

Comparaison des méthodes de classification pour l'identification des noeuds importants dans les graphes dynamiques

Marwan Ghanem

▶ **To cite this version:**

Marwan Ghanem. Comparaison des méthodes de classification pour l'identification des noeuds importants dans les graphes dynamiques. Rencontres jeunes chercheurs en RI, Mar 2019, Lyon, France. hal-02085267

HAL Id: hal-02085267
https://hal.science/hal-02085267

Submitted on 30 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Comparaison des méthodes de classification pour l'identification des nœuds importants dans les graphes dynamiques.

Marwan Ghanem

Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6,
F-75005 Paris, France
marwan.ghanem@lip6.fr

RÉSUME. *De nos jours, nous nous intéressons à la détection d'entités importantes, ceci peut être des mots-clés importants dans un document ou Twitter, ou des individus importants dans un réseau de mouvement. Nous pouvons modéliser ces données sous la forme d'un graphe dynamique et utiliser des métriques de centralité telle que la centralité de proximité temporelle. Malheureusement, cela peut être coûteux. Dans ce travail, nous comparons la précision de plusieurs méthodes de classification supervisée, les unes par rapport aux autres, à la détection de ces nœuds importants. Sur seize jeux de données de natures différentes, nous montrons que ces méthodes réussissent à différencier les nœuds importants de nœuds insignifiants. Nous montrons également que prendre en compte la nature des données diminue la qualité de résultats. Enfin, nous examinons le temps du calcul de chacune de ces méthodes contre le temps du calcul de méthodes exact.*

ABSTRACT. *Nowadays, researchers are interested in the detection of important entities in networks; this can be keywords in a document and Twitter or more even super-spreaders in a movement network. One natural way to detect these entities is to use graph theory; each entity is represented by a node in the graph. Afterward, centrality metrics such as Temporal closeness can be applied to detect the important nodes. Nevertheless, this can be computationally expensive. In this work, we examine three basic characteristics that we consider as the basic blocks of Temporal closeness. We utilize those characteristics to show that classifiers are capable to classify the nodes. In addition, we show that taking into account the dataset's nature does not necessarily produce better models. Finally, we compare the computational time of these models against that of Temporal Closeness.*

MOTS-CLES : *Graphe dynamique, Classification, importance, centralité*

KEYWORDS: *Centrality, Classification, Dynamic graphs, Importance*

1. Introduction

Une des applications du traitement automatique du langage naturel (TALN) est l'estimation de l'importance des mots dans un document, afin d'extraire, par exemple, des mots-clés. Ceux-ci peuvent être utile pour classifier des documents ou encore aider un lecteur à décider s'il veut lire un document (Dias *et al.*, 2018). Ils sont généralement sélectionnés par les auteurs, avoir donc des systèmes automatiques pour extraire ces mots à partir du texte serait plus pratique.

Une façon d'extraire les mots-clés des documents est d'utiliser la théorie des graphes. Formellement, un graphe est composé d'un ensemble de nœuds et de liens, représentant les interactions entre les nœuds. Il existe plusieurs manières de présenter un document sous la forme d'un graphe. Par exemple, les nœuds du graphe sont les mots du document et un lien entre deux nœuds représente deux mots adjacents dans le texte (Schluter, 2014; Palshikar, 2007). Plusieurs approches ont été proposées pour détecter les mots-clés avec cette représentation. Dans (Alqaryouti *et al.*, 2018), les auteurs utilisent les graphlets pour détecter les mots-clés. (Palshikar, 2007; Erkan and Radev, 2011; Schluter, 2014) quant à eux, calculent l'importance, c'est-à-dire la centralité, des nœuds dans le graphe.

À notre connaissance les méthodes existantes utilisent des centralités statiques (la centralité de proximité ; la centralité d'intermédiaire ; la centralité de degré). Ils ne prennent pas en considération l'aspect temporel. Or, (Braha and Bar-Yam, 2008; Pan and Saramäki, 2011) montrent que la prise en compte de l'aspect temporel améliore fortement la précision. En effet, un mot en début de texte n'a pas forcément la même importance qu'un mot en fin de texte. Aussi, dans le contexte de réseaux sociaux, un utilisateur ne peut pas diffuser une information avant qu'il ne la reçoive de quelqu'un d'autre. Ici, pour cette raison, nous considérons uniquement des measures des centralités temporelles. En conséquence, chaque nœud a une valeur d'importance qui évolue au cours du temps, autrement dit une valeur de centralité à chaque pas du temps. Ce calcul a une complexité non négligeable qui nous empêche d'appliquer ces méthodes aux données trop volumineuses.

Pour réduire le temps de calcul, (Ghanem *et al.*, 2018) proposent d'utiliser des métriques simples à calculer afin de détecter les nœuds ayant une valeur de centralité de proximité temporelle élevée. Leur approche consiste à classer les nœuds en fonction de deux métriques. Si un nœud est considéré important par ces deux métriques, le nœud est - en théorie - important au sens de la centralité de proximité temporelle. Dans ce travail, nous considérons certaines de ces métriques mais au lieu de les associer de façons heuristiques nous utilisons des méthodes de classification qui exploitent ces métriques en tant que composantes. Nous comparons ces méthodes entre elles et grâce à cette comparaison, nous observons que ces méthodes réussissent à identifier les nœuds importants mais aussi qu'elles sont plus efficaces quand les données d'apprentissage sont de nature variée.

Cet article est organisé de la façon suivante : tout d'abord, nous présentons la centralité qu'on considère, ainsi que les composantes qu'on considère avec les méthodes

de classification (Section 2). Ensuite nous présentons les jeux de données que nous étudions (Section 3). Enfin, nous présentons les résultats de différente méthodes de classification (Section 4), avant de conclure (Section 5).

2. Contexte

Dans cette section, nous présentons la centralité temporelle que nous considérons dans ce travail, ainsi que les trois composantes que nous utilisons avec les algorithmes de classification.

2.1. *Centralité de proximité temporelle*

Dans (Magnien and Tarissan, 2015), les auteurs représentent les données sous la forme d'un réseau dynamique $G = (V, E)$ où V est l'ensemble des nœuds , où E est l'ensemble des liens de la forme (u, v, t) tel que $u, v \in V$ et où t est une étiquette temporelle. Avec cette représentation, un chemin temporel de v_0 à v_{k+1} consiste en :

– un temps de départ t_s

– une séquence de liens $(v_0, v_1, t_0), (v_1, v_2, t_1), \ldots, (v_i, v_{i+1}, t_i) \ldots, (v_k, v_{k+1}, t_k)$

tels que :

– $t_0 > t_s$

– $t_i < t_{i+1} \forall i, i = 0..k-1$

La durée de ce chemin est égale à $t_k - t_s$. Ce chemin est considéré comme étant le plus court chemin s'il a la plus courte durée parmi tous les chemins de v_0 à v_{k+1} qui commencent à t_s. On note $d_{t_s}(v_0, v_{k+1})$ la durée correspondante, appellée la distance temporelle. S'il n'y a pas de chemin de v_0 à v_{k+1} qui commence à t_s, nous considérons alors que $d_{t_s}(v_0, v_{k+1}) = \infty$. Les auteurs définissent la *temporal closeness* d'un nœud u à l'instant t par :

$$C_t(u) = \sum_{v \neq u} \frac{1}{d_t(u, v)}.$$

Notons que cette définition nécessite que la centralité soit calculée à chaque pas de temps t, ce qui est très coûteux en terme de calcul. C'est pourquoi par la suite, nous calculons la centralité de proximité temporelle de chaque nœud toutes les $\mathcal{I}$ secondes seulement. À chacun de ces instants, les nœuds sont classés entre eux afin d'identifier les nœuds qui sont importants à chaque instant. Pour cela, nous définissons une plage de rangs pour laquelle nous estimons que les nœuds sont importants. Cette plage correspond aux 25% des rangs les plus élevés. Autrement dit, pour un réseau avec n nœuds, cela concerne les nœuds ayant un rang supérieur à $\lfloor n * 0.75 \rfloor$. Puis, nous définissons une valeur (Dur_{top}) qui correspond à la durée pendant laquelle un nœud est important. Nous considérons qu'un nœud est présent dans cette plage entre l'instant

où nous calculons la centralité jusqu'à l'instant suivant. Plus formellement , soit un nœud u et $R(u) = (r_i)_{i=1...k}$ une séquence de rangs de u, nous définissons Dur_{top} la valeur :

$$Dur_{top}(u) = \mathcal{I} \cdot |\{i \leq k-1, r_i \geq \lfloor n * 0.75 \rfloor\}| .$$

Finalement, les 25% de nœuds qui ont les valeurs les plus élevées sont considérés comme importants. Pour identifier ces nœuds, nous considérons trois composantes , détaillées ci-dessous.

2.2. *Composantes*

Parmi toutes les composantes possibles, nous nous concentrons sur trois composantes que nous pensons être les éléments de base de la centralité proximité temporelle.

Pour chaque nœud u, nous considérons :

– Durée d'activité : la différence entre l'étiquette temporelle de la dernière et première interaction de u.

– Nombre de liens : le nombre de liens auxquels u participe.

– Fréquence : la différence moyenne entre l'étiquette temporelle de chaque paire de liens consécutifs.

Nous considérons ces trois composantes car un nœud qui est longtemps actif et qui interagit souvent avec d'autres nœuds, est proche dans le temps des autres noeuds. Nous classons ensuite les nœuds entre eux en fonction de ces trois composantes [1]. Il en résulte trois valeurs continues que nous utilisons pour classifieur si un nœud est important ou insignifiant. Dans ce travail, nous étudions sept algorithmes de classification : *k-nearest neighbors* , *Linear SVM*, *Radial basis function kernel*, *Neural Network*, *Adaboost*, *Naive Bayes* et *Random Forest*.

3. Données

Pour mieux tester les méthodes de classification, nous avons considéré seize jeux de données de natures différentes : échange de courriels, co-occurences, réseaux sociaux et réseaux de contacts.

– Échange de courriel :

- Enron (Shetty and Adibi, 2005) : contient $47\,088$ courriels échangés entre 151 employés pendant trois ans. Pour chaque courriel nous avons l'expéditeur, le destinataire et la date d'expédition.

1. Les nœuds sont triés par ordre croissant.

- Radoslaw (Michalski *et al.*, 2011) : contient $82\,876$ courriels échangés entre 168 employés d'une compagnie de taille moyenne pendant neuf mois en 2010. Pour chaque courriel, nous avons l'expéditeur, le destinataire et la date d'expédition.

- DNC (Kunegis, 2013) : une fuite de données qui contient les courriels échangés entre les membres du Comité National-Démocrate (DNC). Ce jeu de données contient $39\,264$ courriels échangés de septembre 2013 à mai 2016. Nous n'avons gardé que l'expéditeur, le destinataire et la date d'expédition.

- UC (Opsahl and Panzarasa, 2009) : $60\,000$ messages échangés entre les étudiants de l'université de Californie, Irvine.

– Réseaux sociaux

- Facebook (Viswanath *et al.*, 2009) : contient l'activité de $8\,977$ comptes Facebook pendant 1 an, du 31 décembre 2015 au 31 décembre 2016. Un lien (u, v, t) signifie que u a écrit sur le mur de v à l'instant t. Les $8\,977$ utilisateurs ont écrit $66\,153$ posts.

- Bitcoins (Kumar *et al.*, 2016) : après s'être échangé des bitcoins, deux utilisateurs peuvent évaluer le niveau de confiance de l'autre. Chaque lien (u, v, t) signifie que u a évalué v à l'instant t. Le jeu de données contient 3783 nœuds et $24\,186$ liens.

- BitcoinsOTC (Kumar *et al.*, 2016) : un jeu de données de même nature que Bitcoins, mais extrait d'une autre plateforme d'échange de bitcoin. Le jeu de données contient $5\,881$ nœuds et $35\,592$ liens.

- Fb-forum (Opsahl, 2011) : contient $33,720$ messages diffusés par 889 étudiant l'université de Californie, Irvine. Chaque message diffusé équivaut à poster un message sur le mur d'un utilisateur sur Facebook.

– Co-occurrence

- HashTags : enregistre les tweets de comptes associés à des groupes terroristes. Chaque nœud représente un hashtag et deux hashtags sont liés s'ils sont cités dans un même tweet. Par conséquent, un tweet avec plusieurs hashtags génère plusieurs liens. Le jeu de données contient 3048 hashtags et $100\,429$ liens pendant 22 jours.

- Articles : chaque article publié contient un ensemble de mots-clés. Dans ce jeu de données chaque nœud représente un mot-clé. Chaque lien (u, v, t) représente deux mots clés qui apparaissent sur le même article à l'instant t. Le jeu de données contient 2902 nœuds et $571\,877$ liens pendant 15 ans.

- LeMonde (Grasland *et al.*, 2016) : les articles du journal Le Monde sont collectés en utilisant un flux RSS. Les noms de pays apparaissant dans les titres des articles sont récupérés. Chaque nœud représente un pays et chaque lien (u, v, t) représente deux pays u,v qui apparaissent sur le même titre au temps t. Il contient 144 nœuds et 4373 liens sur une durée d'environ une année et demie.

- Herald (Grasland *et al.*, 2016) : à partir du journal The New Zealand Herald. Les noms de pays apparaissant dans les titres des articles sont récupérés. Il contient 182 nœuds et 14019 liens sur une durée d'environ une année et demie.

– Réseaux de contact :

- RollerNet (Tournoux *et al.*, 2009) : représente les contacts physiques entre 62 participants lors d'une sortie en rollers à Paris en août 2006. Il contient 403 834 contacts entre les participants répartis sur environ trois heures.

- Reality (Opsahl and Panzarasa, 2009) : réprésente les contacts de proximité de 96 étudiants à partir de leurs téléphones portables. De septembre 2014 à mai 2015, 1 063 063 liens ont été enregistrés fois.

- Taxi (Bracciale *et al.*, 2014) : ce jeu de données contient le mouvement de 131 taxi. Chaque lien (u, v, t) correspond à deux taxis qui sont proches. Sur une durée de huit heures, 85 732 liens ont été enregistrés.

- Primary (Gemmetto *et al.*, 2014) : il contient 60 611 contacts entre 242 étudiants et enseignants sur une durée de huit heures.

Les statistiques de ces jeux de données sont détaillées dans le tableau 1.

Datasets	$\|V\|$	$\|E\|$	**Durée**	**Nature**
Enron	151	47 088	3 ans	Courriel
Radoslaw	168	82 876	9 mois	Courriel
DNC	1891	39 264	2, 5 ans	Courriel
UC	1899	59 835	6 mois	Courriel
Facebook	8 977	66 153	1 ans	Sociaux
Bitcoins	3 873	24 186	5, 2 ans	Sociaux
BitcoinsOTC	5 881	35 592	5 ans	Sociaux
Fb-forum	899	33 720	6 mois	Sociaux
HashTags	3 048	100 429	22 jours	Co-occurrence
Articles	2 902	571 877	15 ans	Co-occurrence
LeMonde	144	4 373	1, 5 ans	Co-occurrence
Herald	182	14 019	1, 5 ans	Co-occurrence
RollerNet	62	403 834	3 heures	Contact
Reality	96	1 063 063	9 mois	Contact
Taxi	131	85 732	8 heures	Contact
Primary	242	60 611	8, 6 heures	Contact

Tableau 1 – Nombre de nœuds $|V|$, Nombre de liens $|E|$, Durée, Nature.

4. Expérience

Pour classifier les nœuds, nous considérons trois approches possibles. En premier lieu, pour chaque jeu de données (testing), nous considérons les quinze autres jeux de données comme nos données d'apprentissage (Méthode A). Ceci est différent du cas habituel de 25% de test et 75% d'apprentissage, mais nous considérons que ce cas reste le plus proche du cas pratique. Dans un cas réel, nous pouvons avoir accès à un ensemble de jeux de données avec leur vérité de terrain, et vouloir faire le calcul pour un jeu de données, qui est par exemple trop grand. En conséquence, nous pouvons utiliser les jeux de données avec une vérité de terrain pour développer des modèles de

classification. En deuxième lieu, nous tenons compte de la nature des jeux de données. Autrement dit, pour chaque jeu de données, nous considérons les trois jeux de données de même nature en tant que données d'apprentissage (Méthode B). Enfin, nous considérons des données d'apprentissage de natures différentes de celles de validation (Méthode C).

Pour pouvoir comparer ces différentes approches, nous étudions le F1-score de la classe importante. Nous rappelons que le F1-score est defini comme la moyenne harmonique de la précision et du rappel :

$$2 \cdot \frac{\text{précision} \cdot \text{rappel}}{\text{précision} + \text{rappel}}$$

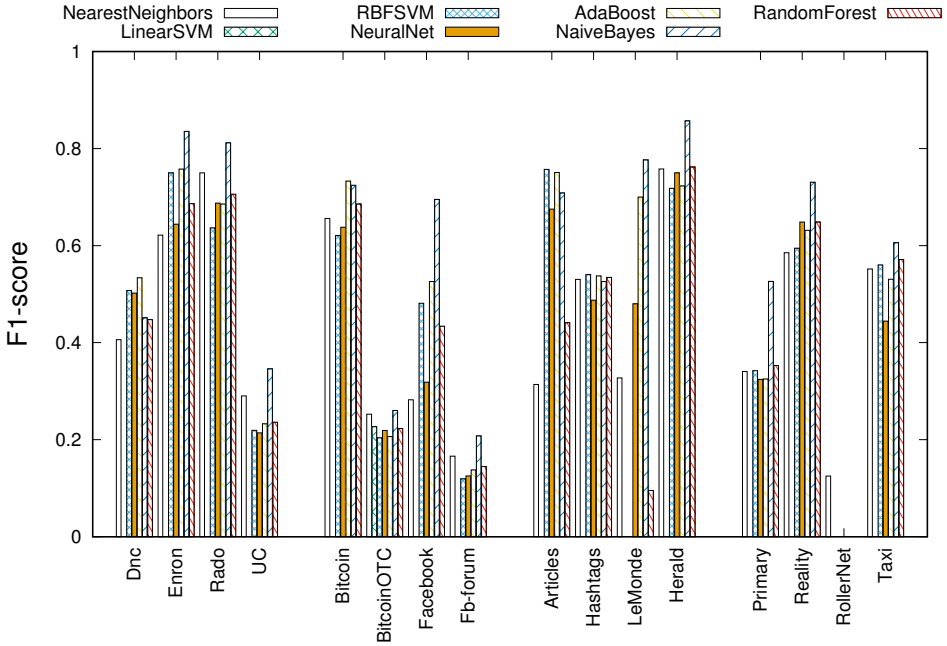


Figure 1 – F1-score pour chaque jeu de données et pour chaque classifieur (Méthode A).

La Figure 1 présente les F1-scores de la méthode A pour chaque jeu de données et pour chaque classifieur. Nous pouvons observer dans certains cas, *e.g.* Rollernet, que les classifieurs n'arrivent pas à identifier les nœuds importants. Ceci est normal, en effet dans ces jeux de données[2] la notion d'importance est absente (Ghanem, 2018), *i.e.* tous les nœuds ont la même importance, ce qui rend la classification impossible.

2. RollerNet , UC , Primary, Fb-forum, BitcoinOTC

Pour les autres jeux de données, nous pouvons observer que les résultats ne varient pas entre les différents classifieurs. Notons que la moyenne des F1-score est de $0,42$ quand nous ne considérons que les jeux de données ayant une notion d'importance. En outre, *Naive Bayes* produit, en moyenne, le meilleur résultat. Il faut savoir, que la moyenne augmente à $0,6$ quand nous excluons les jeux de données sans notion d'importance des données d'apprentissage. Autrement dit, nous considérons seulement 10 jeux de données pour l'apprentissage.

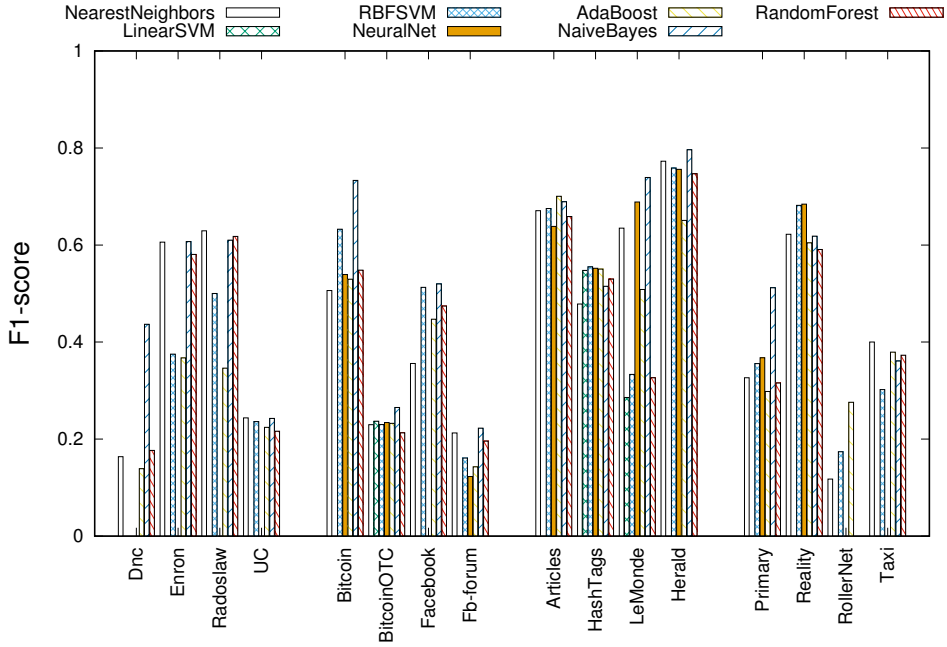


Figure 2 – F1-score pour chaque classifieur (Méthode B).

Pour l'étape suivante, nous étudions comment les résultats varient quand nous prenons en compte la nature des jeux de données. La Figure 2 présente les F1-scores de la méthode B pour chaque jeu de données et pour chaque classifieur, lorsque le jeu de données de validation et ceux d'apprentissage sont de même nature. Nous pouvons remarquer que les résultats sont globalement moins bien que ceux de la méthode A. Par contre, il faut noter que dans le cas de RollerNet, certaines méthodes de classification réussissent à classer les nœuds importants. Nous notons que la moyenne des F1-scores diminue par rapport à celle de la méthode A ($0,42$ à $0,32$). Ce qui était previsible, car nous avons considéré moins de données d'apprentissage.

Enfin, nous considérons la même approche mais avec des données d'apprentissage de natures différentes (méthode C). Pour chaque jeu de données de validation, nous

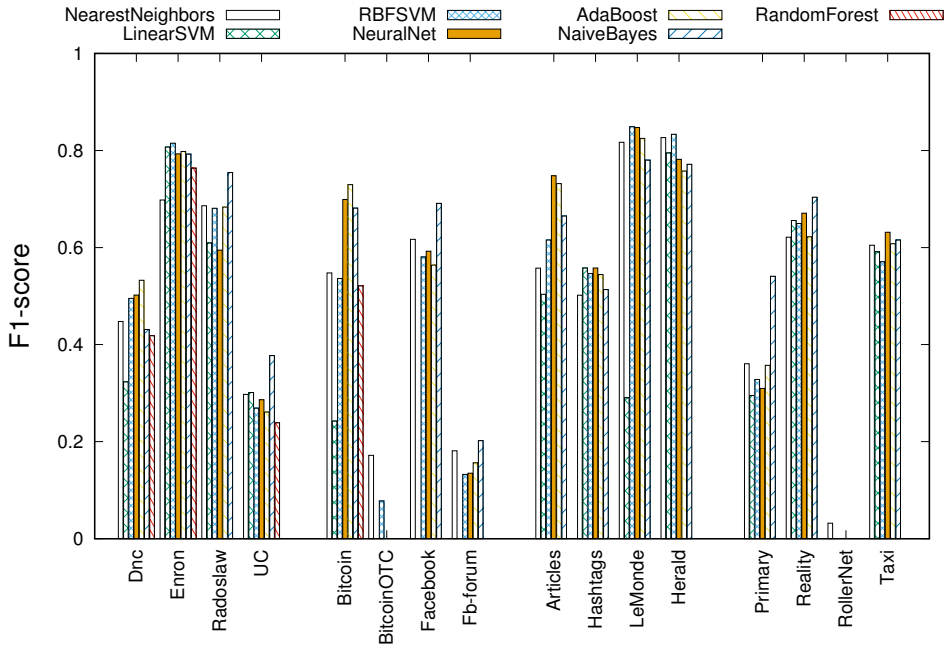


Figure 3 – F1-score pour chaque jeu de données et pour chaque classifieur (Méthode C).

prenons trois jeux de données d'apprentissage de natures différentes[3]. Nous avons considéré plusieurs combinaisons différentes, et nous étudions la moyenne des F1-scores. La Figure 3 présente la moyenne des F1-scores pour chaque méthode de classification et pour chaque jeu de données. Nous commençons par les jeux de données de courriel, nous pouvons observer que les F1-scores restent relativement proche par rapport à ceux obtenus avec la méthode A. On peut noter que *Native Bayes* a un F1-score de $0,8$ pour Enron. Pour les jeux de données de co-occurrence, nous pouvons observer que les résultats restent relativement élevés. Enfin, nous observons pour les jeux de données des réseaux sociaux et mouvements que les résultats diminuent pour certains classifieurs et augmentent pour d'autres. Nous notons que la moyenne de F1-scores augumente par rapport à celle de la méthode A : $0,52$ par rapport à $0,42$. Toutefois, la moyenne est beaucoup plus élevée par rapport à celle de la méthode B : $0,52$ par rapport $0,32$. Nous pouvons constater qu'en choisissant des données d'apprentissage de natures différentes, nous arrivons à obtenir de meilleurs résultats. De plus, éviter de prendre des jeux de données sans notion d'importance comme données d'apprentissage augmente la qualité de classification. Finalement, en prenant moins de

3. Le même nombre de jeux de données que celui de méthode B.

données, la moyenne de F1-scores diminue très peu. Le Tableau 2 présente la moyenne de F1-scores pour les différentes méthodes.

Méthode	Nombre de jeux de données d'apprentissage	Moyenne de F1-score
A	15	0, 42
A(bis)	10	0, 60
B	3 (Même nature)	0, 32
C	3 (Nature différente)	0, 52

Tableau 2 – Méthode, Nombre de jeux de données d'apprentissage, Moyenne de F1-score

Ensuite, nous étudions le temps de calcul de chacune de ces méthodes par rapport au temps du calcul complet de la centralité de proximité temporelle. Notons que, pour la méthode traditionelle, le temps de calcul dépend de la taille du jeu de données et peut donc être de l'ordre de la seconde comme de plusieurs jours. A contrario, les méthodes de classification présentées dans ce travail prennent environ une seconde quelque soit la taille du jeu de données.



Figure 4 – Nuage des noms des pays mentionnés dans le jeu de données LeMonde. La taille est proportionnelle à l'importance du mot. La couleur vert indique l'importance du mot d'après le classifieur. Vert : Importante ; Rouge : Sans importance.

Enfin, nous observons les mots indiqués comme étant que importants par la centralité de proximité temporelle ainsi que *NaiveBayes* dans le jeu de données LeMonde. Dans la Figure 4, nous pouvons observer que les mots[4] importants d'après la centralité

4. Les noms de pays ont été raccourcis pour plus de visibilité.



(a) 22 juin 2014 (b) 10 novembre 2014 (c) 2 décembre 2014

Figure 5 – 3 Nuages des noms des pays mentionnés dans le jeu de données LeMonde pour 3 jours différents. La taille est proportionnelle à l'importance du mot.

de proximité temporelle sont aussi souvent importants d'après *NaiveBayes*. Notons, que FRA et USA sont les deux mots les plus importants, cela n'est guère surprenant de la part d'un journal français.

Grâce à la centralité de proximité temporelle, nous étudions l'évolution d'importance pour les mots globalement importants. La figure 5 présente trois nuages de noms pour trois instants à trois jours différents. Dans les trois jours, nous pouvons observer que certains pays sont toujours importants comme la France (FRA) ou les États-Unis (USA). Par contre, nous pouvons observer que le 22 juin 2014 (Figure 5a) l'Égypte est particulièrement importante, ceci est lié à une visite du secrétaire d'État américain. Le 10 novembre 2014 (Figure 5b), nous observons que la Corée du sud (KOR) est particulièrement importante (ce jour). Ceci est lié à un article mentionnant la Corée du sud et la Chine. Enfin, le 22 décembre 2014 (Figure 5c) la république de Cuba est la plus importante, et ceci est lié à un accident de bus transportant des personnes françaises. Dans tous ces cas, nous pouvons voir qu'un pays devient important quant il est mentionné conjointement à la Chine ou la France. Sans prendre en compte l'aspect temporel, ces phénomènes s'estompent et deviennent difficiles à détecter.

5. Discussion et conclusion

Dans cet article, nous nous sommes intéressés à une métrique de centralité qui peut être utilisée pour détecter les mots-clés dans un document ou encore un diffuseur dans un réseau social. Cette métrique est relativement coûteuse et en conséquence l'analyse de jeux de données très volumineux n'est pas envisageable. Pour cette raison, nous avons réalisé une étude statistique simple sur différentes méthodes de classification. Nous avons montré que ces méthodes réussissent à détecter les nœuds importants. Aussi, nous avons observé que prendre en compte la nature du jeu de données ne produit pas forcément de meilleur résultats. Pour ces raisons, nous pensons que des techniques de classification peuvent être utilisées pour extraire les mots-clés dans des documents de grand taille, ou encore plus généralement pour détecter les entités centrales dans un graphe.

Ce travail ouvre plusieurs perspectives. En premier, nous pouvons analyser des documents grâce à ces modèles et valider les résultats par rapport aux méthodes basées sur la théorie des graphes mais qui utilisent des métriques de centralité statique. Enfin, nous avons vu que certains jeux de données n'ont pas une notion claire d'importance et que ces jeux de données diminuent les F1-scores, nous pensons que des méthodes comme l'analyse en composantes principales peuvent être utilisées pour détecter ces jeux de données d'une manière systématique. Finalement, il serait intéressant de voir si ce phénomène se retrouve aussi dans les données textuelles.

Remerciements

Ce travail est financé par le programme HET20 FETPROACT 2016-2017 de la Commission européenne, dans le cadre de la subvention 732942 (ODYCCEUS), par l'ANR (Agence nationale de la recherche française), dans le cadre de la subvention ANR-15-CE38-0001 (AlgoDiv), de la de France et son programme FUI21 dans le cadre de la subvention 16010629 (iTRAC).

6. Bibliographie

Alqaryouti O., Khwileh H., Farouk T., Nabhan A., Shaalan K., « Graph-Based Keyword Extraction », *Intelligent Natural Language Processing : Trends and Applications*, Springer, p. 159-172, 2018.

Bracciale L., Bonola M., Loreti P., Bianchi G., Amici R., Rabuffi A., « CRAWDAD dataset roma/taxi (v. 2014-07-17) », , Downloaded from `https://crawdad.org/roma/taxi/20140717`, July, 2014.

Braha D., Bar-Yam Y., « Time-dependent complex networks : dynamic centrality, dynamic motifs, and cycles of social interaction », *in* T. Gross, H. Sayama (eds), *Adaptive networks : Theory, models and applications*, Springer, p. 38-50, 2008.

Dias C.-E., Halbeck D. E., Guigue V., Gallinari P., « RNN et modèle d'attention pour l'apprentissage de profils textuels personnalisés », *CORIA*, 2018.

Erkan G., Radev D. R., « LexRank : Graph-based Lexical Centrality as Salience in Text Summarization », *CoRR*, 2011.

Gemmetto V., Barrat A., Cattuto C., « Mitigation of infectious disease at school : targeted class closure vs school closure. », *BMC infectious diseases*, vol. 14, n^{o} 1, p. 695, December, 2014.

Ghanem M., Les centralités temporelles : étude de l'importance des nœuds dans les réseaux dynamiques, PhD thesis, Sorbonne Université, 2018. Type : Thèse de Doctorat – Soutenue le : 2018-10-05 – Dirigée par : Magnien, Clémence – Encadrée par : TARISSAN Fabien.

Ghanem M., Magnien C., Tarissan F., « How to exploit structural properties of dynamic networks to detect nodes with high temporal closeness », *CTW18 : Cologne-Twente Workshop on Graphs and Combinatorial Optimization 2018)*, Paris, France, 2018.

Grasland C., Lamarche-Perrin R., Loveluck B., Pecout H., « L'agenda géomédiatique international : analyse multidimensionnelle des flux d'actualité. In L'Espace Géographique », *L'Espace géographique*, vol. 45, n° 1, p. 25-43, 2016.

Kumar S., Spezzano F., Subrahmanian V., Faloutsos C., « Edge weight prediction in weighted signed networks », *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, IEEE, p. 221-230, 2016.

Kunegis J., « KONECT – The Koblenz Network Collection », *Proc. Int. Conf. on World Wide Web Companion*, p. 1343-1350, 2013.

Magnien C., Tarissan F., « Time Evolution of the Importance of Nodes in Dynamic Networks », *Proceedings of the International Symposium on Foundations and Applications of Big Data Analytics (FAB), in conjunction with ASONAM, 2015.*, FAB '15, ACM, New York, NY, USA, p. 1200-1207, 2015.

Michalski R., Palus S., Kazienko P., « Matching Organizational Structure and Social Network Extracted from Email Communication », *Lecture Notes in Business Information Processing*, vol. 87, Springer Berlin Heidelberg, p. 197-206, 2011.

Opsahl T., « Triadic closure in two-mode networks : Redefining the global and local clustering coefficients », *Social Networks*, 2011.

Opsahl T., Panzarasa P., « Clustering in Weighted Networks », *Social Networks*, vol. 31, n° 2, p. 155-163, 2009.

Palshikar G. K., « Keyword extraction from a single document using centrality measures », *International Conference on Pattern Recognition and Machine Intelligence*, Springer, p. 503-510, 2007.

Pan R. K., Saramäki J., « Path lengths, correlations, and centrality in temporal networks », *Physical Review E*, vol. 84, n° 1, p. 016105, July, 2011.

Schluter N., « Centrality measures for non-contextual graph-based unsupervised single document keyword extraction », *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, vol. 2, p. 455-460, 2014.

Shetty J., Adibi J., « Discovering important nodes through graph entropy the case of Enron email database », *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05*, ACM Press, New York, New York, USA, p. 74-81, August, 2005.

Tournoux P. U., Leguay J., Dias de Amorim M., Benbadis F., Conan V., Whitbeck J., « The Accordion Phenomenon : Analysis, Characterization, and Impact on DTN Routing », *Proceedings of the 28rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, IEEE, p. 1116-1124, 2009.

Viswanath B., Mislove A., Cha M., Gummadi K. P., « On the Evolution of User Interaction in Facebook », *Proc. Workshop on Online Social Networks*, p. 37-42, 2009.