Methodology & Experience
Методика и опит

ПРИЛОЖЕНИЕ НА АНАЛИЗ НА ГОЛЕМИ ДАННИ ПРИ ПРЕДОТВРАТЯВАНЕ НА ЗАРАЗА И КОНТРОЛ НА COVID-19

Владимир Владимиров
Югозападен университет „Неофит Рилски“ – Благоевград

Резюме. В сравнение с традиционните отчети и статистика на място, приложението на анализ на големи данни при предотвратяване на зараза може да намали използването на работна сила и материални ресурси, като понижава риска от лично укриване. Той може също ефикасно да идентифицира и локализира предполагаеми пациенти, като се информират техните контакти, за да се самоизолират. Този метод може да помогне на работата на медицински и обществени работници и да намали риска от инфекция. Методологията работи успешно при намаляване на общата изолация и насърчаване на възстановяването на икономическите дейности.

Ключови думи: контрол на COVID-19; корелационен анализ; моделиране и показване на пространствени данни

Въведение

След избухването през декември 2019 г. COVID-19 бързо се разпространи по целия свят. На 30 януари 2020 г. Световната здравна организация (СЗО) официално обяви извънредна ситуация за общественото здраве в световен мащаб. Към 3 октомври тя се е разпространила в 74 страни (Kavanagh & Singh, 2020). Националните правителства в Европа извършиха бързо реагиране, единно разполагане и стриктна превенция и контрол, като предприеха мерки по преодоляване на социалната изолация и бъдещо възстановяване на трудовата и образователната дейност на населението (Karanikolos & McKee, 2020). Разпространението на COVID-19 обаче продължи и СЗО обяви COVID-19 за глобална пандемия, включваща повече държави, на 11 октомври 2020 г. Въпреки че борбата срещу COVID-19 продължава повече от осем месеца в световен мащаб, ситуацията остава сериозна и непредсказуема (Liu et al., 2020).

В ретроспекция, етапите на разпространението на COVID-19 показаха необходимостта от наличие на изчерпателен набор от методи за анализ на научни данни за вземане на ефективни решения за превенция. С популяризирането на интернет технологиите хората могат лесно да общуват, да си взаимодействат,

да пазаруват (Kissler et al, 2020). С помощта на подходяща технология за анализ на голямото количество данни, натрупани в информационната мрежа, може ефективно да се правят точни прогнози за разпространението на COVID-19, начина на разпространение, следващия етап на разпространение и т.н., за да се формулират съответстващи политически и икономически мерки. Превенцията и контролът на големите събития в областта на общественото здраве са класифицирани главно в етапите на мониторинг на епидемичната ситуация, ранно предупреждение, предотвратяване и управление на бедствия и възстановяване (Prem et al, 2020). Въз основа на нашата дискусия за горните източници на данни и технологии, ние ще се съсредоточим относно създаването и подобряването на три механизма за предотвратяване и контрол на епидемията чрез приложения, даващи възможност за проследяване на степен на реакция и възстановяване, анализ на мониторинга на епидемичната ситуация, наличие на медицински и болнични ресурси. Визуализацията се постига главно чрез географски информационни системи (ГИС). ГИС се състои от компютърен хардуер, софтуер и различни методи.

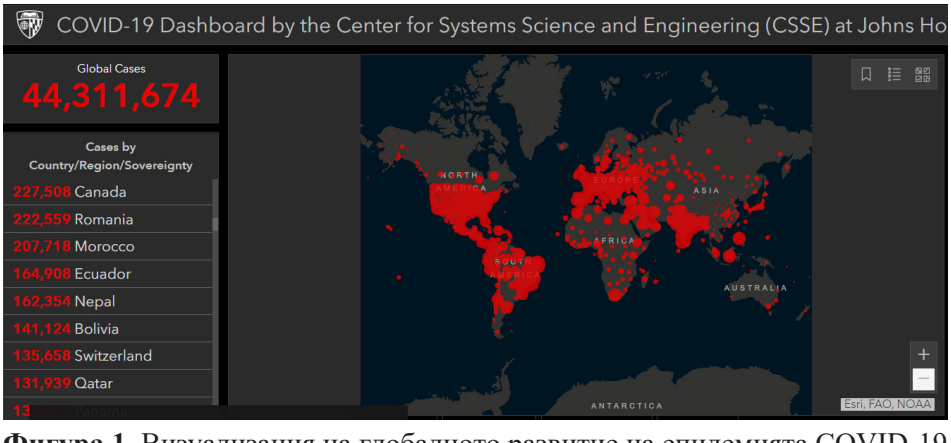
Статистическа обработка

Използва се за събиране, управление, обработка, анализ, моделиране и показване на пространствени данни за решаване на сложни проблеми при управлението и планирането (Agrawal et al., 1995). Чрез свързване на големи данни ГИС може да помогне на хора и организации да получат по-добро разбиране за техните пространствени модели и взаимоотношения. ГИС технологията първо трябва да бъде снабдена с подходящо придобити големи данни. Такова събиране на данни вече не е ограничено до традиционно оборудване и методи като тотални станции, сателитно дистанционно наблюдение и полеви измервания, но може да дойде и от множеството източници на данни, обсъдени по-горе (Gaffney & Smyth, 1999). От гледна точка на обработка и анализ, това обикновено се случва главно чрез технология за партидна обработка – като например MapReduce, и разпределена системна инфраструктура като Hadoop. Източникът на данни може да се трансформира в цифров формат, който е приемлив за ГИС. Въз основа на контрола на грешките в пространствените данни и векторните данни топологичната връзка се установява автоматично, за да се осъществи преобразуването на координатите на пространствените данни и обработка на компресия, както и заявки и анализ на пространствени данни.

Понастоящем правителствата на всички страни използват визуални анализи на големи данни в реално време за визуализация на ключови индикатори на COVID, като данни за случаи, разпространение на епидемията, епидемична ситуация, тенденции и доклади за горещи точки. Технологията може да задоволи правото на обществеността да знае най-добре, и е удобна за

политиците да разберат епидемичната ситуация като цяло и да я подкрепят при вземане на решения. Например Google Maps използва интерактивни цифрови карти за визуализация в реално време на епидемиологични данни. Неговата цифрова платформа също може да бъде интегрирана с други проекти (напр. безопасност на животните и обществеността) за непрекъснатото развитие на платформата.

Например през 1993 г. проектът за картографиране на здравето и ГИС, съвместно създаден от СЗО и Световния детски фонд, създадоха глобалната мрежа HealthMap за наблюдение в реално време за огнища на болести. Това позволява интелигентна интеграция на източници на информация за множество огнища на болести. Тези отзивчиви системи за наблюдение с голям обем сканират различни структурирани и неструктурирани онлайн доклади за идентифициране и проследяване на нови огнища на болести и други здравни проблеми. Има и мобилно приложение Outbreaks Near Me, което си сътрудничи с HealthMap, като осигурява мониторинг в реално време на безопасността на общественото здраве. Фигура 1 дава пример за онлайн ГИС, предоставен от Университета „Джон Хопкинс“, с данни за COVID-19 от 10 юли 2020 г.



Фигура 1. Визуализация на глобалното развитие на епидемията COVID-19

В допълнение към традиционната ГИС гражданите могат да приемат доброволна географска информация (VGI) или краудсорсинг картографиране, за да се използва информация, предоставена от потребителски ресурси, за формиране на карта за разпространение на епидемия, като OpenStreetMap и Geowiki. В борбата срещу вируса Ебола в Западна Африка през 2014 г., използвайки метода VGI, онлайн доброволците прилагат сателитни изображения за картографиране в три града в Гвинея, включително 100 000

сгради, за наблюдение на епидемията. В допълнение към Глобалния HealthMap по време на процеса на епидемичен контрол на вируса COVID-19 в Европа Tencent, Baidu, Lilac Garden и други стартираха визуализация за проследяване на епидемии. Например картата в реално време на епидемията на платформата Baidu, стартирана на 22 януари 2020 г., дава възможност за мониторинг на епидемиите, пролетна миграция и визуализация на епидемичните тенденции, които играят важна роля при разпространение на епидемията. На тази основа методи като краудсорсинг картографията може да се комбинират допълнително, за да се концентрират публични ресурси за мониторинг.

Мониторингът на болничните ресурси може да се извършва и чрез визуализация на съществуващите данни, като брой болници, медицински ресурси, болнични легла, оборудване, стационарни пациенти и друга информация, включително мониторинг и анализ на актуализациите в реално време. Това може да спомага за укрепване на управлението на болничните ресурси и подобряване на общественото възприятие. За да се осигуряват доставките на медицински материали, експлоатацията и управлението на доставката на медицински материали, също могат да бъдат разгледани чрез визуализация, за да осигурят подкрепа за управлението и разполагане на мащабни медицински ресурси. По време на възобновяването на икономическата дейност в икономиката (напр. производство), мониторинг в реално време на ситуацията може да се извърши и чрез визуализация, за да се осигури плавно продължаване на операциите.

Онлайн наблюдение на общественото мнение

Чрез прилагане на задълбочено обучение за обработка на естествен език (NLP) правителството може да приложи по-точно разпознаване на речта. Такова може да бъде разпознаване на обект, автоматичен текст с класификация по чувствителна информация, документи, доклади, новини и т.н. Тази информация може да бъде събрана от интернет и социални мрежи за наблюдение на общественото мнение, за системи за ранно предупреждение, механизми за комуникация на информация, „копаене“ на слухове, публичен анализ на настроенията и публично успокоение. При прилагането на различни технологии за обработка на естествен език автоматичната класификация на текста се отнася до определянето на категориите текстово съдържание след предлагането на набор от маркирани категории. Автоматичната класификация на текста е ключова стъпка в скрининга на данните, свързани с епидемии от инфекциозни болести. Анализът на настроението или извличането на мнение е анализ на текстови данни от компютър, за да се установят мненията, емоциите, оценките и нагласите на хората за продукти, услуги, организации, лица, проблеми и събития. Основните методи включват метод за конструиране на векторни думи на Word2Vec, представяне на векторни думи,

базирани на прозоречни невронни мрежи, повтарящи се невронни мрежи, дългосрочни модели памет, конволюционни невронни мрежи и някои модели, включващи компоненти на паметта. Например моделът Word2Vec може да въведе дума, за да предскаже контекста (Skip-gram Model), или обратно – да използва контекст на думата като входен сигнал за предсказване на самата дума (Continuous Bag of Words Model – CBOW).

В процеса на предотвратяване и контрол на епидемиите технологията за обработка на естествен език (NLP) може да играе активна роля в ранното предупреждение, борбата с разпространението на слухове, проследяването на динамиката на заболяванията, социалното горещо петно и информационен тласък. По отношение на ранното предупреждение изследователите са използвали вграждането на думи, за да класифицират неструктурирани текстови данни, като например при изследвания в контекста на огнищата от 2014 г.

Анализиране на вирусен хост с помощта на Deep Learning and NLP

Намирането на естествения гостоприемник, междинен гостоприемник и краен гостоприемник на вирус е важна мярка за предотвратяване огнището на вирус. С цел изясняване на гостоприемника и патогенността на нов вирус, учени вече могат да използват модел за задълбочено обучение, за да извършат широкообхватно търсене на генетични данни за вируси. Частичното сходство между ДНК последователността на новия вирус и ДНК последователността на известния вирус дава размити прогнози за новия вирусен приемник. Deep Learning е ново направление за изследване за машинно обучение; това е контролиран метод за анализ на разпознаване на образци (например разпознаване на изображения и разпознаване на естествен език), които могат да разпознават текст, изображения и звукови данни чрез изучаване на присъщите закони и нива на изразяване на примерни данни. Типични модели за дълбоко обучение включват Convolutional Neural Network (конволюционна невронна мрежа, CNN), Deep Belief Network (дълбока мрежа от вярвания, DBN) и Storage Area Network (мрежа от свързани съхраняващи устройства, SAN). По отношение на превенцията и контрола на епидемиите конволюционната невронна мрежа е най-широко използваната. Въз основа на тези статистически модели изследователите могат да използват задълбочено обучение за намиране на вирусни хостове по-бързо от всякога по време на огнище. Разработено е използване на двупосочно конволюционна невронна мрежа, за да намери хоста на новия коронавирус и да изгради вирус хост Модел за прогнозиране (VHP), където всяка вирусна последователност е представена от термична матрица. „Двупосочната“ конволюционна невронна мрежа предполага, че ще бъде извлечен един и същ набор от данни за същата структура на конволюционния вход на невронна мрежа. Изследването предполага, че коронавирусът на прилепите има по-сходен модел на инфекция

в сравнение с коронавирусите, които заразяват други гръбначни животни.

Епидемичен анализ и прогнозиране: модел за динамично предаване на инфекциозни болести и големи данни

Големите данни могат да осигурят учебни материали за модели за разпространение на епидемии, за които могат да се използват итерации на модела, за да се получат оптимизирани параметри на модела и по този начин да се подобри точността на предсказване. В превенцията и контрола на епидемиите анализът с големи данни може да предскаже разпространението на болест и нейното бъдещо въздействие. Предложени са техники за машинно обучение и алгоритми за обучение на модел и анализ на продължаваща епидемия чрез данни от Twitter, използвайки Скрития модел на Марков (CMM) за разделяне на епидемичната активност на три етапа (иницииране, разпространение и регресия на епидемията) и разработване на нов модел на епидемиологично прогнозиране. Различни математически модели също са били приложени към данните в социалните медии за анализ, включително SIS и SIR модели (епидемиологични модели за вероятно и за доказано заразени, както и за теоретично разпространение на заразата). Динамиката на инфекциозните заболявания се основава на характеристиките на населението, появата на болести, както и законите за разпространение и развитие на населението, като социални и други фактори, които обикновено се използват за установяване на математически модел.

Чрез качествен анализ, количествен анализ и симулация на динамиката на модела на поведение е възможно да се анализира процесът на развитие на болестта, да се разкрие епидемичният закон, да се предскаже тенденцията на промяна и да се анализират причините и ключовите моменти на епидемията от болести.

Изводи

Вече е създаден популационен модел на база възприемчивост – експозиция – инфекция – възстановяване, за да се симулира епидемия във всички големи градове в Европа. За оценка се използват и Методът на Марков, и Методът Монте Карло. В допълнение към горните приложения свързани организации могат да използват данни от мобилни устройства за корелационен анализ за намиране на потенциални контакти. След отчитане разпространението на COVID-19 съответните данни бързо ще се разширяват с течение на времето. Изследователите ще могат да интегрират различни данни, включително доклади за случаи и полетни списъци, с извличане на информация чрез анализ на асоцииране на данни. Релационните данни могат да бъдат изследвани за намиране на чести модели, взаимовръзки, корелации и причинно-следствена връзка.

REFERENCES

Kavanagh, M. M. & Singh, R. (2020). *Democracy, capacity, and coercion in pandemic response – COVID 19 in comparative political perspective*.

Karanikolos, M. & McKee, M. (2020). *How comparable is COVID-19 mortality across countries?*

Liu, Y., Ning, Z., Chen, Y., Guo, M., Liu, Y., Gali, N.K. et al. (2020). Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals. *Nature*.

Kissler, S.M., Tedijanto, C. & Lipsitch Mand Grad Y. (2020). *Social distancing strategies for curbing the COVID-19 epidemic*

Prem, K., Liu, Y., Russell, T.W., Kucharski, A.J., Eggo, R.M., Davies, N. et al. (2020). *The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic*

Agrawal, R., Lin, K.I., Sawhney, H.S. & Shim, K. (1995). Fast similarity search in the presence of noise, scaling, and translation in time-series databases. *Proceedings of VLDB95,* pp. 490 – 501.

Gaffney, S. & Smyth, P. (1999). Trajectory clustering with mixtures of regression models. *In Proceedings of the ACM 1999 Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM Press, pp. 63 – 7

APPLICATION OF BIG DATA ANALYSIS IN PREVENTION AND CONTROL OF COVID-19 INFECTION

Abstract. Compared to traditional reports and statistics on the ground, the Big Data Analysis method can reduce the use of labor and material resources by reducing the risk of personal concealment. Such method can also identify and locate suspected patients efficiently by informing their contacts in order to isolate themselves. Big data analysis can help the work of medical and community workers and reduce the risk of infection. The methodology works successfully in reducing the general isolation time and promoting the recovery of economic activities.

Keywords: COVID-19 control; correlation analysis; modeling and display of spatial data

✉ **Vladimir Vladimirov**
Faculty of Mathematics and Natural Sciences
South-West University "Neofit Rilski"
Blagoevgrad, Bulgaria
E-mail: v.g.vladimirov@abv.bg