УДК 004.8

К.Ю. Тітов

ХНУРЕ, м. Харків, Україна, kyrylo.titov@nure.ua

ПЕРСОНАЛЬНИЙ НОВИННИЙ ФІЛЬТР, СТВОРЕНИЙ НА ОСНОВІ ТЕХНОЛОГІЙ КОГНІТИВНИХ ОБЧИСЛЕНЬ І СЕМАНТИЧНОЇ ОБРОБКИ ІНФОРМАЦІЇ

У даній статті були досліджені проблеми обробки багатоаспектної новинної інформації із різних джерел, а також методи фільтрації отриманої інформацій. Також були розглянуті методи штучного інтелекту, за допомогою яких можливо вирішити дані проблеми. В статті викладений один із варіантів вирішення проблеми фільтрування новинної інформації за допомогою семантичного порталу та інтелектуальних сервісів від компанії IBM Watson. Метою даної статті є дослідження методів семантичного аналізу новинного контенту з різних джерел на основі перетворення інформації в онтологічну форму подання, а також розробка методу отримання та обробки новинної інформації.

SEMANTIC WEB, OWL, RDFS, XML, IBM WATSON, ФІЛЬТРАЦІЯ, НОВИНИ, ОБ'ЄКТИВНІСТЬ, НЕУПЕРЕДЖЕНІСТЬ

Титов К.Ю. Персональный новостной фильтр, созданный на основе технологий когнитивных вычислений и семантической обработки информации. В данной статье были исследованы проблемы обработки многоаспектной новостной информации из разных источников, а также методы фильтрации полученной информаций. Также были рассмотрены методы искусственного интеллекта, с помощью которых можно решить данные проблемы. В статье изложен один из вариантов решения проблемы фильтрации новостной информации с помощью семантического портала и интеллектуальных сервисов от компании IBM Watson. Целью данной статьи является исследование методов семантического анализа новостного контента с различных источников на основе преобразования информации в онтологическую форму представления, а также разработка метода получения и обработки новостной информации.

SEMANTIC WEB, OWL, RDFS, XML, IBM WATSON, ФИЛЬТРАЦИЯ, НОВОСТИ, ОБЪЕКТИВНОСТЬ, БЕСПРИСТРАСТНОСТЬ

Titov K.Yu. Personal news filtering based on cognitive computing and semantic technologies. In this article the problems of processing multidimensional news information from different sources and methods of filtering the received information were investigated. Also, the methods of artificial intelligence which help to solve these problems were considered. The article outlines one of the approach to resolve the problem of filtering the news information using the semantic portal and intelligent services from IBM Watson. The purpose of this article is to investigate the methods of semantic analysis of the news content from various sources based on the transforming the information to the ontological form of representation as well as to develop a method for solving the problem of the obtaining and processing the news information.

SEMANTIC WEB, OWL, RDFS, XML, IBM WATSON, FILTERING, NEWS, OBJECTIVITY (UNBIASED)

Вступ

У сучасному світі користувачі можуть отримувати новини із багатьох різноманітних джерел, як класичних - газети, журнали тощо, так і сучасних таких як електронні видання, новинні сайти або портали. Але зараз інформаційні потоки дуже відрізняються від колишніх, а саме об'ємами інформації, що надходить. Тисячи журналістів та репортерів збирають та обробляють інформацію по всьому світу, але усі вони різні. Деяки сумлінно виконують свою журналістську діяльність, деякі заробляють цим гроші, а деякі можуть бути частиною пропагандистської машини тієї чи іншої країни, організації тощо, тим саме впливаючи на соціальні, політичні та економічні аспекти життя. Саме тому сьогодні витрачаються багато коштів за для пошуку варіантів вирішення питання надання користувачам доступу до об'єктивної інформації, зменшення впливу пропаганди всередині країні та з боку інших країн.

Зараз читачі не мають можливості отримувати новини за персональними вподобаннями, рекомендації по ключовим словам не беруться до уваги, але деякі компанії, які займаются дослідженнями та розробкою систем та сервісів на основі штучного інтелекту, можуть предоставляти сервіси, за допомогою яких ми зможемо отримувати семантичний опис статті або автора. За допомогою цієї інформації користувач зможе отримувати найбільш цікаві для себе новини.

Обробка, аналіз та фільтрація таких величезних масивів новинних даних за допомогою методів штучного інтелекту є актуальною темою для сучасного світу.

1. Методи та технології отримання та обробки новинної інформації

Найперші фільтруючі системи обговорювалися Гансом Луном у корпорації IBM у 1958 році. Було визнано що в майбутньому значно збільшаться об'єми обміну електронною інформацією. В той час вчені запропонували систему бізнес-аналізу для направлення відповідної інформації користувачам та групам у промислових, наукових та

державних організаціях. Проте основна відмінність між інформаційно-пошуковою системою та системою фільтрування полягала в автоматичному поширенні інформації, що є однією із головних їх відмінностей. Система складалася з механічного створення профілів інтересів для кожної окремої людини або групи користувачів, а IR-техніка відігравала важливу роль у виявленні нової інформації, яку слід було розповсюджувати.

Проблема фільтрування була популярною протягом 60-х та 70-х років як вибіркове розповсюдження інформації. Підтримуючи сучасну освіту в наукових та технологічних сферах такі компанії як NASA і GM інвестували в дослідження та розробку систем, які розповсюджують відповідну інформацію своїм працівникам. Основна мета вибіркового розповсюдження інформації полягала в тому, щоб забезпечити регулярне сповіщення працівників, коли надходили нові документи, які вважалися цікавими для користувачів, саме там де тематика інтересів була чітко визначена користувачем.

В той час як системи фільтрування інформації були розроблені для електронних листів, зокрема для фільтрування електронної пошти через спам, з'явилася нова проблема пов'язана з перевантаженням інформації, пов'язана з впровадженням концепції web 2.0, а саме соціальних мереж, форумів тощо. Такі платформи як Twitter, Facebook та Google, генерують величезну кількість різноманітного контенту щодня.

Сучасні методи фільтрації складаються з наступних модулів:

1. Моделювання інтересів користувачів: тут система визначає інтереси користувача;

2. Аналізатор даних: на цьому етапі система аналізує вхідні дані і представляє їх відповідно до вимоги модуля фільтрації;

3. Фільтрація: система порівнює подання вхідних даних і профіль користувача, що представляє інтерес для визначення релевантності отриманої інформації;

4. Навчання та модифікація: навчання та зміна інтересів користувача з плином часу.

Однією із форм представлення новин є новинні сайти, вони розділяються на тематичні категорії та підкатегорії. Інший варіант це новинні пошукові системи, які дозволяють користувачу шукати новини за термінами що представляють інтерес для користувача. У відмінності від новинних сайтів пошукові системи дозволяють персоналізувати пошук – користувач сам може настроїти свій профіль, вибираючи теми інтересів тим саме підказуючи системі в яких тематичних категоріях слід шукати інформацію. Іноді персоналізація виконується за допомогою методів сумісної фільтрації таки як наприклад google-новини. Така інтерекативність є важливою додатковою функцією для залучення нових користувачів та поліпшення якості надання новинного контенту.

Сучасні рекомендаційні системи мають низку недоліків, однією із найважливіших є відсутність адекватної моделі данних. Профіль користувача може бути побудований за його явними та неявними уподобаннями або на основі їх комбінації. Є три основні методи побудови рекомендацій для користувача:

1. На основі контенту – система робить висновки на основі статей, які сподобались користувачу у минулому;

2. Колоборантний метод фокусується не на самому предметі пошуку, а рекомендує єлементи, які сподобались користувачам із подібними вподобаннями;

3. На основі знань, використовують знання не тільки о вподобаннях користувача але й знання о предметах пошуку.

Як тільки утворюється профіль система застосовує одну або декілька методик для створення рекомендацій для користувача. Усі ці методи можуть бути поєднані за для отримання найбільш релевантного результату.

В Голландському університеті Ротердама в межах курсу «Advanced Software Architecture» розробили новинний портал який аналізував rss-стрічки, та конвертуючи інформацію у онтологію надавав користувачам новини згідно із їхніми вподобаннями. Система за для рекомендації новин спочатку моделювала поведінку користувача, аналізуючи та запам'ятовуючи історію перегляду новиних сайтів та rss-стрічок. Далі вона представляла отриману інформацію у онтологічному вигляді. Засновуючись на змодельованому профілі система рекомендувала новинний контент, який був би цікавим для користувача. Основною метою розробки даної системи було виявлення найбільш оптимальної технології для розробки рекомендаційних новинних систем. Згідно із проведеними дослідженнями було виявлено, що системи засновані на онтологіях працюють краще ніж їх аналоги які не використовують онтологічний підхід.

У цій статті ми рекомендуємо застосувати онтологію для опису предметної області «новини» та побудови фільтрів користувача на основі цієї онтології. Онтологія дозволяє зменшити простір пошуку, а також полегшити процес отримання новин за вподобаннями користувача. Однією з важливих переваг онтологічного підходу є те, що він не використовує колоборантну фільтрацію, він застосовує знання о предметах и вподобаннях користувача описаних в онтології. Для опису властивостей новинної статті або її автора у створеної онтології ми використовуємо показники інтелектуального сервісу IBM Watson – Personality Insights [1].

2. Опис інтелектуального сервісу personality insights від IBM Watson

IBM – є однією з найбільш розвинутих корпорацій, яка вже понад сто років очолює технологічний прогрес. Одним з найцікавіших та відомих проектів останніх років став IBM Watson.

Це когнітивна система, яка може навчатися, розуміти та робити висновки. В рамках цього проекту

була розроблена низка інтелектуальних сервісів серед яких є Personality Insights.

Служба IBM Watson Personality Insights представляє собою зручне API, яке дає представлення о характеристиках особистості з соціальних мереж, корпоративних даних або інших цифрових джерел. Сервіс використовує лінгвістичну аналітику для визначення характеристик особистості. Також він може визначати вподобання користувача, для аналітики маркетингу, впливу реклами тощо. Служба дозволяє проводити аналітику великим компаніям для більш повного та глибоко розуміння своїх клієнтів незважаючи на галузь у який вона працює. Завдяки цьому корпорації можуть поліпшувати якість обслуговування та надання послуг, адаптувати свої продукти під цільову аудиторію

IBM провело низку досліджень для того щоб зрозуміти чи можуть характеристики особистості, отримані із соціальних мереж передбачити поведінку та вподобання людей. Отримані результати показали що людина яка схильна до збудження з більшою ймовірність відгукнеться на нові маркетингові кроки, так само як люди із високими показниками таких якостей як скромність, відвертість та дружелюбність із великою ймовірністю будуть поширювати цікаву інформацію серед знайомих. Усі ці дослідження були перевірені та підтвердженні тестовими опитуваннями учасників. Після цього був проведений аналіз понад 600 твітів, який показав, що індивідуальні характеристики, отримані службою Personality Insights, можуть передбачати вподобання користувача із точністю близько 70%.

Personality Insights під силу автоматичний вивід із потенційно-зашумленних соціальних мереж портрети людей, які відображують їх індивідуальні характеристики. У 2016 році журналістка видання NPR Аарти Шахани проаналізувала за допомогую сервіса свої акаунти у соц-мережах та за її словами вона отримала дуже точну характеристику самої себе. Шахані була здивована такій точній оцінці її особистості, тому що її акаунти з twitter та facebook дуже сильно різнилися.

Сервіс Personality Insights розпізнає характеристики особистості з текстової інформації(статті або інші публікації), автором якого є особа яку ми аналізуємо. Спочатку сервіс розмічає текст для створення представлення у n-мірному просторі. Сервіс використовує технологію word-embeding, з відкритим вихідним кодом, для того щоб отримати векторне представлення для усіх слів з тексту. Далі Personality Insights передає це представлення алгоритму машинного навчання, який описує профіль особистості із його характеристиками. Для навчання алгоритму сервіс використовує оцінки, отримані з опитувань, проведених серед тисяч користувачів, а також їх акаунтів twitter.

Важливо, що під час тестів, IBM встановили, що характеристики особи, отримані з текстів, можуть точно передбачити різноманіття характерів людей та їх поведінку. Завдяки цьому ми можемо отримати характеристику на будь-якого автора новинних статей, сюжетів, досліджень, тощо [2].

3. Створення онтології предметної області

Semantic Web можливо розглядати як розширення звичайного Web, в якому документи можуть не тільки переглядатися людиною, але, і забезпечені метаописами, що забезпечує автоматичну обробку цих документів. Технології Semantic Web орієнтовані на обробку контенту на семантичному рівні. Автором цієї концепції є Т.Бернес-Лі - розробник сучасного Web. Дана концепція просувається Консорціумом W3C, який очолює Т.Бернес-Лі.

Технології Semantic Web, дозволяють формалізувати знання про предметну галузь таким чином, щоб вони могли оброблятися і людиною, і комп'ютерною системою. Для цього використовують онтології. Одним з можливих варіантів вирішення проблеми обробки великих масивів неоднорідної і розподіленої інформації новинного контенту є подання інформації в онтологічній формі і подальша обробка семантики отриманої інформації [3].

Під час досліджень я запропонув вирішувати проблему контролю якості наданого новинного контенту за допомогою веб-системи, яка здатна накопичувати, як сам новинний контент, так і мета-контент, який може вказувати на достовірність новинної інформації.

Рішення базується на використанні онтологічного порталу оцінки якості новинного контенту, який будується за прикладом вже існуючої платформи – порталу, який був розроблений в міжнародному проекті Tempus «Національна система забезпечення якості і взаємної довіри в системі вищої освіти (TRUST)» як технічний засіб підтримки і гармонізації процесів з оцінки і забезпечення якості вищої освіти.

Технологію побудови таких порталів створено для прозорого накопичення та обміну інформацією, що має забезпечити можливість широкого та неупередженого контролю за її якістю. Подібні системи посилюють соціальний вплив на оцінку інформації та унеможливлюють контроль за нею лише зацікавленою стороною. Для цього подібні системи будуються за принципами соціальних мереж. Користувачі порталів є головними постачальниками контенту та контролерами його достовірності (механізми соціальної верифікації). Однак, окрім соціальної верифікації контенту, в запропонованому порталі існують механізми і інструменти автоматичного аналізу достовірності наданої інформації.

Портал дозволяє створювати і застосовувати різні системи цінностей у вигляді гнучких багатовимірних показників якості, зважених за ступенем їх важливості для ранжування запиту. Таким чином, кожен користувач може оцінити якість деяких ресурсів з різних точок зору, так званих «систем цінностей» користувача, який робить оцінку [4].

Портал працює відповідно до інформації, що зберігається в її онтологічній базі знань. Онтології використовуються для опису самого порталу: його архітектури і функціональності, а також для відкритого і гнучкого зберігання інформації, що надається користувачами. Важливою особливістю архітектури порталу є його гнучкість, яка досягається за рахунок розподілу описів самого порталу та предметної області на дві окремі онтології:

1. Сервісна онтологія містить допоміжні класи і властивості для системної бізнес-логіки, підтримки реєстрації ресурсів, бізнес-аналітики, рейтингу і т.д. Вона спроектована для використання в якості основної незалежної структури, досить гнучкої для взаємодії з онтологіями, які описують будь-який можливий домен в системі менеджменту ресурсів підтримки моніторингу якості;

2. Доменна онтологія включає в себе:

– ядро (визначає поняття і властивості, які використовуються для оцінки якості);

м шар користувача (який кожна організація може гнучко адаптувати до власних умов або кожен користувач може адаптувати до власних вподобань);

– системи цінностей (яка визначає вагові коефіцієнти для різних показників якості в різних контекстах);

Завдяки гнучкій структурі побудови порталу за рахунок поділу на дві окремі онтології сервісну і доменну портал може бути повністю змінений шляхом простої модифікації доменної онтології. Сервісна онтологія здатна взаємодіяти із онтологіями, які описують будь-яку галузь, відмінну від вищої освіти, в якій також існують численні ресурси, які потребують оцінки (бізнес, виробництво, медицина, медіа).

Основними компонентами новинної онтології є поняття, відносини, екземпляри та аксіоми. Поняття представляють собою набір або клас сутностей у межах новинної області. Кожен клас, визначений в онтології, описує загальні характеристики індивідів. Найбільш фундаментальні поняття відповідають класам, які знаходяться в корені різних таксономічних дерев. Кожен індивід в світі OWL є членом класу owl:Thing. Таким чином, кожен певний клас автоматично є підкласом owl:Thing. Специфічні для даної області кореневі класи визначаються простим оголошенням іменованого класу. Наприклад такі класи як автор, видання, стаття тощо.

Також онтології включають у себе відношення поміж класами або властивостями. Ієрархія класів визначається шляхом вказування що клас є підкласом іншого класу, таким чином клас «публікації автора» має декілька підкласів.

Кореневим класом створеної онтології є «автор публікації», який має три підкласи – характеристики особистості автора, уподобання користувача та ціннісні характеристики. Ці підкласи є основними логічними розділами сервісу Personality Insights, які включають в себе більш детальні характеристики, для опису індивідуальності автора статті. Усі характеристики зображені на рис. 1.

Рис. 1.

4. Аналіз та фільтрація отриманої новинної інформації

Комбінуючи портал та сервісу Personality Insights, який надає зручний API для інтеграції з вашими проектами. Далі завдяки сервісу інтегрованому в портал було проаналізовано близько двадцяти новин які вже були визнані як не відповідаючи дійсності. Отримавши аналіз цих новин вдалося встановити що більшість з них мали достатньо високі показники емоційності, завдяки таким характеристикам як рівень до закликів або імпульсивність виразів у тексті. Достатньо великими були й показники відкритості але тільки та частина яка характеризує авторитарність та творчість висловлювань. Характеристики які описують потреби користувача навпаки були достатньо низьким. Такі показники як пристрасть до свободи, любові, ідеалів або гармонії не набрали навіть и 15%. Стиль написання статей був простим та доступним за для охоплення більшої маси людей. Тобто проаналізовані новинні статті мали низку схожих характеристик і мало індивідуальних відмінностей. Завдяки цьому можливо зробити висновок що технології створення неправдивих новин схожі та мають певні ознаки такі як агресія, велика кількість закликів, упередженість та висока емоційність статей. Наступним кроком було провести аналіз новин декількох новинних агентств таких як CNN BBC та RT і отримати для кожної з них певну характеристику у вигляді діаграми, одна з них зображена на рис. 2, і згодом почати фільтрування новинного контенту саме цих трьох агентств.

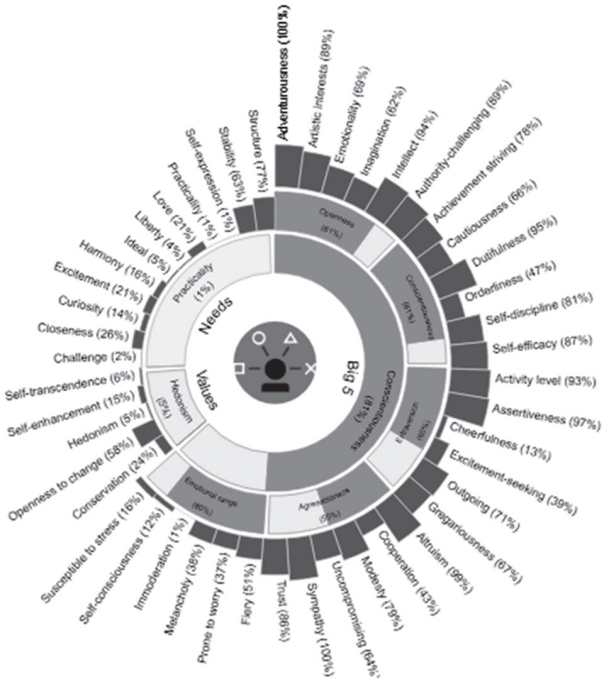

Рис. 2.

Інтегрувавши розроблену онтологію до порталу користувач отримав можливість налаштовувати власні системи цінностей за власними вподобаннями, для цього він повинен вибрати найважливіші для нього характеристики і задаємо поріг проходження. Процес налаштування зображений на рис.3.

Рис. 3.

У порталі для вищої освіти вплив статті (impact) визначається кількістю цитат і категорією журналу. Отже для аналізу публікацій (рейтингу) в ЗМІ ми надали можливість рейтингувати статті з використанням зовнішніх експертів, таких як сервіс від IBM Watson Personality Insights. Це дає можливість отримувати характеристики або ознаки новинних статей та інших публікацій такі як фальш, агресія, емоційність і т.д., з надійного джерела (IBM Watson). Потім користувач фільтрує отриману інформацію в залежності від особистих уподобань за допомогою налаштування фільтрів особистих переваг. IBM Watson - жорсткий фільтр, а портал надає можливість створити власний гнучкий фільтр завдяки персоналізованому настроюванню вагових коефіцієнтів. У свою чергу читачі можуть використовувати завдяки порталу низку фільтрів наприклад – ми хочемо читати новини від автора який має вищу освіту, максимально відкритим згідно персональних характеристик автора та який відповідає моїм ідеалам таким як – свобода та відкритість.

Висновки

У статті було розглянуто можливість заміни онтології, яка описує академічну галузь на онтологію новин на основі інтелектуального сервісу від IBM Watson. Інтегрували сервіс до порталу та провели аналіз близько шести десятків новинних статей. Розроблений сервіс надає зручний інструментарій для обробки та аналізу новинного контенту.

Застосовуючи для реалізації семантичного фільтру новинної інформації інтелектуальні технології ми маємо великі можливості по вдосконаленню сервісу. Наприклад ми можемо розширити онтологію, доповнюючи її новими поняттями. Це допоможе нам оцінювати не тільки статті або авторів, а й компанії, в яких ці журналісти працюють, їх джерела фінансування, заангажованість або кількість неправдивих або необ'єктивних новин

опублікованих цими компаніями. Також можливо підключити додаткові інтелектуальні сервіси IBM Watson такі як Tone Analyzer, завдяки цьому сервісу можливо оцінити в якому емоційному тоні написаний текст статті. Або Discovery News за для автоматизації пошуку та завантаження новин згідно із пошуковими запитами користувача. Для того щоб сервіс мав змогу шукати та аналізувати новини на різних мовах можливо інтегрувати до порталу сервіс Language Translator, який в свою чергу може бути комбінований з іншими сервісами від IBM. Підключення сервісу Visual Recognition надасть можливість аналізувати не тільки текстовий контент але й фотографії зроблені на місці подій. Використовуючи можливості порталу ми зможемо робити висновки об об'єктивності інформації за допомогою експертів. Вибір експертів може спиратися на особисті вподобання користувача – за його особистою системою цінностей. Такий сервіс може стати у нагоді в новій галузі – журналістика даних, але така журналістика включає обробку великого об'єму інформації, що не під силу звичайному журналісту.

В майбутньому, також, планується підключити новинне API або rss-стрічку для автоматизації процессу отримування та обробки новинної інформації.

Список літератури: 1. Personality Insights, documentation [Електронний ресурс]. – Режим доступу: https://console.bluemix.net/docs/services/personality-insights/getting-started.html#getting-started-tutorial **2.** The science behind the service [Електронний ресурс]. – Режим доступу: https://console.bluemix.net/docs/services/personality-insights/science.html#science **3.** H. Wache, T. Vugele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner, «Ontology-based Integration of Information - A Survey of Existing Approaches,» In: Proceedings of IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, 2001, Vol. pp. 108-117. **4.** Terziyan V., Golovianko M., Shevchenko O., Semantic Portal as a Tool for Structural Reform of the Ukrainian Educational System, In: Information Technology for Development, Vol. 21, No. 3, 2015, Taylor & Francis, pp. 381-402.

Resume

Titov K.Y.

PERSONAL NEWS FILTERING BASED ON COGNITIVE COMPUTING AND SEMANTIC TECHNOLOGIES

Background: Now the information flows are extremely different from the previous ones, especially regarding the volume of incoming information. Therefore, processing, analyzing and filtering huge data sets with using the artificial intelligence methods is an actual topic for the modern world.

Materials and methods: This article explores the problems of processing the multidimensional news information from different sources, as well as the methods of filtering the information received.

Results: The artificial intelligence methods were analised from a perspective to apply those for solving the problems of processing large data arrays of heterogeneous and distributed information of the news content. It was ascertained that for the processing of large arrays of news content it is necessary to represent the information in ontological form for more convenient and precise further processing of the semantics of the received information.

Conclusion: The article outlines one of the solutions of the problem of filtering news information using the semantic portal and intelligent services from IBM Watson.

Надійшла до редколегії 20.09.2017