

УДК 004.4

РАЗМЕТКА ДОКУМЕНТОВ ПО СЕМАНТИЧЕСКИМ РОЛЯМ

*Игнатъев Н.А., Абдуллаев К.Д.**n_ignatev@rambler.ru*

Национальный университет Узбекистана имени Мирзо Улугбека,
100174, Узбекистан, Ташкент, ул. Университетская, 4.

Рассматривается разметка документов с учётом семантических ролей. Данные для тематических моделей формируются на основе стандартной гипотезы мешок слов. Вопрос о количестве и уникальности тем при моделировании не является решённым. В качестве причин называется отсутствие единых критериев для оценки качества кластеризации. Были попытки доказательства в форме эксперимента уникальности тем через использование набора критериев-регуляризаторов по результатам группировки документов. Отсутствие такого доказательства объясняется тем, что число групп изначально задается как свободный параметр. Для унификации документов в настоящее время применяются пять типов разметки: метатекстовая, морфологическая, синтаксическая, акцентная и семантическая. В работе используется семантическая разметка через выделение в предложении актантов – именных групп, обозначающих участников ситуации и их семантические роли. Предлагается при определении семантических ролей на узбекском языке проводить унификацию терминов применительно к конкретным предметным областям. Приведён пример разметки документов с использованием семантических ролей и определение их сходства по косинусной метрике.

Ключевые слова: semantic roles, topic uniqueness, document grouping, regularizing criteria.

Цитирование: *Игнатъев Н.А., Абдуллаев К.Д.* Разметка документов по семантическим ролям // Проблемы вычислительной и прикладной математики. – 2024. – № 5(61). – С. 80-90.

1 Введение

Семантический анализ текста включает различные методы и техники для понимания и интерпретации значений слов и предложений в контексте. Разметка — это процесс организации и структурирования текста или других данных, чтобы сделать их более понятными, удобными для восприятия и обработки. В широком смысле разметка применяется в различных областях, таких как разработка веб-страниц, обработка текста, программирование и публикация контента. Разметка является фундаментальным инструментом для организации и форматирования текста и данных в самых разных областях. Ее использование обеспечивает ясность, структурированность и удобство работы с информацией, что делает разметку незаменимым элементом современного информационного мира.

При семантическом анализе текста, как правило, применяются методы кластеризации и классификации. Наиболее полное с точки зрения тематического моделирования совершенство получено при использовании методов кластеризации в модели ARTM [1, 2]. Совершенство как адаптация к реальности достигается за счёт применения нескольких дополнительных критериев – регуляризаторов.

От результатов разметки зависят векторные представление документов и количество тем на которые они разделяются. Вопрос о количестве тем методами кластерного анализа чаще всего пытаются решить через определение параметров дискретного распределения на множестве описаний тем. Большинство исследователей сходятся по мнению в пользу выбора плотности распределения Дирихле.

Проблема устойчивости разбиение документов на темы решается с помощью критериев регуляризаторов. Так в модели ARTM применение регуляризаторов приводит к разделению тем с семантической определённой от фоновых.

Изначально предполагается, что количество тем задается как свободный параметр. Важным для исследователя остается ответ на вопрос как интерпретировать результаты тематического моделирования при разных значениях свободного параметра. Одним из вариантов решения является выбор изначально большого количества тем [3], с постепенным снижением их до определенного минимума.

Существуют предложения о сведении количества оригинальных тем к постоянной величине. Открытыми остаются вопросы:

1. Есть ли методики для анализа сходимости количества тем к определённому, фиксированному числу?

2. Является ли это количество тем уникальным, неслучайным значением?

Возникают проблемы выбора критериев для доказательства существования такой уникальности. Судя по выводам из [3] критерии для доказательства уникальности не разработаны.

Известны множество методов кластерного анализа и классификации объектов. При выборе алгоритмов классификации за основу берется показатель их обобщающей способности. Попытки найти подобный показатель для оценки качества кластерного анализа не увенчались успехом. Теоретическое обоснование отсутствия единого критерия качества кластерного анализа доказывается в теореме Кленберга [4].

Примером семантического анализа документов и понятий с использованием методов классификации является работа Тулиева [5]. Реализация семантического анализа документов основывалась на отношении связанности объектов. Через отношение связанности вычислялись меры компактности документов по классам и парным объединениям классов. Проведено исследование связанности документов по научной классификации, принятой для защиты докторских диссертаций в ВАК Республики Узбекистан. В качестве примера показана семантика термина “физико-математических наук”, используемая при присуждении научных степеней и званий.

Анализ многообразия структур отношений в коллекциях документов на естественном языке (ЕЯ) применяется для добычи новых знаний при тематическом моделировании. Например, в качестве новых знаний рассматривается наличие семантической связанности слов, лежащих в основе разбиения документов на темы, интерпретируемые векторные представления слов для решения проблемы синонимии и полисемии.

Методы машинного обучения на основе векторных моделей с успехом применяются для извлечения знаний из текстовых документов в различных предметных областях. При векторном представлении в виде частотных распределений слов в контекстах считается, что семантически близкие слова имеют близкие векторы. В качестве варианта для тестирования моделей рассматривается ручное формирование наборов пар слов с экспертными оценками близости.

Вопрос о количестве тем и их уникальности [3] предложено решать начиная с чрезмерно большого количества тем, а затем используя регуляризацию, свести боль-

шинство из них к нулю. Предложен регуляризатор способный удалять из модели линейные комбинации существующих тем. Функция, оптимизируемая регуляризатором, имеет дивергенция Кульбака-Лейблера. Используются интуитивные предположения, что количество тем связано с детализацией тем: когда количество тем слишком велико, модель выдает множество мелких тем, похожих друг на друга.

Предметом исследования является разметка документов на узбекском языке на основе семантических ролей. Применяется методология, разработанная в работах профессора К. Воронцова, которая показала высокую эффективность при семантической разметке предложений. Для данной задачи были адаптированы алгоритмы машинного обучения и лексические ресурсы, специфичные для узбекского языка.

Во многих странах отсутствуют государственные стандарты, регулирующие использование терминов и слов в различных предметных областях при переводе. Это приводит к неоднородности терминологии, особенно в условиях, когда количество слов ограничено, а число новых терминов постоянно растёт. Естественные языки непрерывно развиваются, при этом вместе с их развитием расширяется и их понятийная база. На практике должны использовать термины в их правильном контексте, искажение которых может привести к неверной интерпретации ключевых концепций.

Желательно установить государственные стандарты для разных предметных областей, обеспечивающие единую интерпретацию терминов. Это особенно важно в условиях поступательного развития научных исследований и технологий. Четкое и согласованное употребление терминов на конкретном языке способствует их правильному восприятию и распространению. Каждый термин, представленный в единственном варианте для каждой области знаний, должен быть закреплён в стандарте, чтобы не допустить многозначности и двусмысленности его толкования.

Существует множество примеров, когда перевод терминов с иностранных языков на узбекский осуществляется без должного учета их предназначения и контекста. Это приводит к размыванию смысла терминов и потере их исходного значения, что в конечном итоге усложняет их восприятие и применение в предметных областях.

Важнейшей задачей в лингвистическом исследовании является корректный перевод терминов, обеспечивающий точность и однозначность их понимания. Например, на узбекском языке в литературных источниках по информационным технологиям для термина “данные”(data) используют слово “ma’lumot”, несмотря на существование альтернативы в виде слова “berilganlar”, которое точно отражает смысл применительно к компьютерному железу. Аналогично термин “признак”(feature) близок по смыслу к слову “alomat” чем к “belgi”. В перечень примеров отсутствие единого смысла можно включить термин “облачные данные”, которые в переводе озвучиваются как “berilganlar oynasi” или “bulutli ma’lumot” и т.д.

В тематическом моделировании существуют стандартная гипотеза, согласно которой несколько слов, передающих одно и то же значение, должны быть выражены одним словом. Правильный перевод терминов имеет принципиальное значение для сохранения точности и однозначности их толкования в научных и технических документах. Проблема может быть решена с привлечением знаний экспертов из конкретных предметных областей. Эти знания понадобятся для разметки текстов в NLP/ML [6] для:

- распознавание онимов (named entity recognition, NER);
- распознавание частей речи (part of speech tagging, POS);
- выделение тональности онима (sentiment analysis, SA);
- выделение синтаксических связей (syntax parsing);

- выделение семантических ролей (semantic role labeling, SRL);
- выделение текстовых полей данных (slot filling);
- выделение полей в библиографических записях;
- сегментации научных или юридических текстов;
- разрешения анафоры, кореферентности, эллипсиса;

Разметка семантических ролей (SRL): найти в предложении актанты — именные группы, обозначающие участников ситуации и их семантические роли. Примеры семантических ролей, которые приводятся в [6]:

- агенс: одушевлённый инициатор и контролёр действия;
- пациенс: участник, на которого направлено действие;
- бенефактив: участник, получающий пользу или вред;
- адресат: получатель сообщения (может быть бенефактивом);
- инструмент: посредством чего осуществляется действие;
- экспериенцер: носитель чувств и восприятий;
- стимул: источник восприятий;
- источник: исходный пункт движения;
- цель: конечный пункт движения;

Семантический анализ текста включает различные методы и техники для понимания и интерпретации значений слов и предложений в контексте.

2 Постановка задачи

Считается, что задан набор (словарь) $L = (c_1, \dots, c_n)$ из n основ слов естественного языка. В коллекции $R = (D_1, \dots, D_m)$ каждый из m текстовых документов описывается через частоту встречаемости n слов из словаря L .

Разметка документов проводится с учётом семантических ролей. Значение $d_{ij} \in D_i$ вычисляется как количество (доля) вхождений слова $c_i \in L$ относительно общего количества слов из словаря L в документе D_i .

Требуется по множеству пар векторных представлений текстовых документов $\{(D_i, D_j)\}$ выбрать пару с наименьшим значением косинусного расстояния

$$\cos(D_u, D_v) = \frac{\sum_{i=1}^n d_{ui}d_{vi}}{\sqrt{\sum_{i=1}^n d_{ui}^2} \sqrt{\sum_{i=1}^n d_{vi}^2}}.$$

Если значение приближается к единице, то два текста являются похожи ими по смыслу. Ручная разметка документов востребована в качестве прецедентов для машинного обучения алгоритмов. Результаты машинного обучения позволяют автоматизировать процесс разметки.

3 Вычислительный эксперимент

Разметка текстов производилась вручную с помощью перечисленных выше актант. По этим актантам сформирован список слов для вычисления их частот в конкретных документах.

На основе частот встречаемости слов стало возможным анализировать попарную близость документов по смыслу. Обоснованием выбора косинусной метрики в качестве меры расстояния является отсутствие необходимости решать проблему проклятия размерности.

Фрагменты ручной разметки документа по семантическим ролям приводятся ниже.

Документ D1

<источник> Toshkentda</источник> <агенс> Nexia-3</агенс> <пациенс> yonib ketdi</пациенс> <источник> Toshkentning Shayxontohur tumanida </источник> <агенс> Nexia-3 mashinasining </агенс> <источник> motor qismida </источник> <пациенс> yong'in kelib</пациенс> chiqdi. Holat <инструмент> o't o'chiruvchilar </инструмент> tomonidan <пациенс> bartaraf etilgan</пациенс>. <экспериментер> Jarohatlanganlar </экспериментер> yo'q.<источник> 2024 yilning 4 iyun kuni soat 18:30 larda Shayxontohur tumani Nurafshon ko'chasi 15a-uy qarshisida </источник> <агенс> Nexia-3 rusumli transport vositasining </агенс> <источник> motor qismida</источник> <пациенс> yong'in kelibchiqqan</пациенс>. <пациенс> Yong'in </пациенс> tuman <инструмент> FVB xodimlari <инструмент> tomonidan <пациенс> bartaraf etilgan</пациенс>. <пациенс> Yong'in </пациенс> natijasida tan <стимул> jarohati </стимул> olganlar yo'q. Mazkur hodisa yuzasidan <агенс> YHXB </агенс> hamda<агенс> FVB xodimlari</агенс> tomonidan <пациенс> surishtiruv </пациенс> olib borilib, <пациенс> yong'in </пациенс> yuzaga kelishining sabablari va zarar miqdoriga aniqlik kiritilmoqda. "Barcha<адресат> haydovchilardan </адресат> boshqaruvidagi <цель> transport vositasining texnik sozligiga e'tiborli bo'lishlarini so'raymiz</цель>", deyiladi <источник> Toshkent shahar IIBB YHXB xabarida</источник> .

Документ D2

<источник> Samarqandda</источник> <агенс> 9-sinf o'quvchisi</агенс> <инструмент> Nexia</инструмент> bilan urib yuborgan<агенс> 6 yoshli qiz</агенс> <цель> vafot etdi</цель>. <агенс> 9-sinf o'quvchisi</агенс> <адресат> otasiga </адресат> tegishli <инструмент> mashinani </инструмент> <цель> yuvish uchun <цель> ko'chaga olib chiqqan va katta tezlikda harakatlanib <агенс> qizchani</агенс> urib yuborgan.<источник> Samarqand viloyati </источник> <источник> Urgut tumanida </источник> <агенс> 9-sinf o'quvchisi </агенс> <инструмент> Nexia-2 mashinasini </инструмент> boshqarib <агенс> 6 yoshli qizni </агенс> urib yuborgan va <бенефактив> uning </бенефактив> o'limiga sabab bo'lgan. <агенс> Viloyat IIB YHXB axborot xizmatining </агенс> xabar berishicha, hodisa <источник> 2024 yil 2 iyun </источник> kuni sodir bo'lgan. Ma'lum qilinishicha, <источник> Urgut tumani </источник> <источник> Yangihayot mahallasida </источник>, shu tumanda yashovchi <агенс> fuqaroning </агенс> <адресат> 9-sinfda </адресат> o'qiydigan <агенс> farzandi <агенс> <инструмент> Nexia-2 mashinasini </инструмент> <цель> haydab </цель> </источник> ko'chaga </источник> chiqqan. So'ng <инструмент> mashinani </инструмент> katta tezlikda boshqarib <источник> yo'lni </источник> kesib o'tayotgan <агенс> 6 yoshli qizchani </агенс> </цель> urib yuborgan </цель> . Oqibatda <агенс> qizcha </агенс> <стимул> og'ir tan jarohatlari </стимул> olgan va <источник> shifoxonada </источник> <цель> vafot etgan </цель>.<агенс> MUY5 telekanalining </агенс> <пациенс> Markaziy studiya dasturida </пациенс> ko'rsatilgan <адресат> lavhada </адресат> <агенс> o'quvchining otasi </агенс> <инструмент> mashina kaliti </инструмент> <адресат> o'zida </адресат> qolgan paytda <агенс> o'g'li </агенс> <адресат> undan </адресат> so'ramay, <инструмент> mashinani </инструмент> olib chiqib ketganini aytgan.

Документ D3

Bugun bo'lmasa qachon? <агенс> «Grand Pharm Trade» kompaniyasi </агенс> <источник> besh ming </источник> <бенефактив> oilani </бенефактив> <цель> qaramog'iga oldi. </цель> <агенс> Davlatimiz rahbari </агенс> tashabbusi bilan tashkil etilgan <пациенс> «Saxovat va ko'mak» </пациенс> <пациенс> umumxalq harakati </пациенс> tez fursatda tilga tushdi va ko'pchilik tarafidan e'tirof etildi. <агенс> «Grand Pharm» dorixonalar </агенс> tarmog'i ham bu ezgu harakatga qo'shildi va <источник> besh ming </источник> <бенефактив> oilani </бенефактив> murakkab kunlarda <цель> moddiy jihatdan ta'minlashni </цель> o'z zimmasiga oldi. – Bugun yordam bermasak, qachon yordam beramiz? – degan savolni qo'ydi suhbat chog'ida <агенс> «Grand Pharm Trade» kompaniyasi </агенс> <агенс> ta'sischisi </агенс> , <агенс> yetakchi tadbirkorlardan biri Azizxo'ja Abbosxo'jayev </агенс>. Davlatimizning so'nggi 2-3 yil ichida biz <бенефактив> tadbirkorkorlarga </бенефактив> bergan imtiyozlarini sanaydigan bo'lsak, <пациенс> imkoniyatlarimiz </пациенс> naqadar kengayganini anglab yetamiz. Birgina <пациенс> konvertatsiya </пациенс> jarayonining yo'lga qo'yilishi bizning <бенефактив> soha vakillari </бенефактив> uchun qanchalik qulayliklar yaratdi, yechimini kutib turgan <цель> muammolarni bartaraf qildi </цель>. Demak, bugun biz ham bunga munosib javob qaytaradigan kun keldi. Bir narsani alohida va qayta-qayta ta'kidlashni istardim, bu homiylik harakatiga yoki saxovat ishlariga bosh qo'shishga hech kim bizni majburlagani, hatto iltimos ham qilgani yo'q. Yana bir muhim jihat, davlatimiz <пациенс> «Saxovat va ko'mak» umumxalq harakati </пациенс> doirasida qilingan ishlar uchun sarflangan <инструмент> mablag'lar </инструмент> o'rnini turli usullarda to'ldirib bermoqchi. Ammo bu <пациенс> imtiyoz </пациенс> bizni xursand qilsada, turtki bo'lmadi. Qilayotgan <агенс> ehsonlarimizning </агенс> hammasi o'z ixtiyorimizga ko'ra, <цель> xalqimiz roziligini </цель> olish uchun amalga oshirilmoqda. <агенс> «Grand Pharm Trade» kompaniyasi </агенс> ayni paytda <пациенс> yordamga muhtoj </пациенс> bo'lib turgan <источник> besh ming </источник> <бенефактив> oilani </бенефактив> <цель> qaramog'iga oladi </цель> va boshqalarni ham bu harakatga qo'shilishga da'vat qiladi, – deya so'zlarini yakunladi <агенс> Azizxo'ja Abbosxo'jayev </агенс>. Ma'lumot o'rnida: <агенс> «Grand Pharm» dorixonalar </агенс> tarmog'i mamlakatimizda sohaning yetakchi kompaniyalaridan biri hisoblanadi. <агенс> Koronavirus karantini </агенс> boshlangan kunlarda <источник> Kun.uz </источник> bu <агенс> kompaniya </агенс> <пациенс> omborlariga </пациенс> tashrif buyurgan va <пациенс> pande-miya chog'ida </пациенс> eng zarur instrumentlar, jumladan <инструмент> tibbiyot niqoblari, antiseptiklar, dori-darmonlar </инструмент> yetarli miqdorda tayyorlab qo'yilgani hamda yetkazib berilayotganiga guvoh bo'lib qaytgandi.

Документ D4

<источник> Toshkentda </источник> <агенс> fuqarolarni </агенс> <цель> AQShga yuborishni </цель> va'da qilgan <агенс> shaxslar </агенс> ushlendi <источник> Poytaxtda </источник> <цель> xorijga ishga jo'natish </цель> bilan bog'liq </пациенс> firibgarlik </пациенс> holatlari <цель> fosh etildi </цель>. <адресат> Bosh prokuratura huzuridagi departamentning </адресат> <адресат> Uchtepa tuman bo'limi </адресат> tomonidan <агенс> DXX va IIV xodimlari </агенс> bilan hamkorlikda <инструмент> tezkor tadbir </инструмент> o'tkazildi. <адресат> Unda

</адресат>, <агенс> fuqaro G.A. </агенс> <агенс> fuqaro G.B.ning </агенс> <пациенс> ishonchiga kirib </пациенс>, <агенс> 3 nafar fuqaroni </агенс> 42 000 <инструмент> AQSh dollari </инструмент> evaziga <инструмент> tanishlari </инструмент> yordamida <цель> Amerika Qo'shma Shtatlariga yuborishi </цель> hamda <цель> viza </цель> hujjatlarini rasmiylashtirib berishini aytib, oldindan 28 800 <инструмент> AQSh dollarini </инструмент> olgan vaqtida <инструмент> ashyoviy dalillar </инструмент> bilan ushlangan. Shu kabi, <агенс> departamentning </агенс> <агенс> Uchtepa tuman bo'limi </агенс> va <агенс> DXX xodimlari </агенс> hamkorligida yana bir <инструмент> tezkor tadbir </инструмент> o'tkazilgan. Unda, <агенс> fuqaro F.A. fuqaro </агенс> <адресат> M.Ya.ning </адресат> <цель> ishonchiga kirib </цель>, <инструмент> uni tanishlari </инструмент> yordamida <цель> Amerika Qo'shma Shtatlariga borishi </цель> uchun <инструмент> hujjatlarni </инструмент> rasmiylashtirib, <цель> viza </цель> olib berishi evaziga 17 000 <инструмент> AQSh dollarini </инструмент> olgan vaqtida ashyoviy dalillar bilan ushlangan. <пациенс> Mazkur holatlar </пациенс> yuzasidan <источник> Jinoyat kodeksining 168-moddasi </источник> <пациенс> (firibgarlik) </пациенс> va <источник> 28, 211-moddasi </источник> <пациенс> (pora berish) </пациенс> bilan <пациенс> jinoyat </пациенс> ishlari qo'zg'atilib, <пациенс> tergov <пациенс> harakatlari o'tkazilmoqda.

Документ D5

<источник> Qashqadaryoda </источник> <инструмент> mashina </инструмент> <цель> olib berishni </цель> aytib <пациенс> fuqarolarning </пациенс> <пациенс> pullarini </пациенс> <агенс> o'zlashtirgan shaxsga </агенс> <пациенс> jinoyat ishi </пациенс> qo'zg'atildi <агенс> U </агенс> <источник> avtosalondan </источник> qisqa muddatda <инструмент> mashina </инструмент> <цель> olib berishni </цель> aytib, <пациенс> firibgarlik </пациенс> bilan <агенс> ikki kishidan </агенс> <источник> 22000 </источник> <инструмент> dollar </инструмент> olgan. <источник> Qarshi </источник> <источник> shahrida <источник> <адресат> fuqarolarga </адресат> <инструмент> mashina </инструмент> <цель> olib berishni </цель> va'da qilib, ularning <источник> 22000 <источник> <инструмент> dollarini </инструмент> qo'lga kiritgan <агенс> shaxs </агенс> aniqlangan. Bu haqda <агенс> Bosh prokuratura </агенс> huzuridagi <агенс> departament </агенс> xabarbermoqda. Ma'lum qilinishicha, <агенс> departamentning </агенс> <источник> Qarshi </источник> shahri bo'limi tomonidan o'tkazilgan <пациенс> tergovga </пациенс> qadar tekshiruvda <агенс> fuqaro </агенс> <бенефактив> A.Yu. </бенефактив> <агенс> ikki kishiga </агенс> <источник> avtosalondan </источник> shartnoma asosida qisqa muddatda <инструмент> Cobalt </инструмент> va <инструмент> Malibu <инструмент> <инструмент> mashinalarini </инструмент> <цель> olib berishini </цель> aytib, ulardan <источник> 22000 dollarni </источник> <пациенс> firibgarlik <пациенс> yo'li bilan qo'lga kiritgan. Qayd etilishicha, mazkur holat yuzasidan <источник> Jinoyat kodeksining 168-moddasi </источник> <пациенс> (firibgarlik) </пациенс> bilan <пациенс> jinoyat ishi </пациенс> qo'zg'atilib, <пациенс> tergov </пациенс> harakatlari o'tkazilmoqda.

Документ D6

<агенс> ИБВ </агенс> <источник> Toshkentda </источник> <пациенс> narkotik </пациенс> ta'sirga ega <пациенс> kimyoviy </пациенс> инструментlar bilan ishlov

berilgan <цель>niqoblarni</цель> tarqatgan <бенефактив>shaxslar</бенефактив> haqidagi gap-so'zlarni rad etdi. <источник> Toshkent shahrida </источник> <пациенс> narkotik <пациенс> ta'sirga ega <пациенс> kimyoviy <пациенс> instrumentlar bilan ishlov berilgan <цель> niqoblar tarqalgani </цель> haqida gap-so'zlar tarqaldi. <источник> Internet tarmog'ida </источник> <пациенс> audioxabar </пациенс> tarqaldi, unda <бенефактив> noma'lum shaxslar </бенефактив> eshikma-eshik yurib, <цель> niqob tarqatgani </цель> haqida so'z yuritilgan. Aytilishicha, bu <пациенс> mahalliy hokimiyatning </пациенс> yangi <пациенс> tashabbusi <пациенс>. Bildirilishicha, ular <цель> niqob </цель> sizga to'g'ri kelishini ko'rish uchun sizdan <цель> niqobni taqishni </цель> iltimos qilishadi. Mazkur <цель> niqobga </цель> <пациенс> narkotik </пациенс> ta'sirga ega <пациенс> kimyoviy </пациенс> instrumentlar bilan ishlov berilgan. Shundan so'ng ular <агенс> fuqarolarni </агенс> <цель> o'marishi </цель> aytilgan va <бенефактив> begonalardan </бенефактив> <цель> niqob olmaslik <цель> iltimos qilingan. Poytaxt <агенс> ИБВ </агенс> matbuot xizmatima'lumotiga ko'ra, ushbu xabar haqiqatdan yiroqdir. <источник> Yunusobod tumani yoki Toshkent shahrining boshqa tumani hududida </источник> bu kabi <пациенс> holatlar </пациенс> yuz bermagan. "Bu kabi <пациенс> xabarlar </пациенс> tarqalishi to'g'risida <источник> Rossiya Federatsiyasi huquqni muhofaza qiluvchi organlari </источник> <цель> ogohlantirishgan </цель> edi. Bizning <агенс> fuqarolar </агенс> ham ushbu holat go'yoki <источник> Toshkent tumanlaridan </источник> birida ro'y bergan degan holda ba'zi talqinlar bilan bu kabi xabarlarni <цель> tarqatishmoqda </maqsa>. Hurmatli <пациенс> hamshaharlar </пациенс> va Toshkent <пациенс> mehmonlari</пациенс>! Rasmiy <агенс> OAV </агенс>, <пациенс> davlat idoralariga </пациенс> qarashli bo'lmagan har qanday <источник> internet-источникlarda </источник> tarqayotgan xabarlarga <цель> ishonib ketavermang </цель>. <агенс> Ma'lumot </агенс> naqadar <пациенс> ishonchli </пациенс> ekanligini, <пациенс> qayerdan olinganligini </пациенс> <цель> tekshiring </цель>. O'zingizni <пациенс> asrang va salomat </пациенс> bo'ling!" - deyiladi xabarda.

Документ D7

<агенс> Gazeta </агенс> <пациенс> qadriyatmi </пациенс>? - <агенс> Bosma nashrlar </агенс> taqdiri borasidagi munozaralar haqida. <пациенс> Karantinga </пациенс> o'rganib qoldik shekilli, asta-sekin boshqa <пациенс> mavzular </пациенс> - <источник> davlat idoralarida </источник> <инструмент> davlat tilida </инструмент> <цель> ish yuritish </цель>, <агенс> bosma nashrlar </агенс> taqdiri kabi <пациенс> mavzular </пациенс> ham <цель> kun tartibiga </цель> chiqa boshladi. <пациенс> Insoniyat </пациенс> o'zi shunaqa, har qanday sharoitda avval <стимул> noqulaylik his qilsa-da </стимул>, keyin yana o'z shakliga, <пациенс> azaliy mavzulariga </пациенс> qaytadi. <агенс> Bosma nashrlar </агенс> — <пациенс> internet </пациенс> qarama-qarshiligi bizda xuddi shunday, <пациенс> azaliy mavzu </пациенс>. <источник> Milliy </источник> <пациенс> mass-mediani </пациенс> <цель> qo'llab-quvvatlash va rivojlantirish </цель> <агенс> jamoat fondi vasiylik kengashi raisi Komil Allamjonov </агенс> o'z <пациенс> intervyyusida </пациенс> <агенс> bosma nashrlarning </агенс> <пациенс> davri o'tgani </пациенс>, <агенс> matnli axborot </агенс> shakl <цель> almashtirish </цель> vaqti kelganini aytishi hamon, unga <пациенс> hujumlar </пациенс> boshlandi: hatto

<агенс> gazetachilikka </агенс> qarshi bo'lgan <агенс> odamlarniw </агенс> <пациенс> ma'naviyatsiz, kitob o'qimaydigan, madaniy merosimizning dushmani </пациенс> sifatida aks ettirganlar ham bo'lyapti. <агенс> Bosma nashrlar </агенс> o'ziga xos tarzda <пациенс> qadriyat </пациенс> sifatida namoyon qilinmoqda. Aslida, bu <пациенс> bahsning </пациенс> o'z <пациенс> konteksti </пациенс> bor va Komil Allamjonov <агенс> bosma nashrlar </агенс> haqida <пациенс> fikr </пациенс> yuritganda ham, unga berilgan <пациенс> javoblarni </пациенс> ham o'sha <пациенс> kontekst </пациенс> asosida <цель> tushunishimiz lozim. </цель> <пациенс> Asosiy nuqta </пациенс> <агенс> gazetaning </агенс> bir инструмент ekani, инструмент esa, hech qachon <пациенс> qadriyat </пациенс> bo'la olmasligidir. Kun mavzusi: <агенс> Gazetalar </агенс> kelajagi bo'yicha ikki qarash

В приведённом примере из семи документов применяется косинусная метрика для измерения сходства между ними. Использование большинства из известных метрик (например, евклидовой) невозможно из проблемы проклятия размерности.

Косинусная метрика не зависит от длины векторов и лучше отражает различия в содержании текстов, что делает её предпочтительным инструментом для тематического моделирования в задачах с высоко размерными данными.

Результаты использования косинусной метрики к векторному представлению описанных выше 7 документов приводятся в таблице:

	D1	D2	D3	D4	D5	D6	D7
D1	0	0.4252	0.1942	0.1505	0.0887	0.2466	0.0566
D2	0.4252	0	0.0878	0.1299	0.3472	0.2199	0.0681
D3	0.1942	0.0878	0	0.0960	0	0.0944	0.1577
D4	0.1505	0.1299	0.0960	0	0.5775	0.1792	0.0071
D5	0.0887	0.3472	0	0.5775	0	0.1611	0.0252
D6	0.2466	0.2199	0.0944	0.1792	0.1611	0	0.0584
D7	0.0566	0.0681	0.1577	0.0071	0.0252	0.0584	0

Косинусная метрика позволяет определить угловое расстояние между векторными представлениям документов, значение 1.0 указывает на полное совпадение, а 0.0 — на полное различие. Например, $\cos(D3, D5) = 0.0$ означает, что тексты D3 и D5 не имеют общих характеристик, в то время как $\cos(D4, D5) = 0.5775$ показывает значительное сходство между этими текстами.

Косинусная метрика особенно полезна в задачах тематического моделирования, где важно сравнивать тексты по их смысловому содержанию, а не по количеству слов.

Важно отметить, что низкие значения (близкие к 0) свидетельствуют о том, что тексты имеют различный смысловой контент, в то время как высокие значения (близкие к 1) указывают на более значительное пересечение тем.

Косинусная метрика косинуса является эффективным инструментом для кластеризации текстов и определения их тематического сходства.

Однако следует учитывать, что в больших текстовых корпусах, где количество терминов превышает 10 000, эта метрика также может сталкиваться с ограничениями по точности. В таких случаях могут потребоваться дополнительные методы нормализации или уменьшения размерности данных.

4 Заключение

Описана методика разметки документов с использованием семантических ролей. Для определения сходства документов применялась косинусная метрика. Численные значения сходства рекомендуются использовать для выбора алгоритмов классификации и кластерного анализа в тематическом моделировании. Описанная методика востребована для борьбы с фейковыми новостями. Векторные представления размеченных документов необходимо хранить в виде отдельных баз прецедентов с идентификацией новостей. Примерами идентификации являются телефонное мошенничество, брачные аферисты, финансовые пирамиды. С увеличением объема баз прецедентов связана эффективность применения методов машинного обучения для автоматической разметки документов.

Литература

- [1] *Воронцов К.В.* Вероятностные тематические модели (курс лекций, К.В.Воронцов) ВМК МГУ. 1 марта – 2018. URL: [1.http://www.MachineLearning.ru/wiki](http://www.MachineLearning.ru/wiki).
- [2] *Воронцов К.В., Фрей А.И., Апишев М.А., Потапенко А.А.* Тематическое моделирование в BigARTM: теория, алгоритмы, приложения. 14 июнь – 2015.
- [3] *Воронцов К.В., Булатов В.Г., Алексеев В.П.* Determination of the Number of Topics Intrinsically: Is It Possible? 14 June – 2024. <https://arxiv.org/pdf/2406.10402>.
- [4] *Kleinberg J.M.* An Impossibility Theorem for Clustering Jon Kleinberg Advances in Neural Information Processing Systems 15. NIPS – 2002.
- [5] *Игнатьев Н.А., Тулиев У.Ю.* Семантическая структуризация текстовых документов на основе паттернов сущностей естественного языка. Компьютерные исследования и моделирование – 2022. – Т.14 – №5 – С. 1185–1197.
- [6] *Воронцов К.В.* Разметка данных для обучения нейросетевых моделей языка как способ формализации гуманитарных знаний // XVIII научная конференция межрегиональной ассоциации «История и компьютер» «Историческая информатика как Historical Data Science».
- [7] *Tikhomirov M.M.* "Using bert and augmentation in named entity recognition for cybersecurity domain," in Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems, NLDB – 2020. Saarbrücken

Поступила в редакцию 18.10.2024

UDC 004.4

DOCUMENT ANNOTATION BY SEMANTIC ROLES

Ignatev N.A., Abdullaev K.D.

n_ignatev@rambler.ru

National University of Uzbekistan named after Mirzo Ulugbek,

4, University Str., Tashkent 100174 Uzbekistan.

This study examines the annotation of documents with respect to semantic roles. The data for thematic models is formed based on the standard bag-of-words hypothesis. The question of the number and uniqueness of themes in modeling remains unresolved. A key reason cited is the lack of unified criteria for evaluating the quality of clustering. There

have been attempts to prove the uniqueness of themes through experimental evidence using a set of regularization criteria based on the results of document grouping. The absence of such proof is explained by the fact that the number of groups is initially set as a free parameter. Currently, five types of annotation are applied for the unification of documents: metatextual, morphological, syntactic, accentual, and semantic. This paper employs semantic annotation by identifying actants within sentences—noun phrases that denote participants in a situation and their semantic roles. It is proposed that, in defining semantic roles in the Uzbek language, the unification of terms should be conducted in relation to specific subject areas. An example of document annotation using semantic roles is provided, along with a determination of their similarity based on cosine metric.

Citation: Ignatev N.A., Abdullaev K.D. 2024. Document annotation by semantic roles. *Problems of Computational and Applied Mathematics*. 5(61):80-90.

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

№ 5(61) 2024

Журнал основан в 2015 году.

Издается 6 раз в год.

Учредитель:

Научно-исследовательский институт развития цифровых технологий и
искусственного интеллекта.

Главный редактор:

Равшанов Н.

Заместители главного редактора:

Азамов А.А., Арипов М.М., Шадиметов Х.М.

Ответственный секретарь:

Ахмедов Д.Д.

Редакционный совет:

Азамова Н.А., Алоев Р.Д., Амиргалиев Е.Н. (Казахстан), Бурнашев В.Ф.,
Загребина С.А. (Россия), Задорин А.И. (Россия), Игнатъев Н.А.,
Ильин В.П. (Россия), Исмагилов И.И. (Россия), Кабанихин С.И. (Россия),
Карачик В.В. (Россия), Курбонов Н.М., Маматов Н.С., Мирзаев Н.М.,
Мирзаева Г.Р., Мухамадиев А.Ш., Назирова Э.Ш., Нормуродов Ч.Б.,
Нуралиев Ф.М., Опанасенко В.Н. (Украина), Расулмухамедов М.М., Расулов А.С.,
Садуллаева Ш.А., Старовойтов В.В. (Беларусь), Хаётов А.Р., Халджигитов А.,
Хамдамов Р.Х., Хужаев И.К., Хужаеров Б.Х., Чье Ен Ун (Россия),
Шабозов М.Ш. (Таджикистан), Dimov I. (Болгария), Li Y. (США),
Mascagni M. (США), Min A. (Германия), Schaumburg H. (Германия),
Singh D. (Южная Корея), Singh M. (Южная Корея).

Журнал зарегистрирован в Агентстве информации и массовых коммуникаций при
Администрации Президента Республики Узбекистан.

Регистрационное свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

При перепечатке материалов ссылка на журнал обязательна.

За точность фактов и достоверность информации ответственность несут авторы.

Адрес редакции:

100125, г. Ташкент, м-в. Буз-2, 17А.

Тел.: +(998) 712-319-253, 712-319-249.

Э-почта: journals@airi.uz.

Веб-сайт: <https://journals.airi.uz>.

Дизайн и вёрстка:

Шарипов Х.Д.

Отпечатано в типографии НИИ РЦТИИ.

Подписано в печать 30.10.2024 г.

Формат 60x84 1/8. Заказ №7. Тираж 100 экз.

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

No. 5(61) 2024

The journal was established in 2015.
6 issues are published per year.

Founder:

Digital Technologies and Artificial Intelligence Development Research Institute.

Editor-in-Chief:

Ravshanov N.

Deputy Editors:

Azamov A.A., Aripov M.M., Shadimetov Kh.M.

Executive Secretary:

Akhmedov D.D.

Editorial Council:

Azamova N.A., Alov R.D., Amirgaliev E.N. (Kazakhstan), Burnashev V.F.,
Zagrebina S.A. (Russia), Zadorin A.I. (Russia), Ignatiev N.A., Ilyin V.P. (Russia),
Ismagilov I.I. (Russia), Kabanikhin S.I. (Russia), Karachik V.V. (Russia), Kurbonov
N.M., Mamatov N.S., Mirzaev N.M., Mirzaeva G.R., Mukhamadiev A.Sh., Nazirova
E.Sh., Normurodov Ch.B., Nuraliev F.M., Opanasenko V.N. (Ukraine), Rasulov A.S.,
Sadullaeva Sh.A., Starovoitov V.V. (Belarus), Khayotov A.R., Khaldjigitov A.,
Khamdamov R.Kh., Khujaev I.K., Khujayorov B.Kh., Chye En Un (Russia),
Shabozov M.Sh. (Tajikistan), Dimov I. (Bulgaria), Li Y. (USA), Mascagni M. (USA),
Min A. (Germany), Schaumburg H. (Germany), Singh D. (South Korea),
Singh M. (South Korea).

The journal is registered by Agency of Information and Mass Communications under the
Administration of the President of the Republic of Uzbekistan.

The registration certificate No. 0856 of 5 August 2015.

ISSN 2181-8460, eISSN 2181-046X

At a reprint of materials the reference to the journal is obligatory.

Authors are responsible for the accuracy of the facts and reliability of the information.

Address:

100125, Tashkent, Buz-2, 17A.

Tel.: +(998) 712-319-253, 712-319-249.

E-mail: journals@airi.uz.

Web-site: <https://journals.airi.uz>.

Layout design:

Sharipov Kh.D.

DTAIDRI printing office.

Signed for print 30.10.2024

Format 60x84 1/8. Order No. 7. Printed copies 100.

Содержание

Равшанов Н., Шадманов И.

Многомерная математическая модель одновременного тепло- и влагопереноса при сушке и хранении хлопка-сырца на открытых площадках 5

Туракулов Ж.

Численное исследование процесса фильтрования малоконцентрированных растворов через пористую среду 18

Мирзаахмедов М.К.

Математическое моделирование процессов термо-электро-магнитоупругой деформации тонких пластин сложной конструктивной формы 31

Халджигитов А.А., Джумаёзов У.З., Усмонов Л.С.

Новые связанные краевые задачи термоупругости в деформациях 43

Нормуродов Ч.Б., Зиякулова Ш.А.

Численное моделирование уравнений эллиптического типа дискретным вариантом метода предварительного интегрирования 59

Фаязов К.С., Рахимов Д.И., Фаязова З.К.

Некорректная начально-краевая задача для уравнения смешанного типа третьего порядка 69

Игнатъев Н.А., Абдуллаев К.Д.

Разметка документов по семантическим ролям 80

Зайнидинов Х.Н., Ходжаева Д.Ф., Хурамов Л.Я.

Продвинутые модели обработки сигналов в системе умного дома 91

Абдурахимов А.А., Пономарев К.О., Прохошин А.С.

Интеграция методов машинного обучения для раннего обнаружения патогенов в растениях на основе анализа хлорофилла 107

Xalikov A.A., Xurramov A.Sh.

"Pop-Namangan-Andijon" uchastkasining temir yo'l transport tarmog'ida radioaloqa ishonchliligini hisoblashning mantiqiy-ehtimoliy modeli 115

Contents

Ravshanov N., Shadmanov I.

Multidimensional mathematical model of simultaneous heat and moisture transfer during drying and storage of raw cotton in open areas 5

Turakulov J.

Numerical study of the process of filtration of low-concentration solutions through a porous medium 18

Mirzaakhmedov M.K.

Mathematical modeling of thermo-electro-magnit-elastic deformation processes of thin plates of complex constructive form 31

Khaldjigitov A.A., Djumayozov U.Z., Usmonov L.S.

New coupled thermoelasticity boundary-value problems in strains 43

Normurodov Ch.B., Ziyakulova Sh.A.

Numerical modeling of elliptic type equations by a discrete variant of the pre-integration method 59

Fayazov K.S., Rahimov D.I., Fayazova Z.K.

Ill-posed initial-boundary value problem for a third-order mixed type equation . . 69

Ignatev N.A., Abdullaev K.D.

Document annotation by semantic roles 80

Zaynidinov H., Hodjaeva D., Xuramov L.

Advanced signal processing models in a smart home system 91

Abdurakhimov A.A., Ponomarev K.O., Prokhoshin A.S.

Integration of machine learning methods for early detection of pathogens in plants based on chlorophyll analysis 107

Khalikov A.A., Khurramov A.Sh.

The logical-probable model of calculating the reliability of the radio communication in rail transpot network of the "Pop-Namangan-Andijan"plot 115