Liage des données par les systèmes de recommandation intelligents dans une démarche d'optimisation de la qualité des données

Ganda Tandia, Isma Sadoun, Sana Ben Hamida

Liage des données par les systèmes de recommandation intelligents dans une démarche d'optimisation de la qualité des données

Ganda TANDIA Brother France - Université Paris Nanterre - France
Isma SADOUN Brother France - Roissy-en-France
Sana Ben Hamida Université Paris Nanterre - France

Abstract

Une des phases importantes dans une démarche d'optimisation de la qualité des données d'une base est le liage des données. Le liage des données s'intéresse à détecter les descriptions référant au même objet du monde réel (e.g. même personne, même livre) afin de les nettoyer [Saïs and Thomopoulos, 2016]. Ce problème est fréquemment rencontré dans le domaine des ventes indirectes réalisées au travers de revendeurs (cadre de cette étude) retournant des données clients souvent redondantes. L'objectif de ce travail est de fournir un outil performant pour le liage des données basé sur les systèmes de recommandation et l'intelligence artificielle dans un but d'optimisation de la qualité des données. Nous proposons un système hybride combinant un système de recommandation basé sur le contenu avec un système de recommandation basé sur le filtrage collaboratif pour un liage optimal des données clients.

Data Quality Optimization, Recommander systems, Data redundancy detection, Data cleaning

1 Contexte

Les entreprises commerciales faisant appel à des revendeurs font souvent face à un problème de qualité des données concernant leurs bases de données clients liées aux ventes indirectes. C'est le cas de l'entreprise Brother qui a accueilli ce travail. Chaque mois, les revendeurs envoient un rapport qui contient des informations sur les différents clients avec lesquels ils ont pu faire des transactions. Ces rapports contiennent entre autres des informations sur les différents clients. Les données présentes dans ces rapports permettent à Brother de constituer sa base de données client et de réaliser des analyses sur les ventes indirectes menées par leurs revendeurs. Les rapports peuvent être des fichiers xlsx, xls, csv ou json.

Or, chaque revendeur utilise sa propre nomenclature. En plus, un client peut faire des transactions avec différents revendeurs, et apparaitre ainsi dans plusieurs rapports souvent avec des formats différents. Dans de tels cas, il sera considéré comme un nouveau client et va être re-intégré dans la base client. C'est également le cas lorsque des informations sur les coordonnées du client (par exemple l'adresse) changent (i.e. à la suite d'un déménagement). Ces insertions multiples encombrent la base client, rendent les requêtent moins performantes et perturbent le métier. Par ailleurs, ces redondances faussent les rapports d'analyse du service *data analytics*.

Pour améliorer leur efficacité métier, les entreprises dans une telle situation doivent déclencher une démarche d'optimisation de la qualité des données clients par détection et fusion des objets redondants. L'étape de détection des données dupliquées est connue sous le nom *liage* des données [Saïs and Thomopoulos, 201 ou *reconcialiation* des données [Ferrara et al., 2013]. C'est dans dans ce cadre que s'intègre le présent travail. Nous proposons un outil performant pour le liage des données basé sur un système de recommandation hybride. Il combine un système de recommandation basé sur le contenu avec un système de recommandation basé sur le filtrage collaboratif pour un liage optimal des données clients.

2 Concepts de Base

2.1 Le liage de données

Le liage des données (*record linkage*) est étudié depuis plusieurs années essentiellement dans le cadre d'intégration des données. Il a été défini pour la première fois par Newcombde dans un contexte médical [Newcombe et al., 1959]. Son objectif est d'identifier, correspondre et agréger des tuples en double. Actuellement, les cas d'applicaton les plus fréquents se présentent dans le cadre de rachat ou de fusion d'entreprises ou dans la présence de multitude de sources de données comme en bioinformatique ou les librairies digitales.

Selon le contexte, plusieurs modèles de liage de données ont été proposés. La figure 1 présente les étapes définies par Christen en 2008 [Christen, 2008] dans le cadre de fusion de deux bases de données. La première phase dans la majorité des modèles est la préparation des données par nettoyage et standarisation pour faciliter l'étape de comparaison des tuples. La deuxième phase connue sous le nom de *blocking* ou segmentation, consiste à créer des blocks de tuples avec certaines ressemblances syntaxiques, afin de réduire la complexité de l'étape de comparaison. En effet, cette phase permet de réduire le nombre de comparaisons à faire entre les couples de tuples. Cette opération ne se fera donc qu'au sein d'un même block. La pratique la plus connue pour créer ces blocks se base sur le voisinage de certains éléments dans les tuples, definis comme des clés de tri [Hernández and Stolfo, 1998]. Quant à la comparaison des couples des tuples, elle est essentiellement syn-

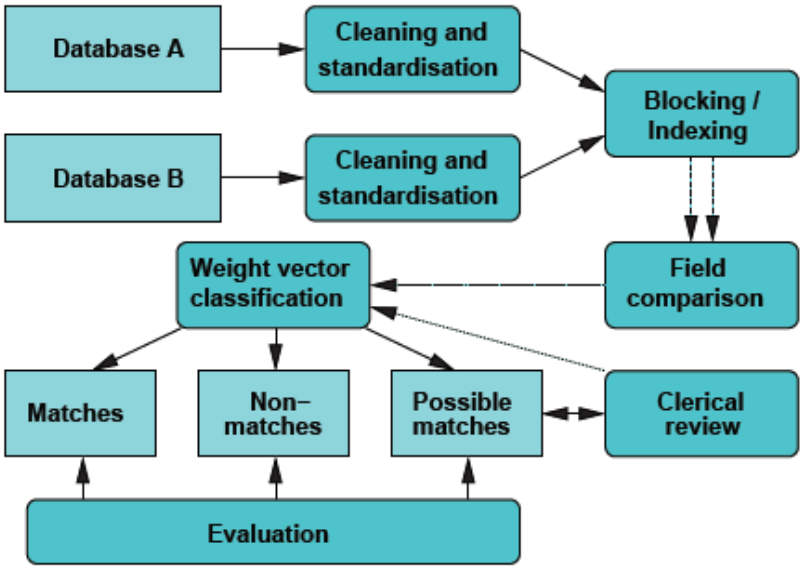

Figure 1: Etapes d'une procédure de Liage (Christen, 2008)

taxique et permet d'estimer la similarité entre les différents éléments des deux tuples. La fonction Longest Common Subsequence (LCS) est souvent utilisée à cette étape. LCS est un algorithme proposé par Allison et al. [Allison and Dix, 1986] et est utilisé pour trouver les plus longues sous-séquences communes dans deux enregistrements. Il a été expérimenté avec succès dans plusieurs contextes tels que liage des données.

Le liage des données a été aussi défini par des méthodes d'apprentissage supervisé. Ces dernières classent des couples d'enregistrements comme identiques (dupliquées) ou pas. Parmi les premières méthodes définies dans cette catégorie on cite la méthode d'Elfeky et al. [Elfeky et al., 2002] basée sur les arbres de décision et celle de Nahm et al. [Nahm et al., 2002] basée sur les *Support Vector Machines*. Cette approche nécessite des données d'apprentissage (*training data*) qui doivent être labellisées à l'avance. La phase de labellisation peut être lourde et coûteuse surtout avec des données complexes.

Dans cette étude, nous proposons une nouvelle stratégie de liage basée sur les sytèmes de recommandation et adaptée essentiellement aux bases de données de personnes (clients, utilisateurs, ...). Comme pour la première approche, elle commence par segmenter les données en clusters de tuples identifiés comme proches syntaxiquement. Cette étape est réalisée grâce à un système de recommandation utilisant une méthode d'apprentissage non

supervisée. La deuxième étape vient compléter la première en permettant de valider les clusters obtenus précédemment grâce à une analyse comportementale.

2.2 Les systèmes de recommandation

La mission principale d'un système de recommandation (SR) est de fournir à un utilisateur des ressources pertinentes en fonction de ses préférences [Negre, 2015].

On distingue trois types de SR: les systèmes se basant sur la popularité (Popularity based filtering), ceux basé sur le contenu et ceux basés sur le filtrage collaboratif. Les systèmes de recommandation basés sur la popularité partent du principe que plus un élément est populaire, plus il y a de chances pour qu'il soit également apprécié par l'utilisateur. La mise en place d'un tel système est assez simple. Il suffit de récolter les mesures sur les choix des utilisateurs et proposer ceux qui ont les meilleurs scores. Ces systèmes ne sont pas utilisés dans notre étude. Nous faisons appels à deux systèmes plus avancés basés sur le contenu et le filtrage collaboratif.

2.2.1 Systèmes basés sur le contenu (*Content based filtering*)

Dans cette approche, les recommandations sont réalisées par rapport aux caractéristiques du produit. Si un individu s'intéresse à un produit, ce système pourra par la suite proposer des produits similaires. Par exemple, si on apprécie un film d'un certain genre, l'algorithme en recherchant des produits possédant des similitudes pourra proposer d'autres films du même genre. Dans ce cas, le point commun entre les produits est le genre. La sélection des points communs peut être réalisée sur plusieurs critères. Dans le cas des films, on peut repérer les similitudes grâce à leur genre, leur réalisateur ou encore leur intrigue(Fig. 2) [Soualah-Alila, 2015].

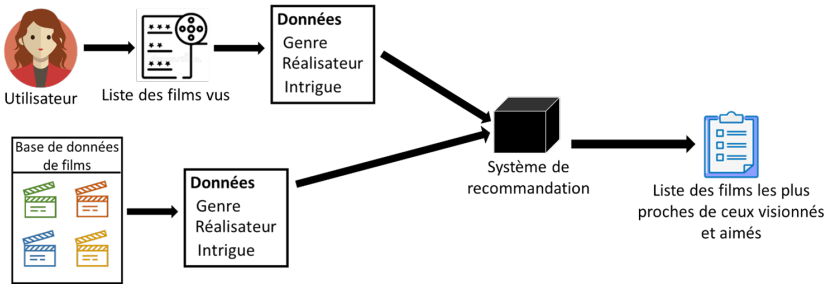

Figure 2: Système de recommandation basé sur le contenu

2.2.2 Recommandation par filtrage collaboratif (*Collaborative filtering*)

La recommandation par filtrage collaboratif est la plus populaire. Ce système se base sur les ressemblances entre les comportements et préférences des individus pour leur proposer des produits qui leur correspondent. Il prend en entrée les préférences des utilisateurs, dans ce contexte cela peut être les films qui ont été mis en favoris. Il va alors rechercher dans la base de données les utilisateurs qui ont mis les mêmes films en favoris et ainsi trouver des profils similaires à celui de l'utilisateur auquel on souhaite faire des recommandations. Ces recommandations seront les films favoris des utilisateurs qui ont un profil similaire (Fig. 3).

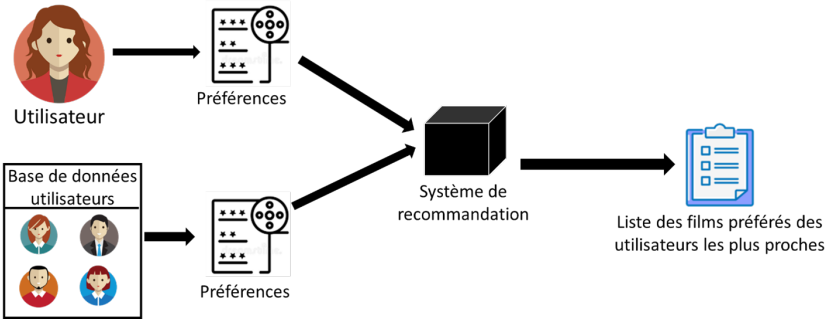

Figure 3: Système de recommandation collaboratif

3 Système de recommandation hybride pour le liage des données

Comme introduit dans la première section, nous proposons l'hybridation d'un système de recommandation basé sur le contenu avec un système de recommandation basé sur le filtrage collaboratif pour un liage optimal des données clients.

3.1 Système de recommandation basé sur le contenu - Clustering des clients à l'aide des k-means

Pour le premier système de recommandation basé sur le contenu, nous proposons une approche d'apprentissage non supervisé basé une technique de clustering par analyse textuelle. Après un premier nettoyage et préparation des données, la segmentation des données est réalisée grâce à la méthode de K-Moyennes (K-Means).

L'utilisation des k-means a pour but de rassembler les clients les plus proches dans les mêmes clusters. L'objectif ici est que toutes instances du

même client soient présentes dans le même cluster, et ce même si certaines informations telles que le nom soient différentes (Fig. 4). Le nombre de clusters donné comme paramètre aux K-Means est défini grâce à la méthode du coude (Elbow method).

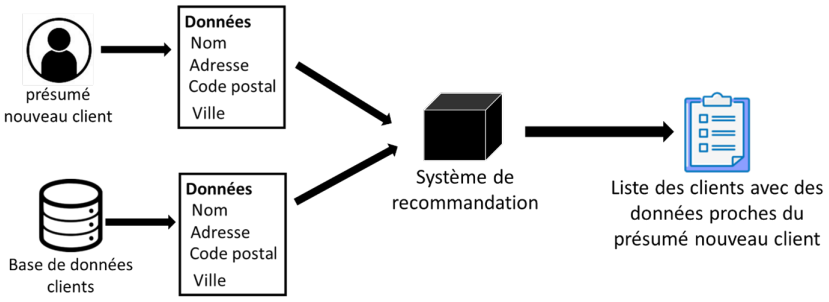

Figure 4: Liage par le contenu

3.2 Système de recommandation collaboratif - Comparaison des habitudes clients

Pour mettre en place le système de recommandation complémentaire basé sur le filtrage collaboratif, nous proposons une approche d'apprentissage supervisé par la méthode KNN. L'apprentissage est d'abord effectué sur le résultat de clustering basé sur la première approche. Le système permet alors de repérer si le comportement d'un client candidat à une insertion dans l abase est proche d'un autre client de son cluster (Fig. 5).

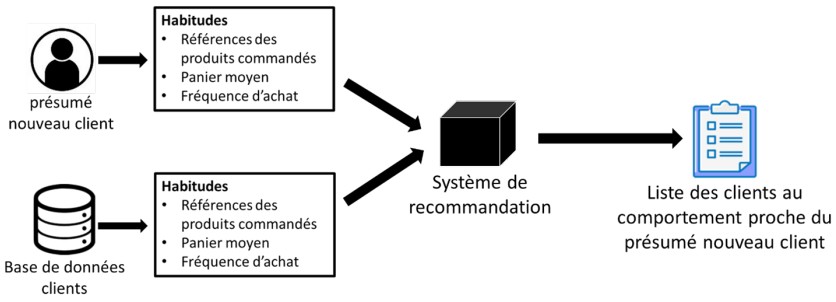

Figure 5: Liage par le filtrage collaboratif

3.3 Hybridation des systèmes

Comme exliqué ci-dessus, le système de recommandation basé sur le contenu est utile pour une détection rapide des données redondantes des clients basée sur leurs informations, telles que leurs noms ou leurs adresses, et éviter ainsi une insertion multiple du même client dans la base de données.

Cependant, cette étape n'est pas suffisante dans le cas d'informations modifiées comme pour le cas de changement d'adresse. Elle est donc complétée par une deuxième phase de liage basée sur un système de recommandation par filtrage collaboratif (Fig. 6). Ce dernier permet quant à lui de rapprocher des clients grâce à leur comportement. Dans le cadre de notre étude, les comportements sont définis par les habitudes d'achat. Cependant, les deux systèmes doivent être utilisés simultanément pour une meilleure performance. En effet, si deux clients de taille similaire ont les mêmes besoins et donc des comportements assez proches, dans ce cas le système pourrait les signaler comme étant un seul et même client sans que ça ne soit le cas. Comme étape finale, le retour des deux approches est analysé afin de suggérer une décision d'insertion ou pas du client.

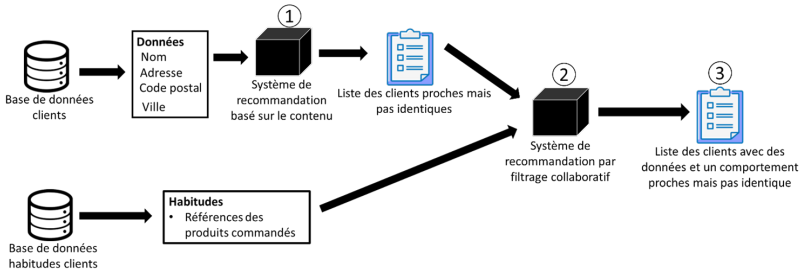

Figure 6: Système de recommandation hybride pour le liage des données

Ce système de recommandation hybride se réalise en plusieurs étapes. Les données utilisées sont celles présentes dans la base de données client. L'utilisation des k-means permet de clusteriser l'ensemble des clients en s'appuyant sur leurs caractéristiques (nom, prénom, adresse, code postal et ville). Ce premier résultat sera constitué de la liste des clients qui sont proches les uns des autres. Les comportements des clients proches va ensuite être analysé par le système de recommandation par filtrage collaboratif. Le résultat final sera donc constitué des clients qui ont de grandes similitudes au niveau de leurs caractéristiques mais un comportement différent. En effet, si des clients sont dans le même cluster et qu'ils ont un comportement identique, ils seront alors considérés par le système comme étant un seul et même client. L'intégration de ce système hybride dans le process général d'intégration des données est illustrée dans la figure 10.

4 Tests et Résultats préliminaires

Pour tester notre système de liage pour la détection de redonances, deux échantillons de données clients fictives sont générés en s'inspirant des données réelles de l'entreprise Brother. Le premier échantillon contenant des informations sur les clients est fourni comme entrée au premier système de recom-

mandation. Le deuxième échantillon contenant des historiques de comportement d'achat est utilisé pour les premiers tests du deuxième système de recommandation.

4.1 Préparation des données

Pour commencer, toutes les observations qui n'avaient pas l'ensemble de leurs informations renseignées ont été retirées du jeu de données. Exécuter l'lagorithme des k-means avec des informations manquantes telles qu'un client qui ne possède pas d'adresse aurait un impact négatif sur son efficacité et la clusterisation perdrait en efficacité. Les données utilisées pour identifier les clients sont principalement des données textuelles. Il faut donc au préalable les transformer afin qu'elles puissent être traitées par l'algorithme des k-means. Ce sont les dummies qui ont été utilisés pour transformer les valeurs textuelles en vecteurs binaires. Afin de pouvoir revenir aux données initiales, nous avons ajouté un préfixe pour chaque colonne.

4.2 Test du système de recommandation basé sur le contenu

Pour tester notre premier algorithme, nous avons généré aléatoirement les valeurs contenues dans la base de données client. Celle-ci contient une trentaine d'observations et possède 4 informations : le nom, l'adresse, le code postal et la ville du client. C'est sur la base de ces informations que le système de recommandation va clusteriser les clients selon leurs caractéristiques. Afin de valider l'algorithme des k-means, les informations de certains clients ont été modifiées, pour certains nous avons modifié l'adresse afin de simuler un déménagement, pour d'autres le nom des clients a subit une légère modification afin de simuler un erreur humaine. Malgré ces changements, l'algorithme a correctement clusterisé tous les clients (Fig. 7).

	Name	Adresse	Code Postal	Ville	Cluster
1	Leannon	53 Sawayn Prairie	69005	Kuvalischester	0
4	Leannon	53 Sawayn Prairie	69005	Kuvalischester	0
14	Leannone	53 Sawayn Prairie	69005	Kuvalischester	0
5	Denesik Ltd	22 Emilia	29166	West Weldon	1
12	Denesik Ltd	22 Emilia	29166	West Weldon	1
16	Ruecker-Jast	9 Monahan Highway	14020	Littelfurt	2
9	Hahn LLC	55 Schroeder View	36223	Marquardtland	3
2	Smith PLC	17 Hoppe Light Suite	5751	Archibaldberg	4
6	Smith PLC	17 Hoppe Light Suite	5751	Archibaldberg	4
15	Luettgen Corp.	80 Mitchell Inlet	86469	TestTown	5
23	Luettgen Inc	80 Mitchell Inlet	86469	Hammesside	5
20	Luettgen Inc	80 rue de l'etoile	93430	Hammesside	5

Figure 7: Résultat des k-means

Le succès de cette opération de clustering repose entièrement sur le nombre de clusters qui a été sélectionné grâce à la méthode du coude. En effet, si le nombre de clusters sélectionné est trop élevé, l'algorithme sera trop précis et les clients similaires n'apparaîtront pas dans le même cluster. Dans le cas contraire, si le nombre de cluster est insuffisant, les clients différents seront rassemblés dans le même cluster. Le succès du clustering est entièrement

dépendant du nombre de clusters il est donc essentiel de trouver le nombre idéal de clusters.

4.3 Comparaison des habitudes des clients

Le premier résultat obtenu suite à l'exécution de l'algorithme des k-means est donné en entrée pour le système de recommandation par filtrage collaboratif. Pour réaliser une comparaison des clients au niveau de leur habitudes nous avons créé une table nommée HabitudesClient. Cette table possède deux colonnes, le nom du client et les références des produits déjà commandées par le client. Pour tester cet algorithme nous nous sommes limités à 4 références représentées par les lettres A,B,C et D. Dans un cas concret, ces références pourraient représenter des consommables noir et blanc ou couleur étant destinés à des imprimantes de type jet d'encre ou laser.

Nous avons agrémenté les résultats des k-means avec la référence du dernier produit commandé par le client. Une jointure a ensuite été réalisée entre ces données et la base de données HabitudesClient. Cette jointure a été effectuée sur le nom des clients. L'algorithme utilisé est présenté dans la figure ci-dessous :

```
Pour chaque cluster :
    HistProduitsCommandes = Historique de tous les produits commandés par les clients de ce cluster

    Pour chaque client de ce cluster :
        Si le dernier produit commandé apparaît dans HistProduitsCommandes
                Le client a le bon comportement
        Sinon
            Le client est proche au niveau des données mais pas au niveau du comportement
            Recommander au gestionnaire de bases de données pour qu'il fasse une vérification
```

Figure 8: Algorithme d'identification des clients grâce à leurs habitudes

Si le dernier produit commandé par le client d'un cluster a déjà été commandé par les autres clients présents dans son cluster, alors nous pouvons tirer la conclusion qu'il s'agit des mêmes clients. Ceux-ci ne feront donc pas partie de la liste de recommandation. Cependant, les clients qui ont commandé un produit qui n'a jamais été commandé par les autres clients de son cluster seront ajoutés à la liste des recommandations étant donné que l'algorithme des k-means a estimé que leurs données étaient proches mais qu'au vu de l'analyse au niveau des habitudes, leur comportement est divergeant. Ces clients-ci nécessitent donc une attention particulière et une analyse humaine plus approfondie. C'est grâce à ses connaissances métier que l'administrateur de données pourra savoir s'il s'agit effectivement du même client ou non.

Avec l'utilisation des systèmes de recommandation, le processus d'intégration des rapports des revendeurs est optimisé (Fig. 10)

Les informations des rapports sont intégrés à la base de données client, le système de recommandation analyse en continu cette base de données. Lorsqu'une possible insertion multiple d'un client est repérée, une notifi-

	Name	Adresse	Code Postal	Ville	Cluster	RefProduit	HistProduitsCommandes
0	Leannone	53 Sawayn Prairie	69005	Kuvalischester	3	A	
1	Leannon	53 Sawayn Prairie	69005	Kuvalischester	3	D	A;D
2	Luettgen Corp.	80 Mitchell Inlet	86469	TestTown	7	C	
3	Luettgen Inc	80 rue de l'etoile	93430	Hammesside	7	D	B;D
4	Luettgen Inc	80 Mitchell Inlet	86469	Hammesside	7	B	B;D

Figure 9: Identification des clients selon leurs habitudes

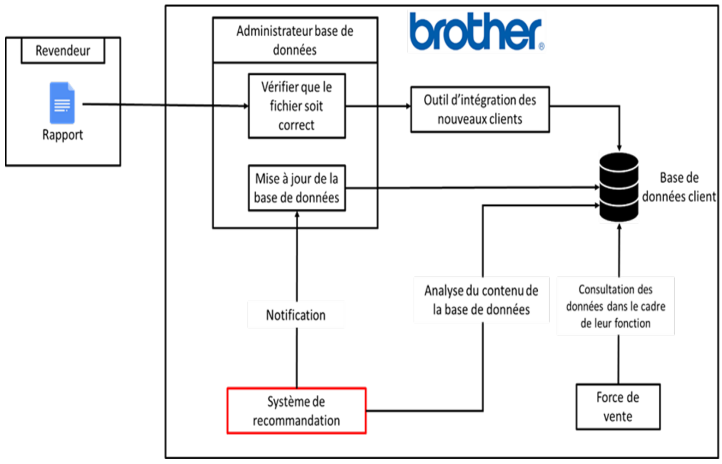

Figure 10: Processus d'intégration des rapports

cation est remontée à l'administrateur. Ce dernier peut alors réaliser une analyse approfondie et ainsi confirmer ou infirmer le fait qu'il s'agisse du même client et modifier les données en conséquence.

5 Conclusion

Pour répondre à la problématique qui concernait le liage des données par les systèmes de recommandation intelligents afin d'optimiser la qualité des données, nous avons vu qu'il était possible de faire appel à un système de recommandation hybride qui fait appel à la fois au système de recommandation basé sur le contenu et le système de recommandation par filtrage collaboratif. Ce système hybride permet de recommander au gestionnaire de base de données les clients qui ont potentiellement été dupliqués dans la base de données. Cependant, toutes les modifications des données sont réalisées manuellement, le système ne réalise que des recommandations. L'utilisation simultanée de ces deux systèmes de recommandation permet dans un premier temps de regrouper les clients les plus proches grâce à leurs caractéristiques, puis dans un second temps d'identifier les clients grâce à leur comportement et habitudes. Les systèmes de recommandation les plus efficaces sont ceux qui font appel à des méthodes de machine learning. Pour clusteriser les clients selon leurs caractéristiques c'est

la méthode des k-means qui a été utilisée. Pour l'analyse comportementale, un algorithme permettant de comparer le dernier produit commandé avec l'historique des produits déjà commandés par les clients proches a été utilisé. Cette vérification à deux niveaux permet de réaliser une identification précise des clients et ainsi de repérer les clients qui ont de fortes chances d'avoir été insérés plusieurs fois. C'est en s'appuyant sur cette liste de recommandation et sur ses connaissances métiers que l'administrateur de bases de données pourra vérifier ces données et corriger les informations incorrectes.

Pour finir, il serait bénéfique d'étendre les fonctions du système de recommandation afin qu'il puisse être autonome et corriger de manière autonome les informations de la base de données client en s'appuyant sur les données présentes dans la base Sirene du gouvernement. Cette opération peut être réalisée grâce à l'utilisation d'algorithmes de partitionnement tels que le KNN, cette partie est actuellement en cours de développement.

Toutes les redondances dans ces premiers échantllions ont été détectées. Les premiers résultats sont très encourageants. D'autres extensions et tests sont en cours de réalisation pour compléter notre système et optimiser son comportement avec des données réelles.

References

[Allison and Dix, 1986] Allison, L. and Dix, T. I. (1986). A bit-string longest-common-subsequence algorithm. *Information Processing Letters*, 23(5):305–310.

[Christen, 2008] Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 151–159.

[Elfeky et al., 2002] Elfeky, M. G., Verykios, V. S., and Elmagarmid, A. K. (2002). Tailor: A record linkage toolbox. In *Proceedings 18th International Conference on Data Engineering*, pages 17–28. IEEE.

[Ferrara et al., 2013] Ferrara, A., Nikolov, A., and Scharffe, F. (2013). Data linking. *Journal of web semantics*, 23.

[Hernández and Stolfo, 1998] Hernández, M. A. and Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery*, 2(1):9–37.

[Nahm et al., 2002] Nahm, U. Y., Bilenko, M., and Mooney, R. J. (2002). Two approaches to handling noisy variation in text mining. In *Proceedings of the ICML-2002 workshop on text learning (TextML'2002)*, pages 18–27. Citeseer.

[Negre, 2015] Negre, E. (2015). *Information and recommender systems.* John Wiley & Sons.

[Newcombe et al., 1959] Newcombe, H. B., Kennedy, J. M., Axford, S., and James, A. P. (1959). Automatic linkage of vital records. *Science*, 130(3381):954–959.

[Saïs and Thomopoulos, 2016] Saïs, F. and Thomopoulos, R. (2016). Fusion de données redondantes : une approche explicative. *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des Connaissances, RNTI-E-30:363–368.

[Soualah-Alila, 2015] Soualah-Alila, F. (2015). *CAMLearn: Une Architecture de Système de Recommandation Sémantique Sensible au Contexte. Application au Domaine du M-Learning.* PhD thesis, Université de Bourgogne.