УДК 004.8

Рибалко В.Ю., магістрант, Пазюк Ю.М., доцент, науковий керівник

# ВИКОРИСТАННЯ МЕТОДІВ DATA SCIENCE ДЛЯ ПРОЕКТІВ СОЦІАЛЬНОГО СПРЯМУВАННЯ

*Запорізька державна інженерна академія, кафедра ПЗАС*

**Актуальність теми дослідження**. В епоху цифрових технологій з прогресивним накопиченням та розвитком інформацій- них структур та пов’язаних з ними соціально-гуманітарних преображень, постає закономірне питання про вплив великих потоків даних на методологію та інструментарій продукування знання. Все швидшими темпами початковий фокус сучасних медіа-технологій спрямовується на першу частину спектра дані-інформація-знання. Наука, що оперувала величезним арсеналом інструментарію змушена переорієнтуватись на новий формат структури науки та виробництва знання. Наука під натиском величезних потоків інформації перетворюється на своєрідний цифровий конвеєр продукування знання та технологій. Для оцінки масштабу обсягів цифрових даних Р. Вільямс склав таблицю відповідності між одиницями виміру даних і звичними для людського сприйняття об'єктами, переведеними в цифрову форму. До 2010 р. обсяг всіх створених у світі цифрових даних становить 1,2 зетабайт, що можна представити як стопку DVD-дисків від Землі до Місяця і назад [1].

**Мета роботи –** Розробка застосунку, який на основі реальних даних зможе візуалізувати, оброблювати, сортувати та визначати проблемні місця в різних частинах міста. Програмний продукт повинен, оброблювати різного роду виклики від реальних людей, які зареєструвалися в даній мережі. Користувачі будуть повідомляти про небезпечні місця, випадки аварій, насильства тощо. Сервіс буде оброблювати ці дані і вже з певною базою викликів та повідомлень візуалізувати карту небезпек міста. Що дасть змогу соціальним структурам знайти більш точну причину небезпек в даній частині міста та незабаром вирішити ці проблеми.

**Проблемна ситуація.** Відкритий обмін інформацією, доступний в мережі між дослідниками, має основоположне значення для науки, для збільшення швидкості досліджень і зростання визнання вчених, стає можливим обмін інформацією між дослідниками, який охоплює ряд ініціатив з метою видалення перешкод для до- ступу до даних і опублікованих документів. Медіа-інфраструктурі характерні підходи, парадигма, моделі, орієнтовані на цифрові дані. Так як даних багато, то існує проблематика правильної обробки цих даних з різних джерел.

**Результат дослідження**. Аналіз даних є невід'ємною частиною всіх прикладних досліджень та вирішення проблем в промисловості. Найбільш фундаментальні підходи аналізу даних — візуалізація (гістограми, точкові ділянки, ділянки поверхні, дерево карт, паралельно координовані ділянки і т. д.), статистика (гіпотеза тест, регресія, СПС і ін), видобуток даних (асоціації гірничодобувної промисловості, і т. д.) і методи машинного навчання (кластеризація, класифікація, дерева рішень, і т. д.). Серед усіх цих підходів, візуалізації інформації, або, іншими словами, візуального аналізу даних, є той, який спирається в основному на пізнавальні навички аналітиків, а також дозволяє розкриття неструктурованих дієвих ідей, які обмежені тільки людською фантазією та творчістю. Аналітик не повинен застосовувати різні витончені методи, щоб мати можливість інтерпретувати візуалізацію даних. Візуалізація інформації — це також схема гіпотез, які можуть бути, і, як правило, є попередниками більш аналітичного або формального аналізу на кшталт статистичних гіпотез [2]. Статистична обробка експериментальних даних, результатів розрахунків і математичного моделювання передусім необхідна для представлення інформації у більш компактній формі, зручній для подальшого використання. В даний час все ширше використовують добре розроблений апарат математичної статистики, яка займається методами систематизації, обробки й використання статистичних даних для наукових іпрактичних висновків. Статистична

обробка неминуче пов'язана з втратою інформації, тому при виборі статистичних характеристик важливо глибоке розуміння специфіки конкретних завдань, щоб в концентрованій формі зберігати потрібну інформацію.

Статистична обробка експериментальних даних зазвичай заснована на граничних теоремах теорії ймовірностей і вимагає обчислення оцінок по порівняно простих формулах. Однак для підвищення якості оцінок необхідна величезна кількість даних, і обсяг обчислень може виявитися дуже великим.

Статистична обробка експериментальних даних, результатів розрахунків і математичного моделювання передусім необхідна для представлення інформації в більш компактній формі, зручній для подальшого використання. Вданий час все ширше використовують добре розроблений апарат математичної статистики, яка займається методами систематизації, обробки й використання статистичних даних для наукових і практичних висновків. Статистична обробка неминучепов'язана з втратою інформації, тому при виборі статистичних характеристик важливо глибоке розуміння специфіки конкретних завдань, щоб в концентрованій формі зберігати потрібну інформацію [3].

**Висновки:**

Будь-який аналіз впливу «великих даних» необхідно розпочати з визначення того, що власне означає цей термін, який часто використовується, але зазвичай не всі його розуміють. Він стосується насамперед величезної кількості даних, які постійно збираються за допомогою пристроїв і технологій, таких як платіжні картки та картки лояльності клієнтів, Інтернету та соціальних медіа і все частіше, через датчики Wi-Fi та електронні мітки. Велика частина цієї інформації є неструктурованою – тобто це дані, які не відповідають певній, заздалегідь визначеній, послідовності. Тобто «великі дані» – це наступний етап розвитку даних. Забезпечення якості даних є ключовим фактором. Процес рівно настільки якісний, наскільки якісні дані, які в ньому використовуються. Оскільки автоматична обробка даних є однією з основних прикладних задач кібернетики, то прописування кодів команд із набором даних повинно перетворюватися у візуальне відображення зрозумілого змісту та форми для пересічного користувача статистичної інформації.

Висновки, які можна зробити в результаті статистичних обрахунків, цілком зумовлені якістю вхідних даних, а саме їх повнотою та достовірністю. Проте, досить часто в статистичних розрахунках має місце похибка отриманих показників. Похибки можуть виникати і нагромаджуватись під час побудови алгоритму розрахунку та формування даних у процесі обчислень. Це обумовлено неточністю та неповнотою визначення змісту викликів, невідповідністю між вимогами та фактичним змістом отриманих даних. Розрізняють систематичні та випадкові похибки. Але, якщо дані, які будуть повідомляти громадяни під час виклику будуть достовірні на 100%, то і результат статистики можна буде прогнозувати с набагато більшою ймовірністю.

## Література


1. Наука про дані - [Електронний ресурс] / Режим доступу до даних: https:// uk.wikipedia.org/wiki/Наука_про_дані – 05.10.2018 р. – Заголовок з екрану.
2. Візуалізація інформації - [Електронний ресурс] / Режим доступу до даних: https://uk.wikipedia.org/wiki/Візуалізація_інформації – 05.10.2018 р. – Заголовок з екрану.
3. Статистична обробка - експериментальні дані - [Електронний ресурс] / Режим доступу до даних http://techtrend.com.ua/index.php?newsid=16799 - – 05.10.2018 р. – Заголовок з екрану.