модификације, применити на различитим типовима пројеката. Ипак, од корисника се захтева или *улагање значајних средстава,* или *промена постојећих пословних процеса и навика*, што отежава примену овог приступа за велики број малих и средњих предузећа.

Решења заснована на онтологијама су оптимално решење са становишта могућности закључивања, али ограниченост на појединачне домене и висок степен ангажовања експерта у имплементацији, значајно отежавају њихову примену. Иако су за сада углавном ограничена на академске кругове, трендови

пораста количине неструктуираних података и развоја алгоритама за аутоматско издвајање информација, указују на могућност шире применљивости овог приступа у будућности.

У следећим поглављима описане су основне карактеристике предложеног приступа за издвајање релевантних концепата на пројекту, у виду *графа значајних фраза*. Треба нагласити да и ова структура *захтева интерпретацију* од стране крајњег корисника, али је однос између уложеног труда за успостављање система и труда при закључивању у току коришћења *повољнији*.

4 Предложени приступ за издвајање знања – претпоставке и ограничења

У истраживању се претпоставља да се *знање* може представити као *скуп детектованих концепата* из различитих категорија, *повезаних предефинисаним релацијама*, насталих из више *текстуалних извора* током времена. Предлаже се *аутоматски поступак за издвајање* значајних концепата из *неструктуираних* и *полуструктуираних* текстуалних извора. Поред тога, предлаже се и *начин њиховог организовања* у репрезентацију погодну за *визуелизацију* и *извођење знања* интерпретацијом од стране експерта. Одабир постојећих и дефинисање нових алгоритама и хеуристика учињен је тако да се предложена решења могу имплементирати у окружењу комплексних инвестиционих пројекта.

4.1 Уведене претпоставке

Предложени приступ за издвајање и организацију значајних концепата подразумева да у току животног циклуса пројекта важе следеће претпоставке:

1. *Значајан* део информација потребних за успешно управљање пројектом *налази се у документима* који циркулишу на пројекту;
2. Текстуални документи на великим пројектима су често *вишејезични*, садрже *делове који се понављају* у већем броју докумената и *нису у довољној мери описани мета-подацима*;
3. Постојећи софтвери за управљање пројектима *не подржавају* аутоматско издвајање знања из неструктуираних извора;
4. Значајни концепти су у тексту представљени као *секвенце речи – значајне фразе*;

5. *Статистичке мере* за одређивање *корелације* између речи у документу, или корпусу, чине *основу* за одређивање концепата у тексту, *независно* од језика;
6. Концепти се могу повезати у семантички богатије структуре на основу *сличности семантичког контекста* у коме се појављују, *без употребе* претходно дефинисаног експертског знања;
7. Погодна *визуелна репрезентација* значајних концепата, заједно са одговарајућим алатима за увид, *олакшава* проналажење скривених образаца и трендова;
8. *Графовске базе података* су погодније од релационих, за препознавање скривених образаца из података код којих је изражена *повезаност*.

4.2 Особине предложеног решења

Суштинска особина предложеног процеса за аутоматско издвајања релевантних концепта из неструктуираних текстуалних извора је да не зависи од претходно дефинисаног експертског знања (или да зависност буде минимална). На тај начин, поступак постаје применљив на различите домене проблема. Под релевантним концептима подразумевају се *значајне фразе*, дефинисане као *секвенце* од две или више речи, које су издвојене статистичким мерама за одређивање корелације (глава 5). Предложени су начини за комбиновање ових мера, као и поступак за отклањање *неинформативних* фраза које се јављају као последица понављања текстуалних образаца на пројекту (поглавље 5.3). Поред значајних фраза, могуће је издвојити релевантне податке који се увек јављају у истим обрасцима (нпр. датуми, поглавље 5.5). Издвојене фразе су повезане, на основу семантичких контекста у којима су се јавиле, релацијама које *не укључују* дефинисање претходних правила (глава 7). Добијене релације омогућавају формирање графа значајних фраза који се потом складишти у графовску базу података (поглавље 8.4). Графовска анализа, спроведена уз помоћ погодних упита над графовском базом (глава 9), омогућава да се открију комплексни концепти на пројекту који не одговарају простим секвенцама речи – изражено

повезани подграфови (поглавље 9.2). На тај начин, подаци из докумената су организовани у синтетисане информације, *без потребе* за мануелним претраживањем и склапањем појединачних делова информације, расутих по различитим изворима. Корисник тако добија резултате који су прилагођени посматраном проблему и може да их интерпретира у складу са претходним знањем и искуством (поглавље 9.3). Интерпретација је олакшана захваљујући могућности да се за све концепте *могу видети* текстуални контексти појављивања унутар изворних докумената.

Предложени приступ је *трансферабилан* јер не захтева значајне ресурсе и прилагођавања за употребу на различитим пројектима. У потпуности је прилагодљив постојећим пословним процесима корисника у раду са текстуалним изворима. Приступ је *независан у односу на језик* документа јер су поступци за детекцију основних концепата-фраза засновани на статистичким методама, које су универзално применљиве у сваком језичком окружењу.

4.3 Ограничења предложеног приступа и њихово превазилажење

Применљивост решења на различите домене проблема (пројекте) је обезбеђена тиме што се издвојени релевантни концепти аутоматски организују у структуру која не захтева претходну конфигурацију. Иако би постојање онтологије обезбедило потенцијално богатију репрезентацију информација, као и формалну логику за закључивање, такав приступ би био ограничен само на онај домен проблема покривен одговарајућом онтологијом. Репрезентација информација, добијена на предложени начин, се у потпуности ослања на *експертску интерпретацију* којом се, из структуираних релевантних концепата и веза између њих, *изводи* ново знање. Међутим, како у предметном приступу није дефинисан механизам којим се испитује валидност изведених закључака, могуће су грешке у интерпретацији услед *когнитивне пристрасности* (поглавље 9.5). У

поглављу 9.5 дискутује се о мерама којима се умањује могућност погрешне интерпретације добијених резултата.

Аналогно трансферабилности, функционисање у сваком језичком окружењу је обезбеђено тиме што се основни поступак, којим се издвајају значајне фразе, не заснива на методама обраде природног језика. За очекивати је да би фокус на неколико познатих језика, који су на одговарајући начин покривени ресурсима за напредну обраду текста, дао боље резултате од метода које су засноване само на статистичким мерама корелисаности речи. Предложени приступ узима у обзир ову чињеницу и дефинисан је тако да може да укључи одговарајуће језичке ресурсе (поглавље 5.2.2).

5 Аутоматска детекција значајних фраза из текстуалних извора

Релевантни концепти из домена инвестиционих пројеката у грађевинарству су махом комплексни – састоје се од две или више речи (нпр. *steel frame*, *лансирна решетка*, *reinforced concrete column*, *примарни кровни носачи*, *installation of concrete batching plant*, итд.). За аутоматско издвајање релевантних информација на пројекту примењен је поступак издвајања значајних фраза, дефинисан у (Turney 2000). Према (Hasan & Ng 2011), фактори који утичу на комплексност задатка издвајања значајних фраза су дужина документа, конзистентност структуре и корелација тема унутар докумената. Дужи документи генеришу већи простор за претрагу па је поступак издвајања из техничких извештаја или записника са састанака тежи у односу на електронску пошту. Са друге стране, конзистентна структура и корелација тема у техничким документима олакшава поступак издвајања.

У овом поглављу је приказан поступак издвајања релевантних концепата (*значајних фраза*). Значајна фраза *реда n* је дефинисана као низ од *n* суседних речи које означавају један појам. Предложен је поступак трансформације докумената и примене метода којима се проналазе оне секвенце речи који су *кандидати* за значајне фразе. Такође, предложене су технике којима се уклањају *неинформативни кандидати*, као и комбиновање са техникама обраде природних језика (ОПЈ), којима се могу повећати перформансе приказаног поступка.

5.1 Формирање хијерархијске репрезентације текстуалних докумената

Документе из којих је потребно издвојити неструктуирани садржај је могуће посматрати као скуп међусобно независних речи - тзв. *вреће речи* (bag of words). Иако је оваква репрезентација валидна у применама попут класификације текста (Cao & Liang 2011), у предложеном приступу се документ представља као уређена секвенца речи. Да би се олакшало проналажење значајних фраза, секвенце речи подељене су на семантичке целине попут *реченице* или *параграфа*. Документи се, уз помоћ парсера[9], трансформишу из изворних формата, у облик погодан за детекцију семантичких целина. Овај поступак назива се *парсирање*. Том приликом, они се преводе у одговарајућу HTML репрезентацију (слика 5.1 горе) која чува структуру документа. Ограничавање простора за претрагу значајних фраза, на нивоу параграфа, умањује број могућих кандидата, без губитка оних који заиста представљају значајне фразе. Саме параграфе је могуће даље разложити на *реченице* коришћењем одговарајућих синтаксичких правила (слика 5.1 доле). Постојећа решења показују да сегментација на нивоу параграфа (Denoyer et al. 2001), или реченица (Ko et al. 2004) побољшава успешност поступка издвајања информација из текста.

[9] Парсер је софтверска компонента која екстрахује текстуалну садржину из формата попут *.docx*, *.pdf*, *.xlsx* и других, у текстуални формат. У истраживању је коришћен Tika парсер (https://tika.apache.org), са могућношћу трансформисања документа у HTML формат.

```
▼<body>
  ▼<div>
     <h1>Cover Sheet</h1>
    ▶<table>…</table>
    ▼<p>
       "GRAĐEVINSKI FAKULTET UNIVERZITETA U BEOGRADU FACULTY OF CIVIL ENGINEERING
       UNIVERSITY OF BELGRADE PROJECT : NEW FLASH FURNACE AND SULPHURIC ACID PLANT
       CONFERENCE NOTES &G "
     </p>
     <p>&"Arial,Italic"&18&F &"Arial,Italic"&18Page &P of &N </p>
   </div>
  ▼<div>
     <h1>MoM</h1>
    ▼<table>
      ▼<tbody>
        ▼<tr>
          ▼<td>
             "Stručni nadzor je naglasio da obzirom na izloženu problematiku od strane
             izvođača treba sprovesti sledeće aktivnosti:"
           </td>
         </tr>
        ▼<tr>
          ▼<td>
             "Problematika vezana za nadoknadu kašnjenja i ostvarenje navedenog plana
             uslovljena je i aktivnostima izvođača, pa je neophodno i njihovo prisustvo na
             sastancima."
           </td>
         </tr>
        ▶<tr>…</tr>
```


Слика 5.1: Структуираност документа: у HTML репрезентацији (горе), поједине целине раздвојене су одговарајућим лабелама – на пример, садржај ћелије табеле смешта се између *<td>* и *</td>*. Различити семантички контексти у репрезентацији документа – низ речи, параграфи и параграфи са реченицама (доле).

5.2 Издвајање значајних фраза

Претпоставља се да је већина битних концепата из пројектне документације представљена у облику значајних фраза реда два – релевантних *парова речи*[10]. Квалитет предложене репрезентације о којој ће бити речи у глави 7 – *графа међусобно повезаних фраза*, у највећој мери зависи од исправне детекције значајних фраза. Реченице представљају *најмању* семантичку јединицу која дефинише границе за детекцију значајних фраза.

Комплексни међународни пројекти у земљама у развоју обично садрже обиман корпус докумената на језику земље домаћина. По правилу, за већину тих језика не постоје адекватни алати за обраду природног језика (ОПЈ), па се за поступак проналажења значајних фраза *предлаже коришћење статистичких мера* за одређивање међусобне повезаности речи. Ипак, предложени приступ предвиђа и могућност интеграције техника ОПЈ, како би се резултати побољшали.

5.2.1 Мере за одређивање корелисаности пара речи

Већина статистичких мера за одређивање корелације речи x и *y*, из пара (x, *y*), заснивају се на поређењу вероватноће заједничког појављивања речи x и *y* у односу на вероватноћу истог догађаја, под условом да су речи у документу поређане насумично. Мере које су коришћене за потребе овог истраживања приказане су у табели 5.1, заједно са описом величина које учествују у мерама. Поступак одређивања значајних фраза подразумева да се, поштујући границе најмањег семантичког контекста – реченице, издвоје сви могући различити парови из корпуса докумената, те да се они рангирају према некој од мера из табеле 5.1. Првих *n* највише рангираних парова проглашава се значајним фразама.

[10] До краја ове главе, под термином *значајна фраза* подразумеваће се релеватни парови – значајне фразе реда два. О издвајању значајних фраза реда већег од два, биће речи у поглављу 7.2.

Табела 5.1: Мере за одређивање корелације пара речи.

Мера	**Формула**
PMI (Church & Hanks 1989)	$\log \frac{f(x,y)R}{f(x)f(y)}$
PMIsig (Washtell & Markert 2009)	$\text{PMI}\sqrt{\min(f(x), f(y))}$
sPMId (Damani & Ghonge 2013)	$\log \frac{d(x,y)}{d(x) * d(y)/D + \sqrt{\max\big(d(x), d(y)\big)}\sqrt{\frac{\ln \delta}{-2}}}$
Dice (Dice 1945)	$\frac{2f(x,y)}{f(x)+f(y)}$
G^2 (Dunning 1993)	$2\left(f(x,y)\log\frac{f(x,y)P}{f(x)f(y)} + f(x,\bar{y})\log\frac{f(x,\bar{y})P}{f(x)f(\bar{y})} + f(\bar{x},y)\log\frac{f(\bar{x},y)P}{f(\bar{x})f(y)} - f(x,\bar{y})\log\frac{f(\bar{x},\bar{y})P}{f(\bar{x})f(\bar{y})}\right)$
$\boldsymbol{f(x)}$	фреквенција на нивоу корпуса за реч x
$\boldsymbol{f(x,y)}$	фреквенција на нивоу корпуса за пар речи (x, y)
$\boldsymbol{R}$	сума фреквенција свих речи у корпусу
$\boldsymbol{P}$	сума фреквенција свих парова у корпусу
$\boldsymbol{f(\bar{x})}$	сума фреквенција свих речи које нису x
$\boldsymbol{f(x,\bar{y})}$	сума фреквенција свих парова које садрже x и не садрже y
$\boldsymbol{f(\bar{x},y)}$	сума фреквенција свих парова које садрже y и не садрже x
$\boldsymbol{f(\bar{x},\bar{y})}$	сума фреквенција свих парова који не садрже ни x ни y
$\boldsymbol{\delta}$	параметар између [0, 1], у истраживању коришћено 0.5
$\boldsymbol{D}$	број документа у корпусу
$\boldsymbol{d(x)}$	број документа у корпусу са макар једним појављивањем x
$\boldsymbol{d(x,y)}$	број документа у корпусу са макар једним појављивањем (x, y)

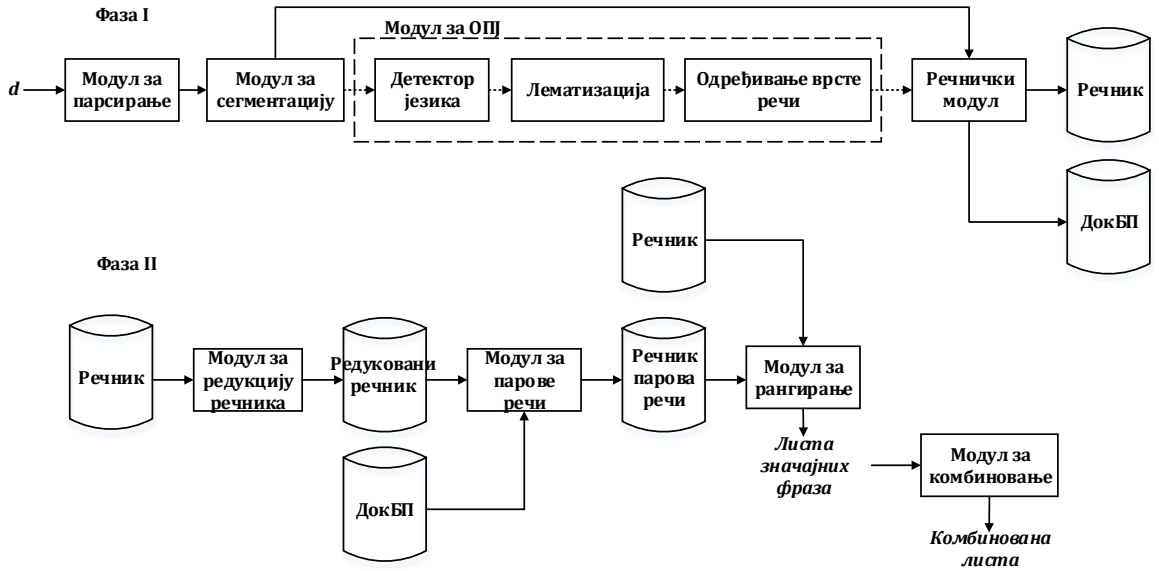
Појединачне мере из Табеле 5.1 преферирају парове речи са различитим фреквенцијама. Како за сваки пар речи важи $f(\mathrm{x}, y) \leq \min(f(x), f(y))$ и $d(x, y) \leq \min(d(x), d(y))$, релације из табеле 5.1 указују да *PMI* и *sPMId* мере фаворизују парове са нижим фреквенцијама, док мере *PMIsig* и G^2 већи степен повезаности дају паровима са вишим фреквенцијама. Мера *Dice* преферира парове у којима обе речи имају сличне фреквенције.

Услед пристрасности статистичких мера према паровима са одређеним фреквенцијама, у истраживању се предлаже приступ који *комбинује* најбоље рангиране значајне фразе добијене применом различитих мера – *Метод комбиноване листе.*

Метод комбиноване листе: формира се листа парова тако што се наизменично попуњава следећим највише рангираним паром из сваке појединачне листе, све док се не одабере претходно дефинисан број парова (значајних фраза реда два). Ако је посматрани пар речи већ додат, јер је био боље рангиран према другој статистичкој мери, листи се додаје следећи најбоље рангирани за текућу меру.

5.2.2 Систем за издвајања значајних фраза

Значајне фразе се издвајају применом протокола у две фазе, пошто се парсирају сви документи из корпуса (Слика 5.2). У *Фази I*, сваки документ пролази кроз *Модул за парсирање* који издваја текст из датотека различитих формата (MS Office, PDF, ...) и формира одговарајућу HTML репрезентацију, којом се чува структура оригиналног документа. Након парсирања, у *Модулу за сегментацију*, документ се трансформише у листу параграфа, где се сваки параграф састоји од листе реченица. Мотив за увођење параграфа, као семантичке јединице у репрезентацији документа, је да се омогући успостављање релације на основу *семантичке блискости* између значајних фраза које се не појављују заједно на нивоу реченице.



Слика 5.2: Систем који имплементира двофазни протокол за издвајање значајних фраза.

Фреквенције различитих речи на нивоу целог корпуса, неопходне за статистичке мере из Табеле 5.1, одређују се у *Речничком модулу*, у коме се инкрементално формира *Речник* различитих речи са одговарајућим фреквенцијама. Исти модул складишти документе у базу процесираних докумената (*ДокБП*).

Приликом примене статистичких мера корелације из Табеле 5.1, треба обратити пажњу на различите морфолошке облике речи (нпр. *оплата*, *оплатом*, *оплате*,...). Последица морфолошке варијације облика речи је *смањење фреквенције* примарног облика речи и повећање броја кандидата за значајне фразе. Под претпоставком да су комплексни инвестициони пројекти по природи углавном вишејезични, предложени приступ поседује могућност додавања *Модула за ОПЈ*, који се састоји од *детектора језика* (ДЈ), компоненте за *лематизацију* (ЛЕМ) и компоненте за *одређивање врсте речи* (ОВР). ДЈ ради на нивоу реченице због могуће вишејезичне природе документа. ЛЕМ своди реч на

канонски облик (нпр. *уговора* - *уговор*), услед чега се добија морфолошки компактнији речник. ОВР класификује речи у граматичке категорије (нпр. *уговор* - именица, *уградио* - глагол, итд.). Класификација по категоријама омогућава формирање правила за дозвољене комбинације речи (нпр. придев-именица, али не придев-придев). Модул за ОПЈ се може конфигурисати тако да се састоји само од ДЈ, ДЈ и ЛЕМ, или ДЈ, ЛЕМ и ОВР, у зависности од тога који језички ресурси су доступни.

Фаза I се завршава када су сви документи у корпусу процесирани. На излазу се добија конструисани *Речник* и документи ускладиштени у одговарајућој бази података (*ДокБП*). *Фаза II* почиње тако што се, у *Модулу за редукцију речника,* уклањају опште речи за одређени језик као што су граматички чланови, везници и предлози (ако је одговарајућа *стоп листа*[11] доступна). Могуће је уклонити и све речи које се појављују у мање од k документа[12]. Наведеним операцијама се формира *Редуковани речник*. Документи из *ДокБП* и *Редуковани речник* се прослеђују *Модулу за парове речи* у коме се конструише речник са свим могућим паровима речи. У овом модулу се, уколико су познате врсте речи (ако се користи ОПЈ), врши редукција према дозвољеним комбинацијама речи. *Речник* и *Речник парова речи* се прослеђују *Модулу за рангирање*, у коме се рачунају мере корелације за све парове према формулама из табеле 5.1. После сортирања, овај модул формира листу од n најбоље рангираних парова за сваку појединачну статистичку меру. *Фаза II* се завршава у *Модулу за комбиновање*, где се формира комбинована листа на начин описан у 5.2.1. Према овом приступу, n најбоље рангираних парова речи из комбиноване листе третирају се као значајне фразе.

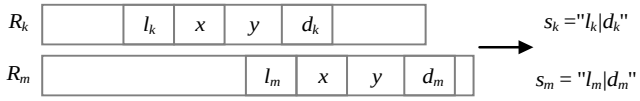
Систем за издвајање значајних фраза реализован је као софтверска компонента у програмском језику Јава.

[11] Честе речи у језику које представљају граматичке односе између других речи - предлози, заменице, помоћни глаголи, чланови и сл. (Luhn 1960).

[12] У свим експериментима коришћено је $k = 2$. Претпоставка је да пар речи који се јавља у само једном документу није довољно информативан, те се не може прогласити значајном фразом.

5.3 Уклањање неинформативних фраза применом ентропије

Већина докумената на пројекту се формира према различитим обрасцима па често садрже исти или веома сличан текст (нпр. заглавља у записницима са састанка, опис поља у захтевима за измену и сл.). Документи могу да садрже и делове текста преузете из других докумената. Парови речи из описаних, мање информативних делова текста, биће боље рангирани према степену корелисаности због повећане фреквенције и тиме неосновано проглашени значајним фразама. Илустративни примери су парови који се односе на стандардне делове електронске преписке између учесника (*unauthorized viewing, disclosure of information*, *This email has been checked for viruses by*, ...).



Слика 5.3. Суседства за пар (*x*, *y*), из реченица R_k и R_m, која се користе за рачунање ентропије пара.

У истраживању је уведена претпоставка да се значајне фразе чешће јављају у различитим семантичким контекстима за посматрани корпус (нпр. *revised steel structure drawings, collision between rolling doors and steel structure, corrosion protection of the steel structure and equipment*, итд.). Предлаже се приступ за уклањање неинформативних фраза на основу *разноврсности суседстава* у којима се појављује сваки пар речи. Нека је (*x, y*) пар детектован у процесу издвајања, и нека је $(x, y)_k$ његово *k*-то појављивање. Суседство се одређује као низ карактера $s_k = "l_k|d_k"$, добијен после спајања суседних речи са леве и десне стране, l_k и d_k (Слика 5.3). Ако се $(x, y)_k$ налази на почетку (крају) реченице, l_k (d_k) се замењује карактером "|". Даље, нека S_{xy} представља скуп свих суседстава пара (*x, y*) и нека се може поделити у *n* група у оквиру којих су сва суседства иста. Ако је укупан

број појављивања (*x, y*) у свим документима једнак N (број елемената скупа S_{xy}), мера *информативности пара* може се израчунати као *ентропија* скупа S_{xy}:

$$E(S_{\text{xy}}) = -\sum_{i=1}^{n} p_i \log_2 p_i \ , \quad \text{где је } p_i = \frac{\text{broj elemenata u grupi } i}{N} \tag{5.1}$$

Из (5.1) следи да је $E(S_{\text{xy}}) = 0$ када се (*x, y*) појављује увек у истом контексту (сва суседства припадају истој групи). Када су сва суседства различита, ентропија је максимална и износи $E(S_{\text{xy}}) = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$.

Да би се поправило рангирање парова који се појављују у више информативних контекста, у *Модулу за рангирање* се мера корелисаности речи, добијена према формулама из Табеле 5.1, множи информативношћу пара добијеном из једначине (5.1).

5.4 Експериментална провера поступака за аутоматску детекцију значајних фраза

У овом поглављу су изложене карактеристике корпуса који је коришћен за валидацију предложених метода, описани су извршени експерименти и дати су коментари на добијене резултате.

5.4.1 Експериментални корпус

Предложени приступ за издвајање значајних фраза је тестиран на корпусима докумената са два капитална инвестициона пројекта реализована у Републици Србији. Корпус $К_{\text{Бор}}$ садржи документе са пројекта „*Реконструкција Топионице и изградње нове Фабрике сумпорне киселине*“ у граду Бору, а корпус $К_{\text{Коридор}}$ чине документи са пројекта „*Изградња деонице аутопута на Пан-Европском Коридору X (секција Ниш - Димитровград)*“. У Табели 5.2 су приказане карактеристике оба корпуса. Корпуси се у највећој мери односе на извођење грађевинских радова, са темама као што су динамика радова, технологија изградње, финансирање и контрола квалитета. Потребно је напоменути да су се

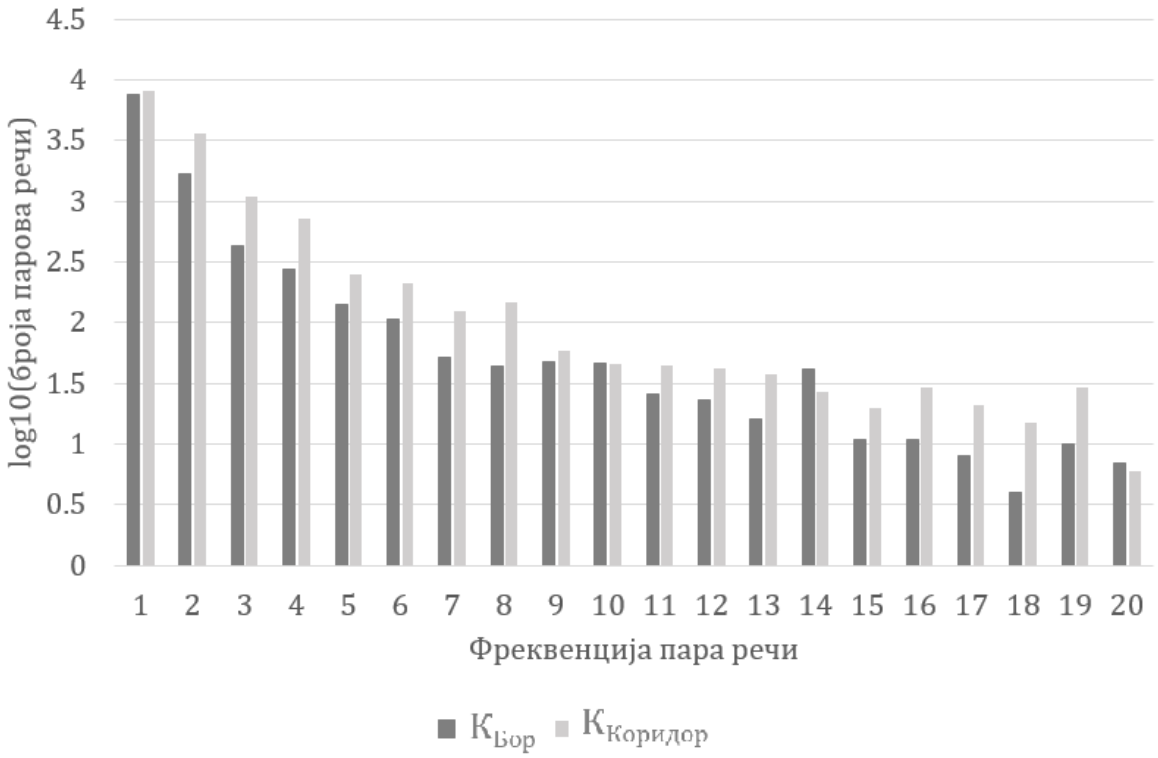
у корпусу $К_{Бор}$, поред терминологије из области грађевинарства, користили изрази из области електро-индустрије и рударства – пројекат је везан за постројења за производњу и прераду руде бакра и производњу сумпорне киселине.

Табела 5.2: Спецификације експерименталних корпуса.

Корпус	**$К_{Бор}$**	**$К_{Коридор}$**
Број докумената	175	79
Типови докумената	варијација, преписка, одштетни захтев, записник са састанка	записник са састанка, недељни извештај, преписка
Период	2011 – 2015	2009 – 2013
Број речи у редукованом речнику	3337	3728
Број реченица	1514	4243
Просечна дужина реченице	17.66	16.18
Стандардна девијација дужине реченице	32.39	11.86

На оба пројекта, инвеститори су јавна предузећа чији је оснивач Република Србија (*Рударско-топионичарски басен Бор*, односно *Коридори Србије*), док су у улози надзора и извођача били и домаћи и страни учесници. Већина документа је била двојезична (српски и енглески). Значајан део документа из корпуса се директно или индиректно бавио различитим аспектима захтева за измене уговорених радова или одштетним захтевима. Корпус $К_{Бор}$ је садржао доста формулара који су имали краће описе поља и дужи садржај, услед чега је значајно повећана стандардна девијација дужине реченице. Дужина реченица у корпусу $К_{Коридор}$ није варирала у већој мери јер је већина документа по форми била ближа техничким извештајима и писмима него формуларима. Поред већег броја речи у редукованом речнику, у корпусу $К_{Коридор}$ је идентификован већи број кандидата за значајне фразе (слика 5.4). Приказана дистрибуција показује да се највише

могућих парова речи појављује само једанпут у корпусу, што је последица *изражене морфолошке варијације* речи.



Слика 5.4: Дистрибуција броја различитих парова речи, у зависности од броја појављивања у корпусу. Број различитих парова дат је на логаритамској скали.

5.4.2 Експертска листа значајних фраза

Два грађевинска инжењера, који су били активно укључени на различитим активностима на оба пројекта и који су били упознати са текстуалним корпусима, одабрани су као експерти који ће их обележити на основу свог претходног знања. Они су, након активног увида у корпусе[13], обележили најважније значајне фразе реда два. Коначна експертска листа је добијена спајањем појединачних листи

[13] Корпуси $К_{Бор}$ и $К_{Коридор}$ представљају репрезентативан узорак пројектне документације. Целокупна документација на пројектима садржи више десетина хиљада докумената.

(експерти су анализирали документе независно). Посматрани пар речи је проглашаван за значајну фразу ако је у стању да *сумаризује поруку* или значајан део поруке саопштене реченицом. Више значајних фраза је могло бити одабрано из једне реченице, у зависности од дужине реченице и њене комплексности. Критеријум за селекцију (Слика 5.5) односио се на процену степена значаја пара за цео корпус: *глобално значајан* (увек одабран), *локално значајан* (одабран по процени) и *неважан* (одбијен).

"In August 2012 as a result of Engineer **verbal instruction**, the Contractor submitted **change order** request for **additional costs** related to **proposed changes** in the **thermal insulation** of the original façade."

Слика 5.5: Критеријум експерта за одабир пара речи као значајне фразе: глобално значајан (подебљано), локално значајан (подвучено).

У складу са предложеном методологијом, све одабране значајне фразе имају додатни услов да морају да се појаве у *макар два* документа из корпуса. Експерти су идентификовали 449 значајних фраза за $К_{Бор}$, и 515 за $К_{Коридор}$.

5.4.3 Резултати експеримента за издвајање значајних фраза

Експерименти су спроведени како би се извршила:

- Провера успешности појединачних мера за одређивање корелисаности из табеле 5.1 и њихово поређење са методом комбиноване листе (поглавље 5.2.1);
- Поређење успешности мера за одређивање корелисаности, без и са укљученим приступом који користи ентропију суседстава за редукцију неинформативних фраза (поглавље 5.3);
- Ефекти примене техника ОПЈ на успешност процеса издвајања, ако су одговарајући језички ресурси доступни (поглавље 5.2.2);

- Поређење модела за аутоматско издвајање са најбољим резултатом, са посебно креираним приступом за издвајање који користи *претходно експертско знање* из домена управљања инвестиционим пројектима;
- Капацитет значајних фраза реда два да опишу релевантне информације на пројекту.

За сваку експертску листу која се састоји од *n* значајних фраза, свака тестирана метода је генерисала *n* најбоље рангираних парова речи. Добијени парови су поређени са фразама из експертске листе, а *прецизност* методе је израчуната као однос између броја препознатих значајних фраза из експертске листе и броја *n* (табела 5.3).

Табела 5.3: Прецизност различитих приступа за издвајање значајних фраза: најбољи приступ за сваку меру – подебљано, глобално најбољи приступ без ОПЈ - *, најбољи приступ – подвучено.

Корпус	**Корекција ентропијом**	**ОПЈ**	**Dice**	**G^2**	**PMI**	**PMIsig**	**sPMId**	**Комб. листа**
$K_{Бор}$	Не	без ОПЈ	0.318	0.350	0.229	0.287	0.298	0.325
$K_{Бор}$	Да	без ОПЈ	0.445*	0.421	0.388	0.399	**0.403**	0.432
$K_{Бор}$	Не	ЛД + ЛЕМ	0.334	0.376	0.245	0.303	0.312	0.352
$K_{Бор}$	Да	ЛД + ЛЕМ	**0.472**	**0.457**	**0.430**	0.385	0.412	0.477
$K_{Бор}$	Не	ЛД + ЛЕМ+ ПОС	0.374	0.394	0.272	0.318	0.334	0.396
$K_{Бор}$	Да	ЛД + ЛЕМ+ ПОС	0.445	0.448	0.428	**0.401**	0.392	**<u>0.510</u>**
$K_{Коридор}$	Не	без ОПЈ	0.175	0.338	0.101	0.289	0.150	0.252
$K_{Коридор}$	Да	без ОПЈ	0.320	0.357*	0.353	0.334	0.276	0.355
$K_{Коридор}$	Не	ЛД + ЛЕМ	0.204	0.369	0.138	0.289	0.183	0.282
$K_{Коридор}$	Да	ЛД + ЛЕМ	**0.351**	**0.384**	**0.373**	0.346	**0.322**	0.386
$K_{Коридор}$	Не	ЛД + ЛЕМ+ ПОС	0.208	0.330	0.122	0.287	0.186	0.326
$K_{Коридор}$	Да	ЛД + ЛЕМ+ ПОС	0.334	0.369	0.355	**0.350**	0.293	**<u>0.412</u>**

Када се не примени ОПЈ и корекција ентропијом, најбоље резултате за оба корпуса остварује мера G^2. Све мере су показале лошије перформансе на корпусу $K_{Коридор}$, услед већег речника парова речи.

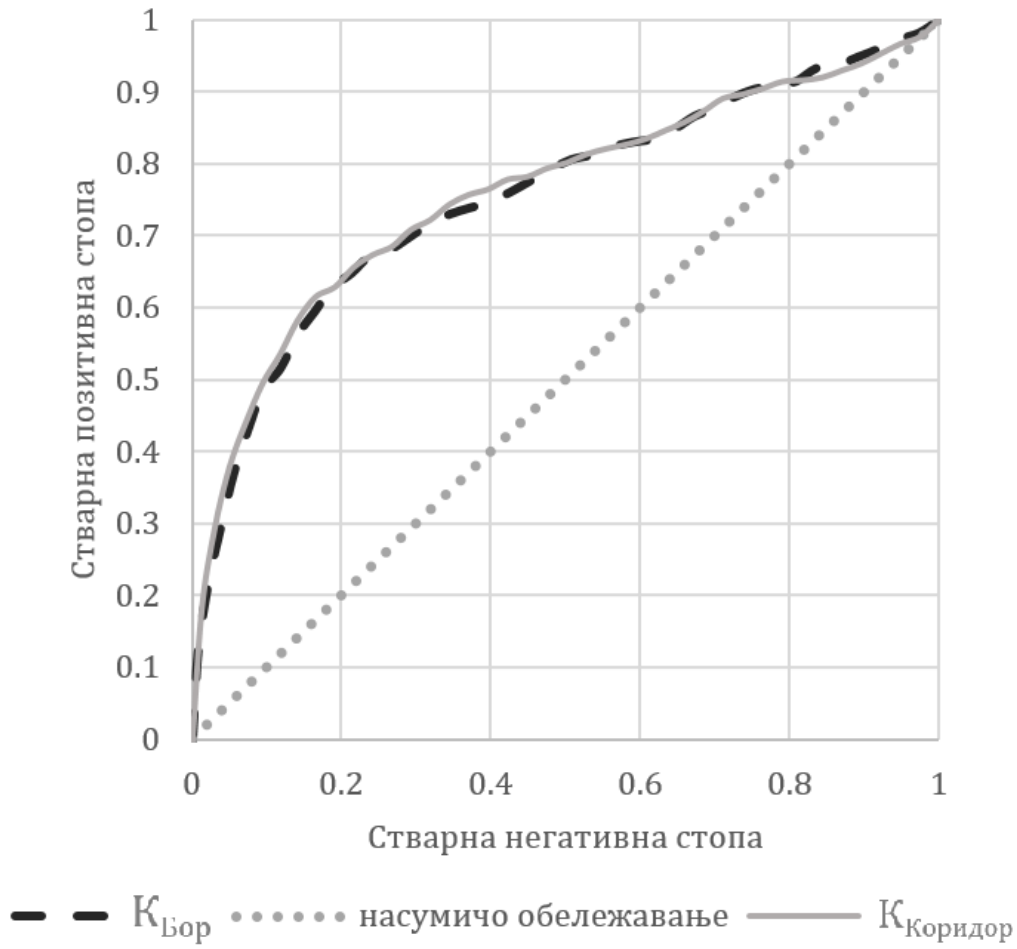
Сви тестирани поступци издвајања су показали да *прецизност расте када се уведе корекција ентропијом*, што потврђује квалитет предложене методе за отклањање неинформативних фраза. Додатни ефекат увођења ентропије је довођење перформанси различитих мера на *исти* ниво, чиме је *доказана иницијална претпоставка* да су фразе које се појављују у *различитим контекстима информативније*.

Све мере су показале побољшање перформанси када се укључе технике ОПЈ, што указује на *значајан допринос језичких ресурса* на поступак детекције значајних фраза. Важно је напоменути да је разлика између најбоље методе без алата ОПЈ (Dice + ентропија за $К_{Бор}$, G^2 + ентропија за $К_{Коридор}$) и глобално најбоље опције (комбинована листа + ентропија + ЛЕМ + ОВР, за оба корпуса) била мања од шест процената. Може се закључити да је предложени приступ *применљив* и за оне случајеве *када језички ресурси нису доступни*.

Оптимални резултат се добија када се примене оба предложена метода – комбиновање листи различитих мера за корелисаност и корекција ентропијом, заједно са ресурсима за ОПЈ. Овај поступак, назван „*комплетна метода*“ ће се користити у свим наредним експериментима.

Утицај величине добијене листе значајних фраза на *одзив система* (израчунат као проценат препознатих експертских фраза), приказан је преко *криве оперативне карактеристике пријемника* (Receiver Operating Characteristics – ROC, (Fawcett 2006)). Крива је добијена варирањем броја најбоље рангираних парова речи (Слика 5.6). ROC крива је дефинисана у координатном систему *стварне позитивне стопе* (True positive rate - TPR), која је еквивалентна одзиву система, и *стварне негативне стопе* (False positive rate - FPR), која представља однос између детектованих фраза које *нису* експертске и свих парова речи из корпуса који *нису* у експертској листи.

		Обележавање експерата	
		значајне фразе (e+)	нису значајне фразе (e-)
Обележавање система	значајне фразе (c+)	слагање система и експерата (e+, c+)	фразе погрешно означене као значајне од система (e-, c+)
	нису значајне фразе (c-)	значајне фразе које систем није препознао (e+, c-)	слагање система и експерата (e-, c-)
		одзив, стварна позитивна стопа $\frac{\sum(e+, c+)}{\sum(e+)}$	стварна негативна стопа $\frac{\sum(e-, c+)}{\sum(e-)}$



Слика 5.6: Перформансе комплетне методе: горе - одређивање стварне позитивне стопе (одзива) и стварне негативне стопе; e/c : експерти/систем; +/- : фраза оцењена као значајна/није значајна. Доле: ROC криве.

Пожељно је да ROC крива има тачку превоја у горњем левом углу (максимална стварна позитивна стопа и минимална стварна негативна стопа). Са слике 5.6 се види да за обе криве важи, да се за одзив од 70% експертских фраза, добија око 25% свих парова речи из речника који *нису* значајне фразе.

5.4.4 Поређење са експертским приступом за обележавање значајних фраза

С обзиром да предложени метод за издвајање може да се конфигурише као *независан у односу на проблем и језик*, поставља се питање његовог понашања у односу на експертски дефинисану процедуру која узима у обзир одређени домен проблема и језик докумената. Под експертски дефинисаном процедуром овде се подразумева поступак који би спровео експерт или група експерата, којим би се, коришћењем претходног доменског знања, формирала листа значајних фраза за неки корпус докумената. Подразумева се да је апсолутно познат домен проблема који се покрива документима, као и њихов језик. У циљу моделирања и верификације описаног поступка, спроведен је експеримент за *Надгледано аутоматско издвајање фраза* (НАИФ), на следећи начин:

- Дефинисани су релевантни текстуални извори из домена управљања пројектима на српском и енглеском језику (уџбеници, стручни речници, стандарди и водичи – списак извора је дат у прилогу 1);
- Формирана је глобална доменски релевантна листа фраза од ставки из индекса или речника;
- Глобална листа је проширена именима учесника на пројекту из адресара, као и називима компанија ангажованим на пројекту (претпоставка је да би експерт користио доступне изворе који су специфични за конкретан пројекат);

- Формирана је листа кандидата од оних елемената глобалне листе који су се појавили најмање у два документа (аналогно граници за број појављивања у документима коришћеној у претходним експериментима).

Добијене листе кандидата за оба корпуса поређене су са експертским листама и израчунати су прецизност и одзив (табела 5.4). Последње две колоне представљају прецизност и одзив за комплетну методу (КМ). Треба приметити да су резултати за КМ нешто лошији од резултата приказаних у табели 5.3, јер је број најбоље рангираних парова речи, n за КМ у овом експерименту, изједначен са димензијом листе кандидата коју је генерисао НАИФ (у претходном експерименту n је било једнако димензији експертске листе). Промена броја кандидата је извршена како би се омогућило валидно поређење два приступа.

Табела 5.4: Поређење НАИФ и комплетне методе (КМ).

Корпус	**К$_{Бор}$**	**К$_{Коридор}$**
Број експертских фраза (n)	449	515
Број кандидата (r)	393	221
Број детектованих фраза (d)	168	129
НАИФ прецизност (d/r)	0.43	0.58
Прецизност КМ	0.47	0.51
НАИФ одзив (d/n)	0.37	0.25
Одзив КМ	0.41	0.22

Комплетна метода је дала *боље* резултате на К$_{Бор}$, уз напомену да би се перформанса НАИФ приступа могла додатно побољшати укључивањем текстуалних извора блиских проблематици која се јављала у К$_{Бор}$**.** Речници из области заваривања и електроинсталација су примери текстуалних извора чији су термини, услед специфичне технолошке природе корпуса К$_{Бор}$**,** били присутни у одређеној мери.

НАИФ поступак је дао *боље* резултате на К$_{Коридор}$ јер су уграђени текстуални извори у највећој мери *покрили теме* из корпуса. Треба имати у виду да је на

резултат КМ утицао параметар који представља број најбоље рангираних парова речи, а који је у овом експерименту одговарао НАИФ поступку (у овом случају 221). Када се постави тако да одговара димензији експертске листе (515 фраза), КМ остварује 0.412 и за прецизност и за одзив (видети табелу 5.3). Ово повећање указује на капацитет КМ да *боље* идентификује релевантне фразе ако се *повећа број* најбоље рангираних парова речи, док сам поступак остаје *независан према домену проблема* покривеног документима. Смањење прецизности, које се јавља када се повећа број парова кандидата, може се делимично избећи увођењем *релација између значајних фраза* и њиховом репрезентацијом као *графа значајних фраза на пројекту*, што ће бити приказано у глави 7. Са оваквим графом, корисник стиче увид у семантичке контексте у којима се јављала посматрана фраза, па има могућност да је филтрира ако није довољно релевантна.

5.4.5 Семантички капацитет фраза реда два

Да би се проверила хипотеза изнета у глави 4 – значајне фразе реда два су погодне да пренесу релевантне информације на пројекту, извршен је експеримент у коме су документи из корпуса *груписани* по сличности. Документи су репрезентовани на два начина: као *скуп речи* и као *скуп значајних фраза реда два*. Циљ је да се провери која репрезентација је погоднија за груписање докумената у кохерентне тематске групе према садржају. Експеримент је извршен на корпусу $К_{Бор}$, те је за верификацију добијених група коришћена постојећа подела докумената из изворног система датотека, а према типовима за тај корпус (*одштетни захтев, захтев за измене уговорених радова, преписка* и *записник са састанака*).

Нека је F_i скуп n најбоље рангираних значајних фраза у документу d_i, одабраних из листе коју је вратила комплетна метода. Слично, нека је R_i скуп од n речи у d_i са највећом *TF-IDF мером* (Robertson 2004). TF-IDF рангира *значај речи* за садржај документа као производ две статистичке мере: фреквентност у

документу (*tf* - term frequency) и инверзна фреквентност по документима (*idf* - inverse document frequency), где је:

$$tf(r,d) = \text{broj pojavljivanja reči } r \text{ u dokumentu } d$$

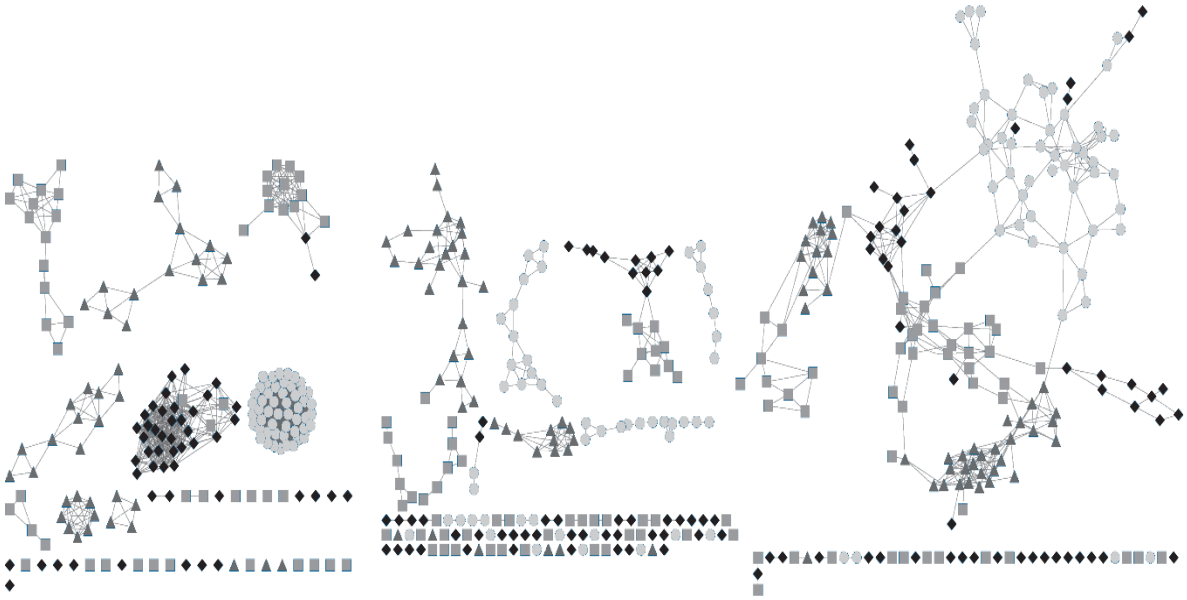
$$idf(r,d) = log\frac{\text{broj dokumenata u korpusu}}{\text{broj dokumenata gde se javlja reč } r}$$

Очигледно, ако се реч више пута појављује у документу, а мање у корпусу, онда је она значајнија за садржај документа. Документи d_i и d_j се проглашавају сличним ако им је Jaccard-ов индекс сличности скупова речи којима су репрезентовани, већи од претходно дефинисане границе t из интервала [0, 1]:

$$\text{sim}(d_i,d_j) = \text{J}(F_i, F_j) = \frac{|F_i \cap F_j|}{|F_i \cup F_j|}, \quad \text{репрезентација значајним фразама} \qquad (5.2)$$

$$\text{sim}(d_i,d_j) = \text{J}(R_i, R_j) = \frac{|R_i \cap R_j|}{|R_i \cup R_j|}, \quad \text{репрезентација речима} \qquad (5.3)$$

Пошто се израчуна индекс сличности за сваки пар докумената, може се конструисати мрежа груписаних докумената у којој чворови представљају појединачне документе (Слика 5.7). Веза између d_i и d_j се успоставља ако је $\text{sim}(d_i,d_j) > t$. За случај када су документи репрезентовани значајним фразама, оптимално груписање је добијено за $t = 0.2$. Овде се може видети да су *захтеви за измену* скоро у потпуности придружени одговарајућој групи (Слика 5.7 лево). Већина *одштетних захтева* груписана је заједно, док су документи из група *преписка* и *записник* формирали више хомогених група, услед веће разноликости тема које се у њима помињу. Одређени број докумената је остао неповезан.



Слика 5.7: Групе докумената у различитим репрезентацијама. Чворови су документи из категорија: *захтев за измену* (кругови), *одштетни захтев* (ромбови), *преписка* (правоугаоници) и *записник* (троуглови). Лево: репрезентација путем значајних фраза, t=0.2. Средина и десно: репрезентација путем значајних речи, t=0.1 (средина) и t=0.08 (десно).

Најбољи резултат, када се користи репрезентација речима (Слика 5.7 средина), указује да чак и за снижену вредност границе индекса сличности (t = 0.1), значајан број докумената остаје изолован. Добијене групе нису одговарале стварним категоријама и ниједна од категорија није потпуно уочљива. Даљим смањивањем границе (t = 0.08), број изолованих докумената је почео да се смањује, али су документи почели да се групишу у јединствену хетерогену групу (Слика 5.7 десно). На овај начин *доказана* је претпоставка да фразе реда два *имају већи семантички капацитет* да пренесу значење докумената, у односу на појединачне речи.

5.5 Аутоматска детекција претходно дефинисаних текстуалних образаца

Досадашњи поступак разматрао је искључиво речи састављене од слова. Бројеви, као и речи које у свом запису садрже бројеве и/или знакове интерпункције, нису разматрани као кандидати за значајне фразе (иако је то било могуће). Разлог за то је природа докумената на пројекту који, поред отвореног текста, садрже различите ознаке, шифре, референце и сл. Међутим, *датуми* и *новчани износи* су примери података који садрже и бројеве и словне карактере, а потенцијално чине део значајних информација.

У овом истраживању, могућ поступак издвајања претходно дефинисаних образаца биће илустрован на примеру *датума*. Проблем детекције датума може се свести на *проналажење задате структуре у низу карактера* (стрингу). Међутим, за разлику од текстуалних извора писаних једним језиком (нпр. новински чланак), вишејезични документи са инвестиционог пројекта садрже датуме различитих формата.

За детекцију датума коришћена су правила заснована на *регуларним изразима*[14] и детектованом примарном језику параграфа (ако је доступан модул за детекцију језика). Регуларни изрази представљају скуп ниски карактера којима је, посебном синтаксом, описана група подтекстова у тексту који задовољавају задати образац. У табели 5.5 су приказане основне ниске карактера и квантификатора регуларних израза.

[14] https://www.regular-expressions.info

Табела 5.5: Примери ниски карактера и квантификатора језика регуларних израза.

Ниска карактера	**Значење**
\d	једна цифра од 0 до 9
\s	сви размаци (*space*, *tab*, *new line*)
[a-zA-z]	сва слова, велика и мала
{n}	елемент пре ознаке {n} мора се јавити тачно *n* пута
{n,m}	елемент пре ознаке {n,m} може се јавити између *n* и *m* пута

На пример, нека је задат формат који генерише датуме попут *3-rd September 2014.* Ако се у тексту пронађе следећа секвенца карактера:

- једна или две цифре (**\d{1,2}**),
- карактер "-" (**-**),
- два словна карактера (**[a-zA-z]{2}**),
- један или више бланко карактера (**\s+**),
- између три и десет словних карактера (**[a-zA-z]{3,10}**),
- један или више бланко карактера (**\s+**),
- четири цифре (**\d{4}**),

она се проглашава датумом (мада то не мора да буде). Последично, регуларни израз за препознавање и издвајање датума наведеног формата је:

"\d{1,2}[a-zA-z]{2}\s+[a-zA-z]{3,10}\s+\d{4}"

У табели 5.6 приказани су формати датума који су били препознавани у оквиру овог истраживања.

Табела 5.6: Различити формати записа датума и одговарајући регуларни изрази.

Формат датума	**Регуларни израз за препознавање**
03-09-2014	"\d{1,2}\-\d{1,2}-\d{4}?"
03-Sep-2014	"\d{1,2}\-\[a-zA-z]{3,4}-\d{4}"
03.09.2014	"\d{1,2}\.\d{1,2}\.\d{4}"
09/03/2014	"\d{1,2}/\d{1,2}/\d{4}"
Sept.3, 2014	"[a-zA-z]{3,4}\.\d{1,2}\,\s+\d{4}"
3rd September 2014	"\d{1,2}[a-zA-z]{2}\s+[a-zA-z]{3,10}\s+\d{4}"
3-rd September 2014	"\d{1,2}-[a-zA-z]{2}\s+[a-zA-z]{3,10}\s+\d{4}"
September 3, 2014	"[a-zA-z]{3,10}\s+\d{1,2}\,\s{0,1}\d{4}"
3. septembar 2014	"\d{1,2}\.\s+[a-zA-z]{3,10}\s+\d{4}"

6 Погодне репрезентације знања

На основу претходно изнетих чињеница, закључује се да постојећи системи за рад са документима, по питању издвајања знања из неструктуираних текстуалних извора, не задовољавају у потпуности специфичности које намеће динамично и комплексно окружење у коме се изводе инвестициони пројекти. Стога се предлаже репрезентација издвојених концепата која, у односу на стандардну текстуалну претрагу, захтева мање труда за генерисање новог знања потребног за доношење одлука.

Различити приступи који се могу искористити као основа за репрезентовање издвојених информација биће разматрани кроз практични пример са реалног пројекта. Наведени пример осликава ситуацију која захтева да се идентификују, разумеју и синтетишу подаци из различитих текстуалних извора.

6.1 Информација записана природним језиком

Природни језик је најопштији медијум за пренос и извођење знања. Он омогућава представљање најкомплекснијих и најапстрактнијих идеја и као такав је незаменљив и широко распрострањен у свим доменима. Управо је способност савладавања и коришћења природног језика једна од главних особина човека као интелигентног бића (Santos 1992).

Највећи део знања на инвестиционом пројекту похрањен је у документима писаним у форми природног језика (Soibelman et al. 2008). Међутим, управо је експресивност природног језика разлог који отежава његову примену за аутоматско издвајање информација и извођење знања: искази у природном

језику могу бити *вишезначни*, *неконзистентни* и *комплексни* за моделовање у рачунарском систему (Jakus et al. 2013).

У наставку је наведен пример једне ситуације са пројекта „*Реконструкција Топионице и изградње нове Фабрике сумпорне киселине*“ у граду Бору. Наведена ситуација описана је исказом, добијеним на основу анализе преписке и записника са састанка (неки појмови и датуми су замењени генеричким називима) – ситуација *Корозија цеви*:

Корозија цеви: Рад на инсталацији цеви у CIGHE области је стопиран због корозије цеви. На састанку С1 је донета одлука да се уграде цеви чија је дебљина зида након уклањања корозије већа од 3.4 мм. Инвеститор К1 је наложио инспекцију цеви од стране квалификоване организације К2 и уклањање са градилишта до договореног датума Д1, оних цеви које не испуњавају дефинисани критеријум.

Репрезентован у форми природног језика, исказ који описује наведену ситуацију је у потпуности разумљив за просечног корисника. Међутим, да би се овакав исказ *формулисао*, неопходан је експерт са постојећим искуством са пројекта, који мора да пронађе одговарајуће документе и синтетише информације из њих. Даље, да би се извлачили закључци и доносиле одлуке на основу приказаног исказа, експерт мора да буде упознат са оним чињеницама које су повезане са наведеном ситуацијом: овако издвојен, исказ је независан од остатка пројекта и корисник мора да се ослони на своје знање и искуство да га правилно интерпретира.

Да би се умањио ефекат презасићености информацијама и побољшао процес доношења одлука на основу чињеница из текста, пожељно је структуирати и ефикасно презентовати релевантне информације крајњем

кориснику. За анализу алтернативне репрезентације, из исказа у природном језику је потребно издвојити најзначајније концепте:

- Компанија: *К1, К2*
- Састанак: *С1*
- Мера: *инспекција цеви*
- Акција: *уклањање са градилишта*
- Критеријум: *дебљина зида након уклањања корозије већа од 3.4 мм*
- Одлука: *уградити* цеви
- Област: *CIGHE*
- Датум: *Д1*
- Догађај: *корозија*
- Материјал: *цеви*
- Активност: *инсталација цеви*
- Статус: *стопирана*

Издвојени концепти и њихови односи биће приказани у наставку, где ће исказ *Корозија цеви* бити представљен кроз различите *репрезентације знања*.

6.2 Својства репрезентације знања

Као основа за одабир одговарајуће репрезентације издвојених информација, размотрене су различите репрезентације знања које се користе за складиштење података, као и правила за закључивање над њима у одређеном проблемском домену. У (Brachman & Levesque 2004), репрезентација знања и закључивање над њом су дефинисани као „област вештачке интелигенције која истражује како се знање може представити симболички и аутоматски обрадити програмима за закључивање“. Процес закључивања се може дефинисати као низ поступака селекције и обраде елемената репрезентације знања којима се изводе закључци о посматраном проблему. Међутим, да би се дефинисала логика закључивања, неопходно је да се експертско знање о проблему (факти, правила,

ограничења, корелације, и сл.) запише *формалним језиком* који омогућава извођење закључака над репрезентацијом. Управо је ограничавање на ужу област неопходно, како би се формализовала постојећа експертска знања и искуства за одређени домен – попут наведених решења из поглавља 3.4, која обрађују документе из области статичких прорачуна, сеизмичке анализе, захтева за информацијом и др.

Како се у отвореном, *вишејезичном* свету документације на пројекту јављају различити типови дисциплина, докумената и формата, дефинисање и одржавање свеобухватне логике за закључивање је *тешко изводљиво*. Међутим, без обзира на недостатак формалне логике за аутоматско закључивање, предложена репрезентација издвојених концепата би требала да буде моделирана тако да омогући што једноставније издвајање чињеница и образаца од стране експерта, као и да има формалне карактеристике постојећих репрезентација знања.

Према (N.A. Stillings et al. 1995), да би репрезентација на одговарајући начин представила посматрани домен проблема, потребно је да поседује следеће особине:

- *Адекватност репрезентације*: способност да представи сво знање од интереса за посматрани домен;
- *Адекватност закључивања*: способност да се манипулацијом структуром репрезентације изводе нове структуре које одговарају новим знањима;
- *Ефикасност закључивања*: способност да се механизам закључивања прилагоди задатим информацијама;
- *Ефикасност аквизиције*: способност да се у репрезентацију инкорпорирају нове информације.

Детаљна листа критеријума које би требала да задовољи репрезентација знања, са становишта практичне примене, је приказана у (Clark 1996):

1. *Експресивност*: језик репрезентације треба да је довољно изражајан како би експерт описао доменске факте;
2. *Природност*: синтакса за рад са репрезентацијом треба да је што ближа природном језику;
3. *Свеобухватност*: кроз репрезентацију треба пронаћи одговоре на највећи могући број питања из домена проблема, уз минималан број некомплетних одговора;
4. *Јасноћа семантике*: изрази у репрезентацији треба да поседују једнозначну и јасно дефинисану семантичку структуру;
5. *Ефикасност*: рад са репрезентацијом треба да буде ефикасан са становишта утрошеног времена и меморије;
6. *Скалабилност*: На перформансе репрезентације не би требало значајно да утиче количина података похрањена у њој;
7. *Тумачење логике закључивања*: способност да се из репрезентације добије поступак којим се одговор на питање изводи;
8. *Интроспективност*: могућност манипулације правилима за извођење знања тако да се на основу њих изводе нова правила;
9. *Енкапсулација знања*: могућност груписања повезаних правила за извођење у концептуалне јединице;
10. *Модуларност*: могућност додавања функционалности репрезентације;
11. *Графички интерфејс*: поседовање одговарајућег графичког окружења за манипулацију знањем;
12. *Портабилност*: могућност преноса на различите платформе.

С обзиром да је основни задатак ове тезе *издвајање* и *структуирање* различитих информација са пројекта из неструктуираних текстуалних извора, у општем случају *нису претходно позната* правила и ограничења за дефинисање логике закључивања. Уместо коришћења претходно познатих формалних правила за закључивање, таква репрезентација би експерту требала да омогући

увид у основне чињенице и обрасце из текста (заступљеност појмова кроз време и по изворима, степен повезаности појмова, и сл.), као и једноставно дефинисање поступака и правила за анализирање издвојених информација. Стога је неопходно одабрати ону структуру репрезентације која даје висок степен *експресивности*, како би издвојене информације верно пренеле поруку записану природним језиком: што је репрезентација *сличнија структури текста* из кога је настала, лакша је за верификацију и интерпретацију од стране експерта.

Поред наведених критеријума, потребно је обратити пажњу и на *комплексност* аутоматског конструисања одговарајуће репрезентације – висок ниво експресивности захтева већу структуираност која може отежати процес аутоматског издвајања.

У даљем разматрању биће разматране репрезентације са становишта применљивости према критеријумима експресивности и комплексности конструисања. Додатни критеријум за одабир репрезентације знања је могућност *погодне графичке репрезентације* која омогућава експерту да, кроз визуелизацију издвојених информација, једноставније идентификује правила и обрасце. Репрезентације које немају подразумевано пресликавање у одговарајућу графичку структуру, као што су *системи засновани на правилима* (Clancey 1983) или *логика првог реда* (Van Emden & Kowalski 1976), неће бити разматрани у овој тези.

6.3 Семантичке мреже

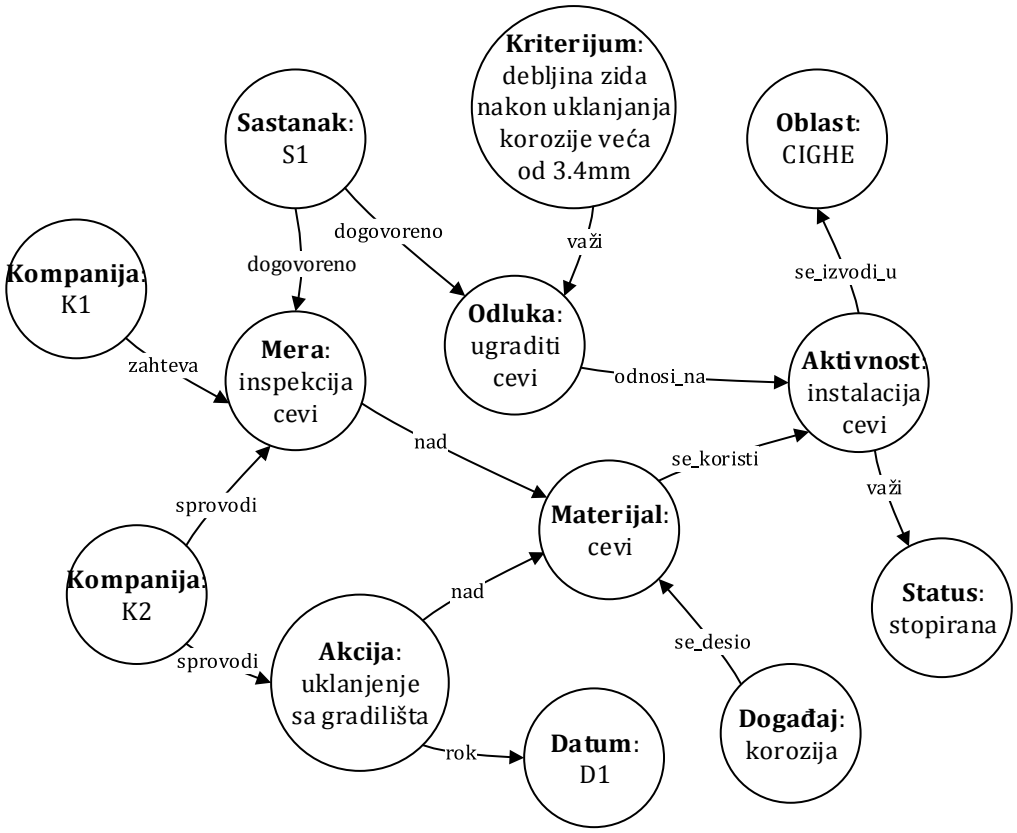
Семантичка мрежа је репрезентација знања у форми *усмереног графа*, где су концепти, објекти или догађаји приказани као чворови, а усмерене линије (гране) које их повезују представљају бинарне релације. Настала је почетком 60-их година 20-ог века са циљем да се семантичке релације између речи моделирају као мрежа (Quillian 1967), како би се симулирао начин на који људи изводе закључке из комплексног текстуалног корпуса. Формално, семантичка мрежа се може дефинисати као скуп чворова и скуп бинарних релација над којима се, за

извођење закључака, може користити *логика првог реда* (Van Emden & Kowalski 1976). Овај декларативни приказ знања може бити потпуно неформалан, а могу му бити придодата и различита формална правила за извођење нових знања. У општем случају, не постоји ограничење на посебне домене па су мреже веома експресивне и погодне за моделирање знања из различитих области. Неке од најпознатијих мрежа настале су из различитих лексичких извора, као што су WordNet или DbPedia (Fellbaum 2012).

Да би се мрежа конструисала, потребно је специфицирати њене основне делове:

- *Лексички*
    - Чворови;
    - Везе;
    - Атрибути (којима се означавају типови чворова и веза).
- *Структурни*
    - Организација чворова и веза у усмерени граф;
    - Придруживање атрибута чворовима и везама.
- *Семантички*
    - Придруживање значења према ентитетима из реалног света, за појединачне чворове и везе.
- *Процедурални*
    - Дефинисање поступака за додавање, брисање, измену и очитавање вредности чворова и веза.

Представљање повезаних чињеница у форми семантичке мреже биће илустровано на примеру који се односи на већ наведену ситуацију са пројекта – *Корозија цеви* (поглавље 6.1). Мрежа, приказана на слици 6.1, конструисана је на основу појмова издвојених из реченица које су записане природним језиком.

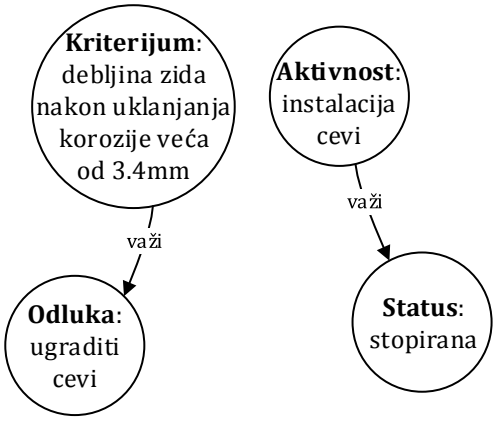


Слика 6.1: Ситуација *Корозија цеви*, репрезентована семантичком мрежом – концепти из реченица су чворови графа, повезани гранама које представљају релације. Приметити да су и чворови и релације различитих типова.

Семантичке мреже су експресивна и флексибилна форма репрезентовања знања јер омогућавају:

- представљање ентитета из различитих домена проблема без ограничења,
- репрезентацију у форми графа, што омогућава коришћење алгоритама за претрагу и закључивање,
- природну визуелизацију повезаних појмова у форми графа,
- једноставно уочавање група елемената који су међусобно више повезани.

Приликом израде и коришћења семантичке мреже потребно је узети у обзир и могуће недостатке. Ако мрежа има мање формалних ограничења и правила за дефинисање веза, повећава се експресивност али и могућност погрешног тумачења репрезентације. Наведено ограничење ће бити илустровано на примеру мреже конструисане за ситуацију *Корозија цеви*. На слици 6.2 приказан је део мреже који за два пара објеката различитог типа има исту релацију „*важи*“. У првом случају *Критеријум* је део *Одлуке*, док у другом случају *Статус* даје стање *Активности*. Иако је додељивање истог имена вези између елемената семантички исправно, јер по значењу глагол *важити* може правилно да се тумачи у оба случаја, извођење знања из овакве репрезентације је ризично јер зависи од корисничког разумевања и тумачења природе релације: неко може претпоставити да је природа везе између оба пара чворова потпуно иста. Могуће решење би било креирање две различите релације, где би се у првом случају веза преименовала у „uslov_za“.



Слика 6.2: Одлуку „*уградити цеви*“ ако важи Критеријум „*дебљина зида након уклањања корозије већа од 3.4 мм*“ (лево). Активност „*инсталација цеви*“ за коју важи Статус „*стопирана*“ (десно).

Додатно ограничење је везано за моделирања *n*-арних релација између чворова (у мрежи су све релација бинарне). Ако је скуп чворова *C* потребно

повезати једном релацијом, неопходно је дефинисање новог чвора који представља релацију, као и релације између њега и осталих чворова из *C*.

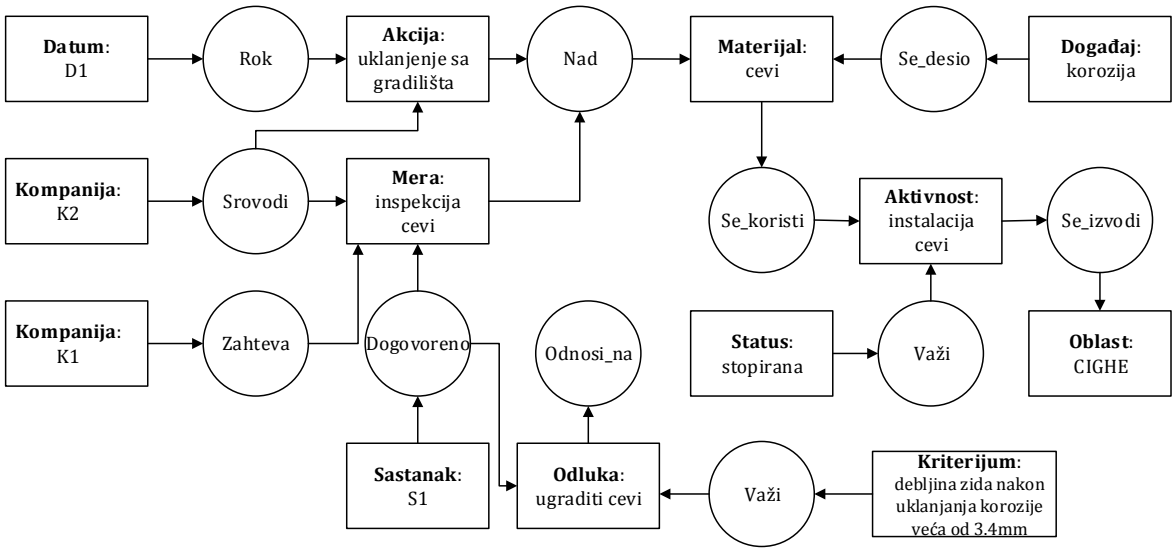
У општем случају, у недостатку формалне семантике, за манипулације мрежом користе се интерне процедуре засноване на логици првог реда. За комплексне проблеме који захтевају виши ниво експресивности, потребно је користити другу репрезентацију над којом је могуће коришћење логике вишег реда.

6.4 Концептуални графови

Концептуални графови су симболичка репрезентација знања заснована на појмовима концепата и *n*-арних релација између њих (Sowa 1992). Настају као последица немогућности класичних семантичких мрежа да репрезентују све специфичности природног језика (нпр. анафора / катафора – коришћење речи којима се референцира претходно / накнадно дефинисани појам). За разлику од мреже, и концепт и релација се представљају као чворови у графу (слика 6.3).

Релациони чворови омогућавају једноставно дефинисање *n*-арних релација – у наведеном примеру, релације *rok* и *se_desio* су бинарне (имају једну улазну и једну излазну везу), док су релације *dogovoreno* или *sprovodi* тринарне.

Ова репрезентација је посебно погодна за закључивање из знања изведеног из текстуалног корпуса (Kamaruddin et al. 2008) јер омогућава извођење комплексних операција над високо структуираном репрезентацијом, уз очување интерпретабилности. Међутим, услед нешто веће комплексности репрезентације, аутоматско генерисање графа представља проблем који ограничава њену ширу примену (Zhong et al. 2011). Аутори у раду наводе да граф поседују одговарајућу формализацију којом се могу описати комплексне структуре издвојене из текста, али да је пракса у већини случајева да се конструишу мануелно. Поред тога, графови заузимају знатно више меморије приликом имплементације на рачунару (и релације су чворови!).



Слика 6.3: Ситуација *Корозија цеви* као концептуални граф: приметити да су поред концепата (правоугаоници) и релације представљене чворовима (кругови). Релације су повезане само са концептима и обрнуто - коначни *бипартитни* граф.

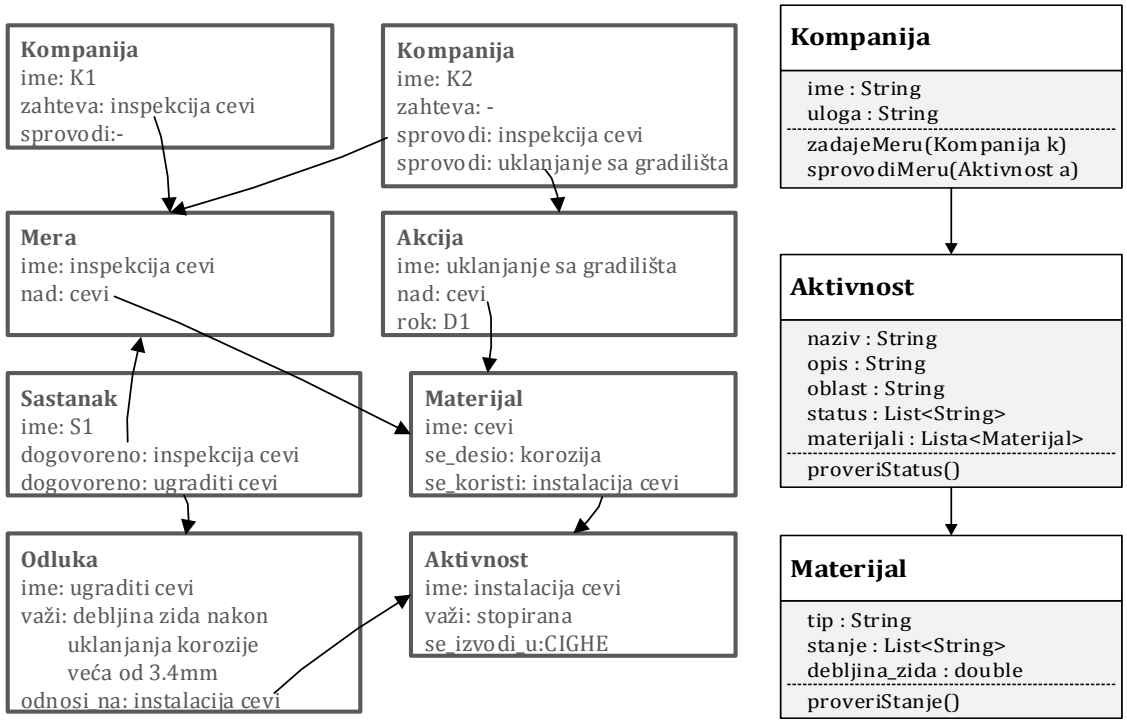
6.5 Оквири и објектно оријентисани приступ

Семантичке мреже и концептуални графови су креирани да би се што боље апроксимирао процес закључивања из природног текста. Алтернативни приступ, на коме се заснивају *оквири* и *објектно оријентисани приступ*, има за циљ уопштено моделирање процеса закључивања у људској меморији (Minsky 1975), где је знање организовано у структуре у којима су концепти хијерархијски уређени и повезани.

Оквир се састоји од *слотова* који могу да садрже карактеристике са конкретним вредностима, процедуре које могу да мењају карактеристике, или референце на друге оквире (Слика 6.4 лево). Различити типови оквира представљају класе. Оквири једне класе могу бити изведени из слотова друге класе (наслеђују њихове карактеристике).

Погодно окружење за примену оквира су *типске ситуације*, када постоји значајан број претходно познатих особина неког концепта или објекта – карактеристике се могу једноставно преносити или рачунати јер оквири природно подржавају наслеђивање, агрегацију и асоцијацију. У општем случају, оквири су погодни онда када је потребно репрезентовати модел са детаљно описаним објектима код којих је за велики број атрибута позната вредност или правила по којима се она одређује.

Парадигма оквира је послужила као основа за дефинисање концепта објектно оријентисаног приступа, који се заснива на описивању *стања* и *понашања* објекта. У овом приступу, *класама* су описани различити типови објеката, где сви објекти из заједничке класе имају исто дефинисано стање и понашање (слика 6.4). У програмским језицима, стање је имплементирано кроз промењљиве које описују објекат, а понашање кроз методе (функције) које се могу обавити над објектом. Као и код оквира, објекти могу да садрже референце на друге објекте.



Слика 6.4: Ситуација *Корозија цеви* као оквир (лево) и класни дијаграм објектно оријентисаног модела (десно). Сваки објекат је посебна инстанца класе са својим стањем, понашањем и референцама на друге објекте.

6.6 Одабир одговарајуће репрезентације знања

На основу наведеног може се закључити да су семантичке мреже и концептуални графови *погоднији* у случају када *претходно није позната* структура концепата који се издвајају из текста. Оквири (објектни приступ) би били преферирано решење када би информације садржане у неструктуираном тексту *имале изражену структуру* (познате типове објеката), хијерархијску уређеност (објекте који се изводе из других објеката) и позната правила интеракције објеката. С обзиром да ће издвојени концепти из текста имати *произвољну*, унапред *непознату* структуру, даље се као подразумеване репрезентације разматрају семантичке мреже и концептуални графови.

Поређење карактеристика мреже и графа указује на виши ниво експресивности и структуираности графа, што га чини погоднијим избором са

становишта интерпретабилности од стране експерта. Међутим, како се у предложеном приступу, концепти издвојен из неструктуираног текста доводе у везу само путем бинарних релација, као подразумевана репрезентација *изабрана је семантичка мрежа*. Још једном се напомиње да се истраживање не бави аутоматским резоновањем над репрезентацијом знања јер би то, у случају комплексног света инвестиционог пројекта, било готово немогуће учинити ефикасно. Репрезентација се у истраживању формира тако да буде погодна за различите врсте корисничких упита, као и за визуелизацију резултата. На експерту је да, на основу искуства и претходног знања, резултате доведе у одговарајући пројектни контекст и изведе самосталне закључке.

7 Предложена репрезентација информација

У овој глави описује се графовска репрезентација значајних фраза које су издвојене из докумената на пројекту, према поступку дефинисаном у глави 5. Репрезентација је структуирана тако да одговара семантичкој мрежи дефинисаној у поглављу 6.3. На овај начин се *истичу везе* које постоје између значајних концепата, што олакшава извођење нових знања по визуелној интерпретацији. Помоћу *визуелизације* издвојених образаца, учесници могу да сагледају текуће трендове на пројекту, што их *додатно мотивише* да истражују знање похрањено у неструктуктуираним подацима.

7.1 Одређивање релација између издвојених значајних фраза

Да би се добијене значајне фразе могле искористити за издвајање комплексних концепата, неопходно је успоставити различите типове релација између њих (у следећим примерима релације су подвучене: *local works* are *delayed due to* a *heavy rain*; *Петар Петровић ради за д.о.о. Градња*). Међутим, издвајање релација карактеристичних за различите домене захтевало би *аутоматску категоризацију* значајних фраза у одговарајуће доменске категорије, као и *дефинисање правила* која важе у посматраном домену. У примеру *Петар Петровић ради за д.о.о. Градња*, категоризација би подразумевала да су фразе *„Петар Петровић“* и *„д.о.о. Градња“* препознате као ентитети типа *Особа* и *Компанија*. По категоризацији фраза, правила за успостављање релације *ради за* могла би узети у обзир текстуалне секвенце типа *„Особа* из *Компанија“*, *„Особа* запослена у *Компанија“* и сличне.

Овакав приступ захтевао би значајан труд за дефинисање категорија и правила, која би често зависила од природе пројекта (аналогно приступу са

онтологијама, поглавље 3.3). Додатну потешкоћу представља то што су често, категорије релевантне за поједине ситуације на пројекту, непознате унапред (динамична природа пројекта). Категоризација и формирање правила зависе и од језика, а ресурси за ОПЈ нису доступни за многе језике (попут српског). Због тога се предлаже увођење доменски и језички *независног* поступка за успостављање релација између значајних фраза, заснованог на сличности заједничких семантичких контекста.

Нека $\mathbb{F}$ представља скуп свих значајних фраза који је издвојен из корпуса. Даље, нека су R_i и P_i скупови реченица и параграфа из свих докумената у којима се значајна фраза $f_i \in \mathbb{F}$ појављује. Под претпоставком да је $|A|$ кардиналност скупа A, бинарна релација $r \subset \mathbb{F} \times \mathbb{F}$, именована као *zajedno_sa*, дефинише се са:

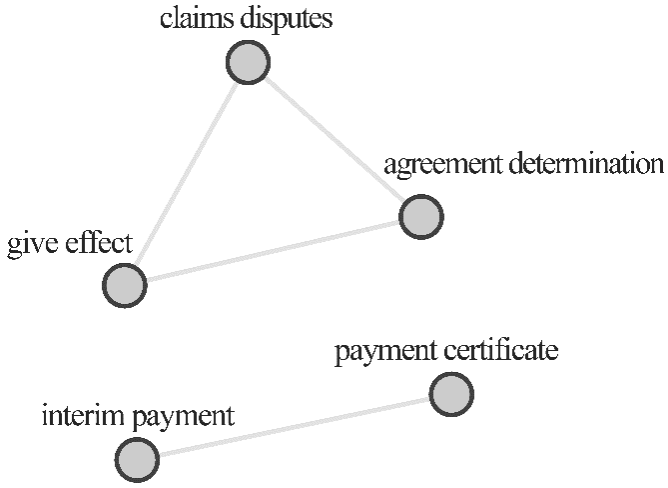
$$r = \left\{ \left(f_i, f_j\right) \;\middle|\; |R_{f_i} \cap R_{f_j}| > 0 \;\wedge\; \frac{|P_{f_i} \cap P_{f_j}|}{|P_{f_i} \cup P_{f_j}|} \geq t \in [0,1] \;\wedge\; |P_{f_i}| < |P_{f_j}| \right\} \qquad (7.1)$$

Ако $\left(f_i, f_j\right) \in r$, онда се релација означава као f_i *zajedno_sa* f_j. Неједнакост $|R_{f_i} \cap R_{f_j}| > 0$ уводи захтев да две значајне фразе *морају* да се појаве заједно *бар у једној* реченици. Међутим, да би се обезбедио виши ниво сличности заједничких семантичких контекста, потребно је да неједнакост $\frac{|P_{f_i} \cap P_{f_j}|}{|P_{f_i} \cup P_{f_j}|} \geq t$, која мери сличност одговарајућих скупова као Jaccard-ијев коефицијент, буде задовољена за претходно дефинисану границу t. Да би се боље описала хијерархијска структура значајних фраза, *смер* релације *zajedno_sa* дефинише се након поређења броја елемената у одговарајућим скуповима: за $|P_{f_i}| < |P_{f_j}|$, релација је усмерена од f_i ка f_j. Релација *zajedno_sa* је усмерена ка општијим (чешћим) значајним фразама. Специјални случај релације *zajedno_sa*, релација именована као *uvek_sa*, настаје када се значајна фраза f_i увек појављује заједно са f_j ($P_{f_i} \subset P_{f_j}$). Ако се обе значајне фразе појављују у истим контекстима ($P_{f_i} = P_{f_j}$), дефинише се бидирекциона релација *isti_kontekst* за f_i и f_j. Треба приметити да је

релација *isti_kontekst* релација еквиваленције која раздваја скуп значајних фраза у класе еквиваленције.

7.2 Конструкција значајних фраза састављених од више речи

Нека је F^2 скуп значајних фраза реда два који је издвојен из корпуса према комплетној методи из поглавља 5.4.3. Поступак детекције фраза вишег реда започиње конструисањем графа $G_{\text{isti\_kontekst}}$, уз помоћ релације *isti_kontekst*, над скупом чворова F^2. Како је *isti_kontekst* релација еквиваленције, добијени граф је *унија раздвојених клика*[15] (класа еквиваленције – слика 7.1). Ако све речи из свих значајних фраза у клики формирају секвенцу дужине *n*, клика представља значајну фразу реда *n* (слика 7.1 доле). Поред клика које представљају значајне фразе вишег реда, граф садржи и клике које то нису. Ове клике садрже значајне фразе реда два које се увек јављају у истим контекстима, а нису део значајне фразе вишег реда (слика 7.1 горе).



Слика 7.1: Клике са релацијом *isti_kontekst*: горе – различити парови речи који чине комплексни концепт; доле – значајна фраза реда три (*interim payment certificate*).

[15] Клика – комплетан подграф неког графа (свака два чвора у подграфу повезана су директном везом).

За детектовање скупа значајних фраза вишег реда $\mathcal{F}^n$ ($n \geq 3$), примењен је *Bron-Kerbosch алгоритам* (Bron & Kerbosch 1973) који детектује све *максималне клике* у неком графу[16]. Да би се разликовале ситуације описане на слици 7.1, као кандидати за значајне фразе реда *n* проглашене су клике које испуњавају следећи услов:

Клика од *m* фраза реда два представља значајну фразу реда *m*+1 ако свака фраза у клики садржи најмање једну реч која се једанпут јавља у другој фрази из клике.

Временска сложеност примењеног поступка диктирана је сложеношћу основног Bron-Kerbosch алгоритма који, за граф од *n* чворова, има експоненцијалну асимптотску сложеност[17] реда $O(3^{n/3})$. Извршавање алгоритма на стандардној радној станици (CPU Intel I5 3.2GHz, 4 језгра, 16GB RAM) за граф од приближно 1500 чворова (фраза реда два) траје око 5 минута. Ово ограничење не представља већи проблем јер се детекција фраза не обавља у време корисничких упита. За већи број почетних фраза може се користити нека од ефикаснијих техника за издвајање максималних клика, која има мању временску сложеност (Saxena & Thakur 2016).

Предложени приступ процењен је од стране експерата који су формирали оригиналне листе значајних фраза реда два. Експерти су добили фразе реда 3+, заједно са семантичким контекстима у којима су се јављале. На овај начин, експерт је могао да процени да ли је издвојена фраза вишег реда релевантна или не. Експеримент је понављан за различит број фраза реда два помоћу кога је конструисан $G_{\text{isti\_kontekst}}$. Прецизност предложеног приступа је приказана у табели 7.1.

[16] Максимална клика је она клика која то више не би била ако би јој се придодао било који преостали чвор из графа (он тада не би био директно повезан са свим чворовима из клике).

[17] Асимптотска временска сложеност алгоритма говори о брзини пораста времена извршавања када величина улазних података тежи бесконачности.

Табела 7.1: Конструисање значајних фраза вишег реда, из почетног скупа F^2, када се варира његова димензија.

Корпус	Број фраза реда два	Број клика кандидата	Број фраза вишег реда	Прецизност
$K_{Бор}$	500	9	4	0.444
$K_{Бор}$	1000	44	17	0.386
$K_{Бор}$	1500	64	21	0.328
$K_{Коридор}$	500	11	7	0.636
$K_{Коридор}$	1000	22	12	0.545
$K_{Коридор}$	1500	53	22	0.415

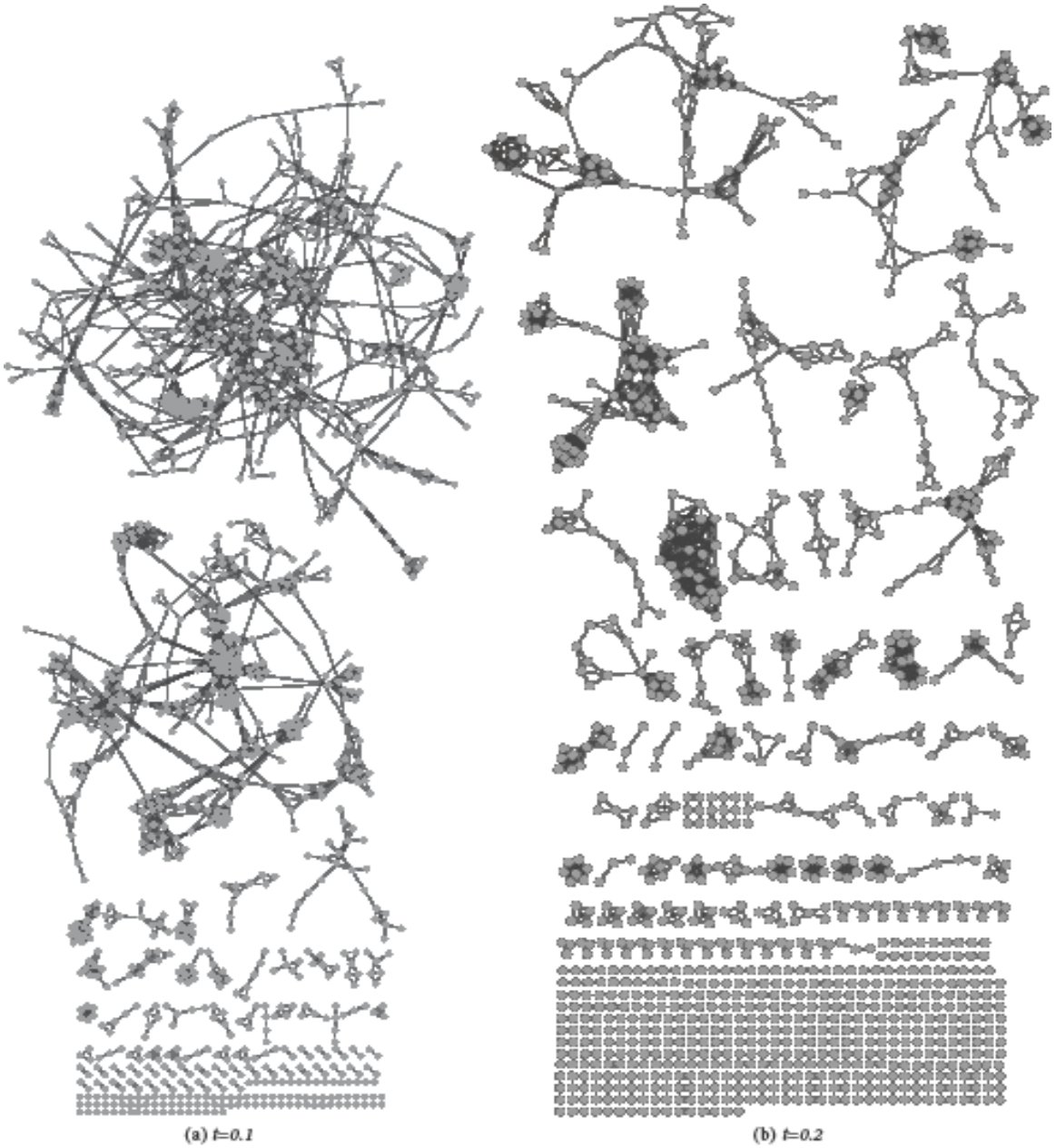
Резултати сугеришу да се из већег броја значајних фраза реда два добија више значајних фраза реда 3+. Међутим, прецизност почиње да *опада* са повећањем броја почетних фраза јер је омогућено формирање већег броја неинформативних клика кандидата. Бољи резултати добијени су за корпус $K_{Коридор}$ који садржи више докумената са дужим, дескриптивним реченицама, што омогућава да се формирају валидни комплексни концепти.

7.3 Граф значајних фраза

Пошто се заврши процес издвајања, *граф значајних фраза G*, који као чворове садржи *фразе свих редова*[18], конструише се уз помоћ релација *zajedno_sa* и *uvek_sa* (поглавље 7.1). Два примера графа са различитим вредностима за параметар који дефинише *минималну јачину релације* (t у једначини (7.1)), конструисана из корпуса $K_{Коридор}$**,** приказана су на слици 7.2. У поређењу са $G_{t=0.2}$, граф $G_{t=0.1}$ садржи значајно више релација, као и мањи број чворова који нису повезани са остатком графа. Два највећа подграфа у оба графа се искључиво састоје од значајних фраза из различитих језика (реченице на српском и

[18] У даљем тексту се под појмом „граф“, или ознаком *G*, подразумева граф значајних фраза свих редова.

енглеском језику). Остали подграфови представљају мање „светове“ који су везани за одређене независне теме, који су такође језички раздвојени.



Слика 7.2: Два графа значајних фраза на пројекту, конструисана на скупу од 1225 значајних фраза: (а) Граф $G_{t=0.1}$, са границом за успостављање релације од 0.1, садржи 2254 релације (б) Граф $G_{t=0.2}$ са границом од 0.2 и 1143 релације.

У даљем тексту, сваки *повезани* подграф (постоји пут између било која два његова чвора) графа *G* ће се сматрати *кандидатом* за неки комплексни концепт (тему). Коначну одлуку о томе да ли подграф представља комплексни концепт доноси корисник са пројекта после интерпретације. Да би се кориснику олакшала интерпретација, сваки чвор је *проширен референцама* на изворне документе (реченице и параграфе), па се могу проверити семантички контексти којима неки скуп фраза припада. О применама ове репрезентације на инвестиционом пројекту биће речи у глави 9.

7.4 Рангирање значајних фраза на основу варијабилности суседства у графу

У поглављу 5.3 приказано је да се применом ентропије, приликом издвајања значајних фраза, могу отклонити неке неинформативне фразе. Овај поступак ефикасно филтрира оне кандидате који се често јављају у *суседству истих речи* у документу (нпр. типски документи). Међутим, недостатак приступа са ентропијом суседстава речи је што не детектује неинформативне контексте који су представљени *различитим речима* са *истим значењем*. Проблем је илустрован на примеру следеће две реченице:

На претходном састанку је договорено да дипл. грађ. инж. Петар Петровић достави потребну документацију у дефинисаном року.

Дипл. грађ. инж. Петар Петровић је у дефинисаном року доставио потребну документацију, као што је договорено на претходном састанку.

Са становишта методе ентропије суседстава речи, сви парови речи издвојени из две приказане реченице били би третирани као да потичу из различитих контекста, што заправо није случај. Имајући на располагању граф значајних фраза као репрезентацију информација, истражене су могућности за прецизнију

детекцију фраза које се јављају у различитим суседствима, али *не у документима*, већ *у графу*[19]. Претпоставка учињена у истраживању је да су фразе, чије је суседство у графу променљиво током времена, значајније (информативније), те да се на тај начин могу рангирати (филтрирати).

7.4.1 Динамичност суседа у графу

За проблем рангирања фраза према променљивости суседства у графу, примењен је поступак изложен у (Goenawan et al. 2016). Поступак се заснива на одређивању мере *динамичности суседа* у графу који *еволуира кроз време*. Претпоставимо да се граф G мења кроз временске тренутке од 1 до n, и да су његове манифестације у тим тренуцима G_1, G_2, …, G_n. Ако се уочи скуп $C = \{c_1, c_2, \ldots, c_m\}$ који чине сви различити чворови из графова G_1, G_2, …, G_n, онда се сваки граф G_k може представити по једном квадратном матрицом суседства $\mathbf{S}^k$, реда m, где је $s_{ij}^k = 1$ ако у графу G_k постоји веза између чворова c_i и c_j. Ако веза не постоји, или ако G_k не садржи чворове c_i и/или c_j, онда је $s_{ij}^k = 0$.

Матрица суседства може се илустровати на примеру реченица, где свака реченица индукује по један граф G_k, а све различите речи из реченица представљају скуп C. У примеру се за две речи сматра да су суседи ако се јављају у истој реченици. Следи пример:

Треба започети истражне радове.	(граф G_1)
Истражни радови су у току.	(G_2)
Који је рок за истражне радове?	(G_3)
Сви започети радови су завршени у року.	(G_4)
Сви започети истражни радови су завршени у року.	(G_5)

[19] Суседство у графу подразумева друге фразе које су директно повезане са посматраном.

На слици 7.3 је приказана матрица суседства $\mathbf{S}^2$, за реченицу представљену графом G_2.

	Радови	треба	започети	Истражни	току	Рок	сви	завршени
радови	1			1	1			
треба								
започети								
истражни	1			1	1			
току	1			1	1			
рок								
сви								
завршени								

Слика 7.3: Матрица суседства $\mathbf{S}^2$ за реченицу *„Истражни радови су у току“*. У приказаној реченици нису разматране стоп речи (*су*, *у*).

Приметити да *i*-та врста матрице $\mathbf{S}^k$ представља суседство чвора c_i у *k*-том графу. У поступку се даље дефинише *матрица просечног суседства* $\bar{\mathbf{S}}$, за све чворове из *C,* чији се елемент $\bar{s}_{ij}$ израчунава као:

$$\bar{s}_{ij} = \frac{1}{n}\sum_{k=1}^{n} s_{ij}^{k} \qquad (7.2)$$

Приметити да *i*-та врста матрице $\bar{\mathbf{S}}$ представља *просечно суседство* чвора c_i у току еволуције графа *G* кроз G_1, G_2, …, G_n. *Динамичност суседа* за сваки чвор c_i из *C* може се израчунати као средња вредност растојања свих његових суседстава од просечног суседства:

$$din_i = \frac{1}{n}\sum_{k=1}^{n}\sqrt{\sum_{j=1}^{m}(s_{ij}^{k} - \bar{s}_{ij})^2} \qquad (7.3)$$

7.4.2 Експериментална провера рангирања значајних фраза према динамичности суседства у графу

Приказани поступак тестиран је на корпусу од 30 докумената издвојених из корпуса $К_{Бор}$. Документи припадају категорији *Записник са састанка*, и покривају период од октобра 2013. до јуна 2014. године. Сваки документ је третиран као посебан граф чији су чворови претходно детектоване значајне фразе. Поступком *комплетне методе* (поглавље 5.4.3) формирана је листа значајних фраза за цео корпус (листа КМ). Значајне фразе из листе КМ рангиране су према мери динамичности суседства, чиме је формирана нова ранг листа (листа ДС).

Да би се тестирао квалитет мере варијабилности суседства за рангирање значајних фраза, три експерта су обележила степен значаја сваке фразе. Пошто је насумице извучен по један параграф из сваког записника (слика 7.4), експерти су оценили значај обележених фраза у параграфу, а које припадају листама КМ и ДС.

Employer stated that the proposal for remedial works on the corrosion protection of ESP will be delivered during the week. The Employer pointed out that the most critical delays are noted on the positions of steel structure prefabrication and installation.

Contract's amendments regarding new agreed mechanical completion date will be discussed on the claim meetings.

Слика 7.4: Примери параграфа за које је обележен степен значаја фразе. Експерти су, имајући у виду контекст параграфа, издвојеним фразама (подвучено) давали оцену *мало*, *умерено* или *веома значајно*.

Фраза се проглашава *веома значајном* ако су се сва три експерта независно определила да је тако оцене - ако је нека фраза два пута оцењена као *веома* и једном као *умерено* значајна, она се не категорише као веома значајна.

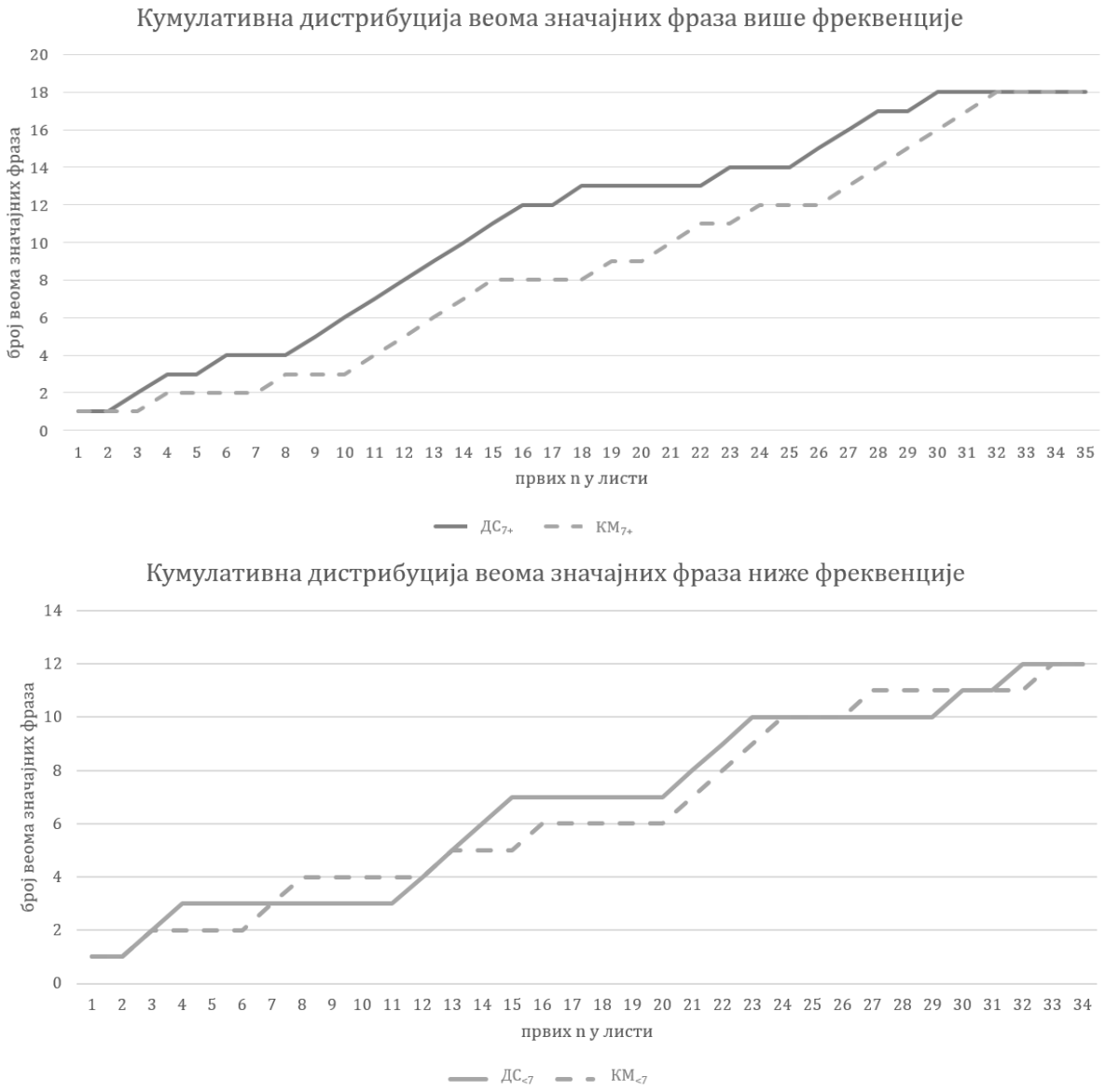
Након обележавања од стране експерта, обе листе фраза (КМ и ДС) су модификоване на следећи начин (табела 7.2):

- редуковане су тако да садрже само фразе које су детектоване у параграфима за валидацију,
- свака листа је подељена на две засебне листе, *према фреквентности фраза* у корпусу записника са састанака. Листе $КМ_{<7}$ и $ДС_{<7}$ садрже фразе које се јављају у мање од седам параграфа, а листе $КМ_{7+}$ и $ДС_{7+}$ садрже фразе које се јављају у седам и више параграфа.

Табела 7.2: Карактеристике листи значајних фраза сортираних према комплетној методи и динамичности суседа.

Листа	**укупно фраза**	**број веома значајних фраза**
$КМ_{<7}$	34	12
$ДС_{<7}$	34	12
$КМ_{7+}$	35	18
$ДС_{7+}$	35	18

Како је након експертског обележавања познато које су фразе веома значајне, може се утврдити како су распоређене унутар сваке од четири листе (слика 7.5). Оптимално рангирање би подразумевало да се, за ранг листу која садржи m веома значајних фраза, све налазе у првих m по значају.



Слика 7.5: Расподела фраза које су експерти обележили као веома значајне. За фразе више фреквенције (листе $КМ_{7+}$ и $ДС_{7+}$), прираштај криве дистрибуције за $ДС_{7+}$ је већи него за $КМ_{7+}$ – *више фраза је боље рангирано* у листи $ДС_{7+}$. За листе $КМ_{<7}$ и $ДС_{<7}$ нема изражене разлике у прираштају криве дистрибуције.

Резултати експеримента показују да накнадно рангирање фреквентнијих фраза, према мери динамичности суседства, *боље рангира* веома значајне фразе. Експертско обележавање указује и да постоји корелација између значаја фразе и

броја појављивања у више различитих контекста. За нискофреквентне фразе, описани поступак не побољшава перформансе комплетне методе.

Како су експерти додељивали фразама категоријске оцене (*мало*, *умерено* и *веома значајно*), процењена је њихова сагласност на нивоу свих фраза. Узевши у обзир специфичности експеримента (три оцењивача, категоријске оцене), за процену сагласности је коришћен Флајсов капа коефицијент (Fleiss & Cohen 1973). За фреквентније фразе, Флајсов капа коефицијент износи 0.617, док је за фразе које су се ређе појављивале коефицијент износио 0.38. Закључује се да експерти *нису били сагласни* у оцењивању значаја фраза мање фреквенције, па не чуди што ни рангирање на нивоу динамичности суседа, за тај случај није дало уочљиво побољшање.

Предложена репрезентација, у виду графа значајних фраза свих редова, треба да се ускладишти на начин који ће омогућити алате за претрагу, визуелизацију и извођење нових знања. У следећој глави биће речи о системима за складиштење и приступање подацима који могу послужити у ову сврху.

8 Складиштење и приступање репрезентацији значајних фраза

Правила за креирање репрезентације значајних фраза омогућавају да се креира концептуални модел из кога је могуће изводити нова знања. Да би се модел имплементирао, потребно је превести репрезентацију у форму која омогућава да се над њом обављају различите операције. У наставку ће бити описани начини складиштења произвољне структуиране репрезентације информација, са освртом на могућност коришћења у случају полуструктуираних текстуалних формата.

База података представља структуирану колекцију која дозвољава приступ и ажурирање података похрањених у њој. Похрањени подаци могу бити различитих типова као што су бројни, текстуални, логички, темпорални и други. Основно значење појма „база података“ односи се на податке и придружену *шему* (опис података). Шема базе дефинише начин на који су подаци у бази организовани. Шема одређује:

- имена ентитета из модела
- особине које описују ентитете
- типове и домене вредности појединачних особина
- начин на који су ентитети у бази повезани
- ограничења која се односе на вредности и везе између ентитета (*интегритет података* - (Gertz & Lipeck 1995))

Да би се убрзала претрага података у бази, практикује се креирање хијерархијски организованог скупа показивача на податке који се претражују - *индекса*. Ако се подаци организују у колекције по редоследу пристизања, у

општем случају они нису сортирани. Проналажење траженог податка из колекције дужине n представља проблем линеарне претраге реда сложености $O(n)$. Креирање индекса над сортираним подацима омогућава примену алгоритама за претрагу сортираних колекција попут бинарне претраге. Бинарна претрага има ред сложености $O(log(n))$, па се на овај начин постиже значајно убрзање. Код коришћења индекса треба водити рачуна о додатном меморијском простору за његово складиштење. Међутим, иако је то могуће, није препоручљиво индексирати све податке из базе.

Поред концепта база података, битан концепт је и *Систем за управљање базом података* (СУБП[20]). СУБП је софтвер за креирање и управљање базом. Он омогућава да клијент (друга апликација или корисник) ускладишти, претражи и ажурира податке у бази. СУБП омогућава извршавање *трансакција* над подацима. Трансакција се дефинише као јединица посла коју чине једна или више операција над подацима (креирање, додавање, брисање и ажурирање). У општем случају, трансакција представља сваку измену у бази и пожељно је да има атомски карактер – све операције које чине трансакцију морају се успешно извршити, или се не сме извршити ниједна.

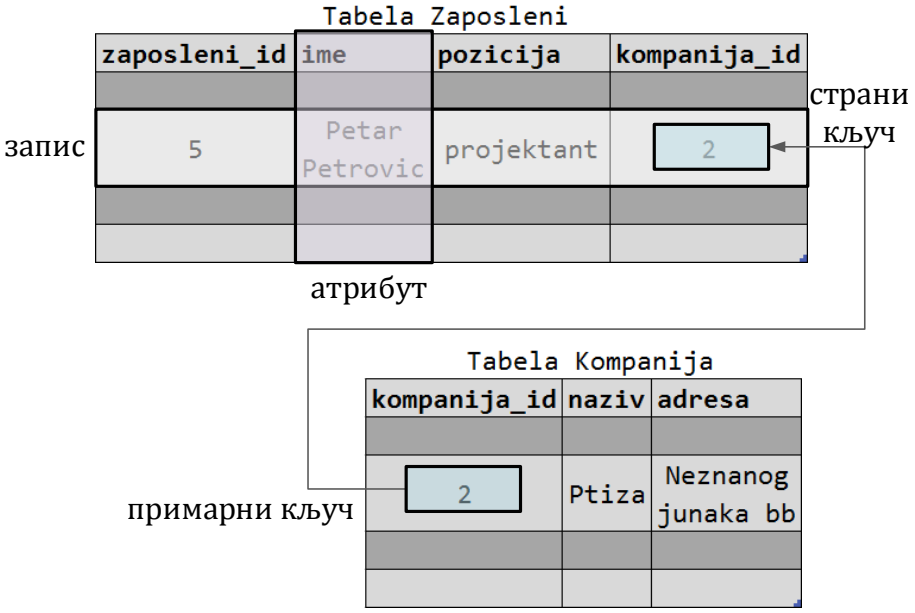
У зависности од начина за складиштење, приступање и манипулацију подацима у бази, СУБП се могу класификовати у више типова. У наставку ће бити приказане *основне* карактеристике база коришћених у докторској дисертацији - *релационих* и *графовских*.

8.1 Релационе базе података

Релационе базе података се заснивају на концепту складиштења података у табелама (слика 8.1). Табела представља апстракцију којом се моделирају ентитети из стварног живота (нпр. *Запослени*, *Машина*, *Пројекат*, итд.). Посматрани ентитет је описан скупом *атрибута* који говоре о његовим

[20] Енглески израз је Data Base Management System (DBMS)

особинама (нпр. за табелу *Запослени* атрибути могу бити *име*, *адреса*, *позиција*, *плата,* и друге). У зависности од типа података којима је атрибут представљен, дефинисани су различити домени вредности атрибута. У наведеном примеру атрибути *име*, *адреса* и *позиција* могу бити произвољни текстови, док атрибут *плата* може да се представи као реални број. Редови табеле представљају записе о ентитетима, а колоне чувају вредности одговарајућих атрибута за све записе (Codd 1983).



Слика 8.1: Основни приказ структуре релационе базе података.

Сваки запис у табели има јединствену идентификацију у виду *примарног кључа* кога чине један или више атрибута. Атрибути који имају различите вредности за све записе и, при томе, немају недодељених вредности (NULL), представљају *кандидате* за примарни кључ. Постојање примарног кључа омогућава да се ентитети различитог типа повежу, чиме се моделирају различите релације између ентитета. Табеле А и Б повезују се тако што се сваком запису из табеле А, додаје одговарајући кључ из табеле Б (примарни или кандидатни кључ у Б, а *страни кључ* у А, слика 8.1). Доделом *страног кључа* не сме се нарушити интегритет релационе базе - сваки страни кључ мора да показује на јединствени

постојећи запис из табеле коју референцира. Спајање табела помоћу кључева омогућава да се подаци из различитих табела не преклапају – *нормализација података*.

СУБП код релационих база најчешће користи SQL[21] упитни језик за приступање и манипулацију података похрањених у бази (Date & Darwen 1997). SQL је декларативни језик заснован на релационом рачуну – наводи се опис жељених података, без потребе да се дефинише начин добијања.

Релационе базе погодне су за структуиране податке са јасно дефинисаном *временски непроменљивом* шемом. Непроменљива шема са издиференцираним подацима којима су прецизно дефинисане карактеристике, омогућава извршавање комплексних трансакција које у потпуности чувају интегритет података.

8.2 NoSQL базе података

Непроменљива шема у релационим базама показала се као недовољно ефикасна за моделирање података који се појављују у великом обиму, разноврсним форматима и који описују временски променљиве ентитете (Mohan 2013). Да би обрадиле такве податке, велике Веб-оријентисане компаније (Google, Facebook, Amazon и сл.) су интензивно почеле да користе базе података са *променљивом шемом*, у којима се подаци не похрањују у фиксним табелама, већ у документима са динамички променљивом структуром - NoSQL[22] базе података. NoSQL системи представљају дистрибуиране, не-релационе базе предвиђене за складиштење података већег обима и њихову обраду у паралели, на великом броју заменљивих сервера (Moniruzzaman & Hossain 2013).

Са порастом обима података доступних на Веб-у почетком 21-ог века, NoSQL базе постају шире распрострањене. Први познатији типови NoSQL база су:

- Google-ова BigTable колонска база (Chang et al. 2006);

[21] SQL - Structured Query Language
[22] NoSQL - Not only SQL

- Amazon-ова Dynamo кључ-вредност база (DeCandia et al. 2007).

Поред наведена два типа, постоји више врста NoSQL база прилагођених раду са различитим структурама података, од којих су најпознатије:

- Документ база,
- Графовска база,
- Објектно оријентисана база,
- XML база.

На овом месту треба указати на одређена погрешна тумачења разлика релационог и NoSQL концепта. Они нису међусобно искључиви, нити је у општем случају један приступ бољи од другог. Прихватање једног концепта не значи да одређене особине другог не могу да се примене; у последњих неколико година појавили су се хибридни системи који комбинују особине оба типа – NewSQL базе (Moniruzzaman 2014).

Приликом пројектовања базе података треба узети у обзир контекст проблема и одабрати одговарајућу имплементацију. У табели 8.1 приказане су кључне разлике релационог и NoSQL концепта.

У NoSQL бази клијент може динамички да дода нове или обрише постојеће атрибуте на нивоу записа. На овај начин, поред СУБП-а, и клијент сноси одговорност за правилно функционисање базе јер мора да обезбеди одговарајућу интерпретацију података у овом случају. У супротном, може да дође до неконзистентности у раду са NoSQL базом.

Табела 8.1: Поређење особина релационих и NoSQL база података

	Релациона база	**NoSQL база**
Записи	Сви записи имају исте атрибуте.	Записи могу имати различите атрибуте.
Шема	Фиксна шема - мора бити дефинисана пре почетка коришћења базе. Ако се промени током рада, постојећи подаци морају се прилагодити промени.	Динамичка шема која се може мењати у току коришћења базе. Ако се промени током рада, постојећи подаци *не морају* се прилагодити промени.
Нормализација	База је обично нормализована (нема редундантних података).	База је у општем случају денормализована (има редундантних података).
Складиштење	Табеле са предефинисаним типовима података.	Различите структуре података попут табеле са променљивим бројем колона по врстама, документа, графа, табеле кључ-вредност, и других.

Приликом обављања упита, због променљиве шеме, у NoSQL приступу захтева се провера текуће структуре података, што може да успори извршавање трансакције.

У денормализованој NoSQL бази, подаци о једном запису се налазе на једном месту, услед чега им се брже приступа. Последица је постојање редундантних података који су придружени различитим записима, што захтева више меморије за складиштење базе и успорава трансакције у којима се ти подаци ажурирају. У општем случају, денормализација NoSQL базе имплицира да се, у односу на релациону базу, подаци *брже читају* и *спорије ажурирају*. Овај сценарио примерен је за ситуације у којима се користе велике количине података које се по природи не ажурирају, већ читају или анализирају (нпр. подаци о продаји у току једне године на нивоу супермаркета; мерења неких физичких величина у току времена). Слично *важи и за документе на инвестиционом пројекту* – када су једном размењени између учесника, документи се не модификују (уобичајено је слање нове верзије документа).

У табели 8.2 приказани су различити сценарији на основу којих је препоручљиво изабрати одговарајући концепт базе података.

Табела 8.2: Критеријуми за одабир типа базе података.

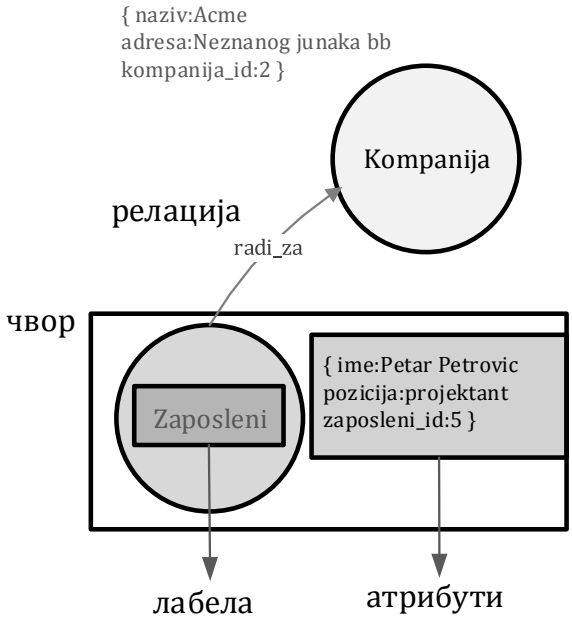
Релационе базе	NoSql базе
Опис података познат је унапред	Опис података није до краја познат или се често мења
Претходно познате непроменљиве релације између ентитета	Релације међу ентитетима креираће се динамички
Значајни интегритет и конзистентност базе	Значајна скалабилност базе
Често ажурирање података	Ретко ажурирање, често читање података
Угњеждене или комплексније трансакције	Једноставније трансакције

Због очекивано високе међусобне повезаности значајних фраза издвојених из докумената на пројекту, у наставку ће детаљније бити обрађене графовске базе података, као посебно погодне за записе са *великим* бројем *међусобних* веза.

8.3 Особине графовских база података

NoSQL графовске базе су прилагођене за складиштење и манипулацију над подацима организованим у форми графа (слика 8.2). Ентитети који се моделирају представљени су различитим типовима чворова, инстанце ентитета (записи из релационе табеле) чворовима, а релације између ентитета као *директне* везе између чворова. Чворови и везе могу имати атрибуте који их додатно описују. Последица овакве организације података је једноставно дефинисање упита за проблеме специфичне за теорију графова, као што су проналажење *најкраће*

путање између чворова (Dijkstra 1959), или рангирање Веб страница (Brin & Page 1998).



Слика 8.2: Основни елементи графовске базе података.

Кључна разлика између релационих и графовских база према (Urma & Mycroft 2015) је *складиштење суседства без индекса*, што значи да сваки чвор графа садржи *референцу* на суседне чворове. Складиштење суседства без индекса је могуће јер су релације у графовској бази ускладиштене тако да се могу идентификовати независно у односу на друге релације и ентитете. Да би се пронашли повезани чворови, није неопходно комбиновати податке из различитих табела по истом кључу, као што се чини у релационим базама. Ова особина графовске базе омогућава једноставно задавање и брже извршавање упита у којима је неопходно испитати велики број веза између чворова, или извршити обилазак графа, јер се *директно* приступа релацијама ускладиштеним на диску.

Релациона табела, којом се моделира веза са атрибутима између табела *A* и *B*, садржи записе који су једнозначно дефинисани страним кључевима из *A* и *B*.

Последично, да би се прикупили одговарајући подаци из различитих табела, потребан је упит *придруживања*[23] који представља комбиновање вредности из колона различитих табела. Са порастом броја табела које садрже велики број записа, значајно расте време потребно за њихово повезивање у упиту придруживања. Овакви упити у релационим базама могу се убрзати коришћењем проточне обраде упита и сличним техникама паралелизације. Са друге стране, складиштење свих релација као засебних ентитета, у графовској бази захтева већу меморију.

У репрезентацији описаној у глави 7, постоји велики број веза које се мењају кроз време – корпус се, у току животног циклуса пројекта непрестано повећава). Због тога је за њено складиштење одабрана графовска база података Neo4j.

8.4 Neo4j графовска база података

Neo4j је графовска база података имплементирана у окружењима Java и Scala. Развијена је од стране компаније Neo Technology, 2007. године. Према бази знања о системима за управљање базама података[24] из септембра 2017. године, рангирана је према популарности као 21. од свих база података и као *прва* међу графовским базама. Доступна је у верзијама Community Edition (GPL лиценца) и Enterprise Edition (комерцијална лиценца). Главне особине Neo4j базе су:

- База се састоји из два основна елемента – *чвора* и *релације*. Чворови и релације могу садржати атрибуте, док чворови морају да садрже бар једну *лабелу* (описни тип чвора); Релације *морају* бити *усмерене*, а могу бити и рефлексивне (релација почиње и завршава се у истом чвору);
- Приликом складиштења, у сваком чвору се чувају референце на суседне чворове, што значи да један корак претраге графа има ред сложености $O(1)$;

[23] Придруживање - Join
[24] Извор: https://db-engines.com/en/ranking

- За упите над подацима из графа користи се декларативни језик *Cypher;*
- База у потпуности подржава трансакције (сагласна је са ACID[25] принципима).

Једно од првих поређења перформанси графовске Neo4j базе у односу на релациону базу приказано је у (Vukotic et al. 2015). У описаном експерименту су анализирани односи у друштвеној мрежи, чији модел је, поред графовског, репрезентован и као релациони у MySQL окружењу. За одређивање пријатеља за све особе, коришћен је упит укрштеног придруживања који представља Декартов производ вредности из две табеле (комбиновање сваког реда једне са сваким редом друге табеле). Када је повећавано растојање на коме се налазе две особе (проналажење пријатеља мојих пријатеља итд.), растао је и број потребних укрштених придруживања, те се значајно повећало време извршавања упита.

У експерименту из (Vukotic et al. 2015), анализирана је база са 1 000 000 особа од којих је свака имала 50 пријатеља у просеку. У табели 8.3 су приказана времена потребна за проналажење пријатеља на растојањима од два до пет. Може се уочити да се упит за особе на растојању два (пријатељи пријатеља) извршава за приближно исто време (као последица оптимизације релационе базе за коришћење индекса приликом упита придруживања). Међутим, приликом понављања упита за већа растојања, потребно време у релационој бази значајно се повећало, док се за суседе на растојању пет, упит није извршио ни после сат времена!

[25] Atomicity - атомичност, Consistency - конзистенција, Isolation - изолација, Durability – трајност (Gray & Reuter 1993)

Табела 8.3: Време извршења упита укрштеног придруживања у релационој и графовској бази података (Vukotic et al. 2015).

растојање	MySQL [s]	Neo4j [s]	враћених записа
2	0.016	0.01	2500
3	30.267	0.168	125000
4	1,543.51	1.359	600000
5	> 1 сат	2.132	800000

Експерименти из више извора показују да се упити који приступају великом броју релација брже извршавају у графовској бази (Vicknair et al. 2010; Hölsch et al. 2017; Joishi et al. 2016). Са друге стране, упити са великим бројем нумеричких операција брже се извршавају у релационој бази. Треба напоменути да резултати у значајној мери *зависе од конфигурације* оба система, па је потребно обезбедити исте услове тестирања, о чему није било речи у (Vukotic et al. 2015).

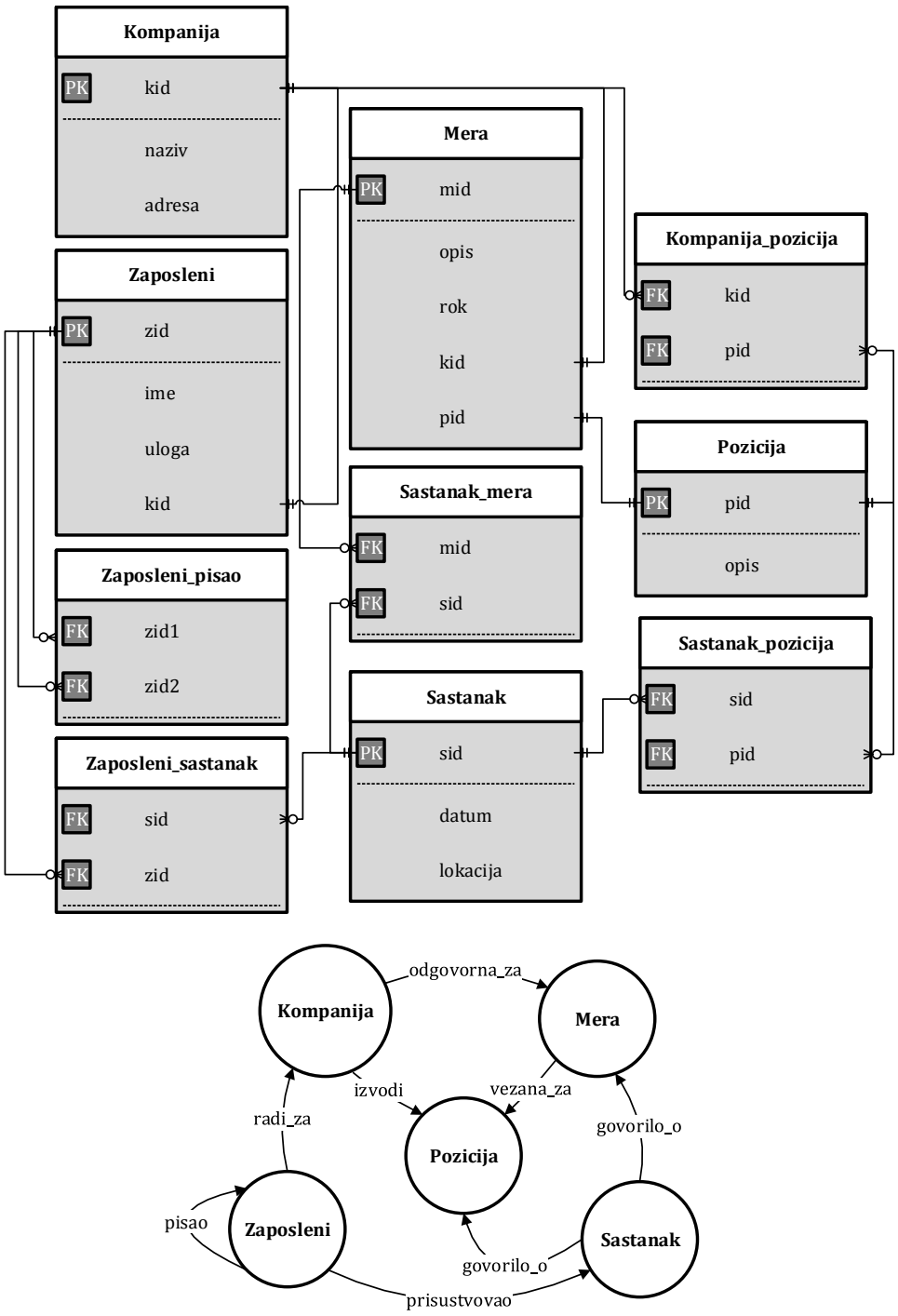
Разлика у извршавању и задавању трансакција у SQL и Cypher упитним језицима је детаљно анализирана у (Hölsch et al. 2017). Аутори су у експерименту користили базу која моделира универзитетску установу. Модел је имплементиран у Neo4j и MySQL базама, за које су дефинисани група аналитичких упита и упита по структури. Аналитички упити рачунали су параметре из графовске анализе (средишњост, централизација и степен чвора), и *брже* су се извршавали у релационој MySQL бази. Према ауторима, могући разлог односи се на потребу да се приступи сваком запису табеле / чвору графа како би се параметри израчунали. Тада до изражаја долази архитектура релационе базе која је *оптимизована* за *секвенцијални пролазак* кроз табелу. Са друге стране, упити по структури, где су задати обрасци за претрагу по релацијама између ентитета, *брже* су се извршавали у графовској бази. Разлика је посебно изражена када је у релационој бази било неопходно вишеструко извршавање упита укрштеног придруживања.

8.5 Поређење релационих и графовских база података на погодном примеру

Да би се илустровале специфичности рада у графовској бази, за потребе овог истраживања креиран је модел којим се репрезентују пословни процеси и интеракција учесника карактеристични за инвестиционе пројекте у грађевинарству. Исти концептуални модел похрањен је у Neo4j и PostgreSQL[26] базама података (слика 8.3). Односи између ентитета у моделу су следећи:

- Сваки запослени ради за једну компанију;
- Запослени из различитих компанија су присуствовали састанцима;
- На састанцима се говорило о позицијама радова (земљани радови, постављање оплате и сл.), као и мерама које треба предузети за одређене позиције (повећати број радника, изменити пројектну документацију и сл.);
- За позицију може бити одговорно више компанија;
- Свака мера се односи на конкретну позицију и за њу је одговорна једна компанија;
- Запослени су међусобно обављали коресподенцију електронском поштом.

[26] Будући да је опште позната и широко распрострањена, PostgreSQL релациона база није описана у овом раду. За детаљни опис видети https://www.postgresql.org .



Слика 8.3: Шема за релациону (горе) и графовску базу података (доле). Лабеле PK и FK се односе на примарни и страни кључ у релационој бази. Приметити колико је графовска шема *природнија* и *једноставнија* за разумевање.

У базама су, за потребе експеримента, извршени следећи упити:

***Упит 1**: За све компаније које су одговорне за мере на позицијама које изводе, приказати име компаније и опис позиција и мера*

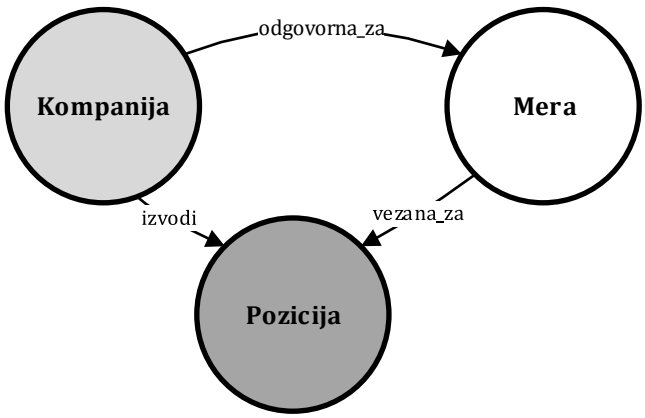
Cypher:

```
match (k:Kompanija)-->(m:Mera)-->(p:Pozicija)<--(k)
return k.naziv,m.opis,p.opis
```

SQL:

```
select k.naziv, m.opis, p.opis
from kompanija k, mera m, pozicija p, kompanija_pozicija kp
where k.kid = m.kid and k.kid = kp.kid
and kp.pid = p.pid and m.pid = p.pid
```

Упит илуструје повезивање ентитета у језику Cypher: тражена тројка ентитета је затворена референцирањем компаније *k* на позицију *p* (у супротном би биле приказане све позиције за које су одређене мере – слика 8.4). Уочити *једноставност* Cypher упита у односу на SQL упит.



Слика 8.4: Упит проналажења тројки чворова (записа) који су међусобно повезани.

***Упит 2**: Приказати датуме пет састанака на којима је било највише учесника.*

Cypher:

```
match (s:Sastanak)<--(z:Zaposleni)
return s.datum, count(*) as broj
order by broj desc limit 5
```

SQL:

```
select t.nu, s.datum from
     (select count(zid) as nu, sid
     from zaposleni_sastanak group by sid) t, sastanak s
where t.sid=s.sid
order by t.nu desc limit 5
```

Упитом се илуструје агрегирање у Cypher-у, уз помоћ функције count(*) која враћа број релација између чворова *Састанак* и *Запослени*, што је овде еквивалентно броју учесника на састанку. За овај упит у SQL-у, прво је креирана привремена табела са бројем учесника по састанцима. Она је накнадно повезана са табелом *Састанци* како би се добили тражени датуми. Уочити да је Cypher упит *једноставнији* у односу на SQL упит.

***Упит 3**: За сваког запосленог приказати запослене са којима је повезан преписком до растојања четири.*

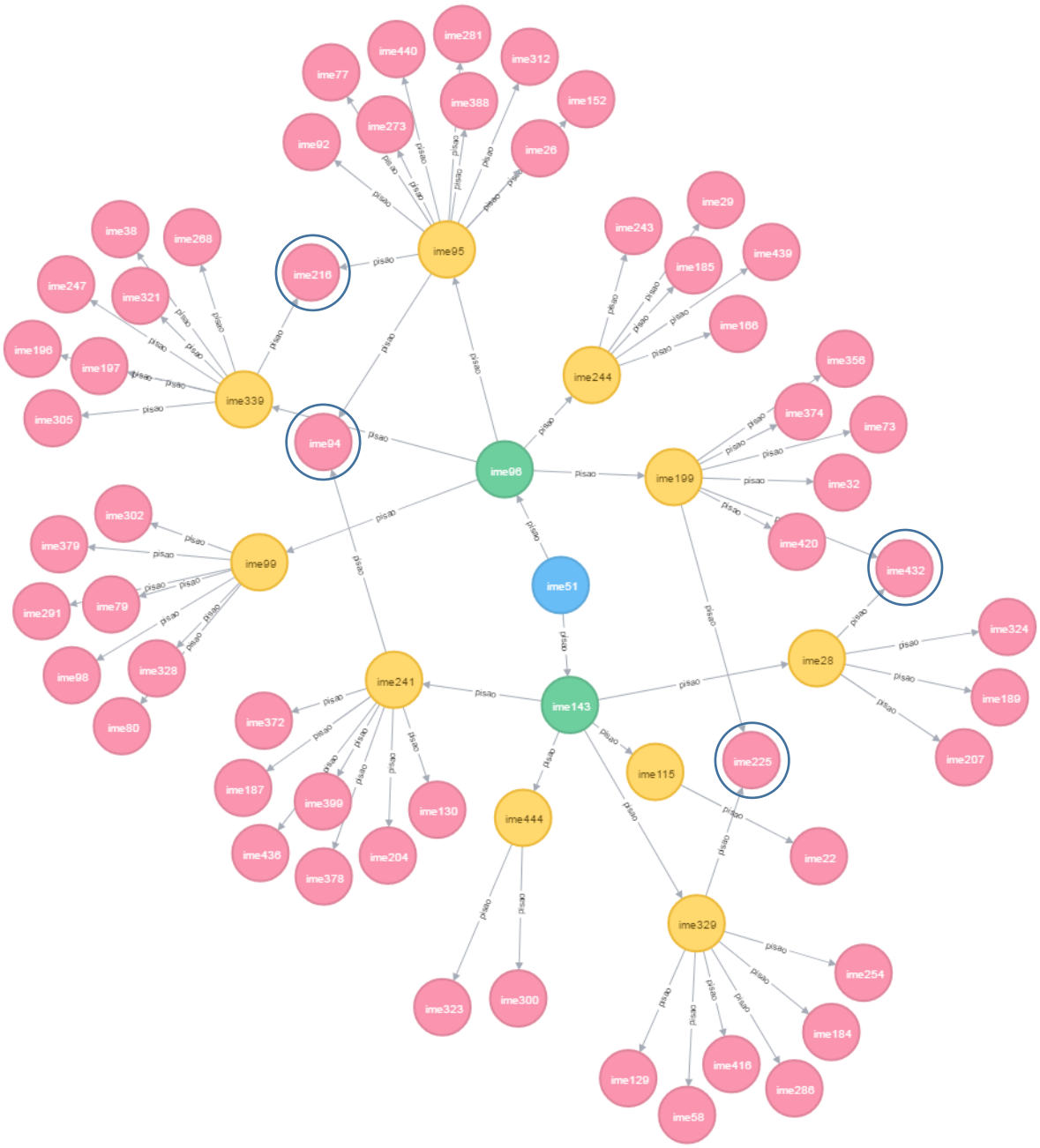
Cypher:

```
match p=(z1:Zaposleni)-[*1..4]->(z2:Zaposleni)
return extract(n in nodes(p)|n.zid)
```

SQL:

```
select 1,zp1.zid1,zp1.zid2, 0, 0, 0 from zaposleni_povezani zp1
union
select 2,zp1.zid1, zp2.zid1,zp2.zid2, 0, 0 from zaposleni_povezani zp1,
zaposleni_povezani zp2
where zp1.zid2=zp2.zid1
union
select 3, zp1.zid1, zp2.zid1,zp2.zid2, zp3.zid2, 0 from zaposleni_povezani
zp1, zaposleni_povezani zp2, zaposleni_povezani zp3
where zp1.zid2=zp2.zid1 and zp2.zid2=zp3.zid1
union
select 4, zp1.zid1, zp2.zid1,zp2.zid2, zp3.zid2, zp4.zid2 from
zaposleni_povezani zp1, zaposleni_povezani zp2, zaposleni_povezani zp3,
zaposleni_povezani zp4
where zp1.zid2=zp2.zid1 and zp2.zid2=zp3.zid1 and zp3.zid2=zp4.zid1
```

Упит илуструје проналажење свих путања које полазе од неког чвора (записа), и које садрже све повезане чворове (записе) до задатог растојања. У приказаном моделу, овај упит показује каква је *структура комуникације* на пројекту, док у општем случају слични упити показују како су повезани ентитети у моделу. На слици 8.5 је приказан резултат извршеног упита у Neo4j окружењу - виде се путање дужине три, од задатог до резултујућих чворова (запослених).



Слика 8.5: Задати запослени (плави чвор, *име51*) је повезан са два запослена на растојању један, са 10 на растојању два и 56 на растојању три. Означени су запослени на растојању три са којима су два или више запослених на растојању два имала преписку.

Упитни језик Cypher подржава рад са *путањама*, што омогућава једноставно задавање упита који захтевају обилазак графа. Границе за дужину растојања задате су параметром релације [*1..4] - максимално дозвољено растојање од почетног чвора је четири. Једноставном изменом параметара могуће је дефинисати путање произвољне дужине.

Приказани SQL упит користи придруживања, по једно за свако повећање растојања од почетног записа. Крајњи резултат се добија унијом повезаних записа на појединачним растојањима. Уочити да је Cypher упит *знатно једноставнији* у односу на SQL упит – за разлику од Cypher упита, комплексност SQL упита *повећава* се са повећањем растојања.

У окружењу Neo4j, резултате упита могуће је приказати табеларно или у виду подграфа (слика 8.5). За задати чвор се могу приказати (или искључити) сви његови суседи, што омогућава интерактивно истраживање добијених резултата. *Интегрисана визуелизација* и интеракција са резултатом упита[27] олакшава закључивање из података у бази: у приказаном примеру уочавају се токови комуникације и преклапања између група запослених у коресподенцији (означени чворови), и без додатне анализе или упита. Код релационог упита, резултат је у виду табеле која би се даље морала проследити у програмско окружење за визуелизацију и аналитичку обраду. У општем случају, окружења за постављање упита и његову визуелизацију и истраживање у релационој бази су *одвојени*. Постоје окружења која то омогућавају[28], али захтевају значајно предзнање корисника за конфигурисање и покретање.

У табели 8.4 су приказана трајања упита у релационој и графовској бази, за проналажење повезаних запослених на различитим растојањима. У коришћеном моделу се налазило 445 запослених и 2317 веза између њих (подаци генерисани из случајне расподеле). Коришћене су ажурне верзије за обе базе података (PostgreSQL 10.1 и Neo4j 3.3.0).

[27] https://neo4j.com/developer/guide-neo4j-browser/ .
[28] https://rickbergfalk.github.io/sqlpad/ .

Табела 8.4: Време извршења упита проналажења повезаних чворова (записа) на задатом растојању, у релационој и графовској бази података.

растојање	PostgreSQL [ms]	Neo4j [ms]	број путања
1	12	61	2317
2	22	141	14177
3	120	772	74987
4	818	4682	385893

Добијени резултати показују да се упит *брже* извршава у релационој бази за *свако* растојање! Време извршавања у графовској бази се може додатно смањити индексирањем чворова или применом посебног програмског интерфејса за обилазак графа[29], у коме се може дефинисати императивни поступак обиласка. Наведене опције нису коришћене јер је учињена претпоставка да се потенцијални корисник *неће* бавити оптимизовањем окружења, већ ће га примарно користити за извођење нових знања. Поред бржег извршавања у приказаном експерименту (што се разликује од резултата у Vukotic et al. 2015), треба имати на уму да дефинисање упита у SQL језику, приказаног у последњем примеру, захтева предзнање корисника, док је приказани упит у Cypher језику значајно *краћи* и *интуитивнији*. Слично, модификација упита би била једноставнија у Cypher језику.

Поред рада са претходно дефинисаном шемом базе, потребно је размотрити ситуацију у којој се врши њена измена. Претпоставимо да је у табели *Запослени* потребно додати атрибуте *адреса*, *телефон* и *електронска пошта*. Ако за сваког запосленог постоји само по једна вредност за наведене атрибуте, они се у релационој бази могу дефинисати као колоне у табели *Запослени*, док у графовској бази они постају атрибути дефинисани на нивоу чвора типа *Запослени*. Проблем се може проширити на случај када је потребно сачувати више

[29] https://neo4j.com/docs/java-reference/current/javadocs/org/neo4j/graphdb/traversal/package-summary.html .

вредности по атрибуту (нпр. фиксни телефон на послу или кући, мобилни телефон итд.). Код релационе базе су тада могућа два решења:

- дефинисање посебне колоне у табели *Запослени* за сваку поткатегорију атрибута,
- креирање нових табела за сваки атрибут (нпр. табела *Адреса*, са пољима *адреса*, *град*, *поштански_број*, *држава* и *запослени_ид*).

У првом случају, у табели *Запослени*, може се догодити да велики број вредности за поједине колоне буде недефинисан (NULL), што доводи до непотребног заузећа меморије. Други случај би дао више једноставнијих табела, но тада су подаци фрагментирани, што компликује упите у бази: ако је један запослени описан са четири табеле, да би се очитали његови подаци потребан је упит са три придруживања. У сваком од наведених случајева, за релациону базу се захтева претходно ажурирање шеме.

Динамичка шема графовске базе дозвољава дефинисање произвољног броја атрибута, као и одговарајућих вредности на нивоу појединачног чвора који су независни од атрибута у осталим чворовима. Даље, упит којим се очитавају подаци о запосленом је једноставан јер се сви подаци налазе у једном чвору.

Уопштено, услед динамичне и променљиве природе процеса на пројекту, пожељно је да додавање нових ентитета, релација и атрибута, као и манипулација над њима, буду што једноставнији. Све горе наведено је могуће обавити у оба типа базе, али је поступак *значајно једноставнији* у графовској бази.

Без обзира на спорије извршавање, *једноставност* и *интуитивност* задавања упита у графовској бази, уз могућност *визуелизације* резултата, дају *веће могућности* за истраживање веза између концепата издвојених из неструктуираних извора. За потребе истраживања приказаног у докторској тези, за складиштење ентитета и релација издвојених из неструктуираног текста, одабрана је графовска база Neo4j. Ентитете и релације карактерише изражена повезаност записа и потреба за динамичком шемом, јер се различити ентитети и везе између њих *постепено* уводе у модел (диктирано непредвидивом природом

пројекта). За складиштење структуираних података, као што су регистар учесника или списак позиција, одабрана је PostgreSQL релациона база података.

9 Примене графа значајних фраза у окружењу инвестиционог пројекта

У овој глави ће бити приказани могући случајеви *анализе* издвојених информација засноване на графу значајних фраза свих редова[30]. Анализа је спроведена на проширеном корпусу од 1836 докумената са пројекта „*Реконструкција Топионице и изградње нове Фабрике сумпорне киселине*“. Документи су класификовани у пет категорија:

- *Одштетни захтев,*
- *Захтев за измену уговорених радова,*
- *Преписка,*
- *Записник са састанка,*
- *Месечни извештај.*

Граф фраза, добијен из проширеног корпуса на начин описан у главама 5 и 7, имплементиран је помоћу графовске базе *Neo4J*. Сваки чвор (значајна фраза) ускладиштен је у бази података, *заједно* са скупом *референци* на оригиналне документе и локације у њима. На овај начин, омогућена је *олакшана навигација* између значајних фраза ка њиховим изворним документима и обрнуто.

За анализу добијеног графа коришћен је упитни језик *Cypher.* У наредним експериментима биће коришћена четири типа упита над графом, развијена за потребе овог истраживања:

- *Упит суседства* – за задату фразу f_i (чвор у графу), враћа све *суседне* фразе (чворове који су директно повезани са f_i);

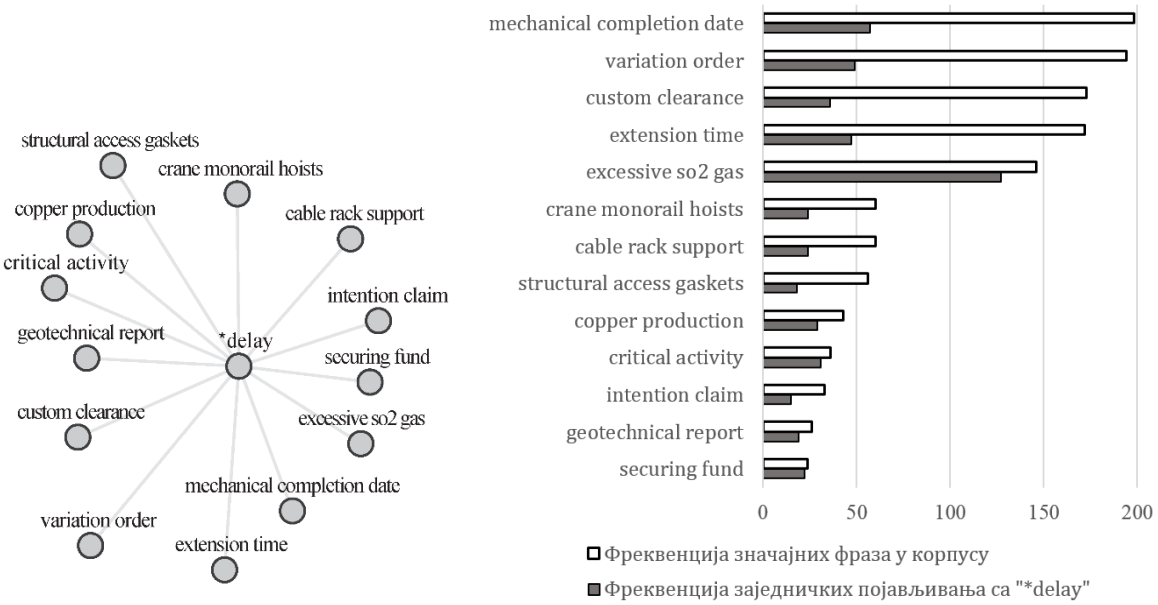
[30] Надаље ће се за термин *значајна фраза* користити само *фраза*.

- *Упит агрегираног суседства* – за задати образац којим се дефинише скуп фраза (нпр. образац "*steel* *" задовољавају фразе "*steel structure*" и "*steel pipe corrosion*"), враћа *унију* суседа за све фразе из скупа;
- *Упит суседства пара* – за задати пар *директно* повезаних фраза f_i и f_j, враћа све фразе које су директни суседи оба чвора;
- *Упит подграфа* – за задати скуп чворова (фраза), враћа *подграф* који, поред наведених чворова, укључује и везе између њих.

Сви претходно дефинисани упити омогућавају преглед оригиналних текстуалних контекста који садрже чворове резултујућих подграфа, што олакшава експертску интерпретацију и омогућава извођење знања.

9.1 Одређивање блиских концепата

Упит агрегираног суседства омогућава да се анализира повезаност групе фраза које се односе на различите концепте са пројекта. Циљ је да се, за задати образац којим се дефинише једна или скуп фраза, одреде фразе које се појављују у истим контекстима, и да се међу њима идентификују оне које су *највише повезане* са задатим обрасцем.



Слика 9.1: Одређивање блиских концепата. Лево: фразе које су суседи за образац "* *delay*". Десно: корелисаност између фраза облика "* *delay*" и њихових суседа.

На слици 9.1 - лево, приказан је пример упита агрегираног суседства за групу фраза чија је последња реч "*delay*", које заједно представљају концептуализацију појма *одлагање*. Упит, поред суседних фраза, враћа број њиховог појављивања на нивоу корпуса, као и суму броја заједничких појављивања са свим фразама које одговарају задатом обрасцу. На основу ова два броја могуће је проценити колико су суседне фразе *корелисане* са задатим упитом (слика 9.1, десно). За задати образац "* *delay*", међу суседним фразама уочавају се корелисане фразе које су очекивано блиске појму одлагања, попут "*critical activity*" или "*extension time*"[31]. Присутне су и фразе попут "*excessive SO_2 gas*" или "*copper production*", које проистичу из саме природе пројекта – ради се о постројењу за производњу и прераду бакра и производњу сумпорне киселине. Међу добијеним фразама се налазе и оне које су специфичне за неку ситуацију

[31] Прегледом текстуалног окружења из кога је издвојена, види се да представља концепт "*extension of time*". Предлог "*of*" одбачен је као реч из стоп-листе.

насталу на пројекту ("*geotechnical report*" или "*custom clearance*"), где се *не може унапред знати* да ће се појављивати у истом контексту са задатим обрасцем.

Наведени тип упита омогућава да се лако идентификују теме од интереса и трендови груписања, што може помоћи у идентификацији ситуација које треба додатно истражити. За одређивање природе заједничких контекста, кориснику су на располагању оригинални текстуални извори, као и упит суседства пара којим се могу приказати сви суседи за две задате, директно повезане фразе.

9.2 Детекција комплексних концепата

Претходни пример илуструје како се може одредити степен повезаности међу различитим фразама. Међутим, за потпуније разумевање семантичког контекста у коме се фраза јавља, било би потребно утврдити да ли је фраза део ширег, *комплексног концепта*. Комплексни концепт се дефинише као *скуп фраза* релевантних за једну апстрактну тему, заједно са *везама* између њих. На нивоу графа фраза, комплексни концепт представљен је једним његовим подграфом. Овако организован подграф[32] може се састојати како од блиских фраза, тако и од оних које немају висок степен повезаности, а појављивале су се у контекстима од интереса за концепт.

Следи пример идентификације комплексног концепта *Корозија цеви*, о коме је већ било речи у поглављу 6.1. Ради се о ситуацији када се у фази извођења јавио *проблем* везан за *корозију цеви* предвиђених за инсталацију у технолошком постројењу *CIGHE* (Cold Interpass Gas Heat Exchanger – Међупролазни измењивач хладног гаса). Предметни концепт („*инсталација цеви у CIGHE постројењу*“) се јављао у више докумената у различитим семантичким контекстима (и документима):

- одређивање нивоа површинске корозије од стране именоване компаније,

[32] У даљем тексту, подграф који репрезентује комплексни концепт називаће се графом комплексног концепта.

- идентификација цеви које нису у значајној мери захваћене корозијом и њихово чишћење,
- замена цеви које су више захваћене корозијом.

Један приступ којим би се могао идентификовати циљни концепт је да се одреде глобално значајне фразе које садрже речи релевантне за њега (у овом случају "*pipe*", "*tube*", "*cighe*", "*цев*", "*корозија*", "*corrosion*" итд.). Резултујући граф би се добио полазећи од глобалних фраза и њихових суседа. Нажалост, услед специфичне природе пројекта, глобалне фразе за овај пример су се употребљавале у различитим контекстима. На пример, упит суседства за фразу "*cighe pipes*" враћа 77 других фраза. Иако су неке од њих релевантне за посматрани концепт, корисник мора ручно да их издвоји јер се значајан број њих односио на друге концепте.

Алтернативни приступ пошао би од идентификације локално значајних фраза које представљају *подконцепте* изведене из предметног концепта. Оне би, заједно са њима суседним фразама, формирале резултујући граф концепта. Међутим, резултат највероватније неће обухватити све релевантне фразе због *непознавања* подконцепата.

У овом поглављу биће предложен *итеративни поступак* којим се могу открити комплексни концепти, уз услов да у резултујућем графу концепта буде *што мање* фраза које нису релевантне за посматрани концепт.

9.2.1 Поступак итеративне конструкције графа комплексног концепта

Резултујући граф G_k, који адекватно описује циљни комплексни концепт k, може се добити помоћу предложене хеуристике за његову итеративну конструкцију:

Иницијализација:

Корисник мануелно дефинише почетни скуп F_0^k који садржи полазне фразе које су релевантне за комплексни концепт k;

Проширивање скупа фраза које гравитирају ка концепту k:

Скуп фраза F_i^k ($i > 0$) добија се проширивањем скупа F_{i-1}^k фразама које *гравитирају* ка концепту k. Нове фразе добијају се извршавањем упита суседства пара за сваки пар директно повезаних фраза из F_{i-1}^k. Нове фразе додају се у F_i^k ако је *више од половине* њихових суседа из F_{i-1}^k (*услов проширивања*). Поступак се понавља све док постоје нове фразе које задовољавају услов за проширивање;

Конструкција графа G_k:

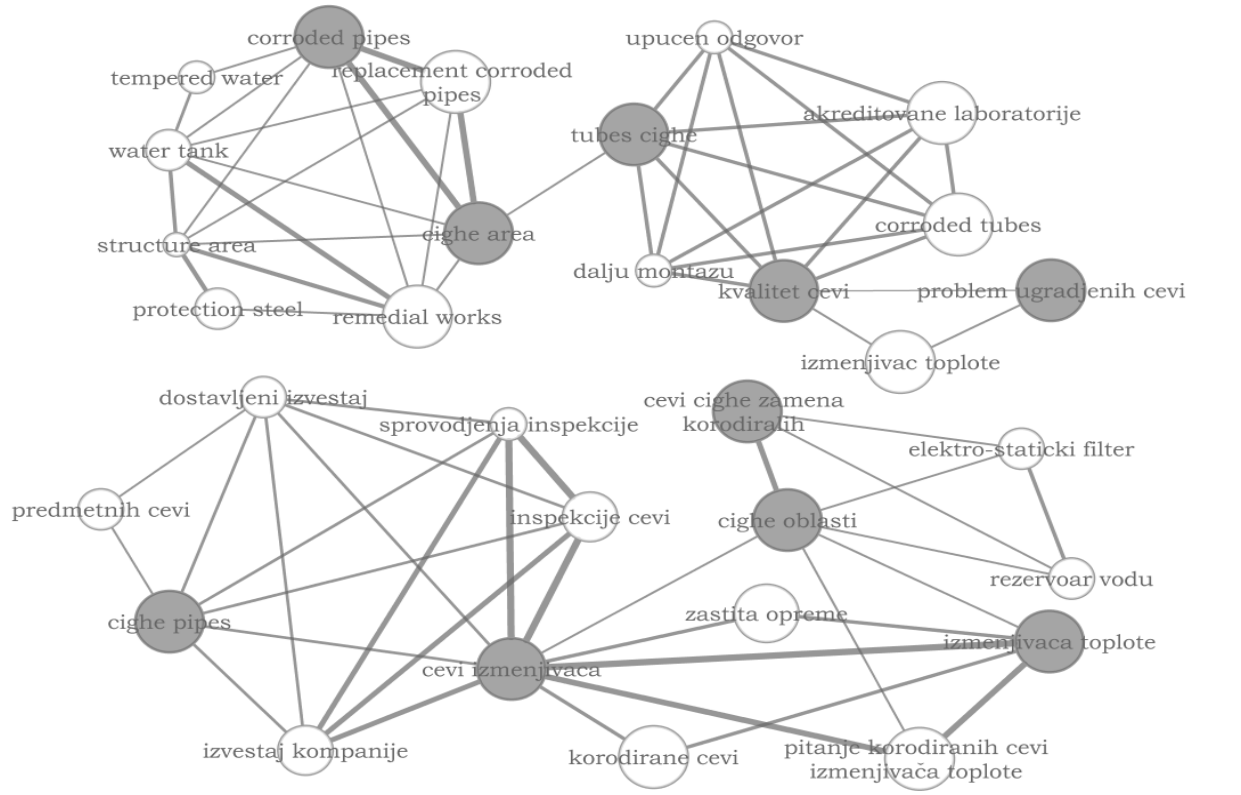
Коначни граф G_k се конструише коришћењем упита подграфа за финални скуп фраза F_n^k.

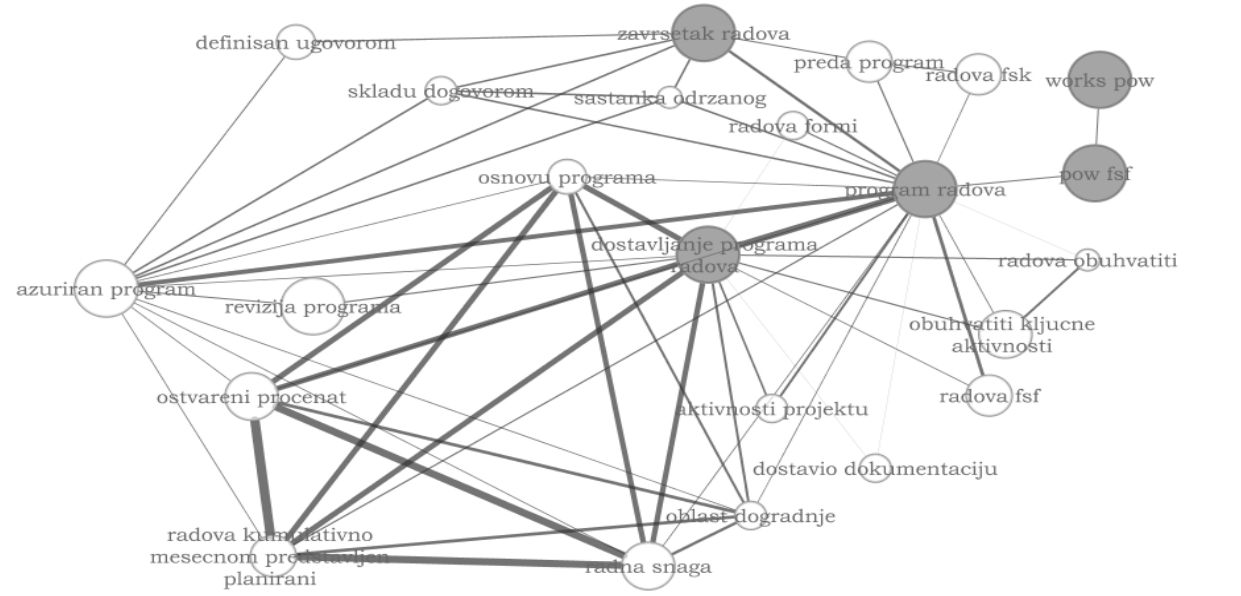
Услов за успостављање везе између две фразе дефинисан у (7.1) и услов проширивања из хеуристике обезбеђују *конвергенцију* предложене процедуре. Резултат поступка у великој мери *зависи од капацитета фраза* одабраних у F_0^k да адекватно опишу концепт k и његове подконцепте. Оптимални резултати се добијају када се за F_0^k одаберу комбинација *глобалних* и *локалних* фраза: глобалне фразе имају више суседа, што даје више кандидатних фраза, док локалне обезбеђују да кандидатне фразе које припадају валидним подконцептима постану део коначног графа G_k. Када резултујући скуп обухвати довољан број локалних фраза, оне „хватају“ неку глобалну која иницијално није обележена као релевантна за k. Њеним укључивањем у G_k, повећава се број могућих кандидата у даљим итерацијама. У општем случају, додавање нових фраза престаје када

престане додавање релевантних глобалних фраза. Сугерише се да корисници у F_0^k не додају најчешће фразе на пројекту (нпр. фразе везане за име пројекта или кључних компанија) јер су то глобалне фразе које имају капацитет да прогласе цео граф значајних фраза за један комплексни концепт. Са друге стране, ако се одаберу само локалне фразе, број кандидата ће се брзо исцрпети и резултат ће обухватити само њихове непосредне суседе.

9.2.2 Примена хеуристике за итеративну конструкцију графа комплексног концепта

Изложени поступак је тестиран на два комплексна концепта: *инсталација цеви у CIGHE постројењу* ($К_{cighe}$) и *програм радова* ($К_{програм}$). $К_{cighe}$ је специфичнији и односи се на конкретне активности везане за радове на једном технолошком постројењу, док је $К_{програм}$ генералнији и циљ му је да обухвати опште информације везане за динамику радова на пројекту. Три експерта оцењивала су резултујуће фразе према релевантности за предметни концепт (1 – мало значајно, 2 – умерено значајно, 3 – веома значајно).





Слика 9.2: Комплексни концепти: горе - *инсталација цеви у CIGHE постројењу*; доле - *програм радова*. Сиви чворови чине почетни скуп F_0^k. Величина кружнице означава просечну оцену значаја фразе. Ширина везе је пропорционална броју заједничких појављивања фраза.

На слици 9.2 су приказани издвојени графови за оба комплексна концепта. Може се уочити:

- $К_{cighe}$ се састоји од већег броја добро издиференцираних подконцепата (груписани чворови). Почетни скуп фраза је задат тако да су подконцепти могли правилно да се формирају;
- Подконцепти из $К_{cighe}$ који представљају клике (видети поглавље 7.2) често су се јављали заједно у реченицама. Термини коришћени за опис једног подконцепта нису коришћени у другим;
- Граф за $К_{програм}$ указује да се овај концепт састоји из јако повезаних подконцепата (тема);
- Три од пет почетних фраза на српском језику генерисале су цео граф за $К_{програм}$. Резултати упита суседства пара, за две почетне фразе на енглеском језику, нису испуниле потребан услов за придруживање резултату;

Приказани поступак се може искористити за *аутоматско обележавање докумената* према заступљености комплексних концепата. Једна од примена би била да се, коришћењем филтера по концептима, олакша текстуална претрага докумената. Међутим, као што је приказано у примерима, неопходно је да корисник поседује разумевање циљног концепта и одабере одговарајући почетни скуп фраза.

9.3 Праћење концепата кроз време

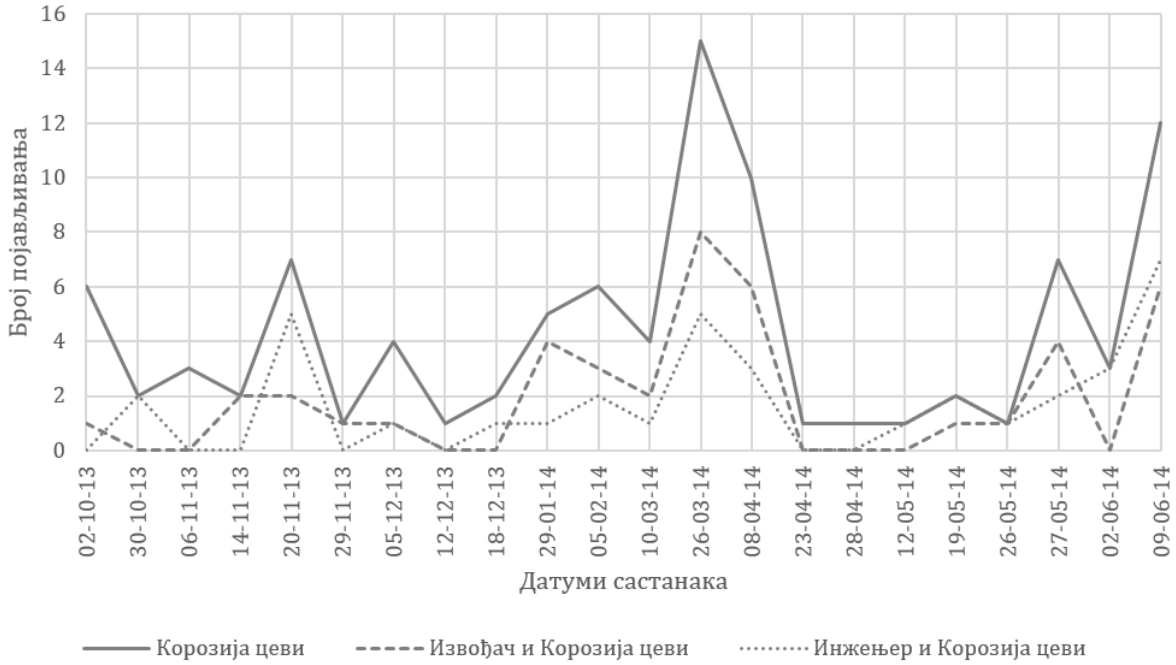
Праћење различитих дешавања на пројекту је једна од кључних активности неопходних за успешно управљање пројектима. На пример, радови на инсталацијама грејања и вентилације се прате кроз записе у различитим извештајима (дневним, недељним), како би се утврдила опасност од кашњења или прекорачења буџета. Поред података из структуираних извора попут програма радова (нпр. гантограм), околности које прате различите активности

се често појављују и у неструктуираним документима (нпр. преписка), *пре* манифестације у структуираном окружењу. Овде се истражује могућност праћења дешавања на пројекту, представљених преко комплексних текстуалних образаца који могу открити релевантне трендове.

Корпус докумената са инвестиционог пројекта је веома динамичан. На пример, стандардна пракса је да, после викенда или празника, учесници међусобно интензивирају комуникацију у којој се помињу актуелна питања. Постојећа пракса која омогућава експертима да прате и решавају горуће проблеме је да се о битним питањима, за која постоје ажурне информације, расправља на *састанцима*. Међутим, ситуација када је неки догађај постао тема састанка значи да је он, условно речено, *већ ескалирао*. У идеалном случају, ако би се идентификовала раније, потенцијално конфликтна ситуација би се лакше решила. На пример, на пројекту је експерт са компетенцијама и знањем да препозна потенцијални проблем. Међутим, проблем се прво манифестује кроз *захтеве за информацијама* који нису у фокусу експерта. Ако експерт није у могућности да *прати све ревизије захтева* на време, постоји опасност да касно уочи будуће жариште.

Предложена репрезентација информација издвојених из неструктуираног текста омогућава да се *сложени текстуални обрасци прате кроз документе*. Могућност праћења једног комплексног концепта илустрована је на примеру *Корозија цеви*. Концепт је детектован у секвенци од 24 документа типа *записник са састанка*, који покривају девет месеци пројектних активности (слика 9.3). Записници су одабрани као тип документа коме се, у предложеном окружењу графа значајних фраза, могу одредити датуми одржавања. Сваком документу је придружен број појављивања концепта, дефинисан као број реченица у којима су се појавиле *минимално две* фразе из концепта. Даље, одређен је број заједничких

појављивања на нивоу реченице за концептима *Инжењер и Извођач*[33], што даје оцену повезаности појединих учесника са посматраним концептом кроз време. Ручно прегледање предметних докумената показује да добијена оцена повезаности одговара стварном степену ангажовања учесника у активностима повезаним са концептом (нпр. Инвеститор *захтева чишћење цеви*).



Слика 9.3: Дистрибуција комплексног концепта *Корозија цеви* кроз време, заједно са фразама које представљају Извођача или Инжењера на пројекту. У периоду највеће заступљености концепта (март и јун 2014. године), на састанцима се дискутовало о резултатима договорених мера санације кородираних цеви, као и о даљим корацима за решавање проблема.

[33] *Инжењер* и *Извођач* као улоге учесника на пројекту дефинисане у (FIDIC 1999). Одговарајући концепти су креирани као скупови фраза које се односе на појединачне компаније (нпр. садрже назив компаније).

9.4 Проширивање графа значајних фраза кориснички дефинисаним ентитетима

У овом поглављу ће бити приказана могућа проширења графа значајних фраза семантички богатијим ентитетима, који се из текста препознају на основу кориснички дефинисаних правила. Под ентитетима се обично подразумевају концепти из реалног света који припадају одређеној категорији (нпр. *компанија*, *особа*, *материјал*, *конструктивни елемент*, *машина*, итд.) и садрже одговарајуће особине (*особа*: име, за кога ради, позиција, ...). Треба истаћи да су овде изабрани они ентитети за чије препознавање је потребан *минималан труд* експерта за дефинисање одговарајућих правила за издвајање. На овај начин, предложено решење остаје на зацртаном курсу: мањи труд за имплементацију на различитим пројектима од онтологија и система за управљање информацијама (слика 3.9).

Постојећи граф проширен је ентитетима типа *Датум*, *Особа*, и *Акција*. Ови ентитети представљају чворове који се *везују* са значајним фразама према релацијама дефинисаним у поглављу 7.1. Ентитети типа *Датум* издвајају се помоћу регуларних израза, према поступку изложеном у поглављу 5.5. Ентитет типа *Особа* је погодан за детекцију унутар неструктуираног текста јер на пројекту постоји *списак учесника* који се може искористити за његово препознавање. Овај списак се код сваке особе може проширити варијантама њеног имена (Mr. Petrović, П. Петровић, ...), па се на тај начин сви појавни облици пресликавају у исту особу.

Акција представља глагол који учесник на пројекту користи приликом дискусије на састанцима (слика 9.4). У овом истраживању акције су издвојене из дела корпуса на енглеском језику - за енглески је била доступна компонента за одређивање врсте речи (слика 5.2). Уколико ова компонента није доступна (српски језик), акције се могу издвојити уз помоћ *регуларних израза*, формираних на основу *списка* најчешће коришћених глагола, са основним варијантама коришћења за сваки глагол. Треба истаћи да се на овај начин пропушта велики број акција, али је то цена која се мора платити због недостатка језичких ресурса.

"*Person1* ***noted*** *that Company1 has also been raising the issue of PoW for the past six months but Company2 still hasn't delivered it and then he moved to the next item on Company2 agenda*"

Слика 9.4: Део записника са састанка у коме је издвојен ентитет типа *Акција* - *noted*. Предуслов за издвајање је да се непосредно пре акције појављује ентитет типа *Особа* (*Person1*).

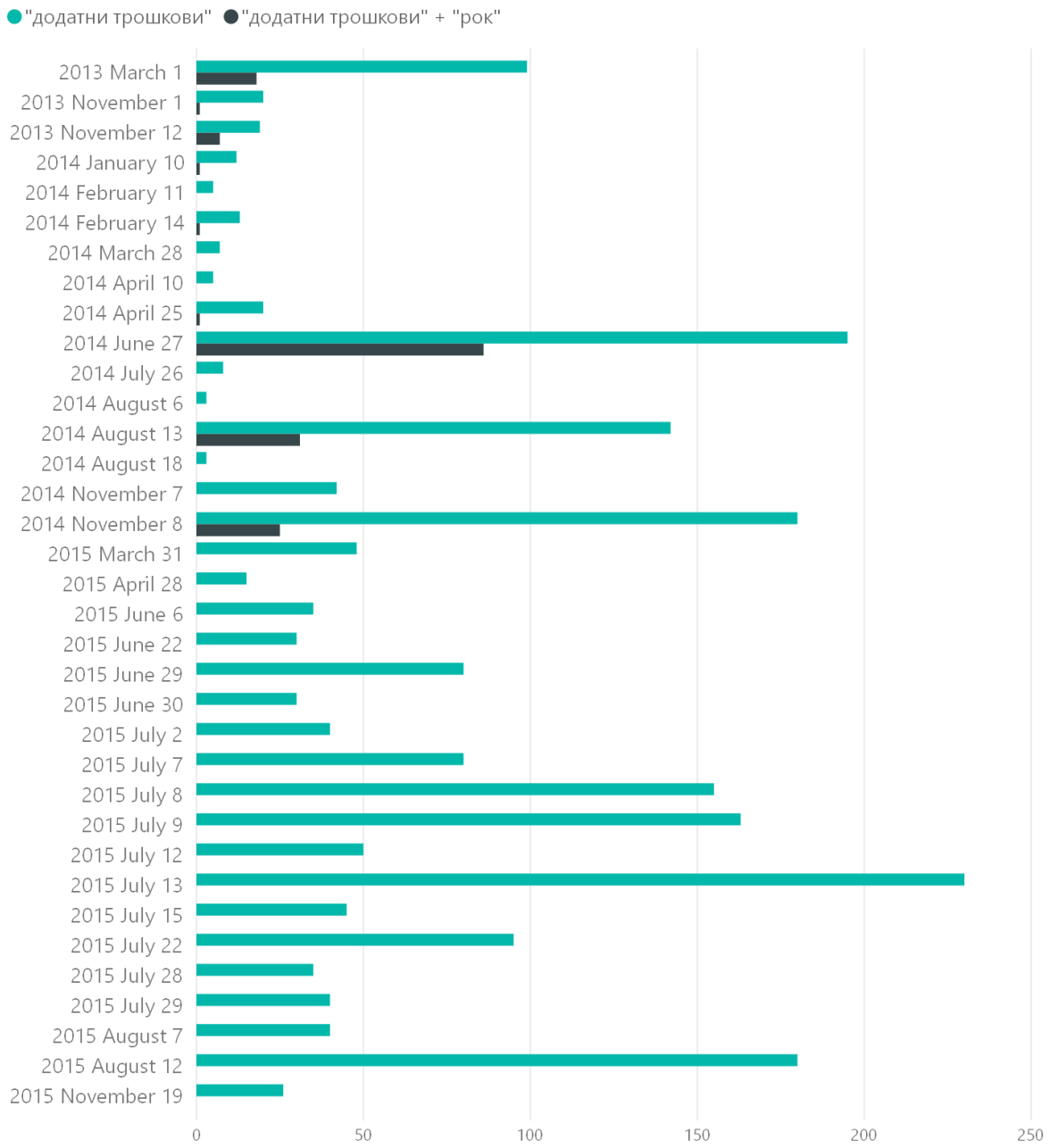
У овом поглављу биће приказана детаљнија анализа информација издвојених из записника са састанака, где увођење различитих типова чворова у графовској репрезентацији може дати одговоре на следећа питања:

- о чему је дискутовано (фразе),
- када је дискутовано (датуми),
- ко је са ким дискутовао (особе),
- како је дискутовано (акције).

9.4.1 Анализа временске расподеле појединих концепта

Директни ефекат увођења додатних типова чворова у репрезентацију је да су контексти у којима се јављају значајне информације боље описани – на пример, за задати образац могу се пронаћи сви датуми са којима се појављивао. Иако се сам датум може односити на различите догађаје (почетак, рок, пресек стања, итд.), његово увођење даје *додатну димензију за навигацију* издвојених информација. На слици 9.5 приказан је број заједничких појављивања ентитета типа *Датум* и групе фраза које описују концепте *додатни трошкови* и *рок*. Број појављивања добијен је коришћењем упита дефинисаних на почетку ове главе.

Увидом у текстуалне контексте заједничког појављивања фраза из концепта *додатни трошкови* и ентитета типа *Датум*, сазнаје се да се највећи део односи на различите *одштетне захтеве* које су учесници испоручивали једни другима (за продужетак рока, додатне трошкове, обрачун пенала, итд.).



Слика 9.5: Број заједничких појављивања ентитета типа *Датум* и група фраза које се односе на концепт *додатни трошкови*. Посебно су издвојена заједничка појављивања концепата *додатни трошкови* и *рок*. Ентитет типа *Датум* садржи временски атрибут на основу кога је обављено сортирање по времену.

Највећи део детектованих датума представља време када је одштетни захтев послат, или датум састанка на коме се о њему расправљало. Са слике се може закључити да су се, у документима у каснијим фазама пројекта, *чешће јављали* одштетни захтеви у којима се помињу додатни трошкови.

Са слике 9.5 се види да су заједничка појављивања концепата *додатни трошкови* и *рок* више изражена у ранијим фазама пројекта. Експерт може поставити питање, због чега се ова два концепта не јављају у заједничком семантичком контексту и у каснијим фазама пројекта? У тражењу одговора на ово питање, може се илустровати интеракција корисника са системом у процесу извођења новог знања. Овде је концепт *рок* дефинисан као подграф фраза у којима се појављује реч „рок“. Увид у фразе концепта *рок*, које се јављају заједно са фразама из концепта *додатни трошкови*, открива подконцепт *продужетак рока за завршетак радова*. Закључује се да су одштетни захтеви у којима се помињу и *продужетак рока* и *додатни трошкови* били више заступљени у ранијим фазама пројекта. Накнадни увид у контексте концепта *додатни трошкови,* који се јављају после 2014. године, показује појављивање других концепата у одштетним захтевима (грешке у пројектовању, оштећење опреме, застоји у раду, …).

9.4.2 Анализа комуникације на пројекту

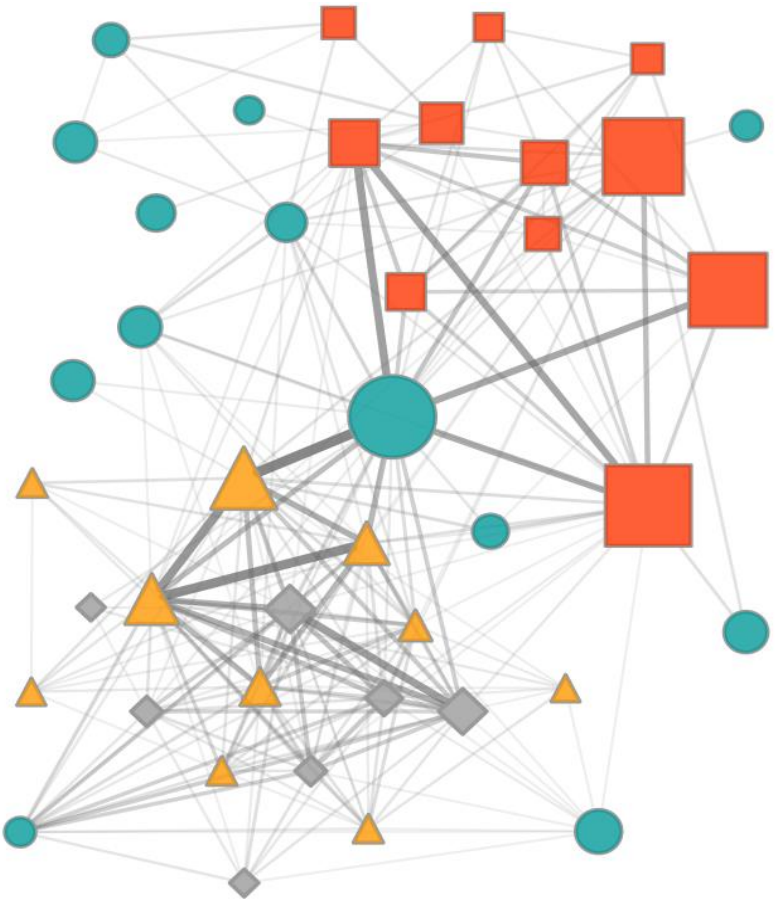
Поступак приказан у претходном примеру могуће је спровести за све комбинације чворова различитог типа. Међутим, посебно је интересантан подграф издвојен само за ентитете типа *Особа,* јер његовом анализом може да се утврди *структура комуникације* на пројекту.

У посматраном корпусу, контексти у којима се заједнички појављују особе најчешће су везани за записнике са састанака. Дискусија учесника бележи се као низ исказа, од којих сваки садржи особу која износи став о некој теми и, опционо, саговорнике којима се директно обраћа (слика 9.6).

***Person1** disagreed with **Person2** that mail from November was ignored and urged him to check his records where he would surely find a reply.*

Слика 9.6: Део записника са састанка. Дискусија о некој теми се бележи као низ исказа у којима учесници наизменично износе ставове.

На слици 9.7 приказан је издвојени подграф са ентитетима типа *Особа*.



Слика 9.7: Структура интеракције учесника на пројекту. Ентитет типа *Особа* садржи атрибут који одређује матичну компанију. На слици су приказани запослени из четири најзаступљеније компаније. Кружница ради за Инвеститора, квадрат за Инжењера, троугао за Извођача, а ромб за Подизвођача. Величина

чвора је пропорционална броју појављивања, а ширина везе броју заједничких појављивања.

Из издвојеног подграфа се може запазити:

- јасно су уочљиви представници Инвеститора и Извођача који за своје компаније представљају главне канале комуникације;
- Инжењер има више представника преко којих тече значајан део комуникације;
- представници Извођача и Подизвођача су више комуницирали међусобно;
- уочљиви су главни токови комуникације на нивоу компанија (Инжењер – Инвеститор; Подизвођач – Извођач; Извођач – Инвеститор).

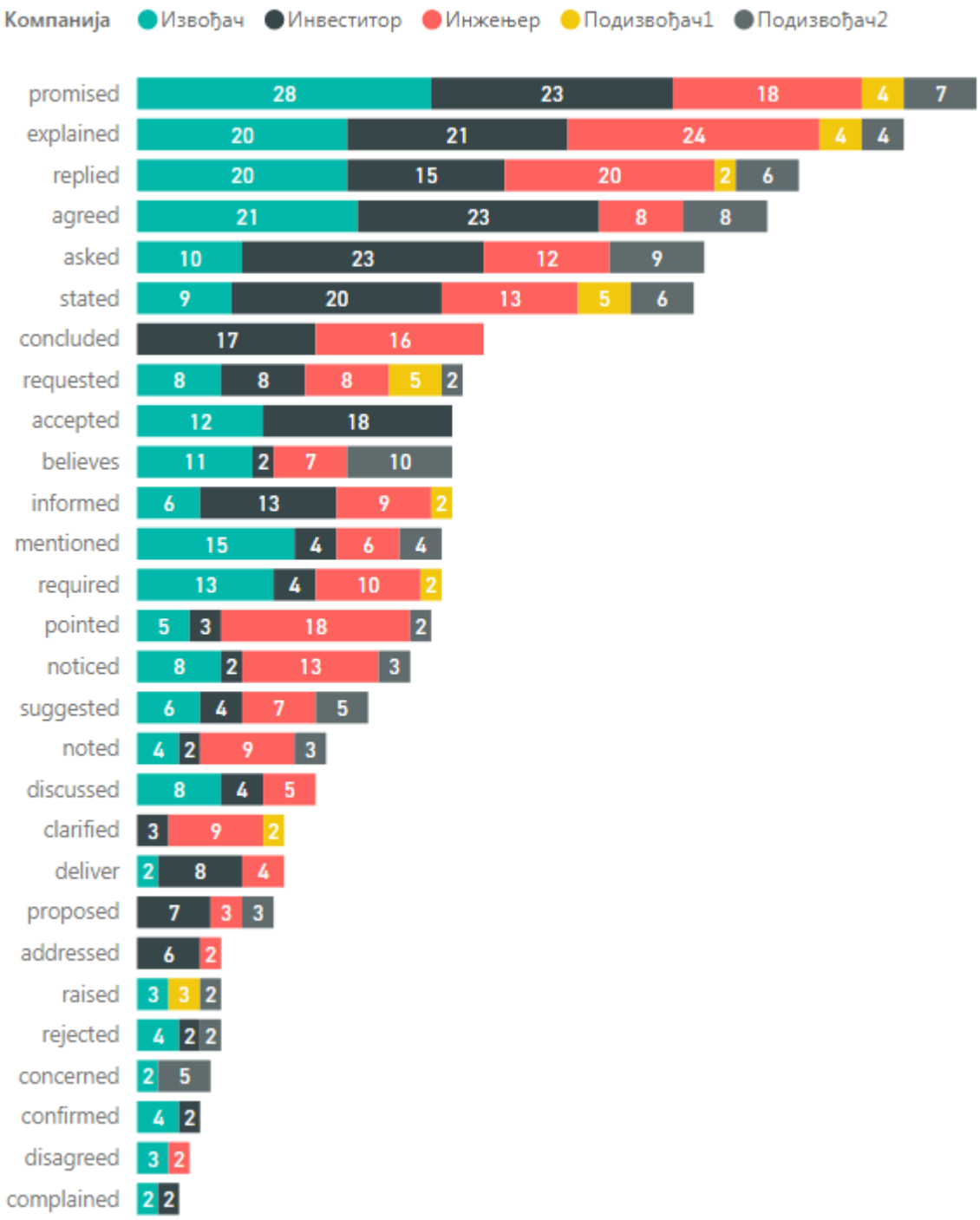
Структура међусобне комуникације на састанцима може *указати* да ли су формирани тимови или канали преписке *валидни*. Приказани подграф може да се прошири фразама које су повезане са учесницима. На овај начин, запосленима се придружују концепти који су на неки начин повезани са њима.

Природа дискусије између учесника на састанку могла би се детаљније описати конструисањем подграфа који садржи ентитете типа *Особа* и *Акција*. Предуслов да *Особа* претходи глаголу који се проглашава *Акцијом* може се искористити за конструисање посебне релације између ова два ентитета. Како је број издвојених акција и припадајућих особа у посматраном корпусу већи, на слици 9.8, уместо подграфа, приказани су сумарни резултати анализе. Зарад јаснијег приказа изостављене су акције са три и мање појављивања, као и две најзаступљеније (*said* и *added*).

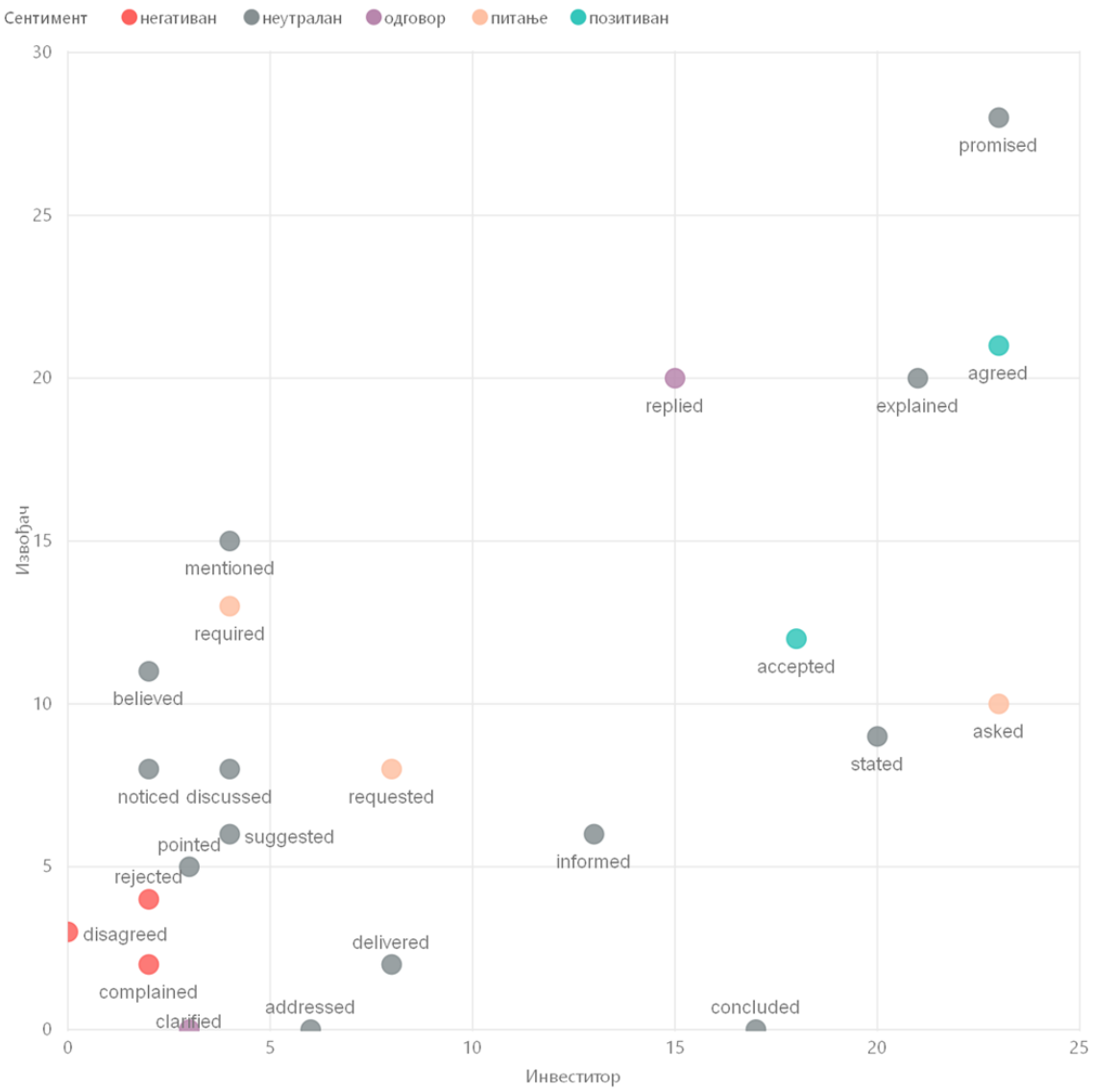
Као што је и очекивано, у дискусији су најзаступљенији представници компанија са кључних позиција (Инвеститор, Извођач, Инжењер). Може се запазити да су се у највећем броју исказа користиле неутралне акције (*pointed*, *stated*, *explained*, *concluded*, итд.). Посебно су интересантни резултати који се односе на негативни (*rejected*, *complained*, *disagreed*) или позитивни (*agreed*,

accepted) *сентимент исказа*. Као посебне категорије акција могу се издвојити и оне које се односе на питање (*asked*, *required*, *requested*), или одговор (*replied*, *clarified*).

Избором одговарајућег графичког приказа могу се упоредно анализирати изнети ставови представника појединачних компанија. На слици 9.9 је приказан тачкасти дијаграм са кога се могу детаљније упоредити обрасци које представници једне компаније користе у комуникацији. Са дијаграма се види да су сви сентименти приближно равномерно распоређени, осим када је у питању акција *asked*, коју су значајно више користили представници Инвеститора (23 пута), у односу на представнике Извођача (10 пута). Број акција које означавају негативни сентимент је мали, што указује да у посматраном корпусу *није било* изражених неслагања.



Слика 9.8: Издвојене акције сортиране према броју појављивања у записницима са састанака (минимални број појављивања већи од 3). Резултати су сумирани за пет најзаступљенијих улога учесника који су их користили.



Слика 9.10: Тачкасти дијаграм за акције које су употребили представници Инвеститора и Извођача. Акције су груписане према сентиментима.

Заједничка анализа фраза и акција, са фокусом на концепте за које су учесници имали највише негативних или упитних контекста, може послужити за идентификацију и праћење оних догађаја који имају потенцијално велики утицај на пројекат.

9.5 Проблем пристрасности експертског тумачења

На примеру случајева коришћења предложене репрезентације уочава се *неопходност сложене интеракције* између корисника и система. Корисник у сваком кораку *интерпретира* добијене резултате и ставља их у контекст *претходног знања* и *окружења* пројекта. Међутим, интерпретација од стране корисника је место где се могу јавити различите *грешке у процесу извођења закључака (когнитивни процес)*. Посебно су интересантни различити видови *когнитивне пристрасности* (Kahneman and Tversky 1974), којима су дефинисана одступања од исправног закључивања. У (Santamarina and Chameau 1989), аутори су указали на неколико когнитивних пристрасности које могу да се доведу у везу са процесом доношења одлука у грађевинарству. Један од закључака је да се експерти често концентришу на чињенице које *иду у прилог највероватнијем могућем закључку* и да занемаре *негативне* чињенице које би га оповргле. Когнитивна пристрасност типа *пристрасност избора*, која настаје када се за анализу користи статистички узорак који није репрезентативан, обрађена је у (Safa et al. 2015). Да би унапредили процес одабира извођача и свели на минимум ефекте лошег узорка, аутори примењују приступ заснован на истраживању конкуренције и разматрају избор извођача различитих карактеристика. Закључивање из историјске базе података једне организације носи опасност од *информативне пристрасности* (Creedy et al. 2010), јер може да садржи вишак непоузданих информација. Могуће стратегије за превазилажење проблема погрешне интерпретације, од којих су неке *делимично примењене* у предложеном решењу, приказане су у Табели 9.3 Приказана решења стављају акценат на примену *различитих алата* за бољи *увид у семантичке контексте* информације, одговарајућу *визуелизацију* и коришћење *више аналитичких приступа* у паралели.

Табела 9.3: Приступи за умањење ефекта когнитивне пристрасности, у процесу закључивања из података.

Опција	Постојеће примене
Разумевање контекста	Представља разумевање конкретне ситуације која чини основу за интерпретацију нове информације и интеракцију са том ситуацијом (Albers 2015). Аутори у (H. Wang et al. 2011) указују да је разумевање контекста у коме се реализује грађевински пројекат важно за управљање пројектом јер помаже да се идентификује релевантна информација за конкретни проблем.
Визуелизација времена	Аутор у (Parsons 2014) истражује особине интерактивних визуелних репрезентација и указује на значај подесивих динамичких погледа на темпоралне информације, што може помоћи да се превазиђу когнитивне пристрасности које настају због статичке визуелне репрезентације.
Интерактивни аналитички интерфејси	У (Chang et al. 2007) имплементирани су вишеструки интерактивни погледи са различитим аналитичким алатима над једним скупом података. Могућност да се виде ефекти акције корисника у више димензија, може значајно да побољша анализу података. У (Tolone 2009), аутор наводи да могућност когнитивне интеракције између различитих интерактивних интерфејса може да поједностави упоредну анализу, планирање и оперативне активности.
Контролисани речници	Представља скуп речи и фраза којима су покривене различите варијанте кључних концепата (у форми скраћеница, синонима, другачијег правописа). Закључивање у репрезентацијама, које садрже кључне речи и фразе, може се унапредити увођењем доменских контролисаних речника значајних концепата (Pollack & Adler 2014; Lee 2010; Isenberg et al. 2017). Тако се умањују двосмисленост и расипање фреквенција речи.
Повратна информација о евалуацији	Компонента која даје повратну информацију о интеракцији корисника са мапом знања, која је део система за одржавање мостова, приказана је у (Tserng et al. 2017). Систем памти обрасце претраге корисника, што побољшава резултате нових претрага. Корисници имају могућност да оцене добијене резултате, што има позитиван и мотивишући ефекат за интеракцију са системом.

У постојећем решењу, корисник *може* да приступи оригиналним изворима из којих су издвојене значајне фразе и њихове релације. Ово омогућава

једноставан увид у сва окружења у којима се посматрани концепт јавља, како на нивоу реченице, тако и на нивоу параграфа или документа. На овај начин, имплементиран је, у извесној мери, принцип *разумевања контекста информације.*

Ако би значајним фразама била придружена одговарајућа временска одредница, било би могуће да се истражи понашање неког концепта кроз време. Ова функционалност би у значајној мери оспособила предложено решење за *визуелизацију темпоралне димензије* посматраног концепта. Конкретна реализација за записнике са састанака дискутована је у поглављу 9.3, где је приказана временска дистрибуција заступљености комплексних концепата.

Поред могућности приказаних у овој глави (детекција повезаних и комплексних концепата, напредни упити по структури графа), увођење додатних *интерактивних аналитичких интерфејса* представља један од приоритета у даљем истраживању и развоју. У том смислу систем би се могао проширити да обухвати следеће аналитичке технике: кластерисање сличних текстуалних контекста (Cheng & Leu 2009), анализа сентимента (Agarwal et al. 2013), (Abbasian-Hosseini et al. 2014), анализа графа (Hossain 2009), динамичка анализа графа (Zhu & Mostafavi 2015) и друге.

Увођењем *контролисаних речника*, на начин приказан у поглављу 9.4, илустровано је *семантичко обогаћивање* постојеће репрезентације. Оно има за циљ да смањи могућност погрешне интерпретације података. Проширивање репрезентације додатним ентитетима (*машине*, *позиције*, *материјали*, …) било би оправдано *само* у случајевима када захтевани труд за дефинисање правила за њихово издвајање не нарушава основну филозофију приступа – лакоћу имплементације и трансферабилност на различите пројекте.

10 Закључна разматрања

За доношење одлука на инвестиционом пројекту најчешће се користе информације похрањене у структуираном и полуструктуираном облику. Међутим, највећи део потенцијално корисних информација је у форми неструктуираних података, најчешће у текстуалном облику, које чине око 80% пословних информација у предузећу. У овој тези предложен је систем за аутоматско издвајање и визуелизацију корисних концепата из неструктуираног текста, који може да послужи као подршка у управљању пројектом.

По питању издвајања знања из неструктуираних текстуалних извора, постојећи системи за рад са документима не могу адекватно да одговоре на све специфичности окружења инвестиционог пројекта. Они су или захтевни за имплементацију на различитим пројектима, или не поседују одговарајуће алате за закључивање из неструктуираних података. У овој дисертацији је предложен нови приступ у издвајању значајних информација из текстуалних докумената, где би однос између уложеног труда за успостављање система и труда при закључивању у току коришћења био повољнији.

Предложени приступ је преносив јер не захтева значајне ресурсе и прилагођавање за употребу на различитим пројектима. Приступ је независан у односу на језик документа јер су поступци за детекцију основних концепата засновани на статистичким методама, које су универзално применљиве у сваком језичком окружењу.

Под претпоставком да је већина битних концепата из пројектне документације представљена у облику значајних фраза реда два (парова речи), за њихову детекцију су примењене различите статистичке мере за одређивање

корелације речи. Након што је уочено да поједине статистичке мере фаворизују различите фреквенције речи, предложена је метода комбиноване листе која има повољније особине у целом фреквентном опсегу. Недовољно информативне фразе су филтриране применом поступка заснованог на ентропији скупа суседстава фраза. Валидност поступка за издвајање значајних фраза је верификована на претходно обележеним корпусима са два различита пројекта. Експерименти показују да прецизност расте када се уведе корекција ентропијом, као и да укључивање техника за обраду природног језика, у поступак издвајања, даје боље резултате. Како се перформансе система, са и без примене посебних језичких ресурса, нису значајно разликовале, може се закључити да је предложени приступ применљив и за оне случајеве када језички ресурси нису доступни. Предложени аутоматски поступак поређен је и са експертски надгледаним приступом који је узео у обзир природу пројектне документације. Експерименти показују приближне резултате за оба приступа, с тим да за предложени приступ није неопходно укључивање претходног доменског знања. Додатно је испитан капацитет фраза, састављених од парова речи, да пренесу значење докумената. Показано је да се документи, представљени фразама, боље групишу по сличности него што би се груписали да су представљени појединачним речима.

Како би се издвојени значајни концепти организовали у одговарајућу структуру, испитане су различите репрезентације знања са становишта експресивности и могућности визуелизације. Пошто ће издвојени концепти из текста имати произвољну, унапред непознату структуру, у којој ће издвојене информације бити повезане бинарним релацијама, као погодна репрезентација су одабране семантичке мреже. За успостављање веза између значајних фраза је предложен доменски и језички независан поступак, заснован на сличности заједничких контекста појављивања. Издвојене фразе организоване су у граф који је искоришћен за одређивање значајних фраза реда већег од два. Том приликом примењен је Bron-Kerbosch алгоритам за проналажење максималних

клика у графу. Дефинисани поступак даје боље резултате у корпусу који садржи више докумената са дужим, дескриптивним реченицама, што омогућава да се формирају валидни комплексни концепти. Са тако добијеним фразама вишег реда, формиран је финални граф релевантних концепата.

Додатно рангирање и филтрирање неинформативних фраза може се обавити на основу динамичности суседа у графу, који се мења у току животног циклуса пројекта. Резултати експеримента показују да мера динамичности суседства боље рангира фреквентније фразе у односу на основни предложени поступак.

Поређење графовске и релационе базе података показало је да се, у релационој бази, упити извршавају брже него у графовској. Графовска база је погоднија за задавање комплексних упита по структури, као и за визуелизацију резултата. Узевши у обзир изражену повезаност фраза, за складиштење концепата и релација издвојених из неструктуираног текста, одабрана је графовска база Neo4j.

Систем је тестиран у интерактивном раду. Приказан је поступак за одређивање блиских значајних концепата на пројекту, као и њихово праћење кроз време. Предложена је хеуристика за одређивање комплексних концепата који покривају одређене теме и који се могу искористити за аутоматско обележавање докумената према њиховој заступљености. Описано је могуће семантичко проширивање предложене репрезентације, увођењем кориснички дефинисаних ентитета попут датума, особе и акције. Могућности проширене репрезентације илустроване су на примеру одређивања интеракције између појединих типова учесника на пројекту. За ту прилику, проширени граф значајних фраза конструисан је из корпуса записника са састанака. На основу свега претходно наведеног, закључак је да се предложени поступак за аутоматско издвајање и визуелизацију значајних концепата из текстуалних извора може успешно искористити као подршка у управљању инвестиционим пројектом.

10.1 Препоруке за даља истраживања

Извођење нових закључака, у интерактивном окружењу предложене репрезентације, се првенствено заснива на правилној интерпретацији од стране крајњег корисника. У истраживању је дискутовано о различитим облицима когнитивне пристрасности, као и мерама за њихово превазилажење. Негативни ефекти пристрасности експертског тумачења умањили би се ако би предложени систем добио додатне функционалности. Могући правци даљег истраживања би били усмерени на проширивање постојећег система тако да обухвата:

- *Визуелизацију времена*, што би омогућило истраживање понашања концепата кроз време;
- *Интерактивне аналитичке интерфејсе*, где би се над резултатима упита примењивали различити аналитички алати, уз могућност да се ефекти акције корисника виде у више димензија;
- *Контролисане речнике*, чиме би се репрезентација семантички обогатила а издвојеним концептима би се умањила двосмисленост;
- *Повратне информације о евалуацији*, где би информације о интеракцији корисника са системом побољшале резултате нових упита.

Даљи рад на предложеном систему обухватио би и истраживање алтернативних приступа за издвајање семантичке структуре из текста. Специјализоване технике, попут векторске репрезентације речи (word embeddings (Mikolov et al. 2013)), би смањиле расипање по фреквенцијама за речи за истим значењем и омогућиле препознавање семантички богатијих веза између група концепата.

Литература

Abbasian-Hosseini, S.A. et al., 2014. From Social Network to Data Envelopment Analysis : Identifying Benchmarks at the Site Management Level. *Journal of Construction Engineering & Management*, 140(8).

Ackoff, R., 1989. From data to wisdom. *Journal of Applied Systems Analysis*, 16, pp.3–9.

Agarwal, A. et al., 2013. SINNET: Social Interaction Network Extractor from Text. *Sixth International Joint Conference on Natural Language Processing*, (October), p.33.

Albers, M.J., 2015. Human–Information Interaction with Complex Information for Decision-Making. *Informatics*, 2(2), pp.4–19.

Björk, B., 2002. The Impact of Electronic Document Management on Construction Information Management. *Conference Proceedings - distributing knowledge in building*, (June), pp.12–14.

Blumberg, R. & Atre, S., 2003. The Problem with Unstructured Data. *DM Review*, 13, p.42.

Boone, H.N.J. & Boone, D.A., 2012. Analyzing Likert data. *Journal of Extension*, 50(2), p.30.

Brachman, R.J. & Levesque, H.J., 2004. *Knowledge representation and reasoning*, Morgan Kaufmann Publishers.

Brin, S. & Page, L., 1998. The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1/7), pp.107–17.

Bron, C. & Kerbosch, J., 1973. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM*, 16(9), pp.575–577.

Caldas, C.H., Soibelman, L. & Han, J., 2002. Automated Classification of Construction Project Documents. *Journal of Computing in Civil Engineering*, 16(October), pp.234–243.

Cao, F. & Liang, J., 2011. A data labeling method for clustering categorical data. *Expert Systems with Applications*, 38(3), pp.2381–2385.

Carifio, J. & Perla, R.J., 2007. Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response

Formats and their Antidotes. *Journal of Social Sciences*, 3(3), pp.106–116.

Chang, F. et al., 2006. Bigtable: A distributed storage system for structured data. *7th Symposium on Operating Systems Design and Implementation (OSDI '06), November 6-8, Seattle, WA, USA*, pp.205–218.

Chang, R. et al., 2007. WireVis: Visualization of categorical, time-varying data from financial transactions. *VAST IEEE Symposium on Visual Analytics Science and Technology 2007, Proceedings*, pp.155–162.

Chassiakos, A.P. & Sakellaropoulos, S.P., 2008. A web-based system for managing construction information. *Advances in Engineering Software*, 39(11), pp.865–876.

Cheng, Y.M. & Leu, S. Sen, 2009. Constraint-based clustering and its applications in construction management. *Expert Systems with Applications*, 36(3 PART 2), pp.5761–5767.

Church, K.W. & Hanks, P., 1989. Word association noms, Mutual Information, and lexicography. *Proceedings of the 27th Annual Conference of the Association for Computational Linguistics*, 16(1), pp.22–29.

Clancey, W.J., 1983. The epistemology of rule-based expert systems --- {A} framework for explanation. *Artificial Intelligence*, 20(1983), pp.215–251.

Clark, P., 1996. Requirements For a Knowledge Representation System: Working Note 10. , pp.1–10.

Codd, E.F., 1983. A relational model of data for large shared data banks. *Communications of the ACM*, 26(1), pp.64–69.

Costa, R. et al., 2013. Facilitating knowledge sharing and reuse in building and construction domain: an ontology-based approach. *Journal of Intelligent Manufacturing*, pp.1–20.

Creedy, G.D., Skitmore, M. & Wong, J.K.W., 2010. Evaluation of Risk Factors Leading to Cost Overrun in Delivery of Highway Construction Projects. *Journal of Construction Engineering and Management*, 136(5), pp.528–537.

Damani, O.P. & Ghonge, S., 2013. Appropriately Incorporating Statistical Significance in {PMI}. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp.163–169.

Date, C.J. & Darwen, H., 1997. *A Guide to the SQL Standard-Addison Wesley*, Addison-Wesley.

Davenport, T.H. & Prusak, L., 1998. *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School Press.

DeCandia, G. et al., 2007. Dynamo: Amazon's Highly Available Key-value Store. *Proceedings of the Symposium on Operating Systems Principles*, pp.205–220.

Denoyer, L., Zaragoza, H. & Gallinari, P., 2001. HMM-based Passage Models for Document Classification and Ranking. *Ecir*, pp.126–135.

Derose, S.J., 1988. Grammatical Category Disambiguation By Statistical Optimization. *Computational Linguistics*, 14(1), pp.1–24.

Dice, L.R., 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3), pp.297–302.

Dijkstra, E.W., 1959. A Note on Two Probles in Connexion with Graphs. *Numerische Mathematik*, 1(1), pp.269–271.

Ding, L.Y. et al., 2016. Construction risk knowledge management in BIM using ontology and semantic web technology. *Safety Science*, 87, pp.202–213.

Dunning, T., 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19, pp.61–74.

El-Diraby, T., 2012. Domain Ontology for Construction Knowledge. *Journal of Construction Engineering and Management*, 139(7), pp.768–784.

El-Gohary, N.M. & El-Diraby, T.E., 2010. Domain Ontology for Processes in Infrastructure and Construction. *Journal of Construction Engineering and Management*, 136(7), pp.730–744.

Van Emden, M.H. & Kowalski, R. a., 1976. The Semantics of Predicate Logic as a Programming Language. *Journal of the ACM*, 23(4), pp.733–742.

Fan, H., Xue, F. & Li, H., 2015. Project-Based As-Needed Information Retrieval from Unstructured AEC Documents. *Journal of Management in Engineering*, pp.1–10.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), pp.861–874.

Fellbaum, C., 2012. WordNet. *The Encyclopedia of Applied Linguistics*.

FIDIC, 1999. *Conditions of Contract for Plant and Design-Build: For Electrical and Mechanical plant, and for Building and Engineering Works, Designed by The Contractor (Yellow Book), 1st Edition*, International Federation of Consulting Engineers.

Fleiss, J.L. & Cohen, J., 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3), pp.613–619.

Gertz, M. & Lipeck, U., 1995. A diagnostic approach to repairing constraint violations in databases. *Informatik-Berichte*, 95(1).

Goenawan, I.H., Bryan, K. & Lynn, D.J., 2016. DyNet: Visualization and analysis of dynamic molecular interaction networks. *Bioinformatics*, 32(17), pp.2713–2715.

Gray, J. & Reuter, A., 1993. *Transaction Processing Concepts and Techniques*, San Mateo, CA: Morgan Kaufmann Publishers.

Gruber, T.R. et al., 1995. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5–6), pp.907–928.

Haksever, A., 2000. A model to predict the occurrence of information overload of project managers. *Proceed. of Int. Conf. on Construction Information*.

Hasan, K.S. & Ng, V., 2011. Automatic Keyphrase Extraction: A Survey of the State of the Art. *Association for Computational Linguistics Conference (ACL)*, pp.1262–1273.

Hölsch, J., Schmidt, T. & Grossniklaus, M., 2017. On the performance of analytical and pattern matching graph queries in Neo4j and a relational database. *CEUR Workshop Proceedings*, 1810.

Horrocks, I., 2013. *Evolution of Semantic Systems*, Springer.

Hossain, L., 2009. Communications and coordination in construction projects. *Construction Management and Economics*, 27(1), pp.25–39.

Isenberg, P. et al., 2017. Visualization as Seen through its Research Paper Keywords. *IEEE Transactions on Visualization and Computer Graphics*, 23(1), pp.771–780.

Issa et al., 2015. *Ontology in the AEC Industry: A Decade of Research and Development in Architecture, Engineering, and Construction*,

Ivković, B. & Popović, Ž., 2005. *Upravljanje projektima u građevinarstvu*, Beograd: Građevinska knjiga.

Jakus, G. et al., 2013. *Concepts , Ontologies , and Knowledge Representation*, Springer Publishing Company, Incorporated.

Joishi, J., Sureka, A. & Delhi, N., 2016. Graph or Relational Databases : A Speed Comparison for Process Mining Algorithm. *ArXiv*, pp.1–22.

Kamaruddin, S.S. et al., 2008. Conceptual graph formalism for financial text representation. *Proceedings - International Symposium on Information Technology 2008, ITSim*, 4, pp.2–7.

Kenneth C. Laudon & Laudon, J.P., 2012. *Management information systems: managing the digital firm*, Prentice Hall.

Kim, B.-G. et al., 2010. Automatic Extraction of Apparent Semantic Structure from Text Contents of a Structural Calculation Document. *Journal of Computing in Civil Engineering*, 24(June), pp.313–324.

Ko, Y., Park, J. & Seo, J., 2004. Improving text categorization using the importance of sentences. *Information Processing and Management*, 40(1), pp.65–79.

Landauer, T.K. et al., 1997. How Well Can Passage Meaning be Derived without Using Word Order ? A Comparison of Latent Semantic Analysis and Humans. *Proceedings of the 19th annual meeting of the Cognitive Science Society*, (January 1999), pp.412–417.

Larsen, B. & Aone, C., 1999. Fast and effective text mining using linear-time document clustering. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*, pp.16–22.

Lee, H.S.P., 2010. Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight. *Scientometrics*, 85(1), pp.65–79.

Lee, S.K., Kim, K.R. & Yu, J.H., 2014. BIM and ontology-based approach for building cost estimation. *Automation in Construction*, 41, pp.96–105.

Li, Y., Bontcheva, K. & Cunningham, H., 2005. SVM Based Learning System for Information Extraction. , pp.319–339.

Lin, H.T., Chi, N.W. & Hsieh, S.H., 2012. A concept-based information retrieval approach for engineering domain-specific technical documents. *Advanced Engineering Informatics*, 26(2), pp.349–360.

Liu, H., Lu, M. & Al-Hussein, M., 2016. Ontology-based semantic approach for construction-oriented quantity take-off from BIM models in the light-frame building industry. *Advanced Engineering Informatics*, 30(2), pp.190–207.

Lu, C.T. et al., 2007. Performance evaluation of desktop search engines. *2007 IEEE International Conference on Information Reuse and Integration, IEEE IRI-2007*, pp.110–115.

Luhn, H.P., 1960. KEY WORD-IN-CONTEXT INDEX. *AMERICAN DOCUMENTATION*, XI(4), pp.288–295.

Ma, Z., Liu, Z. & Wei, Z., 2016. Formalized Representation of Specifications for Construction Cost Estimation by Using Ontology. , 31, pp.4–17.

Mahfouz, T., 2011. Unstructured Construction Document Classification Model through Support Vector Machine (SVM). *Computing in Civil Engineering (2011)*, (413), pp.760–767.

Matthies, B., 2015. What to Do With All These Project Documentations ? – Research Issues in Reusing Codified Project Knowledge. *Proceedings of Pacific Asia Conference on Information Systems*.

Mikolov, T. et al., 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems 26*, pp.1–9.

Minsky, M., 1975. Frame theory. *Theoretical issues in natural language proeessing. Preprints of a conference at MIT; Reprinted in P.N. Johnson-Laird and P.c. Wason*

(Eds.) Thinking: Readings in cognitive science. Cambridge: Cambridge University Press (1977)., pp.355–376.

Mitchell, T.M., 1997. *Machine Learning*, McGraw Hill.

Mohan, C., 2013. History Repeats Itself: Sensible and NonsenSQL Aspects of the NoSQL Hoopla. *EDBT '13 Proceedings of the 16th International Conference on Extending Database Technology*, (March), pp.11–16.

Moniruzzaman, A.B.M., 2014. NewSQL: Towards Next-Generation Scalable RDBMS for Online Transaction Processing (OLTP) for Big Data Management. *International Journal of Database Theory and Application*, 7(6), pp.121–130.

Moniruzzaman, A.B.M. & Hossain, S.A., 2013. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *International Journal of Database Theory and Application*, 6(4), pp.1–14.

Moses, S., El-Hamalawi, A. & Hassan, T.M., 2008. The practicalities of transferring data between project collaboration systems used by the construction industry. *Automation in Construction*, 17(7), pp.824–830.

N.A. Stillings et al., 1995. *Cognitive Science: An Introduction*, Cambridge, Massachusetts: The MIT Press.

Nadeau, D., 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, (30), p.3–26.

Pandit, A. & Zhu, Y., 2007. An ontology-based approach to support decision-making for the design of ETO (Engineer-To-Order) products. *Automation in Construction*, 16(6), pp.759–770.

Parsons, P., 2014. Adjustable Properties of Visual Representations : Improving the Quality of Human-Information Interaction. *Journal of the Association for Information Science and Technology*, 65(3), pp.455–482.

Pearlson, K.E. & Saunders, C.S., 2010. *Managing and Using Information Systems*, WILEY.

Peshkin, L. & Pfeffer, A., 2003. Bayesian Information Extraction Network. *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI'03)*, pp.421–426.

Pietroforte, R., 1997. Communication and governance in the building process. *Construction Management and Economics*, 15(1), pp.71–82.

Pollack, J. & Adler, D., 2014. Emergent trends and passing fads in project management research : A scientometric analysis of changes in the field. *International Journal of Project Management*, 33(1), pp.236–248.

Al Qady, M. & Kandil, A., 2014. Automatic clustering of construction project documents based on textual similarity. *Automation in Construction*, 42, pp.36–49.

Al Qady, M. & Kandil, A., 2010. Concept Relation Extraction from Construction Documents Using Natural Language Processing. *Journal of Construction Engineering & Management*, 136(3), pp.294–302.

Quillian, M.R., 1967. Word concepts: a theory and simulation of some basic semantic capabilities. *Behavioral science*, 12(5), pp.410–430.

Ramakrishnan, C., Kochut, K. & Sheth, A., 2006. A Framework for Schema-Driven Relationship Discovery from Unstructured Text. *Lecture Notes in Computer Science: 5th International Semantic Web Conference (ISWC-2006), Athens, GA, November 6-9, 2006 Proceedings*, 4273, pp.583–596.

Robertson, S., 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5), pp.503–520.

Russell, A.D., Chiu, C. & Korde, T., 2009. Automation in Construction Visual representation of construction management data. *Automation in Construction*, 18(8), pp.1045–1062.

Safa, M. et al., 2015. Competitive intelligence (CI) for evaluation of construction contractors. *Automation in Construction*, 59, pp.149–157.

Santamarina, J.C. & Chameau, J.L., 1989. Limitations in decisionmaking and system performance. *Journal of Performance of Constructed Facilities*, 3(2), pp.78–86.

Santos, D., 1992. Natural Language and Knowledge Representation. *INESC Report*, pp.1–9.

Saxena, P. & Thakur, M.D., 2016. Complexity Analysis of Clique. , 5(1), pp.7–15.

Sint, R. et al., 2009. Combining unstructured, fully structured and semi-structured information in semantic wikis. *CEUR Workshop Proceedings*, 464, pp.73–87.

Soibelman, L. et al., 2008. Management and analysis of unstructured construction data types. *Advanced Engineering Informatics*, 22(1), pp.15–27.

Songer, A.D., Hays, B. & C. North, 2006. Multidimensional visualization of project control data. *Construction Innovation*, 4(3), pp.173–190.

Sowa, J.F., 1992. Conceptual graphs as a universal knowledge representation. *Computers and Mathematics with Applications*, 23(2–5), pp.75–93.

Tastle, W.J. & Wierman, M.J., 2007. Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning*, 45(3), pp.531–545.

Tserng, H.P. et al., 2017. The use of knowledge map model in construction industry. *Journal of Civil Engineering and Management*, 16(3), pp.332–344.

Turney, P.D., 2000. Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, 2, pp.303–336.

Urma, R.G. & Mycroft, A., 2015. Source-code queries with graph databases - With application to programming language usage and evolution. *Science of Computer Programming*, 97(P1), pp.127–134.

Vicknair, C. et al., 2010. A comparison of a graph database and a relational database. *Proceedings of the 48th Annual Southeast Regional Conference on ACM SE 10*, p.1.

Voordijk, H., Van Leuven, A. & Laan, A., 2003. Enterprise resource planning in a large construction firm: Implementation analysis. *Construction Management and Economics*, 21(5), pp.511–521.

Vukotic, A. et al., 2015. *Neo4j in Action*, Shelter Island, NY 11964: Manning Publications Co.

Wang, H. et al., 2011. Ontology-Based Approach to Context Representation and Reasoning for Managing Context-Sensitive Construction Information. *Journal of Computing in Civil Engineering*, 25(5), pp.331–346.

Wang, H.-H., Boukamp, F. & Elghamrawy, T., 2011. Ontology-Based Approach to Context Representation and Reasoning for Managing Context-Sensitive Construction Information. *Journal of Computing in Civil Engineering*, 25(October), pp.331–346.

Washtell, J. & Markert, K., 2009. A Comparison of Windowless and Window-Based Computational Association Measures as Predictors of Syntagmatic Human Associations. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore*, (August), pp.628–637.

Welty, C., 2003. Ontology research. *AI Magazine*, 24(3), pp.11–12.

Yurchyshyna, A. & Zarli, A., 2009. An ontology-based approach for formalisation and semantic organisation of conformance requirements in construction. *Automation in Construction*, 18(8), pp.1084–1098.

Zhang, J. & El-Gohary, N., 2015. Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking. *Journal of Computing in Civil Engineering*, 30(2), pp.4015014-1-14.

Zhong, M., Duan, J. & Zou, J., 2011. Indexing conceptual graph for abstracts of books. *Proceedings - 2011 8th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2011*, 3, pp.1816–1820.

Zhu, J. & Mostafavi, A., 2015. Metanetwork Framework for Integrated Performance Assessment under Uncertainty in Construction Projects. *Journal of Computing in Civil Engineering*, 31(1), pp.1–14.

Zhu, Y., Mao, W. & Ahmad, I., 2007. Capturing implicit structures in unstructured content of construction documents. *Journal of Computing in Civil Engineering*, 21(3), pp.220–227.

Прилози

Прилог 1

Релевантни извори на српском и енглеском језику коришћени у поступку надгледаног аутоматског издвајање фраза:

FIDIC (International Federation of Consulting Engineers). (1999). *Conditions of Contract for Plant and Design-Build: For Electrical and Mechanical plant, and for Building and Engineering Works, Designed by The Contractor (Yellow Book)*

PMI (Project Management Institute). (2000). *Construction extension to PMBOK guide*, Newtown Square, PA

Webster, L. F. (1997). *The Wiley dictionary of civil engineering and construction*, Wiley, New York

Kurtz, J.-P. (2004). *Dictionary of civil engineering*, Kluwer Academic Publishers, New York

FIDIC (International Federation of Consulting Engineers). (2008). *Услови уговарања за пројектовање – изградња и кључ у руке* (Сребрна књига), Југословенски преглед, Београд

Живковић, С. (2002). *Грађевински енглеско-српски, српско-енглески речник*, Orion, Београд

Ивковић, Б., Поповић, Ж. (2005). *Управљање пројектима у грађевинарству*, Грађевинска књига, Београд

Биографија

Ђорђе Недељковић рођен је 12.11.1984. године у Сремској Митровици. У Смедеревској Паланци је завршио основну школу и гимназију природно-математичког смера. Грађевински факултет Универзитета у Београду уписао је школске 2003/04. године, а дипломирао је 2009. године на конструктивном смеру, са просечном оценом 8,61 и оценом 10 на дипломском раду. Докторске студије је уписао крајем 2009. год. Све предвиђене испите положио је закључно са октобром 2014. године, са просечном оценом 9.625. Ради као асистент-студент докторских студија на предметима Рачунарско цртање у грађевинарству, Основе програмирања у VisualBasic-у, Основе програмирања у Пајтону и Објектно оријентисано програмирање.

Учествовао је као истраживач на технолошким пројектима „Примена GNSS и LIDAR технологије у мониторингу стабилности инфраструктурних објеката и терена“ и „Истраживање стања и метода унапређења грађевинских конструкција са аспекта употребљивости, носивости, економичности и одржавања“. Сарађивао је у припреми књиге „Основе програмирања у Matlab-u, збирка задатака“ (издање 2010. год.).

Говори, чита и пише енглески језик.

Изјава о ауторству

Име и презиме аутора: Ђорђе Љ. Недељковић

Број индекса: 902/09

Изјављујем

да је докторска дисертација под насловом

Издвајање и визуелизација знања из текстуалних извора за потребе управљања инвестиционим пројектима у грађевинарству

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

Потпис аутора

У Београду, 28.03.2018.

_____

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора:	Ђорђе Недељковић
Број индекса:	902/09
Студијски програм:	Грађевинарство
Наслов рада:	**Издвајање и визуелизација знања из текстуалних извора за потребе управљања инвестиционим пројектима у грађевинарству**
Ментор:	др Милош Ковачевић, ванредни професор Универзитет у Београду, Грађевински факултет

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањена у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива 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 аутора

У Београду, 28.03.2018.

_____

Изјава о коришћењу

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

Издвајање и визуелизација знања из текстуалних извора за потребе управљања инвестиционим пројектима у грађевинарству

која је моје ауторско дело.
Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

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

1. Ауторство (CC BY)
2. Ауторство – некомерцијално (CC BY-NC)
3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)
4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)
5. Ауторство – без прерада (CC BY-ND)
6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 28.03.2018.

_____

1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.