UNIVERSITÉ JEAN MONNET
SPÉCIALITÉ WEB INTELLIGENCE

TSINGHUQ UNIVERSITY
DEPARTMENT OF ELECTRONIC ENGINEERING

STAGE EN ENTREPRISE: STAGE DE FIN D'ÉTUDE

Rapport de Stage
Année 2013-2014

Fouille de Donnée dans la domaine de télécommunication

Auteur :
Wenyi WANG

Tuteur de stage en entreprise:
Vice directeur de labo NGN:
yongfeng HUANG

Tuteur de l'université:
Amaury HABRARD

De 20 Février 2014 à 20 Juillet 2014

Table des matières

Remerciements

Tout d'abord, je tiens à remercier Amaury Habrard et tous les enseignants de la Spécialité Web Intelligence de l'Université Jean Monnet, aussi les enseignants de Télécom Saint-Etienne et L'école nationale supérieure de Saint-Etienne, qui m'a aidé lors de ces deux années de étude.

Je remercie également M.Yongfeng HUANG pour avoir accepter diriger cette stage, il m'a beaucoup conseillé, et les discussions que l'on a pu avoir se sont toujours révélées très intéressantes et instructives.

Je souhaite également adresser mes remerciement à Zheng YANG, Lindong WEI et xian WU ainsi que tout les membres du laboratoire de Next generation Network(NGN) pour m'avoir soutenu, encouragé et conseillé tout au long de ce stage.

Je tiens à montrer tout ma gratitude envers toutes les personnes qui ont pu m'aider, m'encourager, me soutenir, me remotiver pendant ces années de travail.

1 Résumé

Pendant ces quatre mois de stage, notre groupe de recherche travaille avec les employés de CMCC (China Mobile Communications Corporation) . Le objectif du sujet est: utilise les technique de Fouille de données, étude les données fournir par CMCC, et trouve les relation entre les données et les défaut du système 4G. Nous avons fait plusieurs tentatives pour trouver les résultats, et on a utilise différents logiciel, j'ai utilisé le R, et mon collègue utilise Mathlab, nous avons utilisé plusieurs algorithme (Clusterring, PCA, Association rules, Ajustement). Mais à la fin, nous avons trouvé que à cause des défaut dans la système d'acquisition, les données ne sont pas correct, et nous ne pouvons pas trouver le résultat comme prévu. Mais les recherches que nous avons fait peut faites-leur savoir comment utilise les technique de fouille de donnée dans la domaine de télécommunication.

2 Introduction

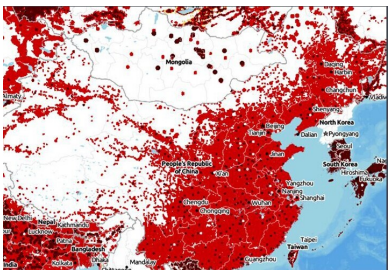
Le 3 avril 1973, M. Mation COOPER le directeur général de la division communication de Motorola, à effectuer un appel téléphonique à Joel ENGEL, son rival et néanmoins confrère chez Belle Labs. c'est la premier appel téléphonique en extérieur, L'idée du téléphone portable devient une réalité.

depuis ce jour, le technique développé très rapidement. dans les 20 dernières années, il y a déjà quatre génération des standards pour la téléphonie mobile, non seulement nous pouvons appeler les autres, les nouvelles technologies et les Smart-phones nous permettons aussi envoyer les message, surfer l'Internet, utiliser le service RTSP(Real Timide Streaming Protocol), et le service VoIP (Voice over Internet Protocole),etc.. les services de communication téléphonique sont devenus un outil très important dans notre vie.

2.1 Introduction du CMCC

Fondé en 3 Septembre 1997, après le regroupement de opérateur des télécommunications en 2008, China Mobile Communications Corporation (CMCC)1(a) est devenu un de trois opérateur des télécommunications en Chine (deux autres sont China Unicom Co., Ltd. et China Telecom). Après plusieurs années de développement, il a construit le plus grand réseau de communications mobiles dans le monde, possède la plus grande base d'utilisateurs dans le monde1(b). En 2013, CMCC a 767 million utilisateurs, 630,2 billion ¥ de revenu, 121,7 billions ¥de revenus net, effectif 197,030.

(a) Logo de China Mobile



(b) Reseau télécommunication

Figure 1 – CMCC

2.2 La crise de CMCC

Mais en même temps, le taux de croissance des nouveaux utilisateur décline de 22,5 % (2006) à moins de 5% 2013 2. Et dans la premier 3 mois, l'entreprise une fois considérés comme la plus rentable de Chine, le taux de croissance des revenu net est 0,3%.

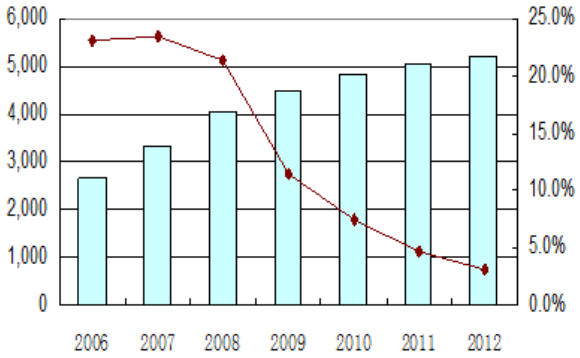
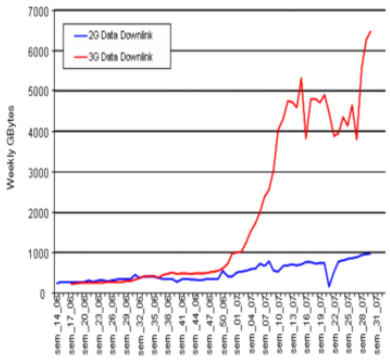
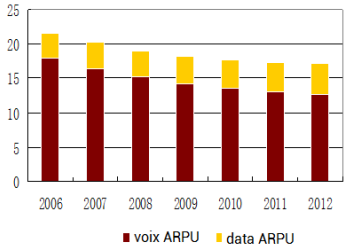


FIGURE 2 – le taux de croissance est décliner

Opérateur des télécommunications Vodafone a fait un étude après il déployé un réseau 3G(the third generation of mobile phone mobile communication technology standards). Comme le réseau 3G permettant des débits (de 2 à 42 Mb/s définis par la dernière génération des réseaux) qui sont bien plus rapides que la génération précédente, par exemple le GSM. Les utilisateur utilisent bien plus souvent le service internet3(a). Comme ils utilisent plus du service internet, le data ARPU (Average Revenue Per User) augment, mais le voix ARPU décline plus rapide que la montant de data ARPU3(b).



(a) Downlink Data Traffic in 2G/3G Network



(b) étude de Vodafone

FIGURE 3 – Vodafone

Mais l'étude de Orange nous montre que si nous pouvons fournir des nouveaux technologies qui a plus haute débit, les utilisateur utiliseront plus souvent le service data. 4

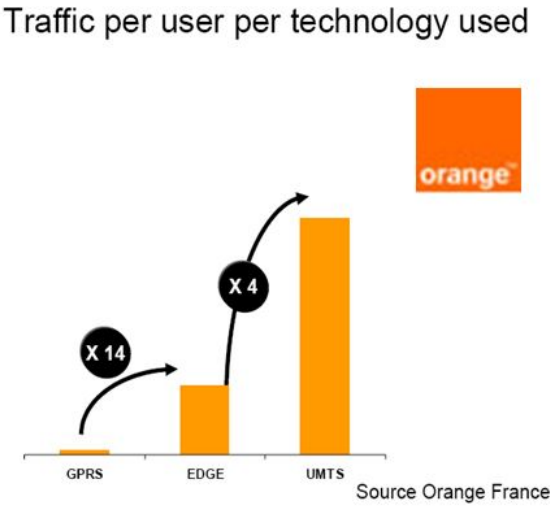


FIGURE 4 – trafic par personne

Des études nous montre nouveaux technologie (comme LTE) peut diminue le prix de revient, qui peut assurer le profit de l'opérateur. Mais déployé les nouveaux matériel coût très cher, en 2009, CMCC dépenser 30 billions ¥en construit les station pour réseau 3G, et à 2014, CMCC a construit 1,5 million stations, à la fin de cet année, il y aura 1,8 million stations, parmi ces station, il y aura 500 mille stations TD-LTE. En ajoutant des équipement 4G, il peut être mis à niveau un station de 3G à 4G. Donc déployé le réseau 4G n'est pas trop cher, selon l'expérience précédente (de 2G à 3G), les utilisateurs iront utiliser plus le service internet, qui peut assurer le profit de l'entreprise.

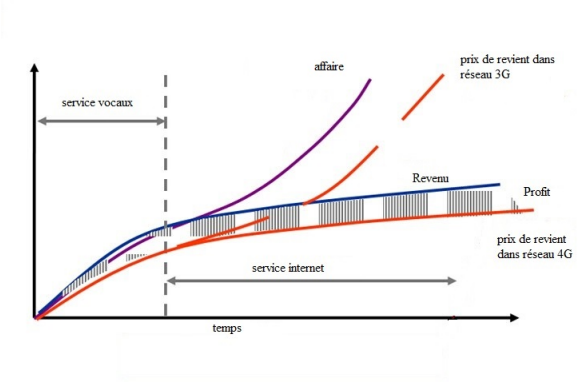


FIGURE 5 – 4G est plus rentable

2.3 L'optimisation du réseau

A part de la évolution des technologies. Un grand enjeu pour les opérateurs est: l'optimisation du réseau télécommunication. Le réseau de communication mobile est très dynamique, la répartition de la densité du trafic est inégale, fréquence très limité, etc. La configuration du réseau état toujours sous-optimal, et la perception de l'utilisateur n'est pas très bien. Donc tous les opérateurs doivent toujours reconfigurer/optimiser/maintien les paramètre du réseau.

Les opérateurs peuvent percevoir les données sur Internet, et utilisent ces informations pour trouver les défauts du système, peut aide l'entreprise optimiser le système.

Mais la optimisation du réseau télécommunication est difficile parce-que: Les technologies d'optimisation de réseau concerné: La technologie de commutation, la technologie sans fil, la configuration et commutation de la fréquence, la signalisation système, l'analyse de trafic, etc. c'est un travail difficile, exiger une meilleure aptitude des employés.

Actuellement, l'optimisation du réseau dépend principalement à la expérience du personnel. Mais des fois les expériences ne sont pas correct. Par exemple, Si l'entreprise besoin de savoir le congestionné d'un station, il faut envoyer les employé avec des équipement pendant les périodes de pointe, mais on ne sait pas si les résultats sont correct 6. En outre, souvent un seul type de donnée ont utilise pour l'analyse et la comparaison

pour optimiser les réseau, plutôt que de trouver un solution d'optimisation basées sur toutes les données liées au réseau (telles que les données statistique de trafic, les données d'essai, etc). Et en raison de l'énorme quantité de données, c'est difficile de traite en temps opportun. il est évident que ce méthode est défectueux. Les défauts du système provoque la satisfaction des utilisateurs inférieure, ce qui a conduit à multiplier.

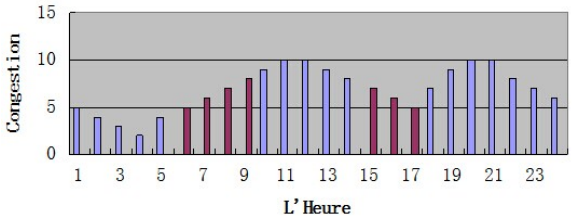


FIGURE 6 – Mesure la congestionné d'un station

Face à des problèmes complexes, les grands entreprises commence utilise les techniques de Fouille de données. Ce technique peut aide l'entreprise faire les décision plus vite et plus précis.

De ce faire, en Juillet 2013, CMCC a lancé ce projet avec quatre laboratoires dans trois université, ils sont Tsinghua University, Shandong University et University of Electronic Science and Technology of China. Le projet inclure trois partiel: Fouille de données, gérés le Clound plateforme et modélisation de l'information dans le système.

2.4 Introduction du laboratoire

De 20 Avril 2014 à 20 Juillet 2014, je fait mon stage chez laboratoire of Next Generation Network Technology & Application (NGN) 7. C'est d'un subordonné de Research Institute of Network And Human-Machine Speech Comunication, Département Ingénierie électronique, Tsinghua University. Le laboratoire se trouve dans la ROHM bâtiment.

FIGURE 7 – Logo NGN

Le principaux axes de recherche sont Théorie des réseaux, Architecture de l'Internet, Traitement de l'information Internet, La recherche dans le domaine de la sécurité Internet, Sentiment analyse, Information hiding, etc.

Mon tuteur professionnel est M. Yongfeng HUANG, vice-directeur de la laboratoire NGN. Dans le laboratoire, il y a cinq groupe, chaque groupe a un docteur et son sujet. dans notre groupes, il y a trois personnes, un étudiant de premier année docteur, un étudiant de M1, et une étudiante de Licence troisième année. On utilise R et Rstudio, et Hadoop aussi.

2.5 Objectif du projet

Dans cet article, nous avons d'abord présente le réseau communication mobile, ensuite je vais décrire l'état de l'optimisation du réseau. Enfin je présente la mise en place de notre programme de recherche.

3 Introduction de l'industrie de la télécommunication

3.1 L'evolution des normes de téléphonie mobile

Depuis 1984, il y a déjà plusieurs standards ont été utilisé par les opérateur dans le monde entier. Voici un tableau de différentes standards mobile en Europe et ses paramétrés 1.

Génération	Acronyme	Description	Débit
1G	Radiocom 2000	Échanges de type voix uniquement	analogique
2G	GSM	Échanges de type voix uniquement	9,05 kbps
2,5G	GPRS	Échange de données sauf voix	171,2 kbps / 50 kbps / 17,9 kbps
3G	UMTS	Voix + données	144 kbps rurale, 384 kbps urbaine, 1,9 Mbps point fixe / -
3.5G ou 3G+ ou H	HSPA	Évolution de l'UMTS	14,4 Mbps / 3,6 Mbps / -
4G	LTE	Long Term Evolution (Données)	150 Mbps / 40 Mbps / -
4G	LTE-Advanced	Long Term Evolution Advanced (Données+voix)	1 Gbps à l'arrêt, 100 Mbps en mouvement / - / -

TABLE 1 – Les différentes générations de téléphonie mobile en Europe

3.1.1 La premier génération

En télécommunication, 1G est la premier génération des standards pour la téléphonie mobile, il s'agit de la première apparition du réseaux de téléphonie mobile, 1G sont des réseaux analogiques, peut échanges de type voix uniquement.

3.1.2 La deuxième génération

2G, la technologie de téléphonie sans fil de deuxième génération, la différence entre le réseaux 1G et 2G est: le signaux radio sur les réseaux 1G sont analogiques, et celle de 2G sont numériques.

Systèmes 2G ont été significativement plus efficaces du spectre permettant de bien plus grand taux de pénétration du téléphone mobile, en plus les données vocales numériques peuvent être compressées et multiplexées beaucoup plus efficacement que les codages de la voix analogique grâce à l'utilisation de codecs différents, ce qui permet plus d'appels à transmettre dans la même quantité de bande passante radio. Et 2G présenté premier foi les services de données pour mobile. Les Technologie 2G permettent les divers réseaux de téléphonie mobile de utiliser des services tels que le SMS et MMS. Tous les message de texte envoyés au delà de 2G sont chiffrés numériquement, ce qui permet le transfert de données de telle sorte que seul le destinataire peut recevoir et lire.

Réseaux 2G ont été construits principalement pour les services téléphoniques et de transmission de données lent (défini dans les documents de spécifications IMT-2000).

Réseaux 2,5G, on le qualifie souvent de le General packet Radio Service ou GPRS, est une norme pour la téléphonie mobile dérivée du GSM et complémentaire de celui-ci, permettant un débit de données plus élevé. Le 2,5 indique que c'est une technologie à mi-chemin entre le GSM (deuxième génération) et l'UMTS (troisième génération). Le GPRS est une extension du protocole GSM : il ajoute par rapport à ce dernier la transmission par paquets. Cette méthode est plus adaptée à la transmission des données. En effet, les ressources ne sont allouées que lorsque des données sont échangées, contrairement

au mode « circuit » en GSM où un circuit est établi - et les ressources associées - pour toute la durée de la communication. Le GPRS a ensuite évolué au début des années 2000 vers la norme EDGE également optimisée pour transférer des données et qui utilise les mêmes antennes et les mêmes fréquences radio.

3.1.3 La troisième génération

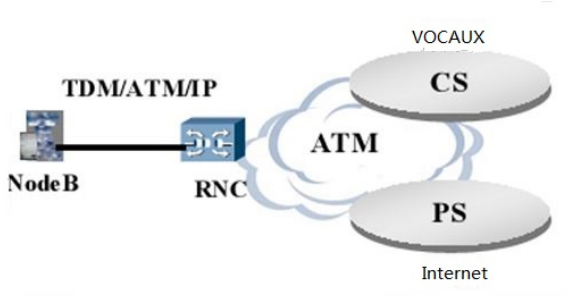
La troisième génération (3G) des normes de téléphonie mobile. Elle est représentée principalement par W-CDMAmm, CDMA2000, TD-SCDMA et WiMAX. Elle permettant des débits de 2 à 42 Mb/s qui sont bien plus rapides qu'avec la génération précédente. Grâce à l'utilisation des règles de classement utilisateur, et les bandes de fréquences supérieures rendant la capacité du réseau augmenter.

Dans les différentes standard 3G et ses prédécesseur, ils utilisent le domaine CS (Circuit Switch) pour les services vocaux, et le domaine PS (Packet Switch) s'occupe des services de données 8(a).

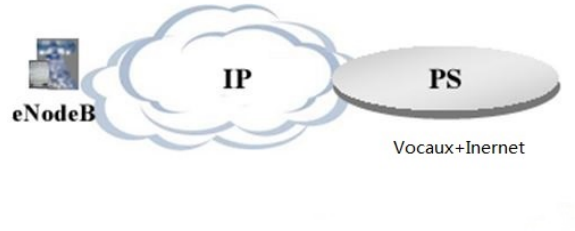
3.1.4 La quatrième génération

La quatrième génération des standards pour la téléphonie mobile, succédant à la 2G et la 3G, en théorie, elle permet de transmette de données à des débits supérieur à 100 Mb/s.

Une des particularité de la 4G est sa EPC (Evolved Packet Core) est basé sur IP, et il n'y a plus de mode commuté (le 'Circuit Switched Domain' qui s'occupe le service vocaux dans les standard précédant), ce qui signifie que les services vocaux transmis sur l'internet 8(b).



(a) Réseau 3G et ses prédécesseur



(b) Réseau 4G

FIGURE 8 – Structure des réseaux

3.2 Le réseau LTE

Le LTE (Long Term Evolution) est l'evolution la plus récent des normes de CDMA 2000, TD-SCDMA, GSM. La norme LTE. La technologie LTE été considérée comme une norme de troisième génération '3.9G', et la 'vraie 4G', appelée LTE-Advanced été reconnu par l'UIT comme une technologie 4G en 2010.
LTE a deux branche: LTE-FDD (Frequency-Division Duplex Long Term Evolution)et LTE-TDD, (Time Division Duplex Long Term Evolution)les deux standards sont similaire 9. En 2011-2012, les réseaux LTE-TDD sont commercialisés sous l'appellation 4G par CMCC un Chine.

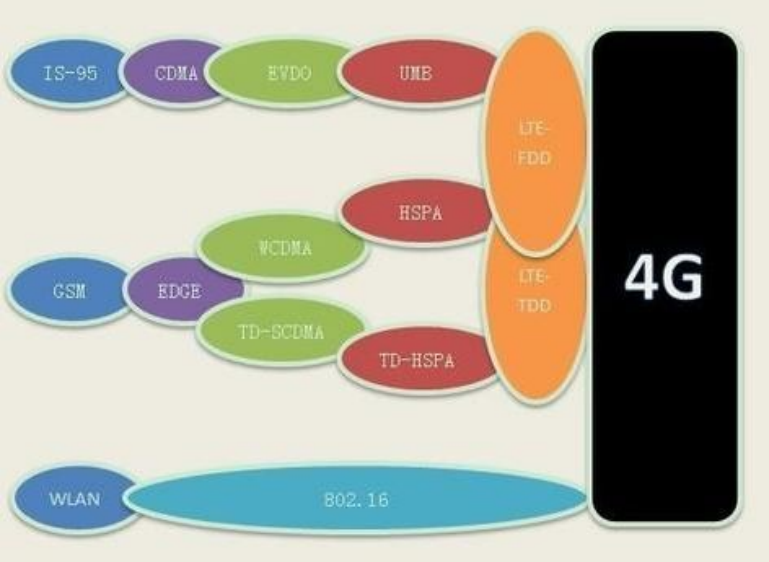


FIGURE 9 – l'evolution des standard

3.2.1 La structure du réseau LTE

Le réseau 4G contient 2 partie: eNodeB (le station), EPC (Evolved Packet Core) qui contient MME, S-GW, P-GW et HSS [10] [2].

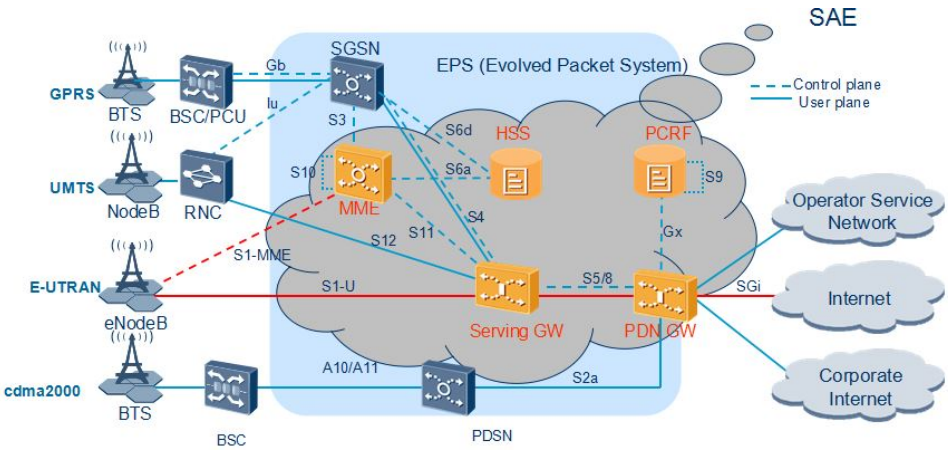


FIGURE 10 – la structure du réseau

Part	Fonction
MME	L'authentification des utilisateurs et la gestion des clés, Cryptage de la couche NAS, Gestion de la liste TA, Sélection P-GW ou S-GW
Service Gateway	Compression d'en-tête IP, Routage de paquets et la transmission, La commutation entre eNB, Facturation des utilisateurs porteur
PDN Gateway	L'allocation des adresses IP de UE, l'accès aux fonction de gestion de réseau externes, Facturation en service
HSS(Home Subscriber Service)	Stockée données de utilisateur associées au service
PCRF	Roaming

TABLE 2 – la fonction du chaque partie

Entre deux E-UTRAN, il y a l'interface X2, l'interface S-11 se trouve entre S-GW et MME, E-UTRAN et S-GW échange les données par l'interface S1-U 11et il échange les donnée par l'interface S1-AP avec MME 11, MME et HSS utilise l'interface S6A, et l'interface S5/8 entre S-GW et P-GW, Gx entre PCRF et P-GW. En mettant des capteur en les interfaces, les opérateurs et les fournisseurs d'équipement peuvent collecter les données de signalisation, et utilisent ces données pour trouver les défauts du système.

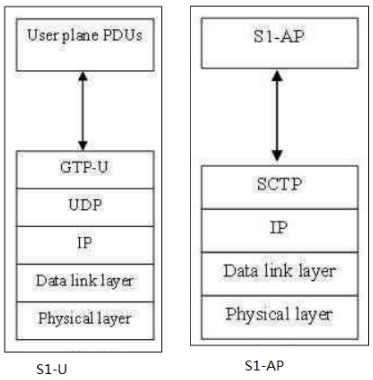


FIGURE 11 – Interface S1

4 Introduction des données

Après quelque semaines de négociation avec les employés de différents départements de la CMCC, ils nous ont fourni deux versions de données, et ses spécifications du format. Nous avons trouvé que CMCC n'a pas de accès direct aux données, et la spécification fourni par CMCC n'est pas correct, et il y a des erreur dans les données fourni par les fournisseur d'équipement.

Ils nous ont envoyé 11 dossiers, chaque dossier correspond à un service. les services sont 'rtsp', 'dns','mail', 'ftp', 'http-wap', 'mms', 'p2p', 'realtimecom', 'VoIP' et les données de signalisation entre E-UTRAN et MME 'S1AP-NAS'12.

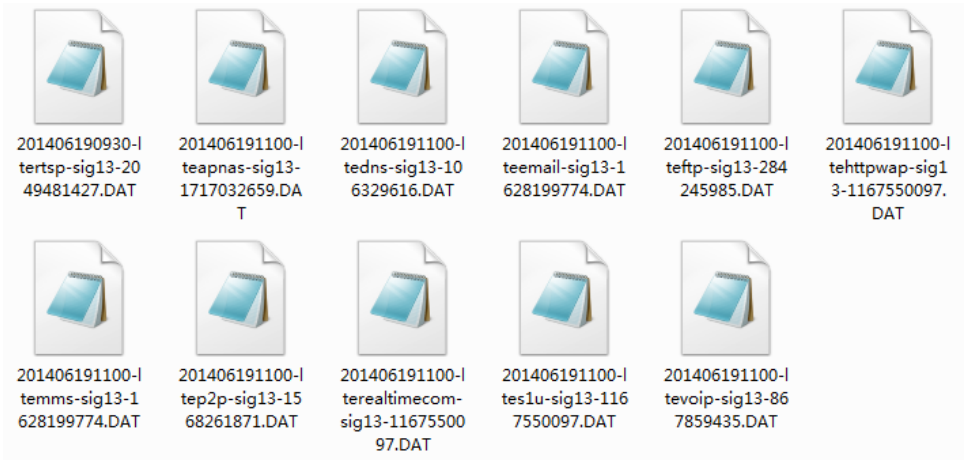


FIGURE 12 – les dossiers de données

Et nous avons trouvé que pour les services comme 'VoIP' et 'RTSP', ils sont très peu de données 3. Donc nous avons décidé de utiliser le donnée du service 'HTTP'.

L'interface	Nombre de ligne
S1-AP	240
RTSP	35
DNS	272562
Maill	44
FTP	71
HTTP-WAP	50854
MMS	193
P2P	515
Realtimecom	2082
S1U	89759
VoIP	28

TABLE 3 – les dossiers de données

4.1 Prétraitement de données

Dans le dossier de HTTP, il y a 50854 lignes, chaque ligne a 76 attributs 13, les données sont collectent dans 20 minutes.

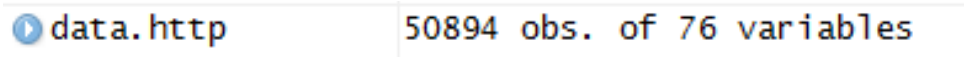


FIGURE 13 – les données du service HTTP

En analysant des données, nous avons trouvé des erreur, et le fournisseur nous a confirmé que ces sont les défaut de leur système 4G. Par exemple, ils sont chiffré les données en utilisant le codage BCD, et Un 14

Deux 15

IMSI	IMEI	MSISDN	M.TMSI	IpType
			o ❖|1	175453935
				1
				1
				1
				1
				1
❖w hw2		hQp❖ p❖|	1	176899993
			H❖ ❖|1	176918718

FIGURE 14 – erreur du codage BCD

UpPac	DownPac	UpDisPac	DownDisPac	UpRePac	DownRePac
6	5	0	0	0	0
10	19	0	0	0	0
5	4	380	0	0	0
8	17	0	7448	40	6384
3	6	0	2128	0	0

FIGURE 15 – Défaut de la système

5 partie1

5.1 Une sous section

On peut mettre des mots en *italique*, en PETITES MAJUSCULES ou en largeur fixe (machine à écrire).

Voici un deuxième paragraphe avec une formule mathématique simple : $e = mc^2$.

Un troisième avec des « guillemet français ».

5.1.1 Écrire en anglais

Do you speak French? Does anybody here speak french?

5.2 Lites

- Liste classique ;
- un élément ;
- et un autre élément.

1. Une liste numéroté
2. deux
3. trois

Description C'est bien pour des définitions.

Deux Ou pour faire un liste spéciale.

5.3 Références

Voici une référence à l'image de la figure 16 page 18 et une autre vers la partie 6 page 20.

On peut citer un livre [LPP] et on précise les détails à la fin du rapport dans la partie références.

5.4 Note de bas de page

Voici une note[1] de bas de page. Une deuxième[2] déclarée différemment. La même note[2].

5.5 Figure

FIGURE 16 – BlogHiko | taille original

1. Texte de bas de page
2. Il a deux références vers cette note

FIGURE 17 – BlogHiko | 50% de la largeur de la page

6 Citation Wikipédia

LaTeX est un langage et un système de composition de documents créé par Leslie Lamport en 198312. Plus exactement, il s'agit d'une collection de macro-commandes destinées à faciliter l'utilisation du « processeur de texte » TeX de Donald Knuth. Depuis 1993, il est maintenu par le LaTeX3 Project team. La première version utilisée largement, appelée LaTeX2.09, est sortie en 1984. Une révision majeure, appelée LaTeX2 epsilon est sortie en 1991.

Le nom est l'abréviation de Lamport TeX. On écrit souvent LaTeX, le logiciel permettant les mises en forme correspondant au logo.

Du fait de sa relative simplicité, il est devenu la méthode privilégiée d'écriture de documents scientifiques employant TeX. Il est particulièrement utilisé dans les domaines techniques et scientifiques pour la production de documents de taille moyenne ou importante (thèse ou livre, par exemple). Néanmoins, il peut aussi être employé pour générer des documents de types variés (par exemple, des lettres, ou des transparents).

Conclusion

Pour conclure, avec LaTeX on obtient un rendu impeccable mais il faut s'investir pour le prendre en main.

Références

[REF] auteur. *titre*. édition, année.

[LPP] Rolland. *LaTeX par la pratique*. O'Reilly, 1999.