УДК 004.8
DOI: 10.25140/2411-5363-2017-3(9)-138-143

Кирило Тітов

ВИКОРИСТАННЯ ТЕХНОЛОГІЙ КОГНІТИВНИХ ОБЧИСЛЕНЬ І СЕМАНТИЧНОЇ ОБРОБКИ ІНФОРМАЦІЇ ДЛЯ ФІЛЬТРАЦІЇ НОВИН ЗА ПЕРСОНАЛІЗОВАНИМИ ВПОДОБАННЯМИ КОРИСТУВАЧА

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

Постановка проблеми. *У цій статті досліджено проблеми обробки багатоаспектної, об'єктивної, новинної інформації із різних джерел, а також методи фільтрації отриманої інформації.*

Аналіз останніх досліджень і публікацій. *Останнім часом у багатьох європейських країнах виникає все більший інтерес до розробки методів боротьби з цієї проблемою. Багато європейських політиків, такі як, наприклад, федеральний уповноважений з проведення виборів Дітер Заррайтер (ФРН), стурбований таким станом речей: «Громадяни та ЗМІ повинні з особливою обережністю реагувати на новини у ході цієї передвиборчої кампанії. Слід знати, що робляться спроби ними маніпулювати».*

Виділення не вирішених раніше частин загальної проблеми. *З розвитком інтернет-технологій ми отримали доступ до різної інформації, у тому числі і новинної, але останнім часом виникла проблема отримання не просто інформації, а фактів, що мають конкретне підтвердження.*

Мета статті. *Метою цієї 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.*

Виклад основного матеріалу. *У статті викладено один із варіантів вирішення проблеми фільтрування новинної інформації за допомогою семантичного порталу та інтелектуальних сервісів від компанії IBM Watson. Ми рекомендуємо застосувати онтологію для опису предметної області «новини» та побудови фільтрів користувача на основі цієї онтології.*

Висновки і пропозиції. *Розглянуто можливість заміни онтології, яка описує академічну галузь, на онтологію новин на основі інтелектуального сервісу від IBM Watson. Розроблений сервіс надає зручний інструментарій для обробки та аналізу новинного контенту.*

Ключові слова: *Semantic WEB; OWL; RDFS; XML; IBM Watson; фільтрація; новини; об'єктивність; неупередженість. Бібл.: 4.*

Постановка проблеми. 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 нині інформаційні потоки дуже відрізняються від колишніх, а саме об'ємами інформації, що надходить. Тисячі журналістів та репортерів збирають та обробляють інформацію по всьому світу, але усі вони різні. Деякі 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 для пошуку варіантів вирішення питання надання користувачам доступу до об'єктивної інформації, зменшення впливу пропаганди всередині країні та з боку інших країн.

Сьогодні читачі не мають можливості отримувати новини за персональними вподобаннями, рекомендації за ключовими словами не беруться до уваги, але деякі компанії, які займаються дослідженнями та розробкою систем та сервісів на основі штучного інтелекту, можуть надавати 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.

Аналіз останніх досліджень і публікацій. 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 становлять інтерес для користувача. На відміну від новинних сайтів пошукові системи дозволяють персоналізувати пошук – користувач сам може налаштувати свій

профіль, вибираючи теми інтересів, тим самим підказуючи системі, у яких тематичних категоріях слід шукати інформацію. Іноді персоналізація виконується за допомогою методів сумісної фільтрації, таких як, наприклад, google-новини. Така інтерекативність є важливою додатковою функцією для залучення нових користувачів та поліпшення якості надання новинного контенту.

Виділення не вирішених раніше частин загальної проблеми. Сучасні рекомендаційні системи мають багато недоліків, одним із найважливіших є відсутність адекватної моделі даних. 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 в минулому.

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

3. На основі знань – використовують знання не тільки про вподобання користувача, але й знання про предмети пошуку.

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

У Голландії в університеті Ротердама в межах курсу «Advanced Software Architecture» розробили новинний портал, який аналізував rss-стрічки та конвертуючи інформацію в онтологію, надавав користувачам новини згідно із їхніми вподобаннями. Система для рекомендації новин спочатку моделювала поведінку користувача, аналізуючи та запам’ятовуючи історію перегляду новинних сайтів та rss-стрічок. Далі вона представляла отриману інформацію в онтологічному вигляді. Засновуючись на змодельованому профілі, система рекомендувала новинний контент, який був би цікавим для користувача. Основною метою розробки такої 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.

Мета статті. 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 про предмети і вподобання користувача, описані в онтології. Для опису властивостей новинної статті або її автора у створеної онтології ми використовуємо показники інтелектуального сервісу IBM Watson –Personality Insights [1].

Виклад основного матеріалу. 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. У межах цього проекту було розроблено багато інтелектуальних сервісів, серед яких є Personality Insights.

Служба IBM Watson Personality Insights являє собою зручне API, яке дає уявлення про характеристики особистості із 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 дослідження 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 з 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 Аарті Шахані проаналізувала за допомогою сервісу свої акаунти у соцмережах 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].

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 та подальша обробка семантики отриманої інформації [3].

Ми пропонуємо вирішувати проблему контролю якості наданого новинного контенту за допомогою веб-системи, яка здатна накопичувати як сам новинний контент, так і метаконтент, 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 у 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 верифіка-

ції). Однак, крім соціальної верифікації контенту, у запропонованому порталі існують механізми й 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 поглядів, 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. ін. Вона спроектована для використання як основної незалежної структури, досить гнучкої для взаємодії з онтологіями, які описують будь-який можливий домен у системі менеджменту ресурсів підтримки моніторингу якості.

2. Доменна онтологія включає в себе:

- ядро (визначає поняття і властивості, які використовуються для оцінки якості);
- 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);
- процеси забезпечення якості (формально визначені внутрішні або крос-організаційні процеси забезпечення якості).

Завдяки гнучкій структурі побудови порталу за рахунок поділу на дві окремі онтології сервісну і доменну портал може бути повністю змінений шляхом простої модифікації онтологій. Сервісна онтологія здатна взаємодіяти з 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 являють 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 у світі OWL є членом класу owl:Thing. Таким чином, кожен певний клас автоматично є підкласом owl:Thing. Специфічні для цієї області кореневі класи визначаються простим оголошенням іменованого класу. Наприклад, такі класи, як автор, видання, стаття тощо.

Також онтології включають у себе відношення між класами або властивостями. Ієрархія класів визначається шляхом вказування, що клас є підкласом іншого класу, отже, 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 у себе більш детальні характеристики для опису індивідуальності автора статті.

Комбінуючи портал та сервісу Personality Insights, який надає зручний API для інтеграції з вашими проектами, ми можемо аналізувати і згодом фільтрувати новинний контент. У порталі для вищої освіти вплив статті (impact) визначається кількістю цитат і категорією журналу. Отже, для аналізу публікацій (рейтингу) у 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 тощо, з надійного джерела (IBM Watson). Потім користувач фільтрує отриману інформацію залежно від особистих уподобань за допомогою налаштування фільтрів особистих переваг. IBM Watson – жорсткий фільтр, а портал – гнучкі вагові коефіцієнти. Для автора публікації головне завдання, щоб його статті мали популярність у читачів, тому вони будуть максимально зацікавлені у проходженні фільтрів сервісу Personality Insights. У свою чергу, читачі можуть використовувати завдяки порталу багато фільтрів: наприклад, ми хочемо читати новини від автора, який має вищу освіту, максимально відкритим згідно з персональними характеристиками автора та який відповідає моїм ідеалам, таким як свобода та відкритість.

Висновки і пропозиції. Ми розглянули можливість заміни онтології, яка описує академічну галузь на онтологію новин на основі інтелектуального сервісу від IBM Watson. Розроблений сервіс надає зручний інструментарій для обробки та аналізу новинного контенту.

Застосовуючи для реалізації семантичного фільтру новинної інформації інтелектуальні технології, ми маємо великі можливості для вдосконалення 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. Використовуючи можливості порталу, ми зможемо робити висновки про об'єктивність 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 обсягу інформації, що не під силу звичайному журналісту.

У майбутньому ми також плануємо підключити новинне API або rss-стрічку для автоматизації процесу отримування та обробки новинної інформації.

Список використаних джерел

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, 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.

References

1. *Personality Insights, documentation.* Retrieved from https://console.bluemix.net/docs/services/personality-insights/getting-started.html#getting-started-tutorial.

2. *The science behind the service.* Retrieved from 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, 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.

TECHNICAL SCIENCES AND TECHNOLOGIES

UDC 004.8

Kirill Titov

THE USE OF COGNITIVE COMPUTING TECHNOLOGIES AND SEMANTIC INFORMATION PROCESSING TO FILTER NEWS ACCORDING TO PERSONALIZED PREFERENCES OF THE USER

Urgency of the research. *Processing, analyzing and filtering such huge arrays of news data using artificial intelligence methods is a topical issue for the modern world.*

Target setting. *In this article we researched the problems of processing multidimensional, objective, news information from different sources and methods of filtering received information.*

Actual scientific researches and issues analysis. *Recently, in many European countries grows interest in developing methods to solve this problem. Many European politicians, such as the federal electoral commissioner Dieter Zarreiter (FRG), are concerned about the state of affairs: «Citizens and the media must be particularly careful about the news during this election campaign. It should be known that attempts are being made to manipulate them».*

Uninvestigated parts of general matters defining. *With the development of Internet technologies, we got access to various information, including the news, but recently, there was a problem of obtaining not just information but facts with concrete confirmation.*

The research objective. *The goal of this article is to study the methods of semantic analysis of news content from various sources, based on the transformation of information into an ontological presentation form, as well as the development of a method for obtaining and processing news information.*

The statement of basic materials. *The article outlines one of the solutions to the problem of filtering news information using the semantic portal and intelligent services from IBM Watson. We recommend using an ontology to describe the subject area of «news» and constructing user filters based on this ontology.*

Conclusions. *We reviewed the possibility of replacing the ontology that describes the academic branch on the ontology of news based on the intelligent service from IBM Watson. The developed service provides a convenient tool for processing and analysis of news content.*

Key words: *Semantic WEB; OWL; RDFS; XML; IBM Watson; Filtering; News; Objectivity (Unbiased).*

Bibl.: 4.

УДК 004.8

Кирилл Титов

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИЙ КОГНИТИВНЫХ ВЫЧИСЛЕНИЙ И СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ИНФОРМАЦИИ ДЛЯ ФИЛЬТРАЦИИ НОВОСТЕЙ ПО ПЕРСОНАЛЬНЫМ ПРЕДПОЧТЕНИЯМ ПОЛЬЗОВАТЕЛЕЙ

В статье исследованы проблемы обработки многоаспектной новостной информации из разных источников, а также методы фильтрации полученной информаций. Также рассмотрены методы искусственного интеллекта, с помощью которых можно решить данные проблемы. 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.

Ключевые слова: *Semantic WEB; OWL; RDFS; XML; IBM Watson; фильтрация; новости; объективность; беспристрастность.*

Библ.: 4.

Тітов Кирило Юрійович – аспірант ХНУРЕ (просп. Науки, 14, м. Харків, Харківська область, 61000, Україна).
Титов Кирилл Юрьевич – аспирант ХНУРЕ (просп. Науки, 14, г. Харьков, Харьковская область, 61000, Украина).
Titov Kirill – Phd student KNURE (14 Nauky Av., 61000 Kharkiv, Ukraine).
E-mail: kirill90titov@gmail.com

Тітов К. Використання технологій когнітивних обчислень і семантичної обробки інформації для фільтрації новин за персоналізованими вподобаннями користувача / К. Тітов // Технічні науки та технології. – 2017. – № 3 (9). – С. 138-143.