N.H. ABROYAN

THE USAGE OF NEURAL NETWORKS AND COMPARISION OF RESULTS AT CLASSIFYING THE REAL-TIME DATA

In recent years, notable success has been recorded in the sphere of machine learning. Particularly, different usages of deep neural networks, including convolutional and recurrent neural networks in image, speech and signal processing are known. However, there is practically no research carried out to find methods for using deep neural networks in elaboration of real-time data of financial transactions. In the frames of this work, the deep neural networks for classifying the real-time data, and also finding new models and methods for analyzing the transaction results are investigated.

Keywords: machine learning, classification, real-time data, convolutional neural networks, recurrent neural networks.

ՀՏԴ 004.22:004.421

Բ.Գ. ԱԹԱՅԱՆ

ԱՄՊԱՅԻՆ ՊԱՀՈՒՍՏԱՎՈՐՄԱՆ ՀԱՄԱԿԱՐԳՈՒՄ ՏՎՅԱԼՆԵՐԻ ՊԱՀՈՒՍՏԱՎՈՐՄԱՆ ՕՊՏԻՄԱԼԱՑՄԱՆ ՄԵԹՈԴ

Ներկայացվում է ամպային պահուստավորման համակարգում տվյալների պահուստավորման օպտիմիզացման մեթոդ, որը հիմնված է գաղտնագրված տվյալների դեդուպլիկացիայի վրա: Համակարգն ապահովում է բազմակի պահուստավորվող տվյալների դեդուպլիկացիա՝ կոնվերգենտ գաղտնագրության շնորհիվ, իսկ եզակի տվյալների անվտագության ապահովման համար օգտագործվում է գաղտնագրության երկրորդ մակարդակը: Համակարգում բոլոր տվյալները խմբավորվում են՝ ըստ տվյալի հայտնիության աստիճանի, որը որոշվում է տվյալը տարբեր օգտագործողների կողմից պահուստավորելու քանակով:

Առանցքային բառեր. ամպային պահուստավորում, կոնվերգենտ գաղտնագրում, դեդուպլիկացիա, բազմաշերտ գաղտնագրում, վեբ կայքերի պահուստավորում:

Ներածություն: Տվյալների քանակը տարեցտարի աճում է գերարագ կերպով, ինչի հետևանքով ստեղծվում են նոր լուծումներ մեծածավալ տվյալների պահպանման համար: Ներկայումս լայն ճանաչում են ստացել ամպային պահուստավորման միջոցները, որոնք թույլ են տալիս օգտագործողներին պահպանել իրենց տվյալները առցանց: Ամպային պահուստավորման միջոցներն ունեն մի շարք առավելություններ՝ ի տարբերություն տվյալների պահուստավորման ավանդական միջոցների: Սակայն ինչպես այլ ոլորտներում, այստեղ նույնպես

գոյություն ունեն որոշակի թերություններ, որոնցից ամենակարևորներից են տվյալների անվտանգության և գաղտնիության խնդիրները:

Ամպային տեխնոլոգիաները լայն կիրառություն են գտել ոչ միայն անհատական օգտագործողների տվյալների պահպանման գործընթացում, այլև այլ ոլորտներում: Այս աշխատանքի շրջանակներում կդիտարկվի ամպային տեխնոլոգիաների կիրառմամբ վեբ կայքերի պահուստավորման օպտիմալացման մեթոդը: Օրեցօր աշխարհում ստեղծվում է մոտ 150000 վեբ կայք[1], որոնց պարունակությունը բազմաբովանդակ է: Որոշ կայքեր ստեղծվում են էլեկտրոնային առևտրի համար կամ զուտ տեղեկատվական բնույթ են կրում, իսկ մյուսները պարունակում են անձնական տվյալներ (բանկային հաշիվներ, հասցեներ և այլն): Ակնհայտ է, որ այդ տեսակի տվյալների կորուստը կարող է հանգեցնել լուրջ հետևանքների: Այդ պատճառով տվյալների պահուստավորումն ունի մեծ նշանակություն և պետք է կատարվի ճիշտ ժամանակին:

Աշխատանքում առաջարկվում է վեբ կայքերի պահուստավորման արդյունավետ մեթոդ, որը հիմնված է տվյալների բազմաշերտ գաղտնագրման վրա, թույլ է տալիս կատարել պահուստավորման գործընթացի օպտիմալացում՝ օգտագործելով կրկնվող տվյալների դեդուպլիկացիան:

Տվյալների սեղմումը և դեդուպլիկացիան տվյալների պահպանման գործընթացի արդյունավետության ապահովման կարևորագույն գործառույթներ են, որոնց ներդրումը թույլ է տալիս պահուստավորման ծառայություններ մատուցող կազմակերպություններին՝ օգտագործելու իրենց պահուստներն ավելի արդյունավետ, և հնարավորություն է տալիս ծառայություն մատուցել ավելի շատ օգտագործողների՝ առանց առկա ենթակառուցվածքի փոփոխության: Տվյալների դեդուպլիկացիան նշանակում է, որ եթե տվյալը պատկանում է մեկից ավելի օգտագործողների, ապա պահուստավորման սերվերում պահվում է այդ տվյալի միայն մեկ օրինակ [2]: Գոյություն ունի դեդուպլիկացիայի չորս տարբեր եղանակ՝ կախված նրանից՝ դեդուպլիկացիան արվում է սերվերայի՞ն, թե՞ օգտագործողի մասում՝ մինչև վերբեռնումը, և, դեդուպլիկացիան արվում է տվյալների բլոկերի՞, թե՞ ֆայլային մակարդակում: Եթե դեդուպլիկցիան կատարվում է օգտագործողի մասում, ապա այդ դեպքում կարելի է նաև շահել պահուստավորվող տվյալների վերբեռնման ժամանակ, քանի որ այն տվյալները, որոնք կրկնօրինակ են, կարելի է չվերբեռնել: Այսպիսով, դեդուպլիկացիայի օգտագործումը թույլ է տալիս պահուստավորման ծառայություն մատուցող ընկերությանը շահել ամպային պահպանման տարածք և ցանցային թողունակություն:

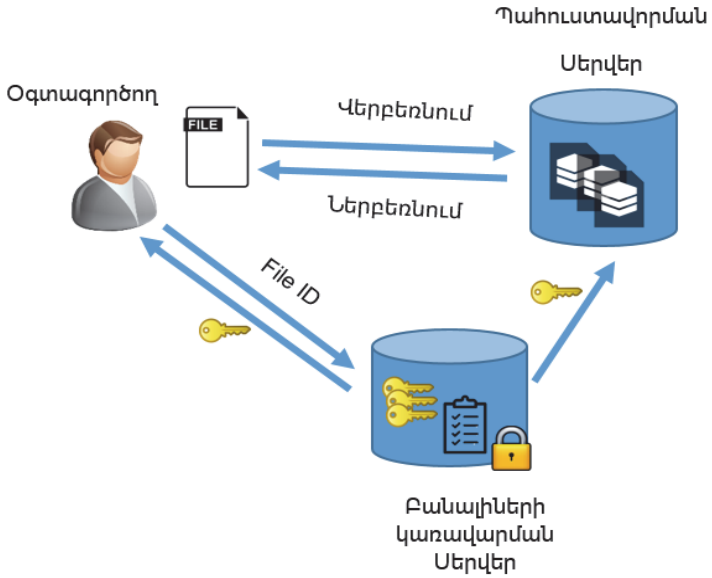
Դեդուպլիկացիայի օգտագործման կարևոր խոչընդոտներից է տվյալների գաղտնիության ապահովումը: Պահուստավորված տվյալների գաղտնագրության օգտագործման դեպքում դեդուպլիկացիան կորցնում է իր արդյունավետությունը: Հստակ կարելի է ասել, որ եթե պահուստների գաղտնագրման համար օգտագործվում է անվտանգ գաղտնագրային ալգորիթմ, ապա դեդուպլիկացիան անհնար է, քանի որ, բացի գաղտնագրման բանալուն տիրապետող անձից, ոչ մեկը չի կարող որոշել՝ երկու գաղտնագրված տարբեր տվյալներ արդյոք համապատասխանում են մեկ բաց տվյալին, թե ոչ:

Դեդուպլիկացիայի վրա հիմնված համակարգերը պահանջում են իրենց միջոցով մշակվող տվյալների տիպից կախված լուծումներ: Ընդհանուր առմամբ, գաղտնագրված տվյալների համար դեդուպլիկացիայի ապահովումը հնարավոր է կատարել՝ օգտագործելով կոնվերգենտ գաղտնագրման ալգորիթմ [3], երբ գաղտնագրման բանալին եզակիորեն որոշվում է գաղտնագրվող տվյալից: Այդ դեպքում, երկու նույն տվյալ միշտ կգաղտնագրվեն նույն բանալիով, և համակարգը կկարողանա որոշել՝ արդյոք երկու տվյալ նույնն են, թե ոչ, և կկատարվի դեդուպլիկացիա: Սակայն այս մեթոդն ունի թերություն, քանի որ կայուն չէ պարունակության գուշակման գրոհների նկատմամբ [4]:

Առաջարկվող մեթոդը: Աշխատանքում դիտարկվել է պահուստավորվող տվյալների տեսակ, որում կարող են լինել մեկ տիպի որոշակի քանակությամբ տարրեր և մեծ քանակությամբ այլ տարրեր: Մասնավորապես, տվյալ տեսակին են համապատասխանում վեբ կայքերը, քանի որ դրանք մեծ մասամբ կազմված են ընդհանուր կայքի աշխատանքը և պարունակությունը կարգավորող մասից՝ CMS (content management system), և կոնկրետ կայքին պատկանող եզակի տվյալներից: Քանի որ CMS-ը ընդհանուր մաս է, ապա պահուստավորման համակարգում հնարավոր է, որ լինեն մի քանի տարբեր օգտագործող, որոնց կայքերի համար օգտագործվում է նույն CMS-ը, օրինակ՝ Wordpress, Joomla [5]: Ներկայացվող մեթոդի հիմնական գաղափարն այն է, որ տվյալները պահանջում են տարբեր մակարդակի անվտանգության ապահովում՝ կախված նրանից, թե որքանով է տվյալը հայտնի, այսինքն, եթե վերածենք դա քանակական ցուցանիշի՝ քանի օգտագործող է միարժամանակ պահպանում և օգտագործում այդ տվյալը: Դիտարկված CMS-ների օգտագործմամբ կայքերի պահուստավորման համար կարելի է շահել CMS-ի ընդհանուր ֆայլերի պահուստավորման ժամանակը և զբաղեցրած ծավալը, իսկ մնացած ֆայլերը կլինեն եզակի՝ օգտագործողների կողմից ստեղծված ֆայլեր, որոնք կպահանջեն ավելի բարձր անվտանգության

մակարդակ: Այս մոտեցումն իրագործվում է՝ օգտագործելով երկմակարդակ գաղտնագրում: Համակարգում գտնվող բոլոր ֆայլերն ի սկզբանե հայտարարվում են ոչ հայտնի և գաղտնագրվում են երկու մակարդակով: Առաջին մակարդակը կոնվերգենտ գաղտնագրումն է, այսինքն՝ ֆայլը գաղտնագրվում է իր պարունակությունից ստացվող եզակի բանալիով: Հաջորդ մակարդակում այն գաղտնագրվում է սիմետրիկ գաղտնագրման ալգորիթմով: Գաղտնագրվելուց հետո ի սկզբանե ոչ հայտնի ֆայլի որոշակի քանակությամբ օգտագործողների կողմից համակարգ վերբեռնվելու դեպքում որոշում է կայացվում, որ տվյալ ֆայլը դարձել է հայտնի, և պետք է կատարել դեդուպլիկացիա: Դա նշանակում է, որ այդ պահին համակարգը պետք է կարողանա վերծանել ֆայլի երկրորդ մակարդակի գաղտնագրումը: Այսպիսի հնարավորություն ստանալու համար օգտագործվում է բանալիների կառավարման վստահելի սերվեր: Բանալիների կառավարման սերվերը ստուգում է օգտագործողի իրավասությունը, նաև՝ արդյոք նշված քանակով օգտագործողներ կան տվյալ ֆայլի հետ կապված ցուցակում, և եթե՝ այո, հանձնում է ֆայլի վերծանման սիմետրիկ բանալին պահուստավորող սերվերին, որն էլ իր հերթին կարողանում է վերծանել ֆայլի գաղտնագրման երկրորդ մակարդակը: Համակարգի աշխատանքի սխեման ներկայացված է նկ. 1-ում: Այսպիսով, երբ ֆայլի բավարար քանակով վերբեռնում է կատարվում, համակարգը կատարում է այդ ֆայլի դեդուպլիկացիա՝ առանց խախտելու այլ ֆայլերի անվտանգության առկա վիճակը: Տվյալի հայտնիությունը համակարգում ձևայնացած կարելի է նկարագրել p փոփոխականով, որը ցույց է տալիս իրավասու օգտագործողների այն փոքրագույն քանակը, ում այդ F ֆայլի վերբեռնումը կհանգեցնի դեդուպլիկացիայի: Համակարգի t շեմը այս դեպքում կնկարագրվի հետևյալ կերպ՝

$$t \geq p: \qquad (1)$$



Նկ. 1. Համակարգի աշխատանքի սխեման

Օգտագործողի կողմից ֆայլի վերբեռնումը և ներբեռնումը պահուստ կատարվում են հետևյալ կերպ.

▪ Վերբեռնում: Ֆայլի բացառիկ նույնացուցիչը կարող է լինել ֆայլային համակարգի նույնացուցիչ կամ հեշ արժեք և այլն) ուղարկվում է բանալիների կառավարման սերվեր և ստուգվում: Եթե այդպիսի նույնացուցիչով ֆայլ արդեն գոյություն ունի, ապա ներկա օգտագործողն ավելացվում է տվյալ ֆայլի օգտագործողների ցանկում, գաղտնագրված կապուղով ստանում է ֆայլի գաղտնագրման իր բանալին, գաղտնագրում է ֆայլը և ուղարկում պահուստային սերվեր: Այս պահին դեռ դեդուպլիկացիա չի կատարվում, քանի որ օգտագործողների ցուցակը դեռ բավարար քանակությամբ օգտագործող չի պարունակում: Եթե այդպիսի ֆայլ գոյություն չունի, ապա բանալիների սերվերում ստեղծվում է տվյալ ֆայլի համար բանալի, բանալին գաղտնագրված կապուղիով ուղարկվում է օգտագործողի կողմ և օգտագործողի կողմում գաղտնագրվում է այդ բանալիով: Այնուհետև գաղտնագրված ֆայլը վերբեռնվում է պահուստային սերվեր:

▪ Ներբեռնում: Ֆայլի նույնացուցիչն ուղարկվում է բանալիների կառավարման սերվեր և ստուգվում: Եթե այդպիսի իդենտիֆիկատոր ունեցող ֆայլ գոյություն ունի, և հարցում կատարող օգտագործողը տվյալ ֆայլն օգտագործողների ցուցակում է, ապա բանալիների կառավարման սերվերը անվտանգ կապուղիով ուղարկում է ֆայլի բանալին հարցում կատարած օգտագործողին,

գաղտնագրված ֆայլը ներբեռնվում է պահուստային սերվերից և վերծանվում օգտագործողի կողմում: Եթե ֆայլը դեդուպլիկացված է եղել, ապա բանալի չի ուղարկվում, քանի որ երկրորդ մակարդակի վերծանման կարիք չկա:

▪ Դեդուպլիկացիա: Բանալիների կառավարման սերվերը պահպանում է բոլոր ֆայլերի համար այդ ֆայլերը պահուստավորած օգտագործողների ցուցակը, և այն պահին, երբ բավարար քանակությամբ օգտագործող ներբեռնում է ֆայլը, բանալիների կառավարման սերվերը որոշում է կայացնում՝ ֆայլի երկրորդ մակարդակի գաղտնագրման բանալիներից որևէ մեկը փոխանակել պահուստավորման սերվերի հետ: Բանալին անվտանգ կապուղով ուղարկվում է պահուստավորման սերվեր, և այն վերծանում է ֆայլը, ֆայլի գաղտնագրված ավելցուկային օրինակները հեռացվում են, և կատարվում է դեդուպլիկացիա:

Եզրակացություն: Ամպային պահուստավորման համակարգում տվյալների պահպանման օպտիմալացման դիտարկված մեթոդը հիմնված է գաղտնագրված տվյալների դեդուպլիկացիայի վրա: Համակարգում բազմակի պահուստավորվող տվյալների դեդուպլիկացիան ապահովվում է կոնվերգենտ գաղտնագրության օգտագործմամբ, իսկ եզակի տվյալների անվտագության ապահովման համար օգտագործվում է գաղտնագրության երկրորդ մակարդակը: Այսպիսով, կատարվում է տվյալների խմբավորում՝ ըստ տվյալի հայտնիության աստիճանի, մասնավորապես՝ քանի օգտագործողի կողմից է այդ նույն տվյալը պահուստավորվել: Եթե տվյալը ժամանակի ինչ-որ պահին դառնում է հայտնի, ապա գաղտնագրության երկրորդ մակարդակը վերծանվում է: Աշխատանքում ներկայացված է օգտագործողի կողմում կատարվող դեդուպլիկացիա, այսինքն՝ տվյալները վերբեռնել պահուստային սերվեր, թե ոչ, որոշումը կատարվում է օգտագործողի կողմում՝ խնայելով համակարգում ֆայլերի զբաղեցված ծավալը, ինչպես նաև ցանցի թողունակությունը:

ԳՐԱԿԱՆՈՒԹՅԱՆ ՑԱՆԿ

1. Most Reliable Hosting Company Sites in September 2017, https://news.netcraft.com/archives/2017/10/05/most-reliable-hosting-company-sites-in-september-2017.html, *Netcraft*, 2017.
2. Understanding Data Deduplication, https://www.druva.com/blog/understanding-data-deduplication, *Druva*, 2009.
3. *US Patent 5778395.* System for backing up files from disk volumes on multiple nodes of a computer network. - 1995.

4. Drew Perttula and Attacks on Convergent Encryption, https://tahoe-lafs.org/hacktahoelafs/drew_perttula.html, 2008.
5. **Mauthe A., Thomas P**., Professional Content Management Systems: Handling Digital Media Assets.- John Wiley & Sons, 2004. ISBN 978-0-470-85542-3.

Б.Г. АТАЯН

ОПТИМАЛЬНЫЙ МЕТОД ХРАНЕНИЯ ДАННЫХ В ОБЛАЧНОЙ СИСТЕМЕ РЕЗЕРВНОГО КОПИРОВАНИЯ

Разработан метод оптимального хранения данных в облачной системе резервного копирования, который основан на дедупликации зашифрованных данных. Дедупликация данных, резервные копии которых были сделаны несколькими пользователями, осуществляется с помощью данных, позволяющих классифицировать совместно используемые файлы и уникальные файлы по уровню популярности. Уровень популярности определяется порогом числа пользователей, которые делают резервную копию одного и того же файла. Описанный подход предоставляет возможность сохранить место облачного хранения и пропускную способность сети, не загружая резервные копии файлов, которые являются общими.

Ключевые слова: облачное резервное копирование, конвергентное шифрование, дедупликация, многоуровневое шифрование, резервное копирование веб-сайтов.

B.G. ATAYAN

AN OPTIMAL DATA STORAGE METHOD IN THE CLOUD BACKUP SYSTEM

An optimal method of data storage in the cloud backup system is presented based on the deduplication of the encrypted files. Deduplication of files that are backed up by different users is carried out using the data classification technique that allows to classify shared files and unique files by the popularity level. The popularity level is then identified by the threshold of the number of users that possess and backup the same file. The described deduplication approach helps to save cloud storage and network bandwidth by not uploading backups of files that are shared.

Keywords: cloud backup, convergent encryption, deduplication, multi-level encryption, website backup.