УДК 624.04

К.Ю. Тітов

# МЕТОДИ АВТОМАТИЧНОГО АНАЛІЗУ НОВИННОГО КОНТЕНТУ, НА ОСНОВІ ПЕРЕТВОРЕННЯ ІНФОРМАЦІЇ В ОНТОЛОГІЧНУ ФОРМУ ПРЕДСТАВЛЕННЯ ТА МАЙБУТНЬОЇ ОБРОБКИ ОТРИМАННОЇ СЕМАНТИЧНОЇ ІНФОРМАЦІЇ

*Анотація. У даній статті були досліджені проблеми отримання об'єктивної та багатоаспектної новинної інформації із різних джерел, а також методи обробки отриманої інформацій. Також були розглянуті методи штучного інтелекту, за допомогою яких можливо вирішити дані проблеми. В статті викладений один із варіантів вирішення проблеми отримання об'єктивної новинної інформації за допомогою семантичного порталу, шляхом заміни онтології порталу, яка описує академічну галузь на онтологію новин.*

*Ключові слова: Semantic WEB, OWL, RDFS, XML, пошук, новини, об'єктивність.*

## Вступ

У сучасному світі вплив ЗМІ на повсякденне життя критично зріс. З розвитком інтернет-технологій ми отримали доступ до різної інформації, в тому числі і новинної, але останнім часом, виникла проблема отримання не просто інформації, а фактів, що мають конкретне підтвердження. Переважна більшість інформаційних агентств залежать від різних джерел фінансування, як державних, так і приватних, через що інформація, яка надається ними, може бути заздалегідь не об'єктивною. Таким чином, на сьогоднішній день кількість недостовірного новинного контенту зростає і ця обставина безпосередньо впливає на соціальні, політичні та економічні аспекти життя. Така ситуація з впливом недостовірного контенту на суспільне життя складається через те, що дуже важко знайти механізми фільтрації необ'єктивного новинного контенту.

Останнім часом в багатьох європейських країнах виникає все більший інтерес до розробки методів боротьби з даною 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 в ході цієї передвиборчої кампанії. Слід знати, що робляться спроби ними маніпулювати". Він наполягає на ретельних перевірках інформаційного контенту, що публікується, з метою відокремити недостовірну чи сфальшовану інформацію від істинної [1]. Крім політиків адміністрація соціальної мережі facebook так само шукає можливі варіанти боротьби з фальшивими або зміненими новинами, за їхніми словами в найближчому майбутньому будуть реалізовані механізми для боротьби з недостовірними новинами [2]. Ці завдання можуть бути вирішеними саме засобами штучного інтелекту [3].

**Проблеми отримання об'єктивної інформації**

У сучасному світі сильно зріс вплив ЗМІ на повсякденне життя в суспільстві або державі. Зараз інформаційні потоки сильно відрізняються від колишніх і новинні агентства можуть використовувати для публікації новин не тільки друковані видання, але соціальні медіа, такі як facebook, twitter та інші, а так само власні новинні вебсайти. Крім власних друкованих або електронних видань агентства можуть використовувати різні новинні портали, які виступають в ролі майданчиків для розміщення новинного контенту з різних джерел. На таких порталах користувач може знайти кілька варіантів однієї і тієї ж новини, які надані різними ЗМІ, але, на жаль, механізм перевірки контенту на достовірність відсутній. В результаті користувач повинен покладатися тільки на свої власні аналітичні здібності, аналізуючи ту чи іншу новинну інформацію.

Для отримання новин агентства застосовують різні методи - від використання власних репортерів і кореспондентів в тій чи іншій країні або регіоні до копіювання новинного контенту з місцевих видань. Деякі агентства використовують інформацію, яка одержана не від професійних журналістів, а від звичайних очевидців або ж від зацікавлених осіб. У цьому випадку встановити достовірність отриманої інформації буває досить складно.

Іншою, не маловажною, проблемою є отримання, зберігання і обробка великих масивів новинних даних. Раніше інформації було набагато менше, ніж зараз, і журналісти витрачали зусилля, в основ-

ному, на її пошуки. Тепер же ми маємо справу з великими масивами інформації, часто неструктурованою. Зараз журналісти змушені працювати на двох рівнях: аналіз і структуризація інформації, що надходить і подання даних у вигляді зрозумілому користувачеві. Сьогодні новини поширюється в той же момент, коли вони відбуваються. Тому так важливо зібрати якомога більше фактів про події для подальшого аналізу. Завдяки цьому змінюється суть журналістики - важливо не те, хто першим повідомив новину, а хто зміг її пояснити, проаналізувати можливі наслідки і об'єктивно і неупереджено оцінити те, що відбувається.

На жаль, в сучасному світі мало хто ставить перед собою завдання об'єктивного інформування суспільства. На сьогоднішній день новинний контент набув такого характеру, що вже майже неможливо знайти неупереджене викладення фактів, що сприяє створенню ілюзії паралельної реальності, що згубно впливає на всі процеси всередині суспільства, держави, регіону і світу в цілому. Більшість ЗМІ займається інтерпретацією фактів замість їх об'єктивного висвітлення через майже відсутність незалежних інформаційних агентств. Таким чином, формування суспільної думки відбувається в тому фокусі, який створено певними корпораціями, фінансовими групами, політичними і державними діячами, які стоять за тими чи іншими інформаційними агентствами. Тоді як саме правдива інформація є фундаментом для прийняття правильних рішень. Саме тому суспільство критично потребує джерел об'єктивної інформації. З іншого боку, немає явної сили, яка була б зацікавлена в об'єктивному висвітленні тих чи інших подій.

Таким чином, боротьба з недостовірним або зміненим новинним контентом є критичним завданням світового суспільства особливо для країн, що розвиваються.

Новинний контент є об'ємним і таким, що постійно змінюється, отже, нам потрібне рішення, яке зможе впоратися з великими обсягами даних і адаптуватися до швидких змін. Загально визнано, що тільки засоби штучного інтелекту здатні вирішити цю проблему. Одним з таких рішень є метод семантичної фільтрації.

Метою даної статті є дослідження методів семантичного аналізу новинного контенту з різних джерел на основі перетворення інфо-

 ISSN 1562-9945

рмації в онтологічну форму подання, а також розробка методу вирішення проблеми, яка викладена вище.

**Представлення інформації за допомогою семантичного порталу**

Для вирішення проблем із застосуванням технології Semantic Web, часто використовуються онтології, які дозволяють формалізувати знання про предметну область таким чином, щоб вони могли оброблятися і людиною і комп'ютерною системою. Онтологічний опис предметної області дозволяє виконувати автоматичну обробку її контенту. Таким чином, онтології визначаються як ключова технологія для застосування і розвитку Semantic Web. Крім того, важливою характеристикою сучасних семантичних методів є їх адаптивність - можливість застосування в різних областях (медицина, політика, освіта, трафік і т.д.).

Одним з можливих варіантів вирішення проблеми обробки великих масивів неоднорідної і розподіленої інформації новинного контенту може бути перетворення інформації в онтологічну форму подання і подальша обробка отриманої семантичної інформації [4].

У процесі дослідження проблеми було вивчено один з можливих варіантів вирішення проблеми контролю якості наданого новинного контенту. Рішення базується на використанні порталу, розробленого на основі семантичних технологій, в якості платформи для прозорої взаємодії декількох суб'єктів, накопичення та обміну їх інформації та забезпечення якості їх діяльності. Портал був розроблений в рамках міжнародного проекту Tempus 516935 "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 якості освіти.

Для посилення соціальної функції порталу та уникнення контролю залежною стороною, система побудована за принципами соціальної мережі. Користувачі є головними вкладниками, контролерами та вигодонабувачами контенту і функцій порталу. Додаючи інформацію, безпосередньо пов'язану з профілем користувача, користувач реєструє нові ресурси або пов'язує раніше визначені ресурси користувача, що створюють нові відносини. Кожен користувач створює приватний простір з набором ресурсів, які пов'язують його профіль з унікальними значеннями властивостей. Відкритий простір створюється спільно.

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 відносну якість деяких ресурсів з різних точок зору.

Портал працює відповідно до інформації, що зберігається в її онтологічній базі знань. Як правило, ці знання або база знань використовується як сховище тільки для системного контенту. Ми також використовуємо онтології для опису самого порталу: його архітектури і функціональності. Онтологія виступає ядром у вигляді онтологічної бази знань і семантичного АПІ, для обробки онтологій призначених для відкритого і гнучкого зберігання та обробки інформації, що надається користувачами. Важливою особливістю архітектури порталу є його гнучкість, яка досягається за рахунок поділу 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);

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

- система цінностей (яка визначає вагові коефіцієнти для різних показників якості в різних контекстах);

- процеси забезпечення якості (формально визначені внутрішні або крос-організаційні процеси забезпечення моніторингу оцінки 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 із онтологіями, які описують будь-який домен відмінний від створеного під вищу освіту і який також оперує з чис-

 ISSN 1562-9945

ленними ресурсами і потребує процесу забезпечення якості, який базується на оцінюванні ресурсів (бізнес, виробництво, медицина, медіа) [5].

Портал надає технічні засоби для:

- Конструювання напівсуспільного профілю користувача;
- Спільного оформлення фактів;
- Перегляду фактів;
- Підтвердження правильності опублікованої інформації методами соціальної перевірки;
- Об'єднання ресурсів, що залежать від створеної системи цінностей користувача на основі гнучких багатовимірних показників якості;
- Обробка фактів, зареєстрованих на порталі та інформації, отриманої з надійних зовнішніх джерел (побудова рейтингів зареєстрованих ресурсів);
- Збір, аналіз і поширення даних, які можуть бути отримані за допомогою анонімних або цільових запитів, обстежень або шляхом імпорту зовнішніх статистичних даних.

**Застосовані технології**

Новинний контент доволі великий та неструктурований, включає в себе багато різноманітних масивів інформації. Дані отримуються з різних джерел і більшість з них ми повинні використовувати спільно для отримання більш повної картини того що відбувається, та також за для розвитку новинних систем. Розроблена онтологія забезпечує один із варіантів представлення новинного контенту. Онтологія описує новинний контент відповідно до пов'язанної інформації та концепцій. Онтологія була створена із ціллю відповіді на три основні питання: Що? – що робиться за для освітлення подій, Де? – де відбуваються події, Коли? – коли вони відбуваються, ці питання можуть виникнути під час проектування новинних систем тощо.

Для побудови моделі об’єкту, яка є описом того що мається на увазі, ми використовуємо RDFS. RDFS представляє систему типів для RDF, але у порівнянні із ним є більш технологічно просунутим.

Сьогодні найпопулярнішою мовою онтологій є OWL. Фактично OWL є розширенням лексики RDF та також є похідним від мови OIL DAML, але із покращеною системою машинного інтерпретування web-контенту. Саме тому при розробці була обрана саме OWL.

Завдяки тому що RDFS та OWL сумістні, розроблена онтологія буде утримувати RDFS елементи у синтаксисі OWL. Для безпосередньої розробки онтології існує багато редакторів, але для даної роботи був використаний найбільш популярний – Protege [6].

Основними компонентами новинної онтології є поняття, відносини, екземпляри та аксіоми. Поняття представляють собою набір або клас сутностей у межах новинної області. Кожний 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 може бути визначена шляхом вказування що клас є підкласом іншого класу, таким чином клас «публікації автора» має декілька підкласів таких як «стаття», «огляд» або «журналістське розслідування». OWL також може визначати властивості класів, які не дуже відрізняються від властивостей RDFS. Прості властивості можливо визначати за допомогою компонентів owl:ObjectProperty та owl:DatatypeProperty.

Характеристики властивостей роблять данні більш виразними таким чином, що система може робити більш точні висновки. В OWL можливо визначати відношення однієї властивості до іншої. Прикладами властивостей можуть бути компоненти owl:equivalentProperty та owl:inverseOf. Функціональні властивості висловлюють той факт, що властивість може мати не більш одного значення для кожного екземпляру. Припустимо, що окремий екземпляр «автор» має властивість вік – він не може мати більше одного показника віку, але це не означає, що автор не може не вказувати вік. Тому заповнюючи профіль автора ми можемо не вказувати деякі подробиці, але чим їх більше вказано, тим більш повною буде характеристика того чи іншого екземпляру онтології. Типові приклади функціональних властивостей розробленої онтології включають прізвище, вік, місце роботи, назва статті, автор статті і т.д. Розроблений набір властивостей є унікальним і підібраний з урахуванням того, що в подальшому онтологія, з

 ISSN 1562-9945

усіма класами і їх властивостями, буде інтегрована до семантичного порталу.

**Обробка інформації**

За допомогою інструментів порталу ми можемо проводити верифікацію інформаційного контенту згідно із джерелами та рейтингування по заздалегідь виставленим пріоритетам. Одним з таких інструментів є соціальна перевірка. Соціальна перевірка представляє собою набір механізмів які дозволяють громадськості брати участь у підтвердженні інформації. Вони дозволяють проводити кросс-контроль над інформацією, проводити голосування або використовувати особисту аргументацію для прийняття рішень.

На відміну від пошукових систем, де аналізуються вподобання користувача і інформація обробляється статично, за допомогою порталу ми можемо будувати експертні системи цінностей на підставі тих джерел інформації, яким довіряє користувач. Ці джерела ми визначаємо на підставі рейтингування, призначаючи вагові коефіцієнти переважним для нас критеріям. Це дозволяє навчати систему на основі експертної думки. Таким чином створюється експертна система на основі саме експертних знань, наприклад, визнаних експертів – новини і аналітика великих новинних агентств. Звичайні користувачі мають можливість спостерігати за рівнем довіри експертів до тої чи іншої інформації і в залежності від нього визначати її достовірність.

**Висновки**

У статті ми розглянули один із варіантів вирішення проблеми отримання об'єктивної новинної інформації за допомогою семантичного порталу. Розроблений за для оцінки якості української академічної спільноти, портал може бути змінений під будь-яку предметну галузь заміною доменної онтології. В даній роботі розглянута можливість заміни онтології, яка описує академічну галузь на онтологію новин. Як ми з'ясували, портал надає необхідний та зручний, для користувача, інструментарій для обробки та аналізу новинного контенту.

**ЛІТЕРАТУРА**

1. Бундестаг ищет способы борьбы с фейковыми новостями [Електронний ресурс]. – Режим доступу:
http://news.liga.net/news/world/14453387-bundestag_ishchet_sposoby_borby_s_feykovymi_novostyami.htm
2. Facebook Looks to Harness Artificial Intelligence to Weed Out Fake News [Електронний ресурс]. – Режим доступу:
https://www.wsj.com/articles/facebook-could-develop-artificial-intelligence-to-weed-out-fake-news-1480608004?mod=pls_whats_news_us_business_f
3. Fake News? Big Data And Artificial Intelligence To The Rescue [Електронний ресурс]. – Режим доступу:
https://www.forbes.com/sites/jasonbloomberg/2017/01/08/fake-news-big-data-and-artificial-intelligence-to-the-rescue/#4931c36b4a30
4. H. Wache, T. Vugele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hьbner, "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.
5. 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.
6. J. Cardoso, “The Semantic Web: A mythical story or a solid reality?”, In: Metadata and Semantics, Springer US, 2009, pp 253-257.

 ISSN 1562-9945