ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫ

әл-Фараби атындағы Қазақ ұлттық университетінің

ХАБАРЛАРЫ

ИЗВЕСТИЯ

НАЦИОНАЛЬНОЙ АКАДЕМИИ
НАУК РЕСПУБЛИКИ КАЗАХСТАН
Казахский национальный
университет имени аль-Фараби

N E W S

OF THE ACADEMY OF SCIENCES
OF THE REPUBLIC OF
KAZAKHSTAN
al-Farabi Kazakh National University

SERIES
PHYSICO-MATHEMATICAL

3 (343)

JULY – SEPTEMBER 2022

PUBLISHED SINCE JANUARY 1963

PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

«ҚР ҰҒА Хабарлары. Физика-математикалық сериясы».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Меншіктеуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.). Қазақстан Республикасының Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде 14.02.2018 ж. берілген **No 16906-Ж** мерзімдік басылым тіркеуіне қойылу туралы куәлік.
Тақырыптық бағыты: *ақпараттық коммуникациялық технологиялар сериясы.*
Қазіргі уақытта: *«ақпараттық технологиялар»бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
Тиражы: *300 дана.*
Редакцияның мекен-жайы: *050010, Алматы қ., Шевченко көш., 28, 219 бөл., тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Типографияның мекен-жайы: «Аруна» ЖК, Алматы қ., Мұратбаев көш., 75.

«Известия НАН РК. Серия физика-математическая».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Собственник: *Республиканское общественное объединение «Национальная академия наук Республики Казахстан» (г. Алматы).*
Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и общественного развития Республики Казахстан **No 16906-Ж** выданное 14.02.2018 г.
Тематическая направленность: *серия информационные коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных ККСОН МОН РК по направлению «информационные коммуникационные технологии».*
Периодичность: *4 раз в год.*
Тираж: *300 экземпляров.*
Адрес редакции: *050010, г. Алматы, ул. Шевченко, 28, оф. 219, тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Адрес типографии: ИП «Аруна», г. Алматы, ул. Муратбаева, 75.

News of the National Academy of Sciences of the Republic of Kazakhstan.
Physical-mathematical series.
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Owner: RPA «National Academy of Sciences of the Republic of Kazakhstan» (Almaty). The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Social Development of the Republic of Kazakhstan **No. 16906-Ж,** issued 14.02.2018
Thematic scope: *series information technology.*
Currently: *included in the list of journals recommended by the CCSES MES RK in the direction of «information and communication technologies».*
Periodicity: *4 times a year.*
Circulation: *300 copies.*
Editorial address: *28, Shevchenko str., of. 219, Almaty, 050010, tel. 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Address of printing house: ST «Aruna», 75, Muratbayev str, Almaty.

NEWS OF THE NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 3, Number 343 (2022), 136-152
https://doi.org/10.32014/2022.2518-1726.143
МРНТИ 28.23.29
УДК 004.89

А.Ж. Картбаев[1], Г.С. Ыбытаева[2*], О.Ж. Мамырбаев[1], К.Ж. Мухсина[1], Б.Ж. Жумажанов[1]

[1]Институт информационных и вычислительных технологий, Казахстан, Алматы;

[2]Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Казахстан, Алматы.
E-mail: ybytayeva.galiya@gmail.com

МЕТОДЫ ФОРМАЛЬНОГО ПРЕДСТАВЛЕНИЯ СУЩНОСТЕЙ В КРИМИНАЛЬНЫХ НОВОСТЯХ ДЛЯ АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ОНТОЛОГИИ ПРЕСТУПЛЕНИЙ

Аннотация. В данной работе мы изучаем методы формального представления сущностей в криминальных новостях в виде онтологии путем определения их свойств и взаимосвязей между ними. Таким образом, разработанная нами онтология может быть использована правоохранительными органами для отслеживания и предотвращения преступной деятельности. Сейчас доступные нам открытые данные из социальных сетей и газетные статьи могут дать много полезной информации о формах преступной деятельности в конкретном месте и личной информации подозреваемых. А также мы проводим анализ наших данных и определяем их сложность, реализуем основные функций нашей системы, проверяем, какие цели и задачи решаются нашей системой. В результате анализа были выделены ключевые классы представления знаний о предметной области, составляющие основную структуру разработанной онтологии, которые мы используем, как ее словарь. Таким образом, была построена вложенная иерархия классов онтологии, представляющая собой совокупность иерархии терминов. В данном проекте мы предлагаем прикладной метод

разработки онтологии преступлений, который сначала использует сбор текста и изображений из новостных статей, затем мы расширяем и обогащаем онтологию, используя соответствующую информацию из известных социальных сетей. Получение семантически обоснованной информации играет важную роль в данном контексте, так как помогает должностным лицам понимать и эффективно работать со своими онтологиями и, в частности, использовать информацию в оптимальных масштабах. После этого нами были предложены способы практического применения построенной предметной онтологии, сформулированы направления последующих исследований.

Ключевые слова: онтология, извлечение информации, семантический анализ, граф знаний, криминальные данные.

А.Ж. Картбаев[1], Г.С. Ыбытаева[2*], О.Ж. Мамырбаев[1], К.Ж. Мухсина[1], Б.Ж. Жумажанов[1]

[1]Ақпараттық және есептеуіш технологиялар институты, Қазақстан, Алматы;
[2]Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Қазақстан, Алматы.
E-mail: *ybytayeva.galiya@gmail.com*

ҚЫЛМЫСТЫҚ ОНТОЛОГИЯНЫ АВТОМАТТЫ ТҮРДЕ ҚҰРУ ҮШІН ҚЫЛМЫСТЫҚ ЖАҢАЛЫҚТАРДАҒЫ СУБЪЕКТІЛЕРДІ РЕСМИ ТҮРДЕ ҰСЫНУ ӘДІСТЕРІ

Аннотация. Бұл мақалада онтология түрінде қылмыстық жаңалықтардағы субъектілерді ресми түрде ұсыну әдістері қасиеттері мен қатынастарын анықтау арқылы зерттеледі. Бұл онтологияны құқық қорғау органдары қылмыстық әрекетті бақылау және алдын-алу үшін қолдана алады. Бүгінде әлеуметтік желілер мен газет мақалаларынан алынған ашық деректер белгілі бір жерде қылмыстық әрекеттің сипаты туралы және күдіктілердің жеке мәліметтері туралы көптеген пайдалы ақпарат бере алады. Сондай-ақ деректер талданып олардың күрделілігі анықталды, жүйедегі негізгі функциялар іске асырылып, қандай мақсаттар мен міндеттерді шешетіні тексерілді. Талдау нәтижесінде дамыған онтологияның негізгі құрылымы болып табылатын пәндік сала туралы білімді ұсынудың негізгі кластары

анықталды. Терминдердің иерархияларының жиынтығы болып табылатын онтология кластарының кірістірілген иерархиясы құрылды. Онтологияны дамыту әдісін ұсынылды. Ол өз кезегінде, қылмыстық жаңалықтар мақалаларынан мәтіндер мен суреттер жиынтығын пайдаланып, танымал әлеуметтік желілердегі тиісті ақпаратты қолдана отырып онтологияны кеңейтеді. Семантикалық бай ақпаратты алу маңызды рөл атқарады, өйткені бұл лауазымды тұлғаларға онтологияны түсінуге және онымен тиімді жұмыс істеуге, сонымен қатар ақпаратты оңтайлы масштабта пайдалануға көмектеседі.

Түйін сөздер: онтология; ақпарат алу; семантикалық талдау; білім графигі; қылмыстық деректер.

A.Zh. Kartbayev[1], G.S. Ybytayeva[2*], O.Zh. Mamyrbayev[1], K.Zh. Mukhsina[1], B. Zh. Zhumazhanov[1]

[1]Institute of Information and Computer Technologies, Kazakhstan, Almaty;
[2]Kazakh National Research Technical University named after K.I. Satpayev, Kazakhstan, Almaty.
E-mail: ybytayeva.galiya@gmail.com

METHODS FOR FORMAL REPRESENTATION OF ENTITIES IN CRIME NEWS FOR AUTOMATIC CRIME ONTOLOGY CONSTRUCTION

Abstract. In this paper, we study methods for formally representing entities in crime news in the form of an ontology by identifying their properties and relationships. The ontology we have developed can be used by law enforcement agencies to track and prevent criminal activity. Todays the open data available to us from social media and newspaper articles can provide a lot of useful information about the patterns of criminal activity in a particular location and personal information of suspects. Also we analyze data and determine its complexity, implement basic functions of our system, and check what goals and objectives our system solves. As a result of the analysis we identified the key classes of knowledge representation about the subject area, which represents basic structure of the developed ontology, which we use as its vocabulary. Then a nested hierarchy of ontology classes was built, which is a set of hierarchy of terms. We propose a method for developing an ontology that uses the collection of text and images from criminal news

articles, then we extend the ontology using relevant information from popular social networks. Extracting semantically rich information plays an important role, because it helps officials understand and work effectively with the ontology and, use the information at optimal scale. After that we suggested cases of practical application of the constructed subject ontology and formulated directions for further research.

Key words: ontology; information extraction; semantic analysis; knowledge graph; criminal data.

Введение. Часто в реальном мире существует огромное количество необработанных данных о преступлениях. Обнаружение знаний в сложных областях может стать проблемой для методов добычи данных, которые обычно ограничиваются представлениями данных, не имея возможности получить доступ к их контексту и значению. Анализировать такой огромный объем данных простым взглядом довольно обременительно, и даже если потратить достаточно времени на их понимание, это не всегда приводит к семантически корректным результатам. Мы выбрали исходные данные криминальных новостей за 2015-2018 года, которые мы сохранили в формате RDF. Следовательно, в нашем исследовании мы реализуем систему поиска с использованием языка запросов RDF, который является одновременно языком запросов и протоколом доступа к данным, используемым для извлечения семантической информации из RDF.

Наша система предоставляет пользовательский интерфейс, в котором сотрудник правоохранительных органов может ввести строку поиска, а базовый API помогает получить и отобразить информацию для сотрудника. Концептуальные структуры, определяющие базовую онтологию, имеют отношение к идее машинного понимания данных в Семантической паутине.

Онтология – это схемы метаданных, предоставляющие контролируемый словарь понятий, каждое из которых имеет четко определенную и обрабатываемую машиной семантику. Определяя общие теории домена, онтологий помогают людям и машинам общаться лаконично, поддерживая обмен семантикой, а не только синтаксисом.

Краткий обзор исследований. В области извлечения информации было проведено достаточно много исследований. Как упоминалось в исследовательской работе (Euzenat и др., 2004) о сходстве между двумя сущностями, определенными между двумя узлами категории X графа, следует двум принципам: оно зависит от рассматриваемой категории;

оно учитывает все признаки этой категории (например, свойства). Пара сущностей, сходство которой оценивается называется якорной парой сравнения, а все пары, которые вносят индивидуальный вклад в расчет общего сходства называются участниками. Агрегирование сходства всех вкладчиков происходит с помощью взвешенной суммы, которая помогает контролировать вклад каждого признака.

Был оценен широкий спектр метрик сходства строк, а также стратегии предварительной обработки строк, такие как удаление стоп-слов и учет синонимов в различных типах онтологий (Gruber, 1995). Представлен набор рекомендаций о том, когда использовать ту или иную метрику, и показать, что оптимальные метрики сходства строк могут сами по себе производить выравнивания, конкурентоспособные с современными подходами в системах выравнивания онтологий. В обзоре по методам сходства текстов (Ristoski и др., 2016) обсуждается несколько подходов для поиска сходства слов, таких как лексическое сходство, семантическое сходство и так далее. Сходство на основе знаний (Cheatham и др., 2013; Qian и др., 2004) – это сходство на основе семантики, которое определяет степень сходства между словами, используя информацию, полученную из семантических сетей. Популярные семантические сети – это «Word Net», а также Natural Language Toolkit (NLTK) для измерения сходства между словами на основе знаний. Сходство на основе знаний также обеспечивает сходство на основе родства слов.

Термин онтология имеет долгую историю в философии, в которой он относится к предмету существования. В контексте управления знаниями онтологией называют общее понимание некоторых областей, которое часто представляется как набор сущностей, отношений, функций, аксиом и экземпляров. Онтологии – это структурированные представления области человеческих знаний, которые состоят из классов, описателей характеристик сущностей в этой области и набора отношений между этими классами.

Методика и материалы. Мы представляем коллекцию из более чем тысячи эталонного набора данных, которые необходимы для преодоления трудности в создании больших графов знаний путем использования сходства сущностей. Эти наборы данных включают данные из собранных нами криминальных газетных новостей, и исследуют сходство, рассчитанное на основе последовательностей данных и их семантических взаимодействий. Наборы данных имеют разный размер и охватывают как минимум три различных вида на разных уровнях полноты аннотации. Для каждого набора данных мы

также делаем расчеты семантического сходства с использованием самых современных репрезентативных мер.

Электронные газеты все чаще читаются пользователями из любого места и в любое время. Газеты являются источником достоверной и своевременной информации. Например, газетные статьи содержат информацию о преступлениях, несчастных случаях, политике, культурных и спортивных событиях. Несмотря на то, что ценная информация доступна в человеко-читаемой форме в газетах и архивах, но программных систем, которые могут извлекать соответствующую информацию и представлять ее, мало, и это представляет значительный интерес для исследователей в области извлечения информации. Таким образом, данный проект направлен на удовлетворение потребности в технологиях извлечения и обобщения информации путем создания концептуальной онтологии информации, собранной из онлайновых статей и социальных медиа. Извлеченная информация из газетных статей формирует базовую онтологию. Релевантная информация из социальных сетей была использована для обогащения онтологии.

Онтология газет создается путем соскабливания текстовых и графических данных из новостных статей в Интернете. Данные были предварительно обработаны и токенизированы для извлечения тегов, которые должны быть представлены в онтологии. Для получения онтологии газет были выполнены следующие шаги. Во-первых, онлайновые новостные статьи были обработаны для извлечения текстовой и визуальной информации. Создается резюме каждой статьи, которое затем подвергается дальнейшей обработке. Для этого был использован инструмент под названием Beautifulsoup, библиотека python для извлечения данных. Наш код далее выполняет автоматическое резюмирование заданной статьи. Автоматическое обобщение – это термин, который относится к извлечению сути документа с помощью программного обеспечения. Основная идея заключается в создании подмножества набора извлеченных данных, которое включает наиболее информативные предложения.

Далее, каждое предложение в резюме, извлеченное с помощью Data Scraping, рассматривается как событие в онтологии. Эти события включаются в онтологию с помощью уникального маркера, который их идентифицирует. Каждое предложение в резюме подвергается тегированию части речи (POS). Для POS-тегирования мы используем инструментарий естественного языка (NLTK). NLTK – это python-фреймворк, используемый для реализации обработки естественного

языка в программах на Python. Он предоставляет простые в использовании интерфейсы к более чем 50 корпорациям и лексическим ресурсам, таким как WordNet, а также набор библиотек обработки текста для классификации, токенизации, стеблирования, тегирования, синтаксического анализа и семантических рассуждений, а также обертки для промышленных библиотек NLP.

В конце мы также распознаем меру сходства изображения с места события с событием. Если в статье нет подписи к какому-либо изображению, заголовок статьи обрабатывается так же, как и подписи. Сходство рассчитывается с помощью синсета (набора синонимов) в WordNet, который дает оценку, основанную на семантическом сходстве различных слов в подписи к изображению и предложении. Все сущности связаны со своим типом сущности. Вот как формируется онтология путем извлечения текста и изображений. Объединенная онтология состоит из данных газет и социальных сетей с событиями, относящимися к любому из событий онтологии газет. Мы решаем, является ли конкретное событие онтологии социальных сетей релевантным или нет, используя методологию, аналогичную предложенной в исследовательской работе по выравниванию онтологий. Мы сравниваем каждый кортеж, полученный после выполнения POS-тегирования на соскобленных данных из социальных сетей, с кортежами из газетных статей для вычисления балла сходства. Для этого каждому типу сущностей были присвоены веса.

$$\{w_{relation} = 0.1;\ w_{person} = 0.25;\ w_{location} = 0.1;\ w_{organisation} = 0.25;\ w_{image} = 0.3\}$$

Пусть w_i представляет собой вес сущности i, а Sim_i ее функцию сходства. Мы вычисляем общий балл сходства для конкретного события онтологии социальных сетей как

$$Sim_t = \sum Sim_i * w_i \qquad (1)$$

который суммируется по всем сущностям, связанным с этим событием, и где $\sum w_i = 1$.

Мы сравниваем Sim_t с пороговым значением. При Sim_t > threshold, событие должно быть добавлено в объединенную онтологию. Порог 0,5 был установлен методом проб и ошибок.

Функции сходства для каждого из атрибутов зависят от типа атрибута. Атрибуты person, location и organization должны иметь сходство строк

(поскольку они являются существительными), в то время как relation должен быть схож семантически (поскольку это глагол). Сходство изображений рассчитывается с помощью сопоставления признаков. Это бинарные функции, т.е. они возвращают 1, если атрибуты совпадают, и 0 – в противном случае.

Мы вычислим сходство строк с помощью косинусного сходства. Косинусное сходство – это мера сходства между двумя векторами пространства внутренних произведений, которая измеряет косинус угла между ними. Косинус угла в 0 градусов равен 1, а для любого другого угла он меньше 1. Если значение больше 0,7 мы заключаем, что строки похожи или одинаковы, и, следовательно, функция сходства возвращает 1.

Мы рассчитываем семантическое сходство двух атрибутов, используя алгоритм Wu&Palmer (Wu и др., 1994). Алгоритм Wu&Palmer рассчитывает родство, учитывая глубину двух синсетов в таксономии WordNet, а также глубину LCS (Least Common Subsumer), используя формулу,

$$\text{Similarity\_score} = (2*\text{Depth(LCS)})/(\text{depth(s1)} + \text{depth(s2)}) \quad (2)$$

Это означает, что 0 < балл сходства <= 1. Оценка никогда не может быть нулевой, потому что глубина LCS никогда не бывает нулевой (глубина корня таксономии равна единице). Оценка равна единице, если два входных понятия одинаковы. Если балл сходства >= 0,6, то слова семантически совпадают.

Общая структура онтологии остается такой же, как и у онтологии газеты. Объединенная онтология содержит все экземпляры онтологии газеты, а также новые атрибуты, добавленные из сопоставленных экземпляров онтологии социальных медиа.

Результаты. Онтологии могут быть использованы для описания реальных объектов посредством процесса семантического аннотирования: объекты связываются с классами онтологии, наиболее подходящими для их описания. Набор сущностей, аннотированных с помощью данной онтологии, составляет граф знаний (Ristoski и др., 2016). Имея такое структурированное представление реальности, можно вычислительно рассуждать над сущностями, упрощая процесс, который был бы гораздо более дорогим и трудоемким, если бы его выполнял человек.

Одной из задач, которая стала возможной благодаря разработке онтологии, является расчет семантического сходства между сущнос-

тями. Мера семантического сходства – это функция, которая, задавая два класса онтологии или два набора классов, описывающих двух людей, возвращает числовое значение, отражающее близость смысла между ними (Seco и др., 2004). На рисунке 1 показано, как преступления представлены их классами и как их можно сравнивать с помощью мер семантического сходства. Точная оценка сходства между парой сущностей зависит от того, насколько хорошо они аннотированы, как в отношении широты (т.е. включения аннотаций для всех аспектов сущности, которые могут быть описаны в области онтологии), так и глубины аннотаций (т.е. выбора наиболее специфических классов онтологии, которые лучше всего описывают сущность) (Минский, 1979; Чень и др., 1983).

Подходы, используемые для количественной оценки семантического сходства, можно различать в зависимости от того, какие сущности они собираются сравнивать: существуют подходы для сравнения двух классов в онтологии и подходы для сравнения двух индивидуумов, каждый из которых связан со своим собственным набором классов. При сравнении классов эти меры могут быть узловыми, то есть изучающими свойства каждого класса, или краевыми, основанными на расстоянии между классами. Однако меры на основе ребер основаны на предположении, что узлы и ребра равномерно распределены по онтологии, что в основном неверно для криминальных онтологий, что делает меры на основе узлов более надежными.

Для расчета семантического сходства для двух сущностей, каждый из которых описывается набором классов, могут использоваться как парные, так и групповые подходы. Парные подходы оценивают сходство между двумя сущностями путем объединения семантического сходства между их аннотируемыми классами. Групповые подходы используют векторные или графовые меры, которые обрабатывают аннотации, взятые вместе как единое целое. Для создания эталонных наборов данных используется одна мера семантического сходства, характерная для каждого подхода.

Matching average «MA» – парный подход, основанный на парной мере, в которой сходство между двумя классами соответствует их среднему значению схожести характеристик. В MA для эвристического расчета парного сходства рассматривается только класс с наилучшим соответствием для каждого класса в каждом наборе классов, описывающих сущностей (т.е. наиболее похожий), который нами задается следующим образом

$$MA(A,B) = \frac{\sum_{c_1 \in C_A} \max sim(c_1,c_2)}{2|C_A|} + \frac{\sum_{c_2 \in C_B} \max sim(c_1,c_2)}{2|C_B|} \quad (3)$$

где уравнение A и уравнение B – сущности, уравнение C – набор классов уравнения, которыми описывается каждая сущность, а $sim(c_1, c_2)$ – наибольшие значение сходства, найденные для уравнения класса c_1, c_2.

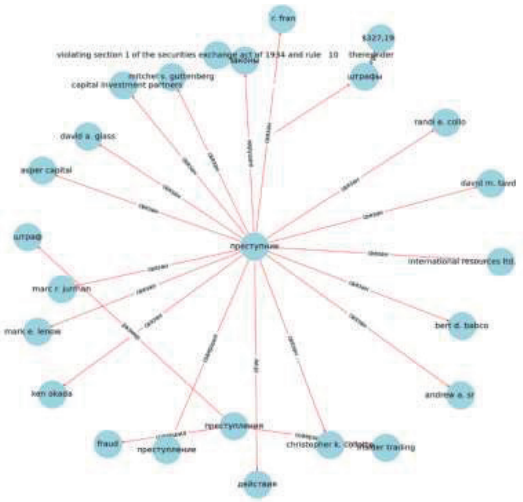

Рисунок 1 – Представление связей в онтологии, разбитые по классам и сущностям

Для создания этих эталонных наборов данных мы разработали общую методологию, разделенную на три этапа. Первый шаг состоит в выборе сущностей графа знаний, которые будут составлять пары в наборах данных. Эти сущности должны быть хорошо охарактеризованы в контексте онтологии, чтобы избежать предвзятости поверхностного аннотирования и иметь достаточно информации для вычисления сходства между ними. Следующий шаг – генерация пар сущностей. При этом мы старались гарантировать широкий диапазон сходства между парами сущностей, от нуля до полной идентичности, чтобы обеспечить репрезентативность пар сущностей. Наконец, отобрав пары сущностей для набора данных, необходимо рассчитать два типа мер сходства: семантическое сходство и приближенные показатели сходства, соответствующие типу сущностей и набору данных (Gruber, 2008).

После классификации методологий мы приступили к извлечению соответствующей информации из этих судебных разбирательств, такой как нарушения, нарушители, меры, принятые в отношении этих лиц, а

также наложенный штраф. Эти данные были сохранены в табличном формате. Вышеуказанные данные также были подготовлены для преобразования в граф знаний, который является вложенным по своей природе.

Для классификации преступлений было реализовано множество алгоритмов, позволяющих точно их идентифицировать. Были выбраны 4 основных класса, в которые были отнесены все документы, это – инсайдерская торговля, незаконное присвоение средств, незарегистрированные брокеры и мошенничество. Многие преступления относятся к нескольким классам, поэтому они были классифицированы соответствующим образом. Были опробованы три подхода. Первый подход заключался в использовании модели без наблюдения для классификации судебных релизов по различным классам. К сожалению, классификация была очень нестабильной и неточной. Кроме того, релизы не могли быть точно отнесены к нескольким классам. Этот подход был отброшен, и мы решили использовать контролируемую модель. Следующий подход заключался в использовании модели классификации текста под наблюдением для классификации релизов о судебных разбирательствах по различным категориям. Было замечено, что модель BERT не смогла правильно классифицировать документы по нужным классам. Модуль классификации текста BERT показал низкие результаты и имел точность 45,89%. Затем было решено увеличить размер обучающих данных, но точность снова осталась прежней. Причиной этого могло быть либо то, что обучающее и тестовое множество были недостаточно большими, либо то, что темы, по которым мы пытались классифицировать, были тесно связаны между собой, и различить их было сложно при недостаточном количестве данных. В конце концов мы решили прочесть достаточно значимое количество документов и выявили определенные закономерности в релизах судебных разбирательств. Это привело нас к использованию регулярных выражений (Kartbayev, 2016). Это был самый надежный подход, который оказался чрезвычайно точным. Релизы были успешно классифицированы на несколько классов с точностью 95% на том же наборе данных, который использовался для модели классификации тем BERT. В результате мы решили использовать regex-парсер для классификации релизов судебных разбирательств на различные преступления.

Для построения графа знаний мы попробовали несколько подходов к определению релевантных сущностей в корпусе. Наш первоначальный

подход состоял в том, чтобы определить субъект и объект документа и найти связь между ними (либо предикат, либо глагольное слово). Было замечено, что такое извлечение было не очень точным, и полученные результаты были неудовлетворительными. Это заставило нас улучшить алгоритм извлечения, и мы решили работать над извлечением SVO (Kartbayev и др., 2018; Kartbayev, 2015). Этот подход подразумевает идентификацию триплетных фраз. Мы определяли субъект и объект фразы, а связь между ними представляла собой глагольную фразу. Этот алгоритм извлечения показал себя значительно лучше, чем наш первоначальный подход. Однако в некоторых случаях связь между сущностями документа была потеряна. Этот недостаток побудил нас использовать концепцию онтологий для представления графа знаний. Мы решили, что поскольку эти документы имеют много сходств между собой, мы можем построить правила онтологии из имеющихся релизов судебных разбирательств. Наша онтология криминальных новостей на данный момент имеет 5 основных классов – Нарушитель, Нарушение, Преступление, Принятые меры, Штраф и Дата. Между различными классами были установлены связи, и это побудило нас использовать вложенную структуру графа знаний вместо стандартных триплетных связей. Затем мы классифицировали классы онтологии.

Полные наборы эталонных данных являются ключом к поиску наиболее эффективных инструментов для конкретного приложения. Существует ряд требований к хорошим эталонным наборам данных, а именно: актуальность, репрезентативность, отсутствие избыточности, масштабируемость и возможность повторного использования. В контексте этих эталонных наборов данных, мера семантического сходства означает, что наборы данных должны включать данные, релевантные для исследуемой области, иметь репрезентативные случаи как с точки зрения метрик сходства, так и их значений, или содержать как положительные, так и отрицательные примеры, чтобы сделать сравнительное исследование между ними более релевантным, должны поддерживать одно и то же исследование в наборах данных разного размера и имеют особую ценность, если они могут быть использованы для разных целей. Репрезентативность имеет особое значение для этих наборов данных, поскольку наборы данных должны обеспечивать сбалансированный срез криминальных сущностей. Коллекция из эталонного набора данных, которую мы представляем, направлена на поддержку крупномасштабной оценки мер семантического сходства на основе сходства графов. Она представляет собой эволюцию по

сравнению с предыдущими усилиями в этой области, как с точки зрения размера, так и разнообразия используемых данных (Akhmetov и др., 2022).

Обсуждение. Большой проблемой при оценке мер семантического сходства является разнообразие исследований, используемых для этого. Меры семантического сходства обычно тестируются на небольшом и контролируемом наборе данных, разработанном только для этого исследования. Такая несистематическая практика оценки может привести к смещению опубликованных результатов, особенно если не сравнивать их с результатами современных мер сходства в тех же условиях, т.е. с использованием точно такой же версии графа и тех же пар сущностей. Более того, отсутствие единой стратегии или, по крайней мере, одинаковых данных, делает результаты этих исследований несопоставимыми между собой.

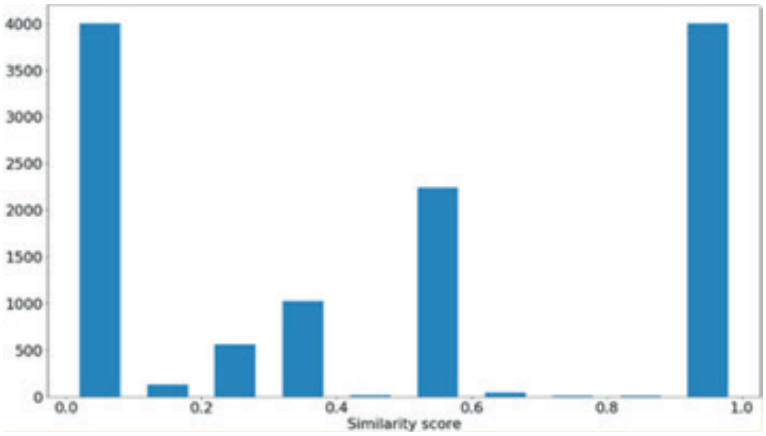

Рисунок 2 – Измерение сходства различных классов данных

Данная работа направлена на решение этих проблем путем предоставления наборов данных с парами сущностей разных видов, аннотированных различными онтологиями и обеспечивающих комбинацию различных проксисходства и нескольких современных мер семантического сходства. Чтобы гарантировать, что меры семантического сходства могут отразить функциональное сходство между сущностями, их значение должно быть хорошо передано в контексте онтологии (Рисунок 2). Это означало выбор сущностей, аннотированных более конкретными классами онтологии (классы с меньшим количеством дочерних классов), поскольку совместное использование одного или нескольких таких классов приведет к более

высокому и значительному семантическому сходству между двумя сущностями. Это было сделано для того, чтобы решить проблему неглубокого аннотирования для мер семантического сходства, в результате чего значения сходства не соответствуют человеческому восприятию из-за неглубоко описанных сущностей.

Кроме того, выбранные наборы данных соответствуют рекомендациям по качеству эталонных наборов данных, а именно: релевантность, репрезентативность, масштабируемость и возможность повторного использования. Хотя, мы предполагаем, что эталонные наборы данных должны быть нередуцируемыми, дублирование наборов данных одного и того же вида, но с разным уровнем заполнения аннотации, может быть использовано для оценки влияния более подробного описания преступлений на производительность мер семантического сходства. Наборы данных на каждом уровне полноты аннотации представляют собой компиляцию всех сущностей в каждом из классов данных.

Репрезентативность была особенно важной характеристикой при разработке этих наборов данных, например, при выборе наказаний, поскольку оценка мер семантического сходства должна проводиться как в сходных, так и в несходных парах сущностей. Кроме того, если эти наборы данных будут использоваться для приложений контролируемого обучения, то эти предикторы выиграют от обучения на более общем наборе данных. Если случаи, используемые для обучения, особенно предвзято относятся к одному признаку, производительность предиктора также будет предвзятой.

Разнообразие в структуре графа и мерах сходства, выбранных для построения этих наборов данных, позволяет предположить, что тестирование одного и того же семантического сходства в различных целевых наборах данных может быть хорошим показателем его способности к обобщению на различные графы, типы сущностей и их применения.

Заключение. В проекте реализован метод извлечения информации из новостных статей и страниц социальных сетей в Интернете и представления ее в интерактивной форме. Мы начинаем с введения различных важных понятий, связанных с извлечением информации и созданием онтологии. Затем мы предлагаем метод для достижения нашей цели. Основа нашего подхода состоит из сбора данных для анализа, затем извлечение сущностей и связей, и, наконец, визуализация информации. Каждый из этих основных этапов включает в себя множество шагов, которые мы объясняем в соответствующих разделах.

Использование данной онтологии было бы чрезвычайно полезно для правоохранительных органов и спецслужб для обнаружения и предотвращения радикализации, гражданских беспорядков и других антисоциальных действий в Интернете. Таким образом, в этом проекте мы обрабатываем множество информации из криминальных новостей, доступные в Интернете, и предлагаем метод разработки онтологии сущностей, и событий, которые связывают эти сущности, обеспечивая тем самым агрегированный обзор информации, представленной в многочисленных источниках.

Благодарность. *Работа выполнена при финансовой поддержке Комитета науки Министерства образования и науки Республики Казахстан (№AP09259309).*

Information about authors:

Kartbayev Amandyk Zhankozhauly – PhD, Institute of Information and Computational Technologies, Almaty, Kazakhstan, a.kartbayev@gmail.com; https://orcid.org/0000-0003-0592-5865;

Ybytayeva Galiya Seitkaliyevna – PhD student, specialty «Management information systems», Satbayev University, Almaty, Kazakhstan, ybytayeva.galiya@gmail.com; https://orcid.org/0000-0002-4243-0928;

Mamyrbayev Orken Zhumazhanovich – PhD, Institute of Information and Computational Technologies, Almaty, Kazakhstan, morkenj@mail.ru; https://orcid.org/0000-0002-8627-1949;

Mukhsina Kuralay Zhenisbekovna – PhD, Institute of Information and Computational Technologies, Almaty, Kazakhstan, kuka_ai@mail.ru; https://orcid.org/0000-0002-8627-1949;

Zhumazhanov Bagashar Zhumazhanovich – candidate of technical sciences, Institute of Information and Computational Technologies, Almaty, Kazakhstan, bagasharj@mail.ru; https://orcid.org/0000-0002-5035-9076.

ЛИТЕРАТУРА:

Akhmetov I., Gelbukh A., Mussabayev R. (2022) Topic-Aware Sentiment Analysis of News Articles. Computacion y Sistemas. – PP. 423-439. (in Eng.).

Euzenat J., Valtchev P. (2004) Similarity-based ontology alignment in OWL-Lite, Proc. 16th European conference on artificial intelligence (ECAI), Valencia, Spain. IOS press, – PP. 333-337. (in Eng.).

Gang Qian, Shamik Sural, Yuelong Gu, Sakti Pramanik. (2004) Similarity between euclidean and cosine angle distance for nearest neighbor queries, Proceedings of ACM Symposium on Applied Computing. – PP. 48-61. (in Eng.).

Gruber T. (1995) Toward principles for the design of ontologies used for knowledge sharing? Human-Computing Studies. – PP. 35-43. (in Eng.).

Gruber T. (2008) Collective knowledge systems: where the Social Web meets the Semantic Web. Journal of Web Semantics. – PP. 4-13. (in Eng.).

Kartbayev A. (2015) (Refining Kazakh Word Alignment Using Simulation Modeling Methods for Statistical Machine Translation. Lecture Notes in Computer Science, Springer. – PP. 421-427. (in Eng.).

Kartbayev A. (2016) Using Kazakh Morphology Information to Improve Word Alignment for SMT. Advances in Intelligent Systems and Computing, Springer. – PP. 351-359. (in Eng.).

Kartbayev A., Tukeyev U., Sheremeteva S., Kalizhanova A., Kalybek Uuly B. (2018) Experimental study of neural network-based Word alignment selection model trained with Fourier descriptors. Journal of Theoretical and Applied Information Technology. – PP. 4103-4113. (in Eng.).

Michelle Cheatham, Pascal Hitzler, Alani H. et al. (Eds.). (2013) String Similarity Metrics for Ontology Alignment, ISWC 2013. – PP. 263-285. (in Eng.).

Ristoski P., Paulheim H. (2016) Rdf2Vec: RDF graph embeddings for data mining. In: Groth P., Simperl E., Gray A., Sabou M., Krötzsch M., Lecue F., Flöck F., Gil Y., editors. The Semantic Web – ISWC 2016. Cham: Springer. – PP. 49-55. (in Eng.).

Ristoski P., Paulheim H. (2016) Semantic Web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics. – Vol. 36, PP. 12-22. (in Eng.).

Seco N., Veale T., Hayes J. (2004) An intrinsic information content metric for semantic similarity in WordNet. Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'04. Amsterdam: IOS Press. – PP. 20-34. (in Eng.).

Wu Z. and Palmer M. (1994) Verb semantics and lexical selection. In Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics. – PP. 133-138. (in Eng.).

Минский М. (1979) Фреймы для представления знаний. – М.: Энергия. – 151 с.

Чень Ч. (1983) Математическая логика и автоматическое доказательство теорем, под ред. С.Ю. Маслова. – М.: Наука. – 360 с.

REFERENCES:

Akhmetov I., Gelbukh A., Mussabayev R. (2022) Topic-Aware Sentiment Analysis of News Articles. Computacion y Sistemas. – PP. 423-439. (in Eng.).

Chen Ch. (1983) Mathematical logic and automatic theorem proving, ed. by S.Y. Maslov. - Moscow: Nauka,. - 360 p. (in Rus).

Euzenat J., Valtchev P. (2004) Similarity-based ontology alignment in OWL-Lite, Proc. 16th European conference on artificial intelligence (ECAI), Valencia, Spain. IOS press, – PP. 333-337. (in Eng.).

Gang Qian, Shamik Sural, Yuelong Gu, Sakti Pramanik. (2004) Similarity between euclidean and cosine angle distance for nearest neighbor queries, Proceedings of ACM Symposium on Applied Computing. – PP. 48-61. (in Eng.).

Gruber T. (1995) Toward principles for the design of ontologies used for knowledge sharing? Human-Computing Studies. – PP. 35-43. (in Eng.).

Gruber T. (2008) Collective knowledge systems: where the Social Web meets the Semantic Web. Journal of Web Semantics. – PP. 4-13. (in Eng.).

Kartbayev A. (2015) (Refining Kazakh Word Alignment Using Simulation Modeling Methods for Statistical Machine Translation. Lecture Notes in Computer Science, Springer. – PP. 421-427. (in Eng.).

Kartbayev A. (2016) Using Kazakh Morphology Information to Improve Word Alignment for SMT. Advances in Intelligent Systems and Computing, Springer. – PP. 351-359. (in Eng.).

Kartbayev A., Tukeyev U., Sheremeteva S., Kalizhanova A., Kalybek Uuly B. (2018) Experimental study of neural network-based Word alignment selection model trained with Fourier descriptors. Journal of Theoretical and Applied Information Technology. – PP. 4103-4113. (in Eng.).

Michelle Cheatham, Pascal Hitzler, Alani H. et al. (Eds.). (2013) String Similarity Metrics for Ontology Alignment, ISWC 2013. – PP. 263-285. (in Eng.).

Minsky M. (1979) Frames for knowledge representation. - M.: Energia. - 151 p. (in Rus).

Ristoski P., Paulheim H. (2016) Rdf2Vec: RDF graph embeddings for data mining. In: Groth P., Simperl E., Gray A., Sabou M., Krötzsch M., Lecue F., Flöck F., Gil Y., editors. The Semantic Web – ISWC 2016. Cham: Springer. – PP. 49-55. (in Eng.).

Ristoski P., Paulheim H. (2016) Semantic Web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics. – Vol. 36, PP. 12-22. (in Eng.).

Seco N., Veale T., Hayes J. (2004) An intrinsic information content metric for semantic similarity in WordNet. Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'04. Amsterdam: IOS Press. – PP. 20-34. (in Eng.).

Wu Z. and Palmer M. (1994) Verb semantics and lexical selection. In Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics. – PP. 133-138. (in Eng.).

МАЗМҰНЫ

СОДЕРЖАНИЕ

CONTENTS

Publication Ethics and Publication Malpracticein
the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethicalguidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any otherlanguage, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Директор отдела издания научных журналов НАН РК *А. Ботанқызы*
Заместитель директор отдела издания научных журналов НАН РК *Р. Жәлиқызы*
Редакторы: *М.С. Ахметова, Д.С. Аленов*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 15.09.2022.
Формат 60х881/8. Бумага офсетная. Печать – ризограф.
17,5 п.л. Тираж 300. Заказ 3.

Национальная академия наук РК
050010, Алматы, ул. Шевченко, 28, т. 272-13-19