Gestion des assortiments de produits dans la grande distribution

Jocelyn Poncelet

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
ÉCOLE NATIONALE SUPÉRIEURE DES MINES D'ALÈS (IMT MINES ALÈS)

En Informatique

École doctorale I2S – Information, Structures, Systèmes
Portée par l'Université de Montpellier

Unité de recherche LGI2P

Gestion des assortiments de produits dans la grande distribution

Présentée par Jocelyn PONCELET
Le 11 Setpembre 2020

Sous la direction de Jacky MONTMAIN

Devant le jury composé de

EL YACOUBI Samira, Professeur, Université de Perpignan	**Rapporteur**
BERTAUX Aurélie, Maître de Conférences (HDR), Université Bourgogne	**Rapporteur**
ABEL Marie-Hélène, Professeur, Université de Technologie de Compiègne	**Examinateur**
DESPRES Sylvie, Professeur, Université de Paris	**Examinateur**
TROUSSET François, Maître Assistant, IMT Mines Alès	**Encadrant de proximité**
MONTMAIN Jacky, Professeur, IMT Mines Alès	**Directeur de thèse**
PECHEUR Nicolas, Chef de projet, TRF Retail	**Invité**
SEGUIN Thierry, Directeur général, TRF Retail	**Invité**

UNIVERSITÉ DE MONTPELLIER

IMT Mines Alès
École Mines-Télécom

PONCELET Jocelyn : *Gestion des assortiments de produits dans la grande distribution ,* Mai 2020

RÉSUMÉ

L'objectif de cette thèse est de proposer un système de recommandations permettant aux grands distributeurs d'améliorer leurs assortiments de produits distribués à travers de nombreux points de vente. Dans ce contexte, la problématique adressée est celle de la planification d'assortiment qui consiste à éliciter les meilleurs produits, *e.g.*, ceux faisant le plus de chiffre d'affaires. Pour ce faire, nous proposons dans un premier temps une comparaison des méthodes pragmatiques mises en place dans l'industrie avec l'état de l'art associé à la planification d'assortiment. Cette comparaison permet de mettre en lumière le problème de transversalité des connaissances utilisées aujourd'hui pour améliorer l'assortiment. Pour pallier ce problème, nous proposons des structures de connaissances propres à la grande distribution. Grâce à ces structures, une méthode Agile d'optimisation de l'assortiment pouvant être intégrée dans un processus d'amélioration continue est formalisée. Cette méthode permet d'intégrer l'expertise humaine, que nous considérons comme indispensable, aux différents leviers actuellement adoptés.

Pour souligner la modularité de notre approche, nous proposons ensuite une analyse sémantique des magasins qui, en plus d'améliorer la précision de nos simulations, permet de définir un nouvel axe d'amélioration de l'assortiment. Cette analyse se base sur les structures de connaissances propres à chaque enseigne et sur les mesures de similarités sémantiques. Enfin, pour perfectionner notre méthode et aller plus loin dans l'exploitation de ces structures, nous proposons une analyse sémantique des consommateurs qui sont les cibles finales de l'assortiment. Cette seconde analyse sémantique permet d'apporter de nouvelles connaissances pour les distributeurs et d'apporter de nouvelles contraintes sur les assortiments. En parallèle de ces contributions scientifiques, différentes applications ont été développées pour souligner l'interopérabilité de nos contributions avec des notions propres à différents types de distributeurs (*e.g.* Alimentaire, Bricolage ...). Ces applications sont présentées dans le manuscrit dans la limite du respect de la confidentialité et de la propriété intellectuelle.

ABSTRACT

The main objective of this thesis is to propose a recommendation system allowing retailers to improve their assortments of products distributed through numerous stores. In this context, the problem addressed is the assortment planning which consists in eliciting the best products, *e.g.*, those with the most turnover. To this end, we first propose a comparison of assortment planning with the pragmatic methods which are commonly used in the industry and the state of the art. This comparison highlights the problem of cross-functionality of the knowledge used today to improve the assortment. To overcome this problem, we propose knowledge structures specific to mass distribution. Thanks to these structures, an Agile assortment optimisation method that can be integrated into a continuous improvement process is formalised. This method makes possible to integrate human expertise, which we deem essential, in the various levers currently adopted.

To underline the modularity of our approach, we then propose a semantic analysis of the stores which, in addition to improving the accuracy of our simulations, allows us to define a new axis of assortment improvement. This analysis is based on knowledge structures which are specific to each brand and on semantic similarity measures. Finally, to perfect our method and go further in the exploitation of those structures, we propose a semantic analysis of the consumers who are the final targets of the assortment. This second semantic analysis allows us to bring new knowledge to retailers and new constraints on assortments. In parallel to these scientific contributions, different applications have been developed to highlight the interoperability of our contributions with concepts specific to different types of retailers (*e.g.* Food, DIY ...). These applications are presented in the manuscript within the limits of respect for confidentiality and intellectual property.

PUBLICATIONS

Ci-dessous la liste des articles publiés au cours de la thèse :

J. Poncelet, P.A. Jean, F. Trousset and J. Montmain. « Semantic Customers' Segmentation. » *In Proceedings of the 6th International Conference on INTERNET SCIENCE - (INSCI 2019)*, pp. 318-325.

J. Poncelet, P.A. Jean, F. Trousset and J. Montmain. « Impact des mesures de similarité sémantique dans un algorithme de partitionnement : d'un cas biomédical à la détection de comportements de consommation. » *In Actes des 26èmes Rencontres de la Société Francophone de Classification (SFC 2019)*, pp. 1-6.

J. Poncelet, P.A. Jean, F. Trousset and J. Montmain. « Semantic Hierarchical Clustering : An Application in the Biomedical Domain. » *In Proceedings of the 19th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2020)*, pp. 1-12.

J. Poncelet, P.A. Jean, M. Vasquez and J. Montmain. « Hierarchical reasoning and knapsack problem modelling to design the ideal assortment in retail. » *In Proceedings of the 18th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2020)*, pp. 1-15.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

1

INTRODUCTION

1.1 CONTEXTE

De la manutention à la vente, en passant par la chaîne d'approvisionnement, la grande distribution forme un éco-système complexe associé à une très grande variété de domaines. Ces dernières années, avec la démocratisation de la vente en ligne, de nouveaux concurrents sont apparus. Pour faire face à la concurrence, pour répondre aux fluctuations de l'environnement économique ou encore aux tendances du marché, les distributeurs doivent en permanence améliorer la performance de leur organisation et de leurs opérations, ce qui les conduit à réorganiser fréquemment tout ou partie de leurs systèmes d'approvisionnement et de ventes [57, 70]. Il s'agit ainsi de pouvoir répondre à des impératifs de qualité, coût, délai, innovation, flexibilité et réactivité qui sont aujourd'hui les leviers majeurs de la performance [13, 14, 72, 77, 100]. Cela suppose une excellente maîtrise des flux financiers, d'information et de décision et une bonne coordination des activités, tant en interne que dans les relations avec leurs partenaires (fournisseurs, sous-traitants, co-contractants...) [1, 57, 71, 140, 150].

Ce problème de réorganisation des systèmes se pose aussi lors de fusions ou lors du passage au commerce électronique (e-économie) ou encore lors de l'implantation de grands systèmes d'information, comme par exemple les ERP (*Enterprise Resource Planning*), les MES (*Manufacturing Execution System*), les SGDT (Système de Gestion de Données Techniques), les SCM (*Supply Chain Management*) ... Pour l'industriel et le consultant, il est dès lors nécessaire de pouvoir s'appuyer non seulement sur des démarches structurées - ou méthodologies d'intervention – mais aussi sur des outils et systèmes d'aide associés pour continuellement améliorer et évaluer la performance des modes d'organisation ainsi que des processus opérationnels [17, 19, 38, 43, 62, 63, 74].

Actuellement, au niveau stratégique, les méthodes d'aide à la décision utilisées se limitent à des considérations financières. Elles sont essentiellement quantitatives et s'en tiennent à une vision globale du système. Au niveau opérationnel, les méthodes théoriques classiques d'optimisation restent dans une logique taylorienne. Elles sont basées sur une fonction de coût monocritère, répondant ainsi de moins en moins à l'évolution de la complexité des systèmes. Par conséquent, pour satisfaire aux besoins d'amélioration et de pérennisation de la performance ainsi que de réactivité du pilotage, la problématique de l'optimisation des décisions se traduit désormais sous la forme d'un problème d'aide à la décision multicritères.

Les stratégies de pilotage pour l'amélioration continue de la performance industrielle doivent intégrer d'une part, l'aspect multicritères de l'évaluation des performances de l'entreprise et, d'autre part, les relations qui existent entre ces performances élémentaires [18, 85, 118]. Orientations et décisions stratégiques peuvent ensuite s'échafauder autour de l'analyse de l'artefact que constitue le système d'indicateurs de performances. Plusieurs modèles et outils pour le diagnostic et l'optimisation de la performance industrielle basés sur des techniques d'analyse multicritères et d'optimisation multi-attributs ont déjà été proposés [15, 16, 114, 138].

Les axes d'amélioration de la performance de la grande distribution sont multiples. Alléger les stocks, optimiser les livraisons, améliorer les systèmes d'information, étendre la collaboration avec les fournisseurs, améliorer le contrôle du couplage de la demande et de la chaîne d'approvisionnement (*Supply Chain*) en interne ... sont autant de chantiers auxquels la grande distribution doit faire face. Si chacun de ces objectifs semblent avoir fait l'objet d'une automatisation locale poussée par le passé, leur amélioration conjointe semble encore souffrir d'une coordination non optimisée des différents services de la grande distribution.

Alors que certaines enseignes comme *Grand Frais* affichent une augmentation significative des ventes, la majorité des enseignes de la grande distribution perdent chaque année des parts de marché. Les solutions novatrices proposées sont nombreuses, mais aucune ne semble être la réponse adéquate au nouveau mode de consommation. Par exemple, Carrefour a proposé une solution [1] basée sur une modification de leurs vitrines avec rayonnage sous forme "d'Univers" représentant des demandes identifiées (*e.g.* Produits Bio, Produits du monde etc.). Cependant, non seulement les résultats ne sont pas concluants, mais ils soulignent très clairement la difficulté à cerner et à satisfaire pleinement la demande client [46].

Les cinquante dernières années ont vu apparaître de nombreux bouleversements des méthodes de vente et de consommation qui se sont ainsi propagés à de nombreux domaines autres que l'alimentaire. Grâce à la massification, les maîtres mots ont longtemps été la baisse des prix et l'augmentation de l'offre. Les consommateurs profitaient d'un pouvoir d'achat toujours plus important grâce à une croissance permanente. Néanmoins, un changement majeur s'est introduit dans les habitudes d'achats. Aujourd'hui, nous sommes témoins d'un changement de tendance, où, les méthodes de consommation s'émancipent des méthodes de vente définies par les grands distributeurs.

Le constat aujourd'hui est que la satisfaction d'une demande fluctuante passe par l'identification d'avantages durables. La survie des enseignes dépendra principalement de leur capacité à fidéliser les clients. Pour cela, elles doivent être en mesure d'identifier les produits

1. Carrefour Planet

adéquats, nous amenant à la question de l'assortiment des magasins, c'est-à-dire des produits proposés à la vente.

Dans ce contexte, TRF Retail, une société éditrice Software As A Service dédiée aux métiers de la distribution, propose des solutions à très forte valeur ajoutée facilement connectées aux solutions d'exécution devenues des commodités comme les ERP ou les solutions BI (*Business Intelligence*). TRF Retail a mis au point un système de notation, d'alertes et de recommandations sur la performance des produits de la grande distribution. Depuis 2010, le laboratoire de recherche du LGI2P l'accompagne dans la formalisation mathématique de cet outil de supervision basé sur la notion d'index de performance produit. Les travaux réalisés dans le cadre de cette thèse s'inscrivent dans cette collaboration et visent à étudier la problématique de l'assortiment dans la grande distribution.

1.2 PROBLÉMATIQUE

L'objectif de cette thèse est de proposer un système de recommandation permettant aux grands distributeurs d'améliorer leurs assortiments de produits distribués à travers de nombreux points de vente. De façon plus précise, la problématique adressée est celle de la planification d'assortiments qui consiste à éliciter les meilleurs produits.

Notre première contribution repose sur la modélisation du problème d'assortiment sous la forme d'un problème d'optimisation combinatoire de type "sac à dos". Des analyses sémantiques sont utilisées pour apporter une nouvelle vision de la similarité entre magasins ou clients afin d'améliorer les suggestions d'optimisation. Pour enrichir cette première contribution, nous proposons un système d'expression de contraintes permettant aux enseignes d'exprimer leur stratégie. Ces contraintes, indispensables pour réduire la complexité du problème, correspondent à notre seconde contribution. Enfin, nous proposons différents scénarios d'amélioration de l'assortiment répondant aux contraintes industrielles existantes. Avec l'aide et l'expertise de TRF Retail, des interfaces aidant les grandes enseignes à prendre des décisions ont pu être développées. La formalisation de ces différents scénarios composent notre troisième contribution.

Ce manuscrit s'organise comme suit. Le chapitre 2 expose la dualité des travaux de recherche et des travaux industriels. Dans ce chapitre, l'état de l'art répondant à notre problématique ainsi que les limites associées sont développés. Le chapitre 3 formalise des structures de connaissances pouvant être exploitées par les grands distributeurs pour pallier certaines de ces limites. Une analyse sémantique des consommateurs est proposée afin d'illustrer l'intérêt d'exploiter ces structures pour l'estimation de similarité et proposer de nouvelles perspectives pour la segmentation de clients. Le chapitre 4 définit ce que nous considérons comme l'assortiment optimal, de plus nous

proposons une approche sémantique permettant de modéliser formellement la problématique associée à son identification.

L'intérêt des analyses sémantiques pour améliorer les recommandations et réduire la complexité de notre problématique est mise en lumière. Le chapitre 5 introduit différents scénarios d'améliorations d'assortiment permettant de répondre aux besoins industriels identifiés. Des interfaces de recommandations développées conjointement avec TRF Retail sont présentées.

Enfin, une conclusion sur nos contributions et les perspectives associées sont discutées dans le chapitre 6.

2

DÉFINITION ET PROBLÉMATIQUE DE L'ASSORTIMENT DE PRODUITS

Ce chapitre propose un état des lieux de l'assortiment de produits dans l'industrie et dans la recherche. La dualité entre les approches industrielles ou académiques est forte. Tandis que les chercheurs proposent des modèles mathématiques avancés, les industriels préfèrent une approche plus pragmatique qui leur permet de gérer de très nombreux produits au travers d'un vaste réseau de magasins. Nous formalisons les principales notions associées à l'assortiment dans le monde industriel. Les produits et les magasins qui sont des notions élémentaires sont brièvement introduits. Ensuite, les deux notions primordiales que sont la structure marchandise et l'assortiment gigogne sont explicitées.

Nous présentons également la divergence qu'il existe entre les contributions scientifiques et la réalité industrielle à laquelle sont confrontés les grands distributeurs grâce à l'état de l'art portant sur notre problématique de planification d'assortiments.

Ce chapitre est organisé de la façon suivante. Dans la section 2.1, nous présentons les notions nécessaires à l'industrie. La section 2.2 présente l'état de l'art associé à la problématique de planification d'assortiments. Enfin, la section 2.3 souligne la dualité des approches et leurs limites.

2.1 GESTION INDUSTRIELLE DE L'ASSORTIMENT

Une enseigne ou un distributeur peut être considéré comme un ensemble d'articles vendus dans un ensemble de magasins. Évidemment, cet ensemble de produits est incroyablement grand, tout comme peut l'être l'ensemble de magasins (*cf.* Tableau 2).

TABLE 1 – Statistiques indicatives pour trois clients différents de TRF Retail (**k** = milliers)

Type de Distributeur	Nombre de produits différents	Nombre de magasins
Alimentaire	120 **k**	2 000
Bricolage	140 **k**	32
Coopérative Agricole	600 **k**	250

Pourtant, deux magasins d'une même enseigne semblent souvent identiques. Ceci nous amène à nous poser 2 questions : Comment les distributeurs font-ils pour planifier leurs assortiments ? Comment choisissent-il les produits à mettre en magasin ? Pour comprendre cela, nous décrivons dans cette section les éléments nécessaires aux grands distributeurs pour conserver une homogénéité dans l'assortiment proposé aux consommateurs.

2.1.1 *Les produits & les magasins*

Le produit est l'élément de base de la distribution ! Par définition, un assortiment est l'ensemble des produits proposés à la vente. A chaque produit est associé un ensemble d'indicateurs clés de performance, appelés *KPI* (*Key Performance Indicator*). Ces indicateurs, dont les plus connus sont comptables, permettent de définir une première notion de performance des produits. Les plus utilisés sont :

— le chiffre d'affaires
— la marge brute
— la marge nette
— le prix de vente
— le prix d'achat principal
— le prix d'achat pour chacun des fournisseurs
— la prévision des ventes
— le stock en quantité et en valeur
— la démarque en quantité et en valeur
— le nombre de jours pendant lesquels un produit était en rupture
— le nombre de jours avant rupture
— la distribution numérique (nombre de magasins possédant l'article)

Ces KPI "élémentaires", lorsqu'ils sont agrégés, permettent d'en définir de nouveaux plus "complexes" qui sont aujourd'hui des informations indispensables pour la prise de décision dans la grande distribution. Par exemple, les parts de marché qui sont définies à partir du chiffre d'affaires d'une enseigne comparé au chiffre d'affaires des concurrents. Les ventes et le chiffre d'affaires restent à l'unanimité considérés comme étant la vision pragmatique de la performance dans l'industrie. Notons tout de même que chaque enseigne exploite tout ou partie des KPI économiques génériques en plus de KPI spécifiques adaptés à leur cœur de métier.

En plus de ces indicateurs, chaque produit possède un ensemble de caractéristiques : une catégorie de produits (*e.g.* Soda, Légumes

...), un code barre EAN (*European Article Numbering*), un UPC (*Universal Product Code*), une marque (*e.g. Coca-Cola*), une TVA, un SKU (*Stock Keeping Unit*) ... Contrairement aux KPI, ces caractéristiques sont généralement statiques et communes à l'ensemble des magasins d'une même enseigne. Certaines caractéristiques sont mêmes partagées par l'ensemble des distributeurs localisés dans une même zone géo-économique. Par exemple, dans la Zone Euro, les produits possèdent un code barre générique EAN commun. Parmi les caractéristiques les plus courantes, nous retrouvons :

— le Nom ou libellé de l'article (*e.g.* `Bouteille Coca-Cola 1L5`)
— la Catégorie de produits (*e.g.* `Soda`)
— l'Unité de besoin (*e.g.* `Coca-Cola`)
— la Saisonnalité du produit (*e.g.* chocolats de Pâques)
— la Gamme de produit (*e.g.* Haute gamme, Premier Prix ...)
— le Fournisseur
— Le pays exportateur
...

Ces informations propres à chaque distributeur servent surtout à la chaîne de distribution (*supply chain*) et peuvent être différentes d'un magasin à l'autre. Par exemple, deux magasins éloignés auront deux fournisseurs différents même s'ils appartiennent à la même enseigne. Deux formes de commerces sont distinguées :

— Le commerce intégré : le point central du réseau de magasins distribue par l'intermédiaire de points de vente. Les points de ventes appartiennent à l'enseigne et sont dirigés par des salariés du groupe ou des gérants.
— Le commerce organisé : les points de ventes sont détenus par des indépendants. Ces points de vente peuvent être des franchisés, des coopératives etc.

Remarque : dans le cadre des travaux présentés dans ce mémoire, nous nous intéressons principalement au commerce intégré. Néanmoins, la planification de l'assortiment dans un commerce organisé consiste principalement à utiliser l'optimisation horizontale formalisée dans la section 5.3. Les notions de centralisation et de massification de l'assortiment ne seront pas prises en considération.

Les plus grandes enseignes sont confrontées à la gestion de *réseaux de magasins* colossaux à travers le monde. Et certaines continuent de s'étendre en s'implantant dans toujours plus de pays. Ces magasins sont généralement regroupés d'après leurs différentes caractéristiques, notamment :

— la taille
— le format (*e.g.* hypermarché, supermarché, supérette ...)

— la localisation (*e.g.* par pays, région ...)
— le statut juridique (*e.g.* Intégré, Franchisé, Coopérative ...)
— le cœur de métier (*e.g.* Bricolage, Alimentaire, Pharmaceutique, Vestimentaire ...)

Ces sous-ensembles de magasins se font appeler des *clusters*. Plusieurs types de clusters sont utilisés pour effectuer des actions groupées tout en contrôlant l'impact. Par exemple, une enseigne peut vouloir ajouter aux magasins de la région "Sud" des produits de plage (*e.g.* parasol, râteau ...) pendant la saison estivale pour satisfaire les attentes saisonnières des clients. Dans le même ordre d'idée, elle peut en parallèle ajouter des produits de type `Crème glacée` dans tous ses hypermarchés. Ces clusters de magasins sont indispensables aux industriels pour optimiser la prise de décision et la *Supply chain*. Un repère commun est indispensable pour que les décisions puissent être appliquées par l'ensemble des magasins.

2.1.2 *Structure marchandise*

Les magasins d'une même enseigne partagent tous un même référentiel permettant de "ranger" les produits appelé structure marchandise. Il s'agit d'un arbre dont les feuilles sont les produits. Il existe différents niveaux de catégories reliées par des liens "is-a". Par exemple, le produit `Coca-Cola 1L5`) ("is-a") `Soda` est lié par héritage ("is-a") à `Liquides` (*cf.* Figure 1).

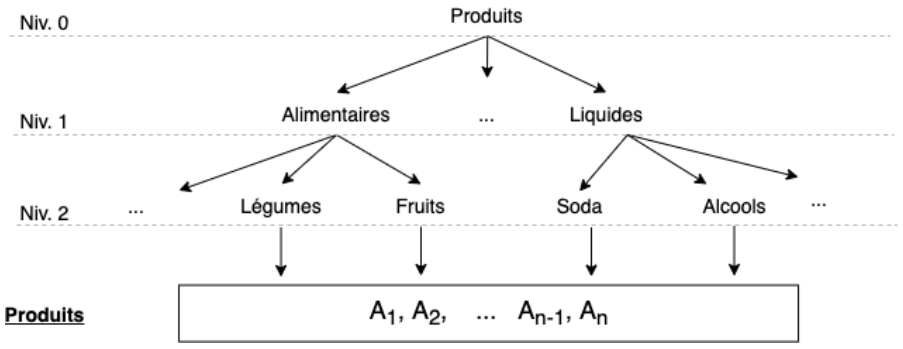


FIGURE 1 – Exemple illustratif d'une structure marchandise

Plus la catégorie de produits se situe en profondeur dans la structure, plus celle-ci est spécifique. Par exemple `Soda` est plus spécifique que `Liquides` dans la Figure 1. Le nœud racine de la structure représente une abstraction virtuelle de tous les produits de l'enseigne. Les structures marchandises sont généralement définies de façon à ce que la plus spécifique des catégories, autrement dit, les nœuds directement parents des feuilles, représentent une Unité de Vente Consommateur (UVC), aussi appelée SKU (*Stock Keeping Units*), ou des unités de besoin (UB).

Le code SKU, ou UGS (Unité de Gestion de Stock) est un élément central dans la gestion des stocks. Il correspond au numéro de référence d'un produit dans la chaîne d'approvisionnement d'un distributeur. Il n'existe, à ce jour, aucune standardisation des codes SKU. Les UB peuvent être considérées comme des préjugés : une attente intuitive que les clients peuvent avoir. Par exemple, il y aura toujours des `Clous`, des `Vis` et des `Outils` dans un magasin de bricolage.

Répondre aux unités de besoin est indispensable pour les grands distributeurs. Un client qui ne trouve pas de produits répondant à son besoin se rendra sûrement chez un concurrent. En plus d'une vente perdue, la probabilité que le client change ses habitudes est élevée. L'identification exacte de toutes les unités de besoin reste utopique d'autant plus qu'elles sont différentes pour chaque cœur de métier. Un client s'attendra à trouver une grande variété de `Jouet` dans une enseigne spécialisée alors qu'une gamme bien plus restreinte sera attendue d'une enseigne généraliste. Sauf durant la période hivernale, pendant laquelle, ces enseignes généralistes transforment complètement leur gamme de `Jouet`. Cet exemple permet de souligner l'aspect temporel des unités de besoin et illustre la complexité associée à leur identification. Certains produits ont même fini par devenir des unités de besoin. Les plus remarquables sont le Coca-Cola et le Nutella qui, chez beaucoup de grands distributeurs sont considérés comme des unités de besoin à part entière.

Chaque grand distributeur possède sa propre structure marchandise qui répond à ses besoins. Le nombre de différentes catégories de produits proposées par une enseigne fait référence à la largeur de gamme. Tandis que le nombre de produits associés à une UB ou un SKU fait référence à la profondeur de gamme. Ainsi, chaque enseigne cherche l'équilibre entre la profondeur et la largeur de sa structure marchandise. Une indication de la différence du nombre de catégories d'une enseigne à l'autre est illustrée dans la Tableau 2. Quand certains grands distributeurs alimentaires privilégient la largeur de gamme, d'autres préfèrent se spécialiser dans certaines catégories de produits (*e.g.* Zara, Toys'r'us, Darty ...). Le compromis entre la largeur et la profondeur de gamme associé à l'identification des UB est une problématique fondamentale dans la grande distribution. De nombreuses propositions scientifiques ont été faites surtout pour redéfinir la taxonomie de produits autour des unités de besoin.

Grâce à la structure marchandise, les distributeurs sont alors en mesure de définir l'assortiment commun à tous leurs magasins.

2.1.3 *L'assortiment gigogne*

Pour permettre une gestion centralisée, simplifiée et homogène des produits proposés en magasin, les industriels utilisent ce que l'on

TABLE 2 – Statistiques indicatives pour trois différents clients de TRF Retail (**k** = Milliers)

Type de Distributeur	Catégorie de produits
Alimentaire	2 200
Bricolage	815
Coopérative Agricole	6 290

appelle l'*assortiment gigogne*. L'assortiment gigogne est généralement défini par les *Catégories Managers*. Chaque catégorie manager définit un indice d'importance d'inclusion (appelé *profil assortiment*) à chacun des produits. Ainsi, une structure hiérarchique partagée par les catégories émerge définissant des sous-ensembles de produits, où chacun est associé à un profil assortiment. Dans l'exemple de la Figure 2, la `Pomme` est le `Fruit` indiqué comme étant le plus important à inclure avec son indice de 1, s'ensuit la `Banane` puis enfin la `Mangue`. Le lien d'inclusion hiérarchique fonctionne de la façon suivante : un magasin qui propose des `Bananes` devra proposer en plus des `Pommes` ; un magasin qui propose des `Mangues` devra proposer en plus des `Pommes` et des `Bananes` ; un magasin peut proposer uniquement des `Pommes`.

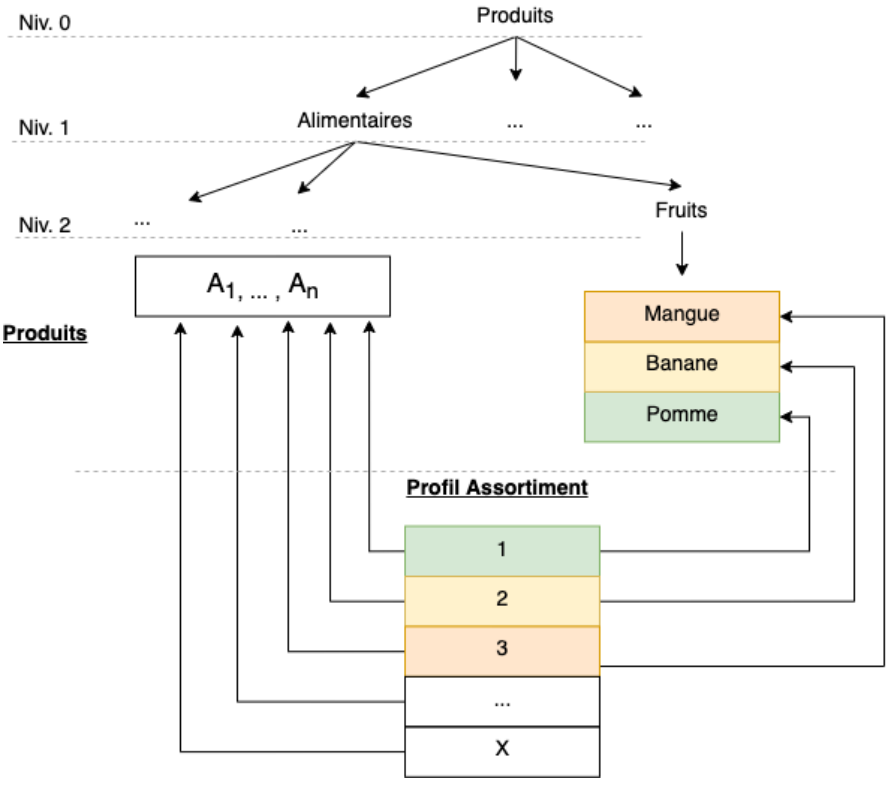


FIGURE 2 – Exemple illustratif de l'assortiment gigogne

A chacun des magasins de l'enseigne est ensuite attribué un profil assortiment. Ce profil assortiment "magasin" est généralement spécifié pour chaque catégorie. Cela permet de personnaliser l'assortiment et améliorer la réponse aux besoins clients. En croisant les profils assortiment associés aux produits et aux magasins les assortiments finaux, peuvent être identifiés. Ainsi en reprenant l'exemple de la Figure 2, un magasin de profil assortiment "2" devra avoir dans son assortiment des `Bananes` et des `Pommes`. Un exemple illustratif plus complet est disponible en Annexe 7.

Les *profils assortiment* permettent de faire émerger une structure hiérarchique qui définit ce que l'on appelle le *cœur d'assortiment*. Nous pouvons représenter le *cœur d'assortiment* de chacun des magasins sous la forme d'un vecteur binaire dont chaque dimension indique la présence ou non d'un article.

A chacun des sous-ensembles de produits sont généralement associées des contraintes de stock et de linéaire. Ces contraintes permettent de contrôler la possibilité de mise en place de l'assortiment dans les magasins. Autrement dit, une petite supérette ne pourra pas inclure tous les produits proposés par l'enseigne de par sa superficie limitée. C'est pourquoi l'identification de sous-ensembles pouvant être inclus dans tous les magasins est indispensable. De plus, les produits associés à des profils assortiment qui seront inclus dans la plupart des magasins doivent être choisis avec minutie. Nous retrouvons généralement les Best-Sellers ou appelés aussi 80/20 (les 20% d'articles qui font à eux seuls 80% du chiffre d'affaires).

Les profils assortiment peuvent étendre la taxonomie de produits comme l'illustre la Figure 3. Cette extension peut être effectuée à tous les niveaux de la structure marchandise dans la mesure ou chaque produit est associé à un unique profil assortiment.

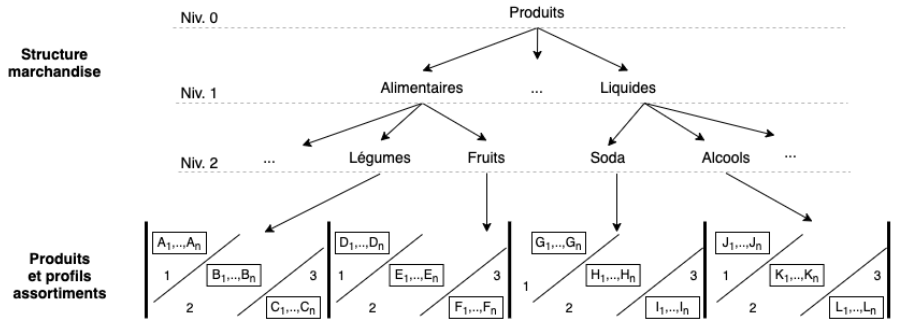


FIGURE 3 – Exemple illustratif d'une structure marchandise étendue avec les profils assortiment

Ces profils assortiment permettent donc d'homogénéiser les produits à travers un vaste réseau de magasins. Cependant, les habitudes d'achats et les besoins des clients varient d'un magasin à un autre. Par exemple, les ventes des `Vins` seront très différentes à Bordeaux

et à Lyon. Pour justement permettre une certaine flexibilité dans la définition de l'assortiment des magasins, certains grands distributeurs utilisent l'assortiment *"marguerite"*. Cette nouvelle définition de l'assortiment gigogne permet de gérer des exceptions dans le *cœur d'assortiment*. Pour cela, un profil d'assortiment optionnel est associé à certains produits qui pourront être choisis pour combler l'assortiment des magasins. Le principal intérêt des produits optionnels est d'offrir l'opportunité aux magasins d'adapter leurs assortiments en fonction de besoins locaux.

En plus des contraintes de stock et de linéaire, les enseignes ajoutent d'autres contraintes visant à véhiculer au mieux l'image de la société. Par exemple, proposer un produit *Haut de gamme* et produit *Premier Prix* a minima pour chacune des unités de besoin, ou encore, proposer tous les articles de la marque du distributeur. Ces contraintes ne sont pas toujours explicites et leur respect va de paire avec leur compréhension. Ainsi, certains magasins d'une même enseigne possèdent, dans certains cas, des assortiments très différents pouvant même être contradictoires avec la volonté de l'enseigne.

L'objectif final de l'assortiment reste d'optimiser les performances des magasins de l'enseigne. Nous revenons au point névralgique de la grande distribution : le *consommateur* ou *client final*. La performance d'un assortiment dépend principalement de sa capacité à satisfaire les exigences des clients. L'identification de ces exigences est une tâche cruciale dans l'optimisation de l'assortiment. Néanmoins, la précision des connaissances sur les consommateurs dépend de la capacité du distributeur à les identifier. Aujourd'hui, la carte de fidélité est la solution la plus commune pour identifier l'ensemble des achats effectués par un consommateur. Cependant, elles ne sont pas systématiquement utilisées et certains clients n'en possèdent tout simplement pas. Les connaissances des clients sont généralement exploitées au travers de CRM (Customer Relationship Management) et servent principalement aux communications marketing et commerciales. Les principales approches consistent à exploiter les CLV (*Customer Life Value*) et RFM (*Revenue Frequency Margin*) des clients de façon à identifier les "bons" clients, qui ne sont finalement que des clients réguliers qui génèrent de la marge et du revenu. Des études empiriques ont démontré que le choix des clients dépend de la perception qu'ils ont de la variété de produits proposés, plutôt que de la variété elle-même. Cette perception peut être influencée par la présence d'articles favoris [23], l'agencement et la disposition des produits en magasin [147] etc.

L'exploitation des connaissances clients reste minime dans la grande distribution néanmoins, des contributions scientifiques proposent d'utiliser ces connaissances pour améliorer l'assortiment. Elles sont détaillées dans la section suivante.

2.2 ÉTAT DE L'ART : PLANIFICATION DE L'ASSORTIMENT

L'assortiment d'un grand distributeur correspond à l'ensemble des produits proposés dans ses magasins. La problématique scientifique associée est celle de la planification de l'assortiment qui consiste à définir ou, à identifier le meilleur ensemble de produits. Dans la littérature, nous remarquons que cette problématique a été abordée de différentes façons. Elle s'est diversifiée selon les résultats des avancées scientifiques mais aussi en fonction des changements des méthodes de ventes ou de l'évolution des habitudes d'achats des consommateurs.

2.2.1 *La genèse*

L'ensemble des produits dans un magasin : l'assortiment, a tout d'abord été vu comme un problème de variété des produits et d'identification des gammes.

La variété des produits et l'identification des gammes ont fait l'objet de nombreuses études dans l'industrie. Principalement pour savoir si les assortiments étaient trop "larges" ou trop "étroits". Les grands distributeurs ont longtemps adopté une stratégie visant à laisser les *Catégories Managers* maximiser les profits de leur catégorie [122] par l'ajout permanent de nouveaux produits et le retrait exceptionnel des produits jamais vendus [7]. Une augmentation significative de la variété des produits était inévitable.

Le principal modèle économique qui étudie la variété des produits repose sur une situation d'oligopole basée sur le modèle Hotelling [66]. Ce modèle revient à considérer que les consommateurs sont répartis uniformément sur un segment de droite et les grands distributeurs choisissent la position de leurs magasins et un prix permettant de maximiser leurs profits. L'utilité ou le rendement des consommateurs vis-à-vis des distributeurs diminue en fonction des prix définis par les distributeurs et de la distance physique entre le distributeur et le consommateur. Finalement, chaque consommateur choisit le magasin qui lui fournit l'utilité maximale. Le modèle Hotelling est illustré dans la Figure 4.

L'objectif des distributeurs est d'identifier le parfait équilibre entre le nombre de magasins, leur implantation et leurs prix ... pour maximiser le bien-être, donc l'utilité des consommateurs, qui en résulte. Les extensions de ce modèle permettent la différentiation des produits. La différenciation horizontale implique que les caractéristiques des produits ne sont pas comparables, d'où la proposition de différentes catégories de produits. Tandis que la différenciation verticale se base sur une préférence des consommateurs pour certaines caractéristiques. Par exemple, la qualité de deux produits comparables (*e.g.* `Coca-Cola` et `Pepsi`) entre dans une différenciation verticale et une meilleure qualité de produit sera toujours préférée par les clients [6, 90].

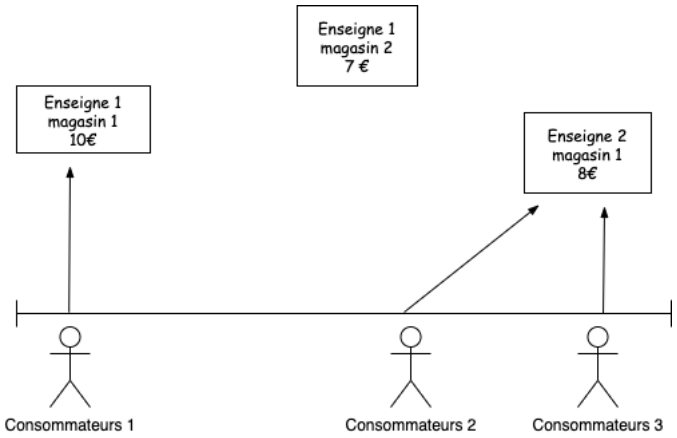


FIGURE 4 – Exemple illustratif du modèle Hotelling

À la problématique de variété de produits se greffe donc la problématique de gamme de produits [115, 117]. Par exemple, considérer une voiture comme un produit à un seul attribut (*e.g.* la taille du moteur), implique que le distributeur doit identifier la taille idéale de moteur en fonction de coûts associés. Une solution proposant une formulation mathématique des coûts permettant d'identifier des gammes pour chaque produit inclut dans l'assortiment a été proposée [42]. C'est ainsi que des notions de produits "bas de gamme" (moins cher), "haut de gamme" (plus cher) sont apparues puisque le coût de revient du produit impactait le prix de vente du produit. Une étude propose d'exploiter la similitude des composants permettant d'identifier des gammes multi-attributs [40]. En 2005, le modèle de Quantité Economique de Commande ou EOQ (*Economic Order Quantity*), défini par Wilson en 1934, est utilisé pour prendre en considération des économies d'échelles [120].

Ces différents modèles étaient les prémices de la planification de l'assortiment. La restriction du nombre de produits, leur positionnement et leur tarification ont, petit à petit, été introduits [30]. Pendant de nombreuses années les grands distributeurs ont élargi leurs assortiments à tel point que des études ont fini par démontrer que réduire l'assortiment ne diminuait plus les ventes [20, 23, 45]. Un assortiment plus large implique des coûts plus élevés. La rationalisation de l'assortiment est devenue nécessaire pour les grandes enseignes qui avaient un manque à gagner plus important en réduisant la variété de produits. De plus, l'optimisation de la chaîne de distribution était obligatoire et cela passait par une gestion des stocks optimale.

Ces changements sont venus apporter des contraintes à la problématique, donnant une nouvelle interprétation à la définition d'un assortiment dans la distribution. Le problème de planification d'assortiments rappelle forcément le problème combinatoire du sac à dos (*Knapsack problem*) [78]. En effet, le fait de choisir un ensemble d'items à mettre dans un sac à dos, avec un coût et une utilité propre pour

chacun de ces items laisse penser que ces deux problématiques sont très proches. Dans la littérature de la distribution, nous retrouvons cette approche sous le nom de problème d'inventaire multi-articles (*multi-item inventory*) [8, 44]. La gestion de plusieurs produits sous contraintes est étudiée avec des modèles utilisant des multiplicateurs lagrangiens [56]. Ces modèles permettent d'identifier des points stationnaires (maximum, minimum ...) des sous-ensembles de produits optimums. Les dernières extensions de ces travaux prennent en considération la notion temporelle des produits (cycle de vie d'un produit) introduisant des pertes éventuelles de ventes [44]. Néanmoins, la substituabilité entre les produits n'est pas prise en considération.

Nous avons tous eu l'expérience d'aller dans un magasin à la recherche d'un produit particulier, de ne pas de le trouver et de se contenter finalement d'un autre produit similaire. C'est ce qui s'appelle la substitution. Le capacité d'un client à prendre un produit de substitution est un paramètre important dans la planification de l'assortiment. Deux produits répondant à une même unité de besoin (UB) peuvent être considérés comme substituables.

D'autres groupes de modèles d'inventaire multi-articles sont donc proposés pour pallier ce problème [119, 123, 133]. Ils considèrent que la substitution des produits n'intervient qu'en cas de rupture. Autrement dit, quand le produit n'est plus en stock. Par exemple, une famille favorisera toujours le lot de produits tandis qu'une personne seule choisira toujours le produit individuel. Ils changeront leurs habitudes uniquement si leur produit favori est indisponible. Donc si le produit qui intéresse le client n'est plus disponible, ce dernier prendra un produit similaire. Ces modèles se focalisent sur des notions de stockage en fonction d'une sélection de produits plutôt que sur la sélection des produits en se basant sur une demande exogène des consommateurs (décrite plus loin dans cette section).

La performance d'un produit est vue comme son utilité pour les consommateurs. L'utilité totale d'un produit est considérée comme la somme de son utilité intrinsèque et l'utilité pondérée des produits qui lui sont substituables [113]. Introduit uniquement pour deux produits, des heuristiques beaucoup plus complexes ont été définies pour proposer des solutions pour N produits substituables [123, 133]. Une première étude suggérant une solution décentralisée du problème a été proposée dans [126]. En 2003, la problématique consistant à distribuer N produits dans un cadre simultanément centralisé et décentralisé (cas de la grande distribution) est définie comme trop complexe pour obtenir une solution possible et explicite [119]. En effet, l'identification d'un sous-ensemble fini de produits parmi des milliers permettant d'optimiser la performance de milliers de points de ventes et satisfaisant des millions/milliards de consommateurs est impossible (np-complet).

La planification de l'assortiment s'est alors décomposée en différentes problématiques. Certaines s'intéressent à la définition de règles d'allocation initiale et de réallocation permettant au distributeur de contrôler la substitution [9]. La solution de réallocation est dans ce cas obtenue en résolvant des problèmes d'optimisation comme celui de tournées de véhicules ou du voyageur de commerce. D'autres se sont focalisées sur la gestion d'allocation d'espace en rayons. Particulièrement pertinente pour les produits à forte rotation (vendus à grande fréquence), l'allocation d'espace en rayons (*Shelf Space Allocation*) a impacté significativement la grande distribution alimentaire. Une étude empirique introduit la notion d'élasticité de l'espace en rayon qui intègre en plus les interactions entre les produits proposés à la vente [36]. L'expérience consiste à estimer les ventes α d'un produit i comme étant :

$$\alpha_i = s_i^{\beta_i} \prod_j s_j^{\delta ij}$$

où, s_i représente l'espace alloué au produit i, β_i son espace "élastique" et δij l'espace "élastique" partagé entre les produits i et j. La maximisation du profit se transforme en un problème de programmation géométrique. Les solutions proposées par ce modèle ont obtenu de meilleurs résultats qu'une allocation de l'espace proportionnelle aux ventes ou aux produits bruts qui eux ignorent les interdépendances entre les produits. Cette étude a permis d'améliorer la précision des prévisions des ventes grâce à la prise en considération de l'impact de l'espace alloué aux produits [21]. Le modèle proposé a été très largement étendu et cette problématique d'optimisation de l'espace en rayon est un thème de recherche à part entière abordé de nombreuses façons [69, 86]. Cependant, ces modèles ne traitent pas explicitement de la sélection des assortiments et ignorent la nature stochastique de la demande des consommateurs. Les modèles les plus récents introduisent les demandes des clients afin d'améliorer la précision des résultats [67]. Cependant, la substitution des produits n'intègre pas la demande client. La nécessité d'intégrer le consommateur dans la définition d'un assortiment a permis l'apparition de nouveaux modèles.

2.2.2 *Le modèle Multinomial Logit*

Ces modèles probabilistes non linéaires considèrent que la substitution et la demande des clients sont des ensembles d'éléments stochastiques combinés à des variables explicatives (modèle Logit ou modèle Multinomial Logit) [105]. Les modèles Multinomial Logit (MNL) se basent sur l'hypothèse fondamentale qu'un consommateur est un maximiseur d'utilité rationnelle. Autrement dit, les consommateurs privilégieront systématiquement les produits qui maximisent

l'utilité totale. Les notations suivantes seront utilisées tout au long de cette section :

— N l'ensemble des produits d'une catégorie, $N = \{1, 2, \ldots, n\}$
— S le sous-ensemble de produits détenus par l'enseigne, $S \subset N$
— λ nombre moyen de clients visitant le magasin par période
— c_j, r_j respectivement le prix d'achat et de vente du produit j

Des notations et indices supplémentaires seront introduits lorsque nécessaire.

La substitution

Nous pouvons distinguer deux types de substitution. Celle engendrée par les ruptures de stock, introduite précédemment et que nous pouvons associer à un *report de ventes*. Dans ce cas là, le distributeur ne perd pas de ventes car le client prend un autre produit. Autrement, il s'agit de la substitution basée sur le changement spontané du consommateur vers une variante de l'assortiment. Dans ce cas précis, le distributeur perd des ventes sur le produit initial. Cette seconde substitution est appelée la *cannibalisation de produits* [102]. Le modèle Multinomial Logit (MNL), couramment utilisé en économie et marketing, considère l'utilité d'un produit j comme étant la combinaison d'une composante déterministe (u_j) et une composante aléatoire (ϵ_j) [84] :

$$U_j = u_j + \epsilon_j$$

Cette utilité est définie pour l'ensemble des consommateurs. La composante déterministe est associée aux choix faits parmi les produits $S \cup \{\emptyset\}$. $\{\emptyset\}$ représente le cas où un client ne prend aucun produit. La composante aléatoire est généralement associée à une variable aléatoire de Gumble, connue également sous le nom de Double-Exponentiel ou d'extrémum généralisée (*Extreme value Type-I*). ϵ_j se caractérise par la distribution suivante :

$$Pr\{X \leq \epsilon\} = e^{-e^{\epsilon/\mu+\gamma}}$$

où, γ correspond à la constante d'Euler-Mascheroni $(0,5772156649)$ et μ est associé au degré d'hétérogénéité des clients, ϵ_j est indépendant d'un consommateur à l'autre. Par conséquent, alors que chaque consommateur a la même utilité attendue pour chaque produit, l'utilité ressentie peut être différente. Ce ressenti peut être dû à l'hétérogénéité des préférences entre les clients ou à des facteurs non observables dans l'utilité du produit à l'individu. Rappelons que nous partons de l'hypothèse qu'un individu choisit toujours le produit avec l'utilité la plus élevée parmi l'ensemble des choix disponibles. Par conséquent, la probabilité qu'un individu choisisse le produit j parmi $S \cup \{\emptyset\}$ correspond à :

$$p_j(S) = Pr\{U_j = \max_{k \in S \cup \{\emptyset\}} (U_k)\}$$

Grâce à la propriété des bornes de la distribution de Gumbel sous maximisation [6], la probabilité qu'un client choisisse un produit j parmi $S \cup \{\emptyset\}$ devient :

$$p_j(S) = \frac{e^{u_j/\mu}}{\sum_{k \in S \cup \{\emptyset\}} e^{u_k/\mu}}$$

Cette propriété permet au modèle MNL d'être un candidat idéal pour modéliser le choix du consommateur. Dans l'industrie du voyage [12], la différenciation des produits [10], le marketing [53] ... les chercheurs ont constaté que le modèle MNL était très utile et très performant pour la modélisation du choix des clients. Pour plus de détails sur le modèle MNL, l'estimation de ses paramètres et sa relation avec d'autres modèles de choix, nous invitons le lecteur à se diriger vers les références suivantes [6, 53, 104, 112]. Les chercheurs ont proposé de nouveaux modèles de planification d'assortiments intégrant cette substitution axée sur le consommateur. Ces modèles se concentrent sur des décisions concernant l'assortiment d'un seul magasin à un instant donné pour proposer une solution optimale [84]. Le modèle van Ryzin et Mahajan [104] formule le problème de planification de l'assortiment comme la maximisation du profit d'une catégorie de produits. Ils définissent le profit d'un produit j ($j \in S$) comme étant :

$$\pi_j(S) = (r_j - c_j)\lambda p_j(S) - C(\lambda p_j(S))$$

où $C(\cdot)$ est la fonction représentant les coûts d'exploitation et d'approvisionnement. Pour refléter les économies d'échelle ils considèrent cette fonction de coûts concave et croissante. L'objectif de leur modèle est d'identifier le sous-ensemble S ($S \subseteq N$) en résolvant :

$$\max_{S \subset N} Z = \sum_{j \in S} \pi_j(S) \tag{1}$$

L'assortiment optimal est donc un équilibre entre l'inclusion d'un nouveau produit et les coûts engendrés par cette inclusion. L'inclusion d'un produit j dans l'assortiment ($S_j = S \cup \{j\}$) correspond à la fonction :

$$h(v_j) = \pi_j(S_j) - (\sum_{k \in S} \pi_k(S) - \sum_{k \in S} \pi_k(S_j))$$

avec, $v_j = e^{u_j/\mu}$ (*cf.* modèle Multinomial Logit), $h(v_j)$ est quasi-convexe sur l'intervalle $[0, \infty)$ [25]. Par définition, cette fonction atteint son maximum aux extrémités de l'intervalle. L'utilité est donc maximale soit en n'ajoutant pas de produit à l'assortiment, soit en ajoutant le produit avec la plus grande utilité. Cette observation conduit au résultat suivant qui caractérise la structure de l'assortiment optimal où les sous-ensembles de produits sont définis par ordre d'utilité décroissante :

$$P = \{\{\emptyset\}, \{1\}, \{1,2\}, \ldots, \{1,2,\ldots,n\}\},$$

$$\forall i, u_i > u_{i+1} \wedge \{u_i, u_{i+1}\} \leq \{u_i, u_{i+1}, u_{i+2}\}$$

L'assortiment optimal est l'un des assortiments ainsi créé [25]. Ce résultat intuitif permet de réduire la complexité du problème de 2^n ensembles à n étant donné que seule la substitution basée sur l'assortiment est prise en compte. À l'origine, ce modèle a montré de bons résultats lorsque le profit π_j d'un produit considère une fonction D de coût associée au modèle de vendeur de journaux [31]. Plus précisément, ils utilisent les coûts d'un modèle de fournisseur de journaux en supposant qu'ils sont distribués selon une loi normale de moyenne λ et d'écart-type δ. La quantité de stock du vendeur de journaux pour un produit j doit être égale au niveau optimal de stock d'un produit (x_j) :

$$x_j = \lambda p_j(S) + z\delta(\lambda p_j(S))^\beta$$

où, z et β sont deux paramètres permettant de contrôler le coefficient de variation de la demande d'un produit j. Ils redéfinissent ainsi la fonction de coût C :

$$C(\lambda p_j(S)) = r_j \frac{e^{-z^2}}{\sqrt{2\pi}} (\lambda p_j(S))^\beta$$

Ce modèle capture le principal compromis entre la variété et l'augmentation des coûts d'inventaire moyens. Le même problème est étudié en proposant une solution tenant compte d'une substitution dynamique dans [106]. Schématiquement, plusieurs produits peuvent se substituer les uns aux autres et le client n'achète rien uniquement si aucun produit de substitution n'est disponible. Ce modèle a été enrichi en considérant : le prix comme partie prenante de la décision du consommateur [101] ; l'incertitude dans la décision des consommateurs [25] ; une répartition plus générale de la demande [84]...

La principale critique du modèle MNL provient de sa propriété d'indépendance des alternatives non pertinentes (IIA) (*Independence of Irrelevant Alternatives*) qui est un axiome des théories décisionnelles utilisé en sciences sociales. Bien que les formulations des IIA peuvent varier, elles ont comme point commun d'essayer de rationaliser le comportement individuel dans une situation d'agrégation de préférences individuelles.

Le modèle Multinomial Logit Imbriqué (*Nested Multinomial Logit*) est introduit [12] pour traiter la propriété IIA. Le modèle est le même sauf que la composante d'utilité non observée est corrélée au choix des consommateurs plutôt que d'en être indépendante. Le modèle Multinomial Logit Imbriqué propose un processus en deux étapes utilisé pour modéliser le choix, par exemple, d'une marque puis d'une unité de besoin. Cela permet d'identifier différentes utilités pour un même produit et un même consommateur.

Pour plus de détails sur le modèle Multinomial Logit Nested, nous invitons le lecteur à consulter [6]. Ce modèle a, par exemple, été utilisé pour modéliser la compétition entre deux enseignes (multi-produits) dans différentes études [6, 26].

2.2.3 *Le modèle de demandes exogènes*

Le modèle de demandes exogènes considère que les hypothèses suivantes caractérisent pleinement le comportement de choix des clients :

1. Chaque client choisit toujours son produit favori parmi N. La probabilité qu'un client choisisse un produit j est notée p_j avec :

$$\sum_{j \in N \cup \{\emptyset\}} p_j = 1$$

2. Si le produit favori j n'est pas disponible, le client choisit son deuxième produit favori k avec une probabilité δ et une probabilité $1 - \delta$ de décider ne pas l'acheter. La probabilité que j soit substitué par k est définie par α_{kj}.

Dans le cas où le second produit favori n'est pas non plus disponible, l'opération se répète. La probabilité que le client n'achète pas $(1 - \delta)$ et la probabilité de substitution peuvent rester les mêmes pour chaque tentative répétée ou bien être spécifiée différemment. Ce modèle considère qu'il n'y a pas de comportement sous-jacent du consommateur. Il est le plus couramment utilisé dans la littérature sur la gestion des stocks de produits substituables. La probabilité de substitution (α_{kj}) est déterminée par une matrice de probabilité. La forme de cette matrice définit le mécanisme probabiliste utilisé pour calculer la substitution entre produits (*e.g.* matrice de substitution adjacente, matrice de substitution proportionnelle, etc.) [84].

L'avantage du modèle de demandes exogènes est de pouvoir différencier des catégories de produits avec des taux de substitution faibles et élevés avec un seul paramètre δ. Chaque matrice suppose des contraintes différentes entre les catégories. Seule la matrice de substitution proportionnelle a des propriétés cohérentes avec le modèle MNL. En effet, c'est la seule matrice de substitution qui considère qu'un magasin ne contient pas tous les produits ($S \subseteq N$). Cela signifie qu'un consommateur qui ne trouve pas son produit favori dans le magasin est plus susceptible d'acheter un substitut à mesure que l'ensemble des substituts potentiels augmente. Le modèle de demandes exogènes suppose ainsi finalement qu'il n'y a plus de tentatives de substitution. Soit le produit de remplacement est disponible dans le magasin et la vente est conclue, soit la vente est perdue. Des études démontrent que limiter le nombre de tentatives de substitution n'est pas restrictif [4, 82].

Si ce modèle s'est autant répandu, c'est parce qu'il dispose de plus de degrés de liberté que le modèle MNL. Étant donné que les options de l'ensemble de choix sont supposées être homogènes, le modèle MNL n'est pas en mesure de saisir les types de substitution adjacente, la substitution d'un produit ou la substitution au sein d'un sous-groupe. Dans le modèle MNL, les taux de substitution dépendent de l'utilité relative des options. C'est à la fois un avantage et un inconvénient pour le modèle MNL.

L'avantage est qu'il permet d'intégrer facilement des variables marketing telles que les prix et les promotions dans le modèle de choix. L'inconvénient est qu'il ne peut pas faire de différence entre le choix initial et le comportement de substitution. Contrairement au modèle MNL, le modèle de demandes exogènes peut différencier les catégories qui peuvent avoir la même demande initiale pour la catégorie mais des taux de substitution différents par le choix des consommateurs. Par conséquent, le modèle MNL ne peut pas traiter les substitutions basées sur l'assortiment et les ruptures de stock de manière différente. En revanche, il est possible d'utiliser des matrices de probabilités de substitution différentes pour les substitutions basées sur l'assortiment et les ruptures de stock dans le modèle de demandes exogènes [84].

Des modèles de planification d'assortiments exploitant le modèle de demandes exogènes ont été proposés. En 1996, N. Agrawal et S.A. Smith ont constaté que les données de ventes dans la distribution s'accordaient très bien avec la distribution binomiale négative (NBD) (*Negative Binomial Distribution*) [3]. Ils proposent d'identifier des bornes inférieures et supérieures au problème. Ces bornes permettent de formuler une heuristique spécifiant le sous-ensemble de produits optimal pour un magasin. Le modèle de Smith et Agrawal considère que le niveau des stocks évolue à chaque nouveau client. Pour un assortiment S donné, une fonction f_j définit le niveau de service de chaque produit. $A_k(S, m)$ est une variable binaire indiquant si le

produit k est disponible pour le client m. La probabilité que le m^{ime} client achète un produit j est notée $g_j(S, m)$:

$$g_j(S, m) = d_j + \sum_{k \notin S} d_k \alpha_{kj} + \sum_{k \in S \setminus \{j\}} d_k \alpha_{kj} (1 - A_k(S, m))$$

où d_j correspond aux taux de demande moyens pour le produit j ($d_j = \lambda p_j$). La complexité de la fonction $g_j(S, m)$ oblige à exploiter les propriétés des distributions binomiales négatives [84]. Ils considèrent que la demande de chaque produit devrait suivre la NBD. Le modèle de Smith et Agrawal rejoint le modèle proposé par van Ryzin et Mahajan en posant un problème d'optimisation visant à maximiser le total des profits d'une catégorie (*cf.* équation 1). Cependant ils ajoutent une fonction de profit. Ce problème d'optimisation se transforme en problème de programmation d'entier non linéaire et l'intégration d'une nouvelle contrainte est triviale (coûts de détention, contrainte budgétaire, espace en rayon ...). Les chercheurs proposent de résoudre le problème en utilisant une approche de relaxation lagrangienne [1] pour supprimer des contraintes complexes en les intégrant dans la fonction objectif suivi d'une recherche unidimensionnelle sur la variable duale qu'il désigne. En théorie de l'optimisation, les problèmes peuvent être vus selon deux perspectives : le problème primal ou le problème dual. La solution du problème dual apporte une borne supérieure à la solution du problème primal. Cependant, les valeurs optimales des problèmes primal et dual ne sont pas systématiquement égales (saut de dualité).

Ce modèle est mis en application et enrichi dans de nombreux travaux de recherche [96, 132, 155]. Notamment au travers du modèle de Kök et Fisher [83].

2.2.4 *La planification de l'assortiment : bilan*

A partir des notions et propriétés précédemment introduites, de nouveaux modèles ont été proposés.

La planification de l'assortiment a pu être guidée par la demande d'emplacement (*locational choice*) [48]. Développé par Hotelling [66] une différence majeure existe avec le modèle MNL. Dans le modèle MNL, la substitution peut se produire entre deux produits quelconques. Tandis que dans le modèle de demande d'emplacement, la substitution entre les produits est localisée à des produits dont les caractéristiques sont proches. Ce modèle s'affranchit de la propriété IIA. Certains ont étendu le modèle de planification d'assortiments

1. La relaxation lagrangienne est une manipulation usuelle en optimisation sous contraintes permettant d'obtenir des bornes pour les valeurs optimales de problèmes d'optimisation combinatoire. Cela repose sur le fait de relaxer une partie des contraintes sous la forme de pénalités combinées linéairement.

proposé par van Ryzin et Mahajan à une chaîne d'approvisionnement décentralisée [64, 87, 88, 148]. L'assortiment résultant appartient toujours aux assortiments définis par utilité décroissante (*cf.* Section 2.2.2) avec cependant une faible variété de produits.

La recherche s'est principalement concentrée sur la planification d'assortiments pour une seule catégorie [84]. Néanmoins, une étude a démontré l'évidente corrélation qui pouvait exister entre des catégories de produits [107]. Cette corrélation peut provenir d'une complémentarité naturelle des catégories (*e.g.* les `Céréales` et le `Lait`) ou en raison de "coïncidences" (*e.g.* les `Bières` et les `Couches`). Les propositions académiques introduisent principalement des connaissances visant à améliorer le processus managérial de la planification d'assortiments dans un environnement à multi-catégories [24, 84].

Les innovations technologiques ont permis la mise en place de chaînes d'approvisionnement hautement réactives et flexibles (*e.g.* Zara). Les transformations industrielles associées ont changé l'état du problème de planification d'assortiments de statique à dynamique. La capacité limitée à réviser les assortiments de produits est devenue problématique tant pour les chercheurs que pour les industriels. Le problème dynamique de l'assortiment a donc été formulé de la façon suivante :
Au début de chaque période, l'enseigne décide quel assortiment doit être proposé. Elle recueille des données sur les produits proposés dans l'assortiment à chaque période. Chaque période dispose d'une contrainte concernant le nombre de produits proposés K. Le problème se rapporte alors au commerce d'exploration contre exploitation. L'enseigne doit décider s'il faut exploiter (optimiser les revenus en fonction des informations actuelles, soit proposer des produits déjà distribués dans au moins un point de ventes), ou bien explorer (en apprendre davantage sur la demande de produits non proposés, ce qui revient à ajouter de nouveaux produits dont la performance reste inconnue) [27, 93].

En résumé, les chercheurs proposent la formulation d'un problème d'optimisation permettant d'élire un ou plusieurs ensembles de produits qui maximisent la "performance" d'un magasin sous certaines contraintes. Les caractéristiques utilisées pour contraindre l'espace de recherche se focalisent principalement sur l'espace linéaire, les UB, les fournisseurs, la marque ... sans prendre en considération les corrélations existantes entre ces caractéristiques. L'optimisation est généralement proposée sur une unique catégorie de produits [84].

2.3 CONCLUSION

Les approches adoptées dans le domaine de la recherche et le milieu industriel sont, à certains égards, très complémentaires. L'industrie

s'est focalisée sur sa stratégie avec une approche globale, tandis que les chercheurs se sont dirigés vers une approche plus détaillée.

Les enseignes considèrent que la planification d'un assortiment commence par la stratégie. En effet, la variété de produits et les gammes proposées définissent la position de l'enseigne dans le paysage concurrentiel. Ce phénomène de stratégie est reconnu par les scientifiques [25] mais à notre connaissance aucun travail de recherche ne l'intègre. Les grands distributeurs ajoutent à la stratégie le rôle de chacune des catégories généralement défini par la matrice BCG (*Boston Consulting Group*). Cette matrice prend en compte le taux de croissance du marché et la part de marché relative à l'enseigne pour "ranger" les catégories de produits en quatre groupes :

— vaches à lait ou trafic (leader sur un marché mature)

— vedettes ou destination (leader sur un marché en croissance)

— dilemmes ou image (challenger sur un marché en croissance)

— poids morts ou service (challenger sur un marché en déclin)

Cette vision, peu étudiée dans la littérature de planification d'assortiments, est largement utilisée dans la littérature marketing.

Les distributeurs reconnaissent à l'unanimité la spécificité de chacune des catégories (*e.g.* `Conserves` et `Produits Frais`) [41]. Les catégories managers sont donc chargés d'exploiter ces spécificités afin d'optimiser la performance de chacune des catégories d'une enseigne. Néanmoins, les catégories sont traitées indépendamment malgré l'évidence de corrélations existantes [5].

Les enseignes sont conscientes de la nature dynamique du problème et les magasins permettant de tester de nouveaux produits sont de plus en plus nombreux. Les résultats obtenus sur ces magasins permettent d'améliorer l'assortiment gigogne de l'enseigne. Même dans les catégories matures, l'introduction fréquente de nouveaux produits oblige à revoir les assortiments.

La personnalisation de l'assortiment au niveau du magasin a peu retenu l'attention des distributeurs à défaut de celle des chercheurs. La littérature de planification de l'assortiment dans la grande distribution propose principalement des solutions logicielles engendrant des assortiment sous-optimaux pour le problème de planification d'assortiments [51].

Aujourd'hui, les distributeurs ont développé des processus qui traitent les complexités du monde dans lequel ils vivent. Néanmoins, seuls quelques catégories managers s'occupent de la problématique de planification d'assortiments qui est un problème NP-complet [51]. Ces processus omettent des variables critiques comme les corrélations entre catégories, l'aspect périodique d'un assortiment et la nécessité d'une personnalisation des assortiments pour chacun des magasins.

Dans ce manuscrit nous proposons, avec le soutien et l'aide de TRF Retail, une méthode Agile d'optimisation de l'assortiment. Cette

méthode repose sur l'existence de structure de connaissances dédiée à la grande distribution. Le chapitre suivant introduit ce type de structures que nous pouvons retrouver dans la grande distribution ainsi que les méthodes permettant de les exploiter.

3

STRUCTURES DE CONNAISSANCES

Dans ce chapitre, nous introduisons les structures de connaissances disponibles chez les grands distributeurs. Nous défendons dans cette thèse l'idée que la définition d'un assortiment mérite d'être traitée au travers de notions plus précises que les seules catégories de produits. Les modèles scientifiques proposés dans la section 2.2 et les processus industriels requièrent, pour répondre aux problématiques d'aujourd'hui, une représentation des connaissances plus formelle, notamment, via les ontologies.

Dans ce chapitre, nous proposons avec le support et l'expertise de TRF Retail, la définition de structures de connaissances dédiées à la grande distribution.

Ce chapitre est organisé de la manière suivante. Dans la section 3.1, nous définissons formellement ce qu'est une ontologie et les mesures de similarité sémantique associées. Dans la section 3.3, nous proposons d'exploiter la taxonomie de produits et les mesures de similarité sémantique. Cette contribution apporte une nouvelle approche éclairant l'identification des habitudes d'achats des consommateurs. Enfin, dans la section 3.2, nous présentons les principales représentations des connaissances spécifiques au domaine de la grande distribution qui sont exploitées grâce au modèle de données de TRF Retail.

3.1 ÉTAT DE L'ART : STRUCTURES DE CONNAISSANCES ET MESURES SÉMANTIQUES

La méthodologie proposée dans ce manuscrit repose sur l'exploitation de représentations formelles des connaissances : les ontologies. Les ontologies sont des structures reposant sur une connaissance experte a priori (*e.g.* structure marchandise). Ces modèles de connaissances sont utilisés en particulier pour identifier et proposer des solutions en lien direct avec les connaissances du domaine. Cette section formalise dans un premier temps les modèles taxonomiques, le langage de représentation des connaissances (OWL) ainsi que certains modèles de connaissances existants et leurs applications. Nous présentons dans un second temps l'état de l'art relatif aux mesures sémantiques qui permettent de définir la notion de similarité entre les concepts des structures de connaissances (*e.g.* ontologies, taxonomies ...) qui sera utilisée, par la suite, dans nos travaux.

3.1.1 *Techniques de représentation des connaissances*

Les taxonomies permettent de structurer les éléments d'un domaine qui présentent des caractéristiques similaires en classes ordonnées et hiérarchisées. Très intuitives, elles ont été initialement proposées en Biologie pour définir des classes (concepts, taxons) regroupant les organismes qui partagent des propriétés communes . Un modèle de connaissances de type taxonomique permet une classification systématique et hiérarchisée de taxons dans diverses catégories selon les caractères qu'ils ont en commun, des plus généraux aux plus particuliers. Par exemple, la classe (ou le concept) `Fruit`, sera plus générale que la classe `Fruit Rouge`, elle-même plus générale que la classe `Fraise`. La sémantique portée par une taxonomie de classes est non-ambiguë, l'interprétation de la relation taxonomique est formellement exprimée par des propriétés/ axiomes, et correspond à la relation d'inclusion considérée en logique formelle.

Les relations taxonomiques ont la particularité d'être i) réflexive, ii) transitive et i) antisymétrique. Ce qui formellement signifie :

(i) Chaque classe est subsumée par elle-même.

(ii) Si une classe x subsume une classe y, qui elle-même subsume une classe z, alors x subsume z.

(iii) L'inclusion d'une classe d'individus x dans une classe d'individus y implique que y n'est pas incluse dans x.

Remarque : nous considérons que les notions de concepts, de classes ou encore de catégories font référence à une notion commune regroupant des objets de même nature.
Formellement, une taxonomie définit un ordre partiel sur un ensemble de concepts C. Cet ordre peut être exprimé par une relation binaire $\preceq$ entre les concepts qui est :

(i) réflexive : $\forall c \in C : c \preceq c$

(ii) transitive : $\forall x, y, z \in C : (x \preceq y \preceq z) \Longrightarrow x \preceq z$

(iii) antisymétrique : $\forall x, y \in C : (x \preceq y \bigwedge y \preceq x) \Longrightarrow x = y$

Une taxonomie peut aussi être représentée par un graphe orienté acyclique qui généralement possède un unique concept abstrait "racine" (RDAG) (*Rooted Directed Acyclic Graph*).

Dans la section 3.2, nous proposons une ontologie d'application destinée à la grande distribution.

3.1.2 *Mesures sémantiques*

Les représentations des connaissances permettent d'exprimer une appréciation consensuelle de la connaissance pour un domaine. Les modèles ontologiques, même dans leur forme taxonomique la plus

simple, permettent d'apprécier des subtilités de domaine importantes dans les analyses qui les exploitent.

Cependant, ces modèles seuls ne sont pas suffisants. Pour être pleinement exploités dans des traitements reposant sur des techniques de raisonnements approchés, ils requièrent d'être associés à l'utilisation de mesures exploitant la similarité des concepts qu'ils décrivent. Par exemple, les couples (`Fraise`, `Banane`) et (`Fraise`, `Salade`), bien que tous les deux associés à la notion de `Fruits & Légumes`, ne sont pas composés de produits reliés sémantiquement avec le même degré d'intensité. Le couple (`Fraise`, `Banane`) fait en effet référence à deux produits intuitivement voisins parce qu'appartenant à une même catégorie (*e.g.* `Fruits`).

Les mesures sémantiques associées aux modèles de connaissances permettent de tirer parti de la structure de connaissances pour comparer des concepts. De nombreuses mesures de similarité sémantique (SSM) (*Semantic Similarity Measures*) exploitant les structures de connaissances ont été définies dans la littérature [58, 61]. Basées sur des fondements ensemblistes, sur la théorie des graphes, ou encore sur la théorie de l'information, ces mesures permettent d'apprécier la similarité de deux concepts (mesures pairwise) ou de deux groupes de concepts (mesures groupwise).

3.1.2.1 *Mesures Pairwise*

Les mesures pairwise permettent d'estimer la similarité entre deux concepts dans une taxonomie. Dans la littérature, trois grandes familles d'approches peuvent être distinguées. La première repose sur la théorie des graphes [60, 131, 154]. La similarité entre deux concepts est fonction de la longueur du chemin qui les relie. Ces approches se basent sur une représentation de la taxonomie sous forme de graphe acyclique dirigé (DAG).

La seconde propose un point de vue ensembliste de l'évaluation de la similarité. Ici, ce sont les ensembles de caractéristiques topologiques partagées et différenciantes qui permettent d'estimer la similarité [152].

Enfin la troisième, inspirée de la théorie de l'information, exploite la notion d'informativité d'un concept et les aspects topologiques de la taxonomie pour mesurer la similarité [75, 97, 134].

Des propositions hybrides ont aussi été étudiées [91]. La section ci-dessous décrit plus en détail ces différentes approches.

Approches basées sur la théorie des graphes

a) Parcours de chemin

Les mesures qui reposent sur la théorie des graphes considèrent le degré d'interconnexion entre les concepts. Généralement, la similarité dépend de la distance entre deux concepts dans la taxonomie. Sur la

base de l'analyse des longueurs des chemins qui relient les concepts, Rada propose une distance taxonomique entre deux concepts u et v pouvant être définie par la recherche du plus court chemin qui les relie $sp(u,v)$ [131] :

$$Dist_{Rada}(u,v) = sp(u,v)$$

Cette distance consiste à trouver un chemin d'un sommet à l'autre de façon à ce que la somme des poids des arcs de ce chemin soit minimale. Une adaptation non linéaire de la distance de Rada a été proposée pour définir la mesure de similarité [91]. Cette similarité (sim_{LC}) dépend d'une normalisation basée sur la profondeur totale de l'ontologie max_depth :

$$sim_{LC}(u,v) = -log(\frac{sp(u,v)}{2 \times max\_depth})$$

b) Ancêtres communs

La notion de plus petit ancêtre commun (LCA) (*Lowest Common Ancestor*) a aussi été étudiée pour définir la notion de similarité [154]. Le LCA de deux nœuds u et v dans un arbre ou un DAG T est le nœud le plus bas (*e.g.* le plus profond) qui a simultanément u et v comme descendants.

L'idée intuitive est que plus le LCA de deux concepts est profond dans la taxonomie plus les concepts (ou produits) comparés sont proches. Reprenons les couples de produits (`Fraise`, `Banane`) et (`Fraise`, `Salade`), le LCA du premier couple est le concept `Fruits` tandis que le LCA du second couple de produits est la catégorie de produits `Alimentaires`. Intuitivement, on souhaite exprimer que les produits (`Fraise`, `Banane`) sont plus proches car le concept de `Fruits` est plus profond dans la taxonomie proposée. Pour rendre compte de ces propriétés, la similarité $sim_{Wu\&Palmer}$ a été formalisée ainsi :

$$sim_{Wu\&Palmer}(u,v) = \frac{2 \times depth(LCA)}{depth(u) + depth(v)}$$

où $depth(x)$ représente la somme des poids des arcs entre le nœud racine et le concept. Ainsi, plus le LCA est profond dans la taxonomie, plus la similarité entre u et v est grande et $sim_{Wu\&Palmer} = 1$ quand $u = v$.

c) Hiérarchie ancestrale

Certaines mesures sémantiques se basent sur les méthodes ensemblistes exploitant les caractéristiques des concepts obtenus à partir

d'un graphe. La similarité entre deux concepts dépend alors du rapport ou de la différence entre les caractéristiques communes et discriminantes. Les concepts, ancêtres, descendants, frères ... sont définis comme les ensembles de caractéristiques comparées. L'héritage le plus couramment utilisé pour caractériser un concept est l'ensemble de ses ancêtres. Nous notons $A(u) = \{v | u \preceq v\}$, comme étant l'ensemble des ancêtres du concept u. La similarité entre deux concepts basée sur l'indice de Jaccard peut ainsi être exprimée [103] :

$$Sim_{CMatch}(u, v) = \frac{|A(u) \cap A(v)|}{|A(u) \cup A(v)|}$$

Pour jouer sur l'asymétrie de la mesure, un paramètre λ ($\lambda \in [0, 1]$) a été introduit pour finalement définir une nouvelle similarité [136] :

$$Sim_{RE}(u, v) = \frac{|A(u) \cap A(v)|}{\lambda \times |A(u) \backslash A(v)| + (1 - \lambda) \times |A(v) \backslash A(u)| + |A(u) \cap A(v)|}$$

La distance taxonomique comme le rapport entre des caractéristiques distinctes et communes a aussi été proposée [11] :

$$Dist_{Batet}(u, v) = Log_2 \left(1 + \frac{|A(u) \backslash A(v)| + |A(v) \backslash A(u)|}{|A(u) \backslash A(v)| + |A(v) \backslash A(u)| + |A(u) \cap A(v)|} \right)$$

Approches basées sur la théorie de l'information

Les approches qui exploitent le Contenu Informationnel (IC) (*Information Content*) reposent sur la théorie de l'information de Shannon [145]. Les mesures qui en sont issues se basent sur la comparaison des contenus informationnels des concepts, l'information partagée et l'information discriminante. Pour évaluer la pertinence d'un concept, il faut calculer son IC. Il est obtenu en calculant la fréquence du concept dans un corpus (ou une base d'observation des concepts). La notion d'IC peut être calculée à partir de la théorie de l'information de Shannon en fonction de la fréquence dans un corpus [134] :

$$IC(u) = -log(p(u))$$

où $p(u)$ représente la probabilité d'occurrence du concept u. Dans cette définition, plus un concept est fréquent, moins il est informatif. Par définition, l'IC augmente à chaque fois que l'on se rapproche des feuilles de l'ontologie (*e.g.*, $u \preceq v \implies IC(u) > IC(v)$). Autrement dit, l'$IC$ d'un concept est toujours supérieur à celui de ses ancêtres puisque les occurrences des hyponymes d'un concept sont ajoutées à

ses propres occurrences [139, 141, 143, 158]. La similarité entre deux concepts est alors définie comme étant fonction de l'ancêtre commun le plus informationnel (MICA) (*Most Informative Common Ancestor*) [135] :

$$sim_{Resnik}(u,v) = IC(MICA_{u,v})$$

L'un des inconvénients de cette proposition est qu'elle ne prend pas en compte la spécificité des concepts comparés. Pour illustrer cette limite, la similarité des concepts (`Fruits`, `Soda`) et (`Alimentaires`, `Soda`) est la même car leur $MICA$ respectif est le concept abstrait `Produits` (*cf.* Figure 1). La profondeur taxonomique des concepts comparés n'est pas explicitement exploitée. Pour remédier à cette limite, plusieurs propositions ont été faites pour intégrer explicitement la profondeur taxonomique. Parmi les plus connues, nous citons :

$$sim_{Lin}(u,v) = \frac{2 \times IC(MICA_{u,v})}{IC(u) + IC(v)} \text{ [97]}$$

$$Dist_{Jc}(u,v) = IC(u) + IC(v) - 2 \times IC(MICA_{u,v}) \text{ [75]}$$

$$Sim_{Nunivers}(u,v) = \frac{IC(MICA_{u,v})}{max(IC(u), IC(v))} \text{ [111]}$$

$$Sim_{PSec}(u,v) = 3 \times IC(MICA_{u,v}) - IC(u) - IC(v) \text{ [129]}$$

$$Sim_{Faith}(u,v) = \frac{IC(MICA_{u,v})}{IC(u) + IC(v) - IC(MICA_{u,v}}) \text{ [130]}$$

$$sim_{IC}(u,v) = \frac{2IC_{MICA}}{IC_u + IC_v} \times \left(1 - \frac{1}{1 + IC_{MICA}}\right) \text{ [94]}$$

Pour plus de détails sur chacune de ces similarités, nous invitons le lecteur à se référer aux références associées.

Approches hybrides

La combinaison des modèles précédents permet d'introduire des similarités qui considèrent plusieurs aspects taxonomiques. Une méthode qui combine le comptage des arcs (théorie des graphes) et la

méthode de l'IC a été proposée dans [91]. La mesure de la similarité au travers de l'information portée par les concepts parents a aussi été étudiée. Ces similarités associent approche ensembliste et approche basée sur la théorie de l'information :

$$Sim_{Mazando}(u,v) = \frac{2 \times \sum_{c \in A(u) \cap A(v)} IC(c)}{\sum_{c \in A(u)} IC(c) + \sum_{c \in A(v)} IC(c)} \text{ [110]}$$

$$Sim_{JacAnc}(u,v) = \frac{\sum_{c \in A(u) \cap A(v)} IC(c)}{\sum_{c \in A(u) \cup A(v)} IC(c)} \text{ [37]}$$

Enfin, [136] propose encore, par exemple, une approche hybride exploitant la profondeur des concepts et [127] présente plusieurs mesures basées sur l'agrégation de mesures existantes. Une approche hybride exploitant la profondeur des concepts est également proposée dans [136]. Enfin, plusieurs mesures basées sur l'agrégation de mesures existantes sont proposées [127].

3.1.2.2 *Mesures Groupwise*

Les mesures groupwise permettent d'estimer la similarité entre deux objets associés chacun à un ensemble de concepts. La littérature propose deux principales approches : directe et indirecte. L'approche directe correspond à une généralisation des approches définies pour la comparaison de deux concepts tandis que l'approche indirecte propose une agrégation des similarités définies sur le produit cartésien des deux ensembles (objets) à comparer. Nous décrivons ces deux approches dans la suite de cette section.

Mesures groupwise directes

Comme nous l'avons précédemment annoncé, les mesures groupwise directes généralisent les mesures pairwise. Par conséquent, nous retrouvons les trois familles d'approches sur lesquelles elles se basent : théorie des graphes, méthodes ensemblistes et théorie de l'information.

Nous définissons $G_T^+(X)$ comme étant le graphe induit par l'union des ancêtres des concepts qui composent l'ensemble X et $C_T^+(X)$ l'ensemble des concepts contenus dans $G_T^+(X)$.

La similarité de deux objets (ici, deux ensembles de concepts) U et V peut dépendre par exemple de la longueur du plus long des plus courts chemins liant un concept c ($c \in G_T^+(U) \cap G_T^+(V)$) à la racine [49] avec une approche basée sur la théorie des graphes. Une approche basée sur la théorie de l'information formalise la similarité entre deux objets U et V comme étant [128] :

$$Sim_{GIC}(U,V) = \frac{\sum_{c \in G_T^+(U) \cap G_T^+(V)} IC(c)}{\sum_{c \in G_T^+(U) \cup G_T^+(V)} IC(c)}$$

Enfin, plusieurs mesures ont été proposées en se basant sur les méthodes ensemblistes. Par exemple [50] :

$$Sim_{UI}(U,V) = \frac{|C_T^+(U) \cap C_T^+(V)|}{|C_T^+(U) \cup C_T^+(V)|}$$

Mesures groupwise indirectes

Les mesures groupwise indirectes agrègent une mesure pairwise de concepts définie sur le produit cartésien des deux ensembles comparés. Comme précédemment introduites, les mesures groupwise indirectes agrègent une mesure pairwise de concepts définies sur le produit cartésien des deux objets comparés. Nous pouvons, par exemple, utiliser la similarité moyenne (sim_{avg}) entre les ensembles de concepts pour définir la similarité finale entre deux objets U et V :

$$sim_{avg}(U,V) = \frac{\sum_{u \in U} \sum_{v \in V} sim(u,v)}{|U| \times |V|}$$

où $sim(\cdot,\cdot)$ est une similarité pairwise quelconque (*cf.* Section 3.1.2.1). D'autres mesures groupwise indirectes plus complexes ont été proposées comme la BMM (*Best Match Max*) ou encore la BMA (*Best Match Average*) :

$$sim_{BMM}(U,V) = \max\left(\frac{1}{|U|}\sum_{u \in U} \max_{v \in V} sim(u,v), \frac{1}{|V|}\sum_{v \in V} \max_{u \in U} sim(v,u)\right) \text{ [142]}$$

$$sim_{BMA}(U,V) = \frac{1}{2|V|}\sum_{v \in V} sim(v,U) + \frac{1}{2|U|}\sum_{u \in U} sim(u,V) \text{ [128]}$$

Pour résumer, les mesures de similarité sémantique sont calculées à partir, a minima, d'une structuration taxonomique des connaissances. Les mesures *pairwise* permettent de calculer la similarité entre deux concepts d'une même taxonomie. Certaines d'entre elles (mesures pairwise) exploitent le contenu informationnel (IC) associé à chacun des concepts, c'est-à-dire la quantité d'information d'un concept et plus un concept est spécifique, plus son contenu informationnel est

grand. Ce sont celles qui nous intéresseront le plus particulièrement. A partir de mesures pairwise, des mesures groupwise peuvent être introduites pour comparer des groupes de concepts associés à des objets [60]. Elle nous seront essentielles pour comparer, par exemple, deux clients ou deux magasins. Il existe plusieurs mesures groupwise, comme il existe plusieurs mesures pairwise et plusieurs définitions pour l'IC. Pour plus d'informations sur les mesures de similarité sémantique nous invitons le lecteur à se référer aux articles suivants [61, 128, 142].

3.2 REPRÉSENTATION DE CONNAISSANCES DANS LA GRANDE DISTRIBUTION

Dans la grande distribution, la représentation de connaissances repose principalement sur la structure marchandise appelée aussi taxonomie de produits (*cf.* Section 1). Cette taxonomie est utilisée dans de nombreuses analyses pour apporter des connaissances plus générales sur l'environnement du domaine. Nous la retrouvons par exemple dans l'analyse de panier moyen [52] ou encore l'étude de fréquence d'achats d'ensembles de produits : règles d'association (*Frequent itemset*) [149]. Même si nous pouvons trouver différentes approches pour définir les catégories servant à regrouper les produits, l'Unité de Vente Consommateur (UVC ou SKU) reste la notion centrale. Les chercheurs travaillent principalement sur la profondeur et la largeur de la structure qui sont, dans le cas de la grande distribution, directement liées aux problèmes de variété de produits et de profondeur de gammes (*cf.* Section 2.2.1).

Certains adoptent des solutions dans lesquels des niveaux intercatégories spécifiés par des experts du domaine et des spécialistes du marketing sont utilisés [159]. D'autres proposent d'ajouter des dimensions à la structure initiale [146]. Nous retrouvons cette structure marchandise au cœur des systèmes de recommandations [34, 68].

La modification de la structure marchandise est proposée pour prendre en considération la sensibilité des clients vis-à-vis des marques [68]. Des chercheurs présentent un framework collaboratif qui exploitent les similitudes à plusieurs niveaux, implicites dans les taxonomies de produits [92]. Les recherches sont aujourd'hui de plus en plus guidées par les acteurs du e-commerce et le Web Semantique [28].

Une partie des scientifiques se tourne vers la définition d'un environnement de système d'information d'entreprise ouvert et flexible [124]. L'émergence de l'informatique orientée services et les architectures web sémantiques sont mises en avant grâce notamment aux nombreux langages de description (OWL, WSDL ...) [89]. L'exploitation de mesure de similarité sémantique (SSM) est reconnue comme un facteur crucial [125].

Les ontologies sont un outil de modélisation utilisé dans de nombreuses applications dans le Web sémantique [137, 144]. D'autres chercheurs formalisent des ontologies considérant les nombreux facteurs environnant les chaînes de ventes et d'approvisionnement [95, 125, 156]. Beaucoup plus applicatifs, ils proposent des solutions de systèmes BI (*Business Intelligence*) permettant une gestion complète des chaînes d'approvisionnement et de vente pour la vente en ligne. Les propositions scientifiques se tournent principalement vers le e-commerce qui est en demande permanente d'innovations.

Contrairement à la grande distribution où quelques géants sont en concurrence frontale, les acteurs du e-commerce sont contraints d'innover pour rivaliser avec une concurrence toujours croissante. Dans ce domaine, beaucoup plus de données sont partagées car il ne s'agit pas d'une guerre des prix mais bien d'une course à l'innovation. La divergence du e-commerce et de la grande distribution a fractionné les approches scientifiques. Aujourd'hui, les problématiques scientifiques abordées pour la grande distribution concernent principalement la chaîne d'approvisionnement au détriment des problématiques annexes qui sont elles traitées spécifiquement pour le e-commerce. La section suivante propose la définition et la formalisation d'une ontologie dynamique pour la grande distribution.

Aujourd'hui, la nécessité d'une structuration de connaissances est indispensable, que ce soit pour améliorer les outils d'analyses comme nous l'avons vu précédemment et les connaissances ou tout simplement permettre une vision globale. L'interopérabilité joue un rôle essentiel dans le monde hétérogène qu'est la grande distribution. La capacité de prendre des décisions efficaces est d'une importance capitale pour la survie d'une organisation. Pour réussir, une entreprise moderne doit tirer parti de toutes les informations disponibles. Néanmoins, la définition d'une structure sémantique des connaissances reste du domaine du fantasque pour l'industrie. Bien que l'intérêt d'une telle structure soit largement démontré par les chercheurs [125] et accepté par les experts du domaine, les coûts associés à la mise en place semblent excessifs.

Pour permettre la création d'une ontologie dédiée à la grande distribution, nous avons travaillé conjointement avec les experts de TRF Retail. La limite majeure à la définition d'une unique ontologie pour la grande distribution vient de l'absence de standardisation. Comme il n'existe aucune nomenclature, chaque enseigne définit la sienne dont la vocation est de simplifier la chaîne d'approvisionnement (*Supply Chain*). Bien que les chercheurs aient proposé de nombreuses ontologies, aucune n'est suffisamment pragmatique pour être utilisée industriellement dans la grande distribution (hors e-commerce).

Grâce au modèle de données de TRF, nous pouvons définir des structures de connaissances, notamment des ontologies d'application [54] dédiées à chaque enseigne [55]. Les ontologies d'application

offrent le plus fin niveau de spécificité, dans notre cas les produits. Les relations entre les ontologies et les modèles de données ont été traitées par les chercheurs sous deux aspects :

— En ingénierie, pour transformer un modèle conceptuel de données en ontologie [73].

— En rétro-ingénierie, pour transformer une ontologie en modèle conceptuel de données [47, 153].

À partir du modèle conceptuel de données de TRF Retail, nous pouvons recréer les liens qui existent entre les différentes notions/éléments appartenant à l'éco-système de chaque grand distributeur (*e.g.* marques, fournisseurs ...).

Pour formaliser l'ontologie d'application, nous utiliserons les notations suivantes. Des notations et indices supplémentaires seront introduits lorsque nécessaire.

S L'ensemble des produits détenus par l'enseigne, $S = \{1, 2, \ldots, n\}$,

K_j L'ensemble des caractéristiques d'un article j, $K_j = \{c_j, r_j, \ldots\}$, avec c_j, r_j respectivement le prix d'achats et de vente du produit j

M L'ensemble des magasins d'une enseigne, $M = \{1, 2, \ldots, m_n\}$

Pour définir une ontologie d'application, trois hypothèses intuitives sont indispensables :

— un grand distributeur possède des produits ($S \neq \emptyset$)

— un grand distributeur a au moins un magasin ($M \neq \emptyset$)

— un grand distributeur dispose de caractéristiques génériques (*e.g.* Chiffre d'affaires) et spécifiques ($K_j \neq \emptyset$) pour chacun de ses produits (S)

De plus, nous partons de l'hypothèse fondamentale que chaque grand distributeur possède sa propre structure marchandise (taxonomie de produits) qui représente implicitement sa stratégie (variété de produits, profondeur de gammes ...).

Nous pouvons alors définir pour chaque enseigne une taxonomie de produits $T = (\preceq, X)$, où X représente l'ensemble des concepts (catégories de produits) et ($\preceq$) l'ordre partiel. $A(y)$ représente l'ensemble des ancêtres du concept y dans la taxonomie T tel que $A(y) = \{x \in X | y \preceq x, y \in X\}$. L'ensemble des concepts (catégories de produits) de la taxonomie T sont liés par des liens "is-a". Nous considérons la structure marchandise comme un RDAG (*Rooted Directed Acyclic Graph*) dont le nœud *root* correspond à un concept qui abstrait tous les produits. Ce nœud est l'unique concept qui n'a que lui-même comme ancêtre. Les concepts sans descendants (sauf eux-mêmes) sont dans notre cas les produits.

A partir de ces informations, nous sommes en mesure de définir une structure minimaliste des connaissances (*cf.* Figure 1) : nous

considérons que cette structure est naturellement accessible pour tous les distributeurs. Cette "base" d'ontologie d'application ne tient compte pour l'instant que des produits (le plus fin niveau), des catégories de produits et des liens "is-a".

Il est impératif de considérer en plus dans cette structure le réseau de magasins (M) qui est partie prenante dans la grande distribution. En considérant les magasins, un nouveau type de liens s'intègre à la structure, le lien "is-in" (*e.g.* un produit "is-in" (est dans) un magasin). L'intégration des magasins nous offre la possibilité d'ajouter la notion conceptuelle de clusters de magasins (*cf.* Section 2.1.1).

La Figure 5 propose un exemple illustratif d'ontologie d'application définie par les éléments physiques que sont les produits et les magasins enrichis par des éléments conceptuels (la Taxonomie de produits et deux types génériques de clusters de magasins, les Formats et les Régions). L'exemple d'ontologie d'application de la Figure 5 ne tient

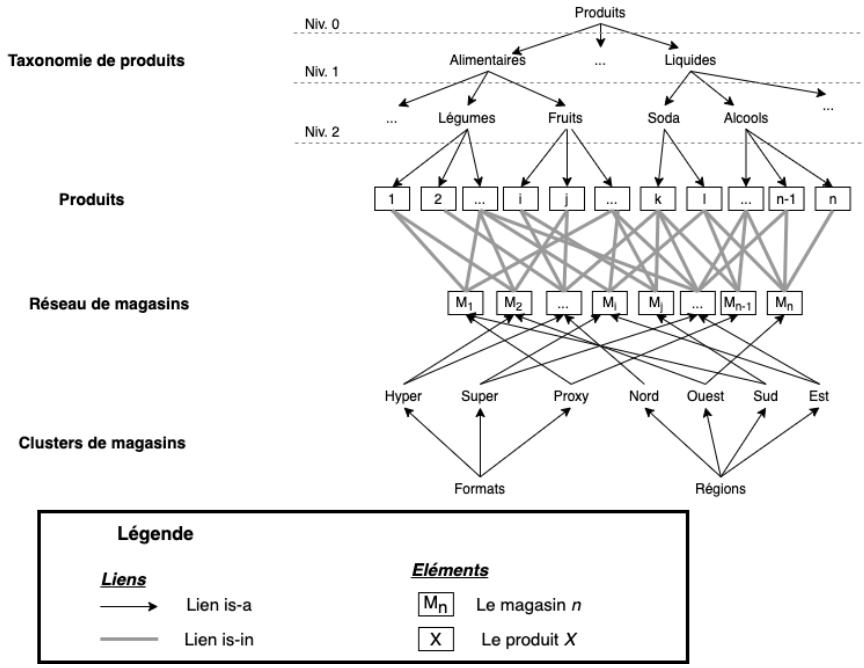


FIGURE 5 – Exemple d'ontologie d'application avec Taxonomie de produits et Magasins

compte que de liens "is-a" et "is-in". Le nombre de liens "is-a" est proportionnel au nombre de catégories de produits dans la structure marchandise ($|X|, T = (\preceq, X)$) et l'ensemble des produits de l'enseigne ($|S|$). Le nombre de liens "is-in" est quant à lui proportionnel au nombre de magasins ($|M|$) et au nombre de produits ($|S|$). Cependant, alors qu'un produit n'est associé ("is-a") qu'à un sous-ensemble de catégories X, il peut être ("is-in") dans la totalité des magasins (M). A titre d'indication, le tableau 3 reprend le nombre de catégories, de produits, de magasins et de liens existants pour les trois clients de TRF Retail introduits dans le chapitre précédent.

TABLE 3 – Statistiques effectives de l'ontologie d'applications *(**k**=milliers, **M**=millions)*

Distributeur	Catégorie	Produits	Magasins	"is-a"	"is-in"
Alimentaire	2 200	120 **k**	2 000	500 **k**	10 **M**
Bricolage	815	140 **k**	32	370 **k**	550 **k**
Coop. Agricole	6 290	600 **k**	250	1.5 **M**	2 **M**

Pour permettre un raisonnement à dimensions humaines sur les concepts obtenus à partir de l'ontologie d'application, nous retirons les produits pour transformer l'ontologie d'application en ontologie de domaine [54]. Une ontologie de domaine est limitée à la représentation de concepts dans des domaines donnés contrairement aux ontologies globales (*Top-Level Ontology*) qui représentent un plus haut niveau d'abstraction formelle. Les ontologies globales sont dédiées à des utilisations générales.

Les ontologies d'application et de domaine sont propres à un domaine spécifique et sont généralement construites de manière ad-hoc pour répondre à un besoin spécifique [59]. Alors qu'une ontologie d'application considère le plus fin niveau conceptuel (*e.g.* un produit), les ontologies de domaine considèrent uniquement les principaux éléments conceptuels. L'ontologie de domaine obtenue en retirant les produits de l'ontologie d'application est représentée sur la Figure 6. Nous remarquons sur la Figure 6 que les liens "is-in" entre les maga-

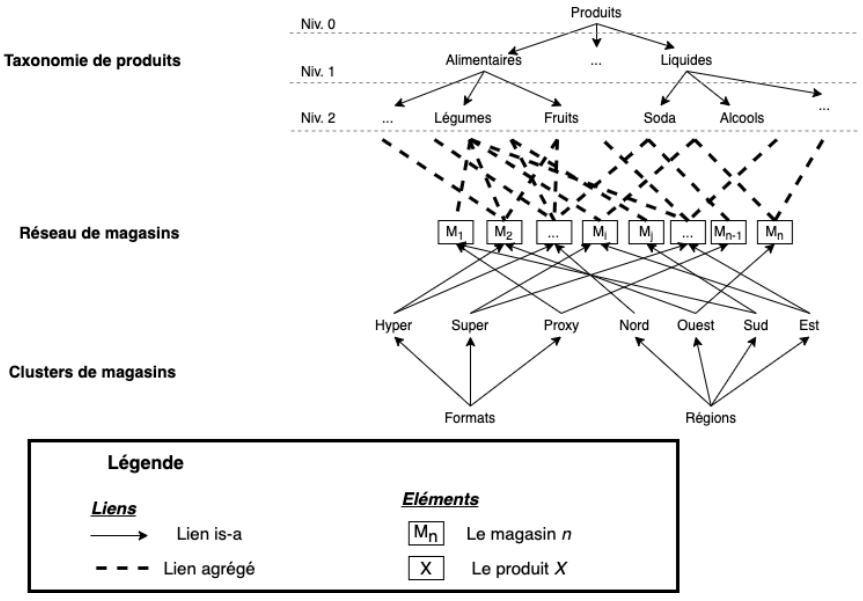


FIGURE 6 – Exemple d'ontologie de domaine avec Taxonomie de produits et Magasins

sins et les produits et les liens "is-a" entre les produits et leur catégorie ont été fusionnés pour créer des liens "agrégés". Autrement dit, les

caractéristiques de produits dans un magasin se sont agrégées sur les concepts de la taxonomie de produits. Ces liens représentant la performance des produits dans un magasin ("is-in") sont implicitement utilisés par les logiciels BI à des niveaux sémantique plus abstrait (*e.g.* catégorie de produits) grâce aux liens "is-a". Par exemple, le Chiffre d'affaires d'un magasin a pour une catégorie x est contenu dans le lien agrégé reliant M_a et la catégorie x. Nous retrouverons évidemment le nombre de produits associés ("is-a") à la catégorie x et présents dans le magasin ("is-in") M_a qui n'est autre que la cardinalité des liens ("is-in") des produits de la catégorie x.

Le manque de standardisation des caractéristiques des produits dans la grande distribution et la concurrence spécifique du domaine font qu'à notre connaissance aucune structuration conceptuelle des connaissances n'a été proposée. Des propositions standardisés ont néanmoins étaient faites pour le e-commerce [99, 109]. Cependant, elles ne semblent pas adaptées aux besoins des grands distributeurs. Pour cette raison, ils ce sont tournés soit vers des modèles de données (*e.g.* TRF Retail, ARTS[1] Retail Data Model, RLDM[2], ADRM[3] Retail Data Environment ...) soit vers des solutions BI (*e.g* [156]) apportant toujours une orientation opérationnelle indispensable. Les notions métiers sous-jacentes connexes sont considérées comme des éléments hétérogènes. Grâce au modèle de données de TRF Retail nous pouvons définir des structures de connaissances dédiées à la grande distribution qui mettent en lumière les connexions existantes entre tous ces éléments hétérogènes de la grande distribution. Ces structures sont, comme nous l'avons souligné, indispensables.

En partant des caractéristiques des produits depuis longtemps maîtrisées par les distributeurs, nous pouvons définir une ontologie d'application et une ontologie de domaine ; autrement dit, réalisables pour n'importe quel type de distributeur répondant aux **hypothèses** (*cf.* Section 3.2). Cette démarche pragmatique est le résultat d'une réflexion combinant expertise industrielle et expertise en représentation des connaissances. Les ontologies formalisées dans notre exemple considèrent uniquement deux notions de la grande distribution : les magasins (et leurs clusters) et la structure marchandise. Cette modeste structure offre pourtant la possibilité d'applications nouvelles, notamment des applications sémantiques précédemment introduites. De plus, elles sont indispensables pour l'expression de contraintes proposée dans le chapitre 5.

1. The Association for Retail Technology Standards
2. The Teradata Retail Logical Data Model
3. Applied Data Resource Management

3.3 CLUSTERING SÉMANTIQUE

Les méthodes permettant de regrouper les objets en fonction de leurs caractéristiques, aussi appelées clustering, sont devenues des outils d'aide à la décision indispensables. Elles permettent d'agréger l'information afin d'aider les preneurs de décision à obtenir une vision globale de leurs données.

Notre contribution proposée dans cette section repose sur l'utilisation de structures de connaissances et l'exploitation de mesures de similarité sémantique pour la segmentation de clients afin d'extraire les habitudes d'achats de consommateurs. Nous proposons de construire des clusters de clients sur la base de similarités sémantiques parce que nous pensons qu'elles rendent mieux compte des similitudes entre comportements d'achats. Une métrique classique sur l'espace des produits pourrait conduire à considérer qu'un client qui achète du Pepsi est tout aussi différent d'un client qui achète du Coca que d'un client qui achète un téléviseur ; alors qu'intuitivement les deux premiers achètent du soda et semblent plus proches l'un de l'autre qu'ils ne le sont du troisième. Deux clients auront très rarement exactement le même panier alors qu'ils fréquenteront souvent un même magasin car ils y trouvent les unités de besoin qu'ils partagent (*e.g.*, ils viennent chercher leurs courses alimentaires de la semaine, même s'ils ne mangent pas la même chose). Le recours aux similarités sémantiques et à la taxonomie de produits va permettre d'abstraire le panier des clients, et de les comparer non plus sur la base des produits qu'ils achètent, mais sur la base des unités de besoin qui leur sont nécessaires. Pour pallier les problèmes d'évaluations de la pertinence des clusters, nous proposons une analogie avec des données du domaine biomédical dont les connaissances sont déjà très structurées. Cette analogie se donne pour objectif de valider le recours aux similarités sémantiques lors d'une tâche de clustering lorsque les dimensions de l'espace sont organisées par un ordre partiel. En effet, pour valider l'approche sur des données consommateurs, il aurait fallu que de grands distributeurs acceptent d'évaluer la pertinence des segments de clients obtenus avec notre approche, ce qui n'était pas possible dans le seul cadre de cette thèse.

3.3.1 *Identification des habitudes d'achats des consommateurs*

Les clients ou consommateurs sont la cible finale de l'assortiment. Identifier leurs besoins et les fidéliser sont des analyses cruciales pour les grands distributeurs. En analysant leurs achats, nous souhaitons faire émerger des comportements collectifs qui permettront d'apporter des solutions toujours plus adaptées à la clientèle de chacun des magasins. L'utilisation des mesures de similarité plutôt que des métriques classiques permet de comparer des clients non plus selon les

produits qu'ils achètent, mais selon les unités de besoin, *i.e.* les classes de produits, qu'ils consomment et viennent chercher dans le magasin : elles permettent de s'abstraire du produit au besoin, d'esquisser un modèle plus riche du comportement du consommateur.

Aujourd'hui, les grands distributeurs reconnaissent l'importance de décisions basées sur la spécificité de leur clientèle. Cependant, ils se concentrent sur l'identification de "bons" clients en considérant principalement les "indicateurs comptables" tels que le chiffre d'affaires, la marge et la fréquence des clients [33, 52, 98, 149]. Néanmoins, ils ne prennent pas en compte la sémantique qui peut être associée aux produits achetés. Par exemple, un client qui achète du poisson et des courgettes est intuitivement plus proche d'un client qui achète de la viande et de la salade que d'un client qui achète des produits ménagers. Les deux premiers font leurs achats alimentaires dans le magasin analysé alors que le dernier lui n'y vient chercher que de la droguerie. La comparaison sur les unités de besoin permet de créer des classes de clients dont la sémantique traduit mieux les modes de consommation. Pour proposer des clusters relatifs à ce raisonnement, nous proposons d'analyser les données récoltées à l'aide de carte fidélité de clients afin d'extraire la sémantique sous-jacente de leurs achats.

3.3.1.1 *La segmentation sémantique de clients*

Dans le domaine de la grande distribution, le regroupement de clients, généralement appelé "segmentation de clients" dans la littérature, consiste à diviser des groupes de clients hétérogènes en fonction d'attributs communs [2, 65, 121].

Habituellement, les attributs suivants sont pris en compte : a) données démographiques (*e.g.* sexe, âge, situation matrimoniale) ; b) données psycho-graphiques (*e.g.* classe sociale, style de vie, caractéristiques personnelles) ; c) données géographiques (*e.g.* zone de résidence ou de travail) ; d) données d'attitude (*e.g.* données recueillies à partir d'enquêtes) ; etc.

De nombreuses approches de segmentation ont été proposées. Elles utilisent généralement les critères RFM - Revenue (chiffre d'affaires), Frequency (fréquence) et Monetary value (marge) - ou CLV - Customer Life Value (valeur de vie du client) - [32, 33, 79] pour ensuite appliquer des méthodes :

— de clustering (sans connaissance préalable sur les classes) [98]

— de classification (avec connaissance préalable sur les classes) (*e.g.* réseaux de neurones, arbres de décision) [32]

— de modèles basés sur des associations (*e.g.* règles d'association, chaîne de Markov) [32, 80]

TABLE 4 – Exemple de matrice de similarité

	Client 1	Client 2	...	Client X
Client 1	1	0.5	...	0.8
Client 2	0.5	1	...	0.6
...	...	...	1	...
Client X	0.8	0.6	...	1

— de prévisions [79] etc.

Bien que notre objectif dans cette analyse consiste également à segmenter les clients, nous souhaitons, exploiter les connaissances dont disposent les enseignes sur leur structure marchandise.

Ce raisonnement d'abstraction sur les comportements d'achats des clients est basé sur les notions de taxonomie (hiérarchie d'abstraction définie par un ordre partiel sur la structure), de contenu informationnel et de mesures de similarité sémantique précédemment introduites. Grâce à l'utilisation de la structure marchandise, les clusters résultants devraient être davantage interprétables par les grands distributeurs car associés à des notions qui leur sont propres.

La section suivante décrit le protocole utilisé sur un échantillon réel de clients.

3.3.1.2 *Expérimentation et analyse des résultats*

Pour présenter les résultats pouvant être obtenus avec un processus de clustering sémantique de consommateurs, nous avons mis en place le protocole suivant. Pour réaliser ce clustering, nous avons utilisé la configuration de SSM suivante : la BMA en mesure Groupwise [142] combinée à la mesure Pairwise définie par Resnik basée sur la notion de contenu informationnel [134] (*cf.* Section 3.1.2). Une matrice (symétrique) de *similarité* entre les différents clients est ainsi définie comme illustré dans le tableau 4. Plus la similarité entre les clients est élevée, plus leur comportement d'achats est similaire.

À partir de cette matrice, nous appliquons une classification ascendante hiérarchique (CAH) en nous basant sur la méthode de Ward qui cherche à minimiser la distance/l'inertie à l'intérieur des clusters (intra-clusters) en maximisant la distance entre les différents clusters (inter-clusters) [116]. Dans cette étude, nous avons choisi d'utiliser la répartition du chiffre d'affaires pondéré par catégorie pour chacun des clients. En effet, nous considérons que la proportion d'argent investit dans une typologie de produits représente la volonté d'un client à favoriser un type de produits plutôt qu'un autre.

TABLE 5 – Statistiques sur les données

	Nombre total	Moyenne par client	Moyenne par ticket de caisse
Nombre de clients	1 025	-	-
Nombre de tickets de caisse	3 692	5.75	-
Nombre de produits vendus	32 552	48.54	8.63
Nombre de catégories différentes	692	25.36	6.78

Les différentes étapes du clustering hiérarchique sont représentées sur la Figure 7. L'expérimentation a été réalisée sur un magasin pari-

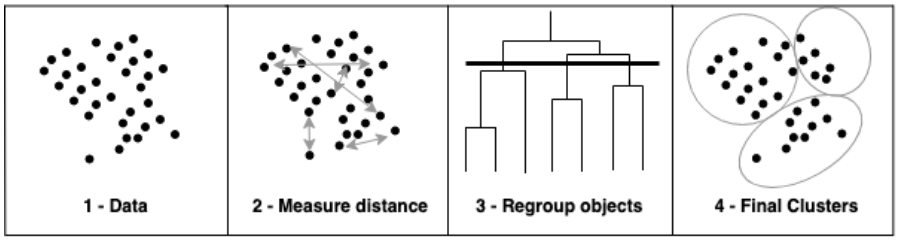


FIGURE 7 – Etapes du processus de classification ascendante hiérarchique (CAH)

sien pour lequel nous avions 32 500 produits vendus pour 1 025 clients (cartes de fidélité).

Le tableau 5 propose des statistiques descriptives du jeu de données utilisé. Nous y retrouvons différentes informations : une vision globale, une vision moyenne par client ainsi qu'une vision moyenne par ticket de caisse. Par exemple, on peut remarquer que les 1025 clients (cartes de fidélité) ont acheté en moyenne 25,36 catégories de produits différentes (*Niveau 1 sur la taxonomie des produits*).

Comme expliqué précédemment, pour rassembler les clients, nous avons utilisé la classification ascendante hiérarchique qui regroupe l'ensemble des clients deux par deux selon leur similarité. Celle-ci dépend de la répartition du chiffre d'affaires pondéré sur la structure marchandise. Autrement dit, deux clients qui dépensent des montants proportionnellement équivalents dans les même catégories de produits seront considérés comme similaires.

Le dendrogramme obtenu est présenté dans la Figure 8. Nous retrouvons en abscisse les 1 025 clients ayant servis à l'étude et en ordonnée la distance séparant les clients. Comme nous utilisons une

similarité normalisée (comprise entre 0 et 1), la distance en ordonnée, d entre deux clients a et b correspond à :

$$d(a, b) = 1 - sim(a, b)$$

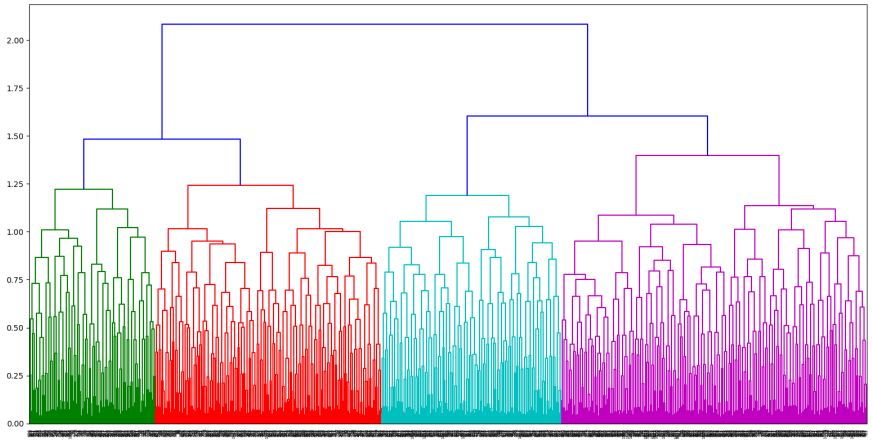


FIGURE 8 – Dendrogramme Clients

L'élément crucial de cette méthode de clustering réside dans l'identification du nombre idéal de clusters. Pour cette expérience, nous avons utilisé la méthode de Ward en plus d'une limite fixe. En effet, les clusters idéaux pourraient être un client par cluster, cependant, cela n'aiderait en rien les grands distributeurs qui tentent d'identifier un ensemble limité de clusters. Nous pouvons considérer qu'ils essaient de définir des "clients fictifs" qui représentent autant de clients que possible sans perdre d'information capitale. Nous avons fait varier le nombre de clusters de 2 à 25. Puis, nous avons analysé celui qui minimise l'inertie intra-classe et maximise l'inertie inter-classes. Dans cette expérience, nous avons obtenu un nombre optimal de **4 clusters**, nommés respectivement "Cluster 1" à "Cluster 4".

La Figure 9 représente la répartition du chiffre d'affaires des clients pour chacune des catégories de produits les plus abstraites dans la structure marchandise (*cf.* Figure 1). Nous pouvons noter que dans cette expérimentation, c'est le chiffre d'affaires qui a été utilisé car il agrège la quantité vendue et le prix de vente. De plus, il s'agit de l'indicateur le plus fréquemment utilisé dans le monde de la grande distribution. Ces informations peuvent être interprétées comme étant le panier moyen des clients de chaque cluster.

Tout d'abord, l'analyse des spécificités des clusters souligne que plus de la moitié des clients (591/1 025) sont regroupés dans le "Cluster 1". Ces individus achètent des produits de toutes catégories. Les

FIGURE 9 – Fréquence d'achats des clusters par catégorie de produits

trois autres clusters "Cluster 2", "Cluster 3" et "Cluster 4" ont respectivement 137, 148 et 149 clients et représentent le même ordre de grandeur, $\approx$ 15% des consommateurs. Cependant, chaque cluster représente une clientèle spécifique. Les clients du cluster "Cluster 2" achètent principalement des produits des catégories `LIQUID`, `SALES GROCERIES` et `NON-DAIRY FEES`. C'est le seul cluster de clients qui n'achètent que des produits provenant de trois catégories de produits.

Les clusters "Cluster 3" et "Cluster 4", quant à eux, regroupent des comportements d'achats opposés. En effet, les clients de "Cluster 3" achètent principalement des produits de la catégorie `DAIRY FEES` tandis que les clients de "Cluster 4" achètent essentiellement des produits de la catégorie `SWEET GROCERIES`. Nous pouvons remarquer que ces deux groupes de clients n'achèteront jamais de produits de la catégorie discriminante du cluster "opposé".

L'objectif de cette analyse est d'identifier des clients similaires en fonction de leur comportement d'achats. Le comportement d'achats d'un consommateur a été modélisé par sa projection sur la structure taxonomique de produits. Ainsi, le clustering sémantique est basé sur la structure marchandise associée aux mesures de similarité appropriées et apporte par conséquent des résultats plus compréhensibles pour les grands distributeurs en permettant par exemple d'associer des catégories de produits plus ou moins abstraites aux clusters de clients. Notons que nous avons utilisé la CAH pour regrouper les clients, mais d'autres méthodes de regroupement et/ou d'autres configurations de mesures de similarité sémantique (IC, Pairwise, Groupwise) peuvent être utilisées. Différentes configurations de SSM sont comparées dans la suite de ce chapitre sur un jeu de données dont les résultats attendu

sont connus. Cela nous permet d'évaluer la pertinence des mesures de similarités sémantique dans le cadre de clustering sémantique.

A partir des résultats de cette analyse, de nouvelles contraintes ou orientations peuvent améliorer les recommandations faites lors de scénarios d'optimisation de l'assortiment. Les actions faites par les catégories manager devraient logiquement porter sur les catégories de produits qui impactent un minimum les consommateurs. Dans les résultats présentés tout laisse supposer que les changements de l'assortiment devraient porter sur les catégories de produits SWEET GROCERIES et DAIRY FEES.

Cette analyse que nous avons illustrée sur un magasin peut être faite sur l'ensemble des clients d'une enseigne pour définir des comportement globaux pour identifier les spécificités de la clientèle à l'échelle du réseau de magasins. Cela permet d'introduire de nouvelles connaissances dans les ontologies ou de renseigner des contraintes associées à la clientèle dans le problème d'optimisation de l'assortiment.

Une comparaison des clusters obtenus dans différents magasins peut permettre d'introduire de nouveaux clusters de magasins dépendants des habitudes d'achats. La principale limite à laquelle nous sommes confrontés reste l'évaluation des clusters obtenus. Leur pertinence a simplement été illustrée par notre lecture et notre interprétation en termes de comportements de consommation. Une validation aurait nécessité le recours à une évaluation experte par des représentants de la grande distribution, ce qui dépasse le cadre de cette thèse. La principale limite à cette évaluation repose sur la subjectivité des évaluateurs, la divergence des points de vue et les attentes que les acteurs de la grande distribution peuvent avoir. Pour toutes ces raisons, nous avons choisi de valider le recours aux similarités sémantiques pour le clustering d'items dans un espace où les dimensions sont organisées par une structure taxonomique dans un domaine où les ontologies et taxonomies sont des structures de connaissances largement adoptées et où la pertinence des clusters pourra être évaluée quantitativement.

3.3.2 *Validation du clustering sémantique*

Pour évaluer et démontrer l'intérêt des approches sémantiques, il est nécessaire d'avoir un jeu de données dont les résultats ont déjà été approuvés dans le domaine. Cependant, à notre connaissance, la grande distribution ne propose pas de benchmarks dans un contexte de concurrence et l'expertise permettant d'évaluer l'ensemble des clusters et des individus à la main aurait constitué un travail trop important de la part des grands distributeurs. Pour pallier ce problème, nous avons reproduit un protocole sur un jeu de données biomédicales en se basant sur l'analogie suivante : un symptôme est à la maladie

ce qu'un produit est au client. Nous nous sommes tournés vers le domaine biomédical puisqu'il met à disposition de nombreux jeux de données, que les structures de connaissances de type ontologies et taxonomies y sont communément admises, et que les résultats attendus des analyses peuvent être vérifiés objectivement. Nous avons utilisé les données proposées dans le travail de [157] qui offre une liste de maladies (objets/clients) caractérisées par des symptômes (caractéristiques/concepts/produits) désambiguïsés par le MeSH, comme illustré dans la Figure 10.

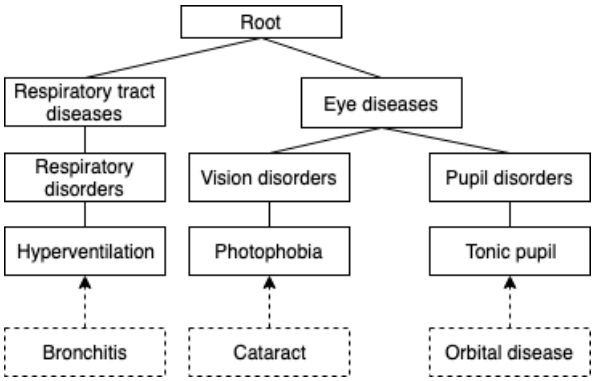


FIGURE 10 – Exemple de relation "is-a" dans le MeSH

Dans cette expérimentation, nous voulons regrouper les maladies en fonction de leurs symptômes associés, *e.g. Cataracte*, *Maladie orbitale* et *Bronchite*, respectivement caractérisées par les symptômes *Photophobie*, *Pupille tonique* et *Hyperventilation*. Plus précisément, les maladies sont déjà classées dans le MeSH par famille, *e.g.*, Maladies visuelles, Maladies Musculaires, etc. Nous voulons regrouper les maladies par rapport à ces classes identifiées au sein du MeSH en se basant sur les symptômes : les classes de maladies sont répertoriées dans le MeSH, nous allons donc évaluer dans quelle mesure nous pouvons retrouver ces classes en faisant du clustering sémantique sur les symptômes. L'évaluation de la méthode repose sur le fait de retrouver (ou non) les ensembles de maladies déjà explicités par les experts du domaine. Pour illustrer l'idée qu'il y a derrière l'utilisation de mesures de similarité sémantique dans le processus de clustering, prenons un exemple de maladies et de leurs symptômes associés, *e.g.* les maladies *Cataract*, *Orbital disease* et *Bronchitis*, respectivement caractérisées par les symptômes *Photophobia*, *Tonic pupil* et *Hyperventilation*. Classiquement, ces trois symptômes seront considérés comme indépendants et constitueront des dimensions distinctes de l'espace métrique de clustering. Pourtant *Photophobia* et *Tonic pupil* sont clairement liées par leur concept père *Eye Symptoms*. Ainsi, intuitivement, on imagine que *Photophobia* et *Tonic pupil* ayant des classes de symptômes en commun seront rapprochées dans le processus de clustering et que l'on retrouvera la Classe Maladies des yeux via le processus de clustering sur les symptômes alors que *Bronchitis* sera placée dans un

cluster où l'on retrouvera les maladies respiratoires (qui ne partagent pas de symptômes avec les maladies des yeux si ce n'est des classes très abstraites). Cet exemple illustre l'importance d'utiliser les liens sémantiques entre symptômes pour obtenir les résultats attendus.

Dans le jeu de données proposé par [157], chaque objet $d \in \mathcal{D}$, où $\mathcal{D}$ est l'ensemble des maladies, est représenté par un vecteur dont les coordonnées correspondent aux symptômes. La valeur $w_{d,i}$ liée à ces coordonnées symbolise la force d'association entre le concept (symptôme) i et l'objet (maladie) d. Plus un symptôme est observé lors du diagnostic d'une maladie, plus ce symptôme est révélateur de la maladie (de la même façon qu'un produit acheté par un client en nombre atteste de l'intérêt pour ce produit). [157] considère la force d'association $w_{d,i}$ basée sur la valeur TF-IDF [76] calculée à partir d'une analyse de corpus de textes.

$$w_{d,i} = W_{d,i} log \frac{N}{n_i}$$

où, N indique le nombre total de maladies dans le jeu de données, n_i est le nombre de maladies où le symptôme i apparaît et $W_{d,i}$ est le nombre de co-occurrences de la maladie d et du symptôme i. À partir de ces vecteurs, les auteurs calculent la similarité cosinus entre tous les objets, les maladies. Cet ensemble de données est utilisé dans notre expérimentation pour illustrer et démontrer l'intérêt des approches sémantiques. Les objets, les maladies, sont déjà regroupés en 26 "groupes ou classes de maladies" distincts au sein du MeSH. L'idée est de retrouver a posteriori ces classes de maladies du MeSH (ce que nous n'avions pas dans le domaine de la grande distribution pour proposer une validation de nos choix) le plus fidèlement possible par un clustering sémantique sur les symptômes. Nous constatons dans la Figure 11 que les clusters "C21", "C22" et "C24" (respectivement "*Désordres d'origine environnementale*", "*Maladies des animaux*" et "*Maladies professionnelles*") sont sous-représentés par rapport aux autres groupes de maladies. Nous proposons dans cette étude de regrou-

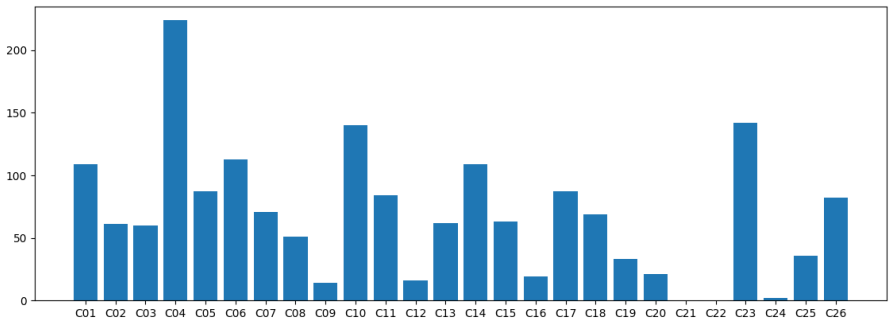


FIGURE 11 – Nombre de maladies par cluster

per les maladies représentées par leurs symptômes pour vérifier a posteriori, à l'aide de mesures appropriées, que les maladies d'un même groupe appartiennent au même cluster. En d'autres termes,

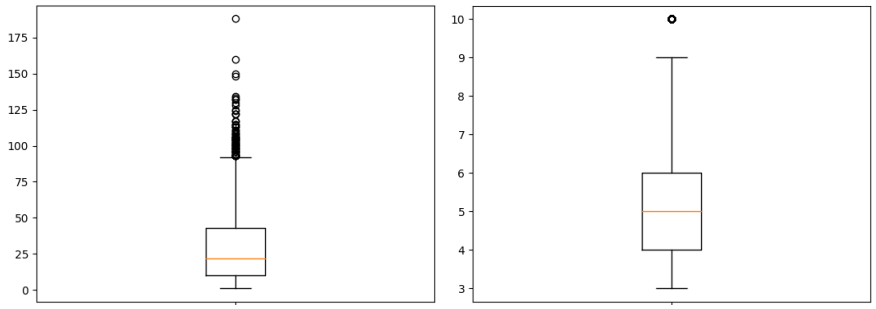
l'idée est de mettre en œuvre différentes configurations de mesures de similarité sémantique en faisant varier les définitions des IC, les mesures pairwise et les mesures groupwise introduites précédemment. A l'aide de ces mesures de similarité sémantique, nous avons construit la matrice de similarité pouvant être exploitée dans un processus de classification ascendante hiérarchique (CAH). Dans cette matrice, nous retrouvons la similarité obtenue entre chacune des maladies, deux à deux, permettant de définir un distance entre elles afin d'apporter par la suite des groupes de maladies "proche". Cette matrice est identique à celle utilisée dans l'expérimentation sur les consommateurs (*cf.* Tableau 4). Nous avons finalement réalisé le processus de clustering pour évaluer la façon dont les clusters résultants pouvaient correspondre aux classes de maladies initiales.

Les clusters résultants ont également été comparés à des méthodes plus classiques que nous pouvons retrouver dans la grande distribution [33, 52, 98, 149]. Pour proposer une évaluation complète, il est indispensable de se comparer aux méthodes déjà exploitées. Pour cela, nous avons utilisé en plus le K-means et d'autres CAH pour lesquels la matrice de distance a été construite avec la distance euclidienne ou la similarité du cosinus. Ces méthodes de clustering ont été choisies car ce sont celles les plus souvent utilisées dans le monde de la grande distribution [52, 98]. Ces "base-lines" (CAH avec distances "usuelles" et algorithme du K-means) sont toutes deux calculées avec des vecteurs TF-IDF ou des vecteurs binaires (présence ou absence de symptômes sans tenir compte des fréquences). Notons que toutes nos classifications utilisent la méthode de Ward qui consiste à minimiser la distance intra-cluster tout en maximisant la distance inter-cluster.

Pour évaluer nos résultats, la F_1-mesure est utilisée pour calculer la précision du partitionnement. La F_1-mesure correspond à la moyenne harmonique de la précision et du rappel [108]. Définissons la précision et le rappel dans notre problème de correspondance entre les clusters et les classes de maladies du MeSH. L'évaluation est calculée par comparaison de toutes les paires de maladies. Si deux maladies d'un même cluster appartiennent à la même classe du MeSH (*e.g.* "*Infections bactériennes et mycoses* [C01]"), elles seront considérées comme "Vrai Positif" (TP). Sinon, si elles ne font pas partie du même cluster, alors qu'elles sont de la même classe du MeSH, elles seront considérées comme "Faux négatif" (FN). Deux maladies de classes différentes (*e.g.* "*Infections bactériennes et mycoses* [C01]" et "*Maladies des yeux* [C11]") seront considérées comme "Vrai Négatif" (TN) si elles appartiennent à des clusters différents et "Faux Positif" (FP) autrement. Nous définissons alors la F_1-mesure comme étant :

$$F_1 = 2 \times \frac{Precision \times Rappel}{Precision + Rappel}$$
$$Precision = \frac{TP}{TP + FP}$$
$$Rappel = \frac{TP}{TP + FN}$$

Pour chaque maladie, seules les caractéristiques (symptômes) les plus spécifiques sont conservées. Par exemple, une maladie caractérisée par les symptômes *Douleur* et *Douleur d'estomac* est réduite à *Douleur d'estomac* pour éviter la redondance d'informations. L'ensemble de données final contient 1738 maladies et 318 symptômes différents. Comme le montre la Figure 12a, les maladies sont globalement caractérisées par 5 à 100 symptômes. La Figure 12b présente la profondeur des symptômes dans le MeSH.



(a) Nombre de symptômes par maladie (b) Profondeur des symptômes

FIGURE 12 – Statistiques sur les données

Les mesures de similarité sémantique ont été calculées au moyen de la *Semantic Measures Library* [4], développée au laboratoire par S. Harispe. Le tableau 6 présente les résultats de la mesure F_1 obtenus sur le jeu de données précédemment détaillé exploité avec les méthodes K-means, la CAH vectorielle (*i.e.*, avec une métrique classique) et la CAH sémantique (*i.e.*, avec les mesures de similarités sémantique). Les vecteurs de symptômes et la matrice de proximité ont été d'une part exploités tels quels (colonne F_1), et d'autre part normalisés (colonne F_1*) pour observer l'impact de la normalisation sur le processus de regroupement. Pour rappel, la mesure de Wu & Palmer *pairwise* ne nécessite pas le calcul d'IC. La configuration des différentes mesures de similarité sémantique évaluées est décrite de la manière suivante : mesures *groupwise*, mesures *pairwise*, contenu de l'information (IC).

Comme nous l'avons expliqué précédemment, la F_1-mesure permet de calculer la performance du processus de regroupement à recouvrer les classes de maladies du MeSH. Les résultats obtenus par cette

4. https ://www.semantic-measures-library.org/sml/

TABLE 6 – Résultats

	F_1-mesure	F_1-mesure*
K-means		
TFIDF	0,113	0,159
Binary	0,117	0,183
Mesures Vectorielles		
TFIDF, Cosine	0,164	0,155
TFIDF, Euclidean	0,110	0,099
Binary, Cosine	0,097	0,120
Binary, Euclidean	0,068	0,072
Mesures de similarité sémantique		
BMA, Lin, IC Resnik Ext	0,107	0,172
BMA, Lin, IC Resnik Ext Norm	0,108	0,170
BMA, Sim IC, IC Resnik Ext	0,109	0,181
BMA, Sim IC, IC Resnik Ext Norm	0,116	**0,196**
BMA, Resnik, IC Resnik Ext	0,101	0,183
BMA, Resnik, IC Resnik Ext Norm	0,114	0,174
BMA, Wu & Palmer	0,094	0,116
BMM, Lin, IC Resnik Ext	0,102	0,124
BMM, Lin, IC Resnik Ext N.	0,100	0,139
BMM, Sim IC, IC Resnik Ext	0,102	0,137
BMM, Sim IC, IC Resnik Ext Norm	0,115	0,155
BMM, Resnik, IC Resnik Ext	0,105	0,126
BMM, Resnik, IC Resnik Ext Norm	0,110	0,153
BMM, Wu & Palmer	0,086	0,101
Min, Lin, IC Resnik Ext	0,094	0,092
Min, Lin, IC Resnik Ext Norm	0,095	0,098
Min, Sim IC, IC Resnik Ext	0,103	0,094
Min, Sim IC, IC Resnik Ext Norm	0,092	0,097
Min, Resnik, IC Resnik Ext	0,118	0,118
Min, Resnik, IC Resnik Ext Norm	0.118	0,118
Min, Wu & Palmer	0,098	0,090
Max, Lin, IC Resnik Ext	0,093	0,074
Max, Lin, IC Resnik Ext Norm	0,093	0,079
Max, Sim IC, IC Resnik Ext	0,095	0,109
Max, Sim IC, IC Resnik Ext Norm	0,112	0,183
Max, Resnik, IC Resnik Ext	0.112	0,182
Max, Resnik, IC Resnik Ext Norm	0,113	0,162
Max, Wu & Palmer	0,099	0,074
Average, Lin, IC Resnik Ext	0,084	0,136
Average, Lin, IC Resnik Ext Norm	0,086	0,141
Average, Sim IC, IC Resnik Ext	0,094	0,141
Average, Sim IC, IC Resnik Ext Norm	0,113	0,136
Average, Resnik, IC Resnik Ext	0.121	0,135
Average, Resnik, IC Resnik Ext Norm	0,137	0,159
Average, Wu & Palmer	0,100	0,110

expérimentation permettent deux constats. La premier repose sur l'impact de la normalisation dans le processus de clustering. En effet, on constate que dans la plupart des cas, la F_1-mesure est meilleure lorsque les vecteurs de symptômes et la matrice de proximité sont normalisés. Ensuite, d'après les résultats du tableau 6, la meilleure configuration de SSM est : la BMA comme mesure *groupwise*, la sim_{IC} comme mesure *pairwise* et le contenu d'information normalisé défini par Resnik. Dans l'ensemble, les autres combinaisons de SSM ont donné de meilleurs résultats que les mesures de base (K-means & CAH Vectorielle) mais la configuration précédente reste significativement plus efficace. Le tableau 7 présente les résultats moyens de la F_1-mesure pour chaque paramètre des configurations. Si la "meilleure" configuration obtenue sur l'ensemble des données était basée sur la mesure *pairwise* sim_{IC}, on remarque que les configurations basées sur la mesure *pairwise* de Resnik obtiennent la meilleure F_1-mesure en moyenne. Ces résultats soulignent l'intérêt des SSM sur les métriques classiques et l'importance du choix de la combinaison de SSM. Contrairement

TABLE 7 – Détails des configurations sémantiques

	F_1-mesure	F_1-mesure*
Contenu Informationnel		
IC Resnik Ext	0,101	0,134
IC Resnik Ext Norm	0,107	**0,144**
Mesure Pairwise		
Lin	0,096	0,123
Resnik	0,114	**0,151**
Sim IC	0,105	0,143
Wu & Palmer	0,095	0,098
Mesure Groupwise		
Average	0,102	0,137
BMA	0,107	**0,170**
BMM	0,103	0,134
Max	0,101	0,123
Min	0,100	0,101
Total	0,137	0,133

à la "meilleure" combinaison obtenue sur l'ensemble des données, on peut voir que, la mesure *pairwise* de Resnik obtient la meilleure $F1$-mesure en moyenne. Les mesures de similarité sémantique ont la particularité de tirer parti des liens existants entre les concepts organisés dans une taxonomie. Les SSM permettent donc de prendre en compte les relations de dépendance "is-a" entre les dimensions de

l'espace. Ainsi, l'intérêt d'utiliser ces connaissances "a priori" dans la métrique de clustering lorsque les dimensions de l'espace sont organisées par un ordre partiel taxonomique a été validé sur cet exemple du domaine biomédical. De plus, lorsqu'une taxonomie est disponible, ces mesures évitent la comparaison de matrices éparses grâce à des relations hiérarchiques "is-a" entre les caractéristiques qui fournissent de nouvelles informations. Afin de permettre la reproductibilité du protocole expérimental, les données et le code utilisé sont disponibles sur Github [5].

L'expérimentation menée sur un jeu de données biomédicales publiques permet de démontrer l'intérêt des approches sémantiques dans le cas où les dimensions du processus de clustering sont organisées au sein d'une taxonomie d'abstraction. Ainsi, les SSM sont des outils d'intérêt pour les grands distributeurs qui disposent d'une structure de marchandise assimilable à une taxonomie car elles vont permettre une comparaison plus fine de deux clients ou deux magasins et par suite fournir des outils de clustering ou de segmentation plus adaptés à la grande distribution que les métriques classiques. Revenons donc en détail sur cette structure marchandise dans la grande distribution.

3.4 CONCLUSION

Dans ce chapitre, nous avons décrit les différentes structures de connaissances existantes dans la grande distribution. De plus, nous avons démontré l'intérêt d'exploiter ces structures notamment à l'aide des mesures de similarités sémantique. Ces mesures nous ont permis d'obtenir plus d'informations concernant les habitudes d'achats des clients.

Grâce aux structures de connaissances présentées dans ce chapitre la formalisation du problème de l'assortiment en un problème d'optimisation combinatoire est maintenant possible. De plus, ces structures sont utilisées dans le chapitre 5 pour aider les enseignes dans l'expression de leur stratégie aux travers de contraintes. Enfin, ces structures de connaissances et les mesures de similarité sémantique vont être utilisées pour adresser le problème de similarité des magasins détaillé dans le chapitre suivant.

5. https ://github.com/PAJEAN/diseases_segmentation

4

FORMALISATION DU PROBLÈME D'OPTIMISATION DE L'ASSORTIMENT

Pour aider les grands distributeurs, la nécessité de travailler directement sur les profils assortiments de chacune des catégories des magasins semble évidente. En effet, ces derniers sont confrontés à la gestion de réseaux de magasins de grande envergure et utilisent un assortiment commun (l'assortiment gigogne) [20]. Dans ce chapitre, nous nous intéressons à l'identification du meilleur assortiment pour un magasin sur la base des profils assortiments sélectionnés pour leurs performances dans des magasins similaires du réseau. La problématique que nous adressons repose sur la notion de similarité entre deux magasins. C'est une notion qui demande à être formalisée : la similarité peut être associée à des considérations sur la localisation des magasins, les produits proposés en rayon, le chiffre d'affaires, le format (*e.g.*, hypermarché, supermarché) ... [84]. Cette notion de similarité permet d'accéder aux concepts de groupement ou cluster de magasins dépendant de certain(s) critère(s) utile(s) pour identifier les produits qui ont les meilleures performances chez les voisins. La notion de similarité entre magasins joue donc un rôle central dans cette contribution.

De plus, comme nous l'avons souligné dans la section 3.1.1, les ontologies associées aux mesures de similarité sémantique vont nous permettre de raisonner de façon intuitive sur la notion de magasins aux fonctionnements semblables (*cf.* Section 3.1.1). Nous proposons donc d'étudier les magasins à l'aide de ces mesures qui, dans notre cas, exploitent des connaissances spécifiques à chaque enseigne, notamment la structure marchandise.

Ce chapitre est organisé de la façon suivante. Dans la section 4.1, nous formalisons un modèle de calcul du meilleur assortiment sous la forme d'un problème d'optimisation combinatoire pouvant s'intégrer dans la méthode Agile proposée dans le chapitre 5. La section 4.2 exploite la taxonomie de produits et les mesures sémantiques afférentes pour proposer une analyse sémantique de la similarité entre magasins qui, d'une part, permet d'améliorer la fiabilité des estimations de gain associées à un changement d'assortiment, informations nécessaires à la résolution du problème d'optimisation combinatoire énoncé pour le calcul du meilleur assortiment ; et d'autre part, accroît les connaissances des distributeurs sur leurs magasins pour une nouvelle vision des clusters dans leur réseau. La section 4.2.2 propose un exemple illustratif et met en avant les avantages de cette nouvelle contribution.

4.1 OPTIMISATION DES PROFILS ASSORTIMENT DES MAGASINS

Dans ce manuscrit nous avons insisté sur l'impact des points de vente dans la grande distribution dès la présentation de la problématique de l'assortiment optimal. C'est pourquoi, il est indispensable de proposer une solution orientée magasin et profil assortiment cette fois ci pour compléter notre étude.

Pour rappel, les profils assortiment sont définis :

— Pour chacun des produits.

— Pour chacun des magasins et chacune de leurs catégories (le niveau des catégories est défini par l'enseigne).

Nous nous focalisons sur la seconde définition pour proposer aux grands distributeurs une solution industrialisable qui puisse prendre en considération tous les éléments liés à la problématique de l'assortiment optimal.

Cette section se décompose en deux parties. La première formalise la problématique générale en un problème d'optimisation combinatoire. La seconde présente une méthode permettant d'améliorer l'estimation des gains liés à un changement d'assortiment en se basant sur les performances de magasins similaires : ces estimations de gain constituent des informations nécessaires à la résolution du problème d'optimisation.

4.1.1 *Problème d'optimisation*

Les notations suivantes seront utilisées tout au long de cette section :

— Ω un magasin d'une enseigne

— F_i la i^{me} catégorie de produits définie par l'enseigne (*e.g.* `Soda`) parmi celles permettant de définir les profils assortiment (*cf.* Section 2.1.3)

— $S^{k_i}(F_i)$ l'ensemble de produits de la catégorie F_i correspondant au profil assortiment $\leq k_i$ (par construction ils appartiennent à tous les profils assortiment inférieurs à k_i)

Des notations et indices supplémentaires seront introduits lorsque nécessaire.

Formellement, pour chaque catégorie, une hiérarchie de sous-ensembles de produits $S^{k_i}(F_i) = 1..\mathcal{K}$ au sens de la relation d'inclusion (*i.e.* $S^{k_i}(F_i) \subset S^{k_i+1}(F_i)$) est définie et un magasin peut choisir son assortiment parmi ces sous-ensembles $S^{k_i}(F_i)$. Cette suite d'ensembles permet de définir un assortiment gigogne monodimensionnel pour une catégorie de produits.

Par exemple, le premier niveau de profil assortiment pour la catégorie `Soda` pourrait être `Coca-Cola 1.5L`; le deuxième `Coca-Cola 1.5L + Lipton 2L + Orangina 1.5L`; le troisième `Coca-Cola 1.5L + Lipton 2L + Orangina 1.5L + Schweppes 1.5L` et ainsi de suite.

Donc, les produits détenus par un magasin ne peuvent être que l'un des sous-ensembles de produits de cet ensemble fini de profils assortiment $S^{k_i}(F_i) = 1..\mathcal{K}$ défini a priori par l'enseigne. Le nombre de produits est minimal lorsque $k_i = 1$ et maximal lorsque $k_i = \mathcal{K}$.

En pratique, l'assortiment gigogne peut être bidimensionnel. La Figure 13 présente le véritable assortiment gigogne d'un grand distributeur. Des indicateurs alphanumériques quelconques sont utilisés pour définir les différents sous-ensembles d'articles et, la relation d'inclusion est la même que pour l'assortiment gigogne monodimensionnel présenté jusqu'à présent. L'inclusion fonctionne de droite à gauche (*e.g.* $V \subseteq 1 \subseteq A$) et de bas en haut (*e.g.* $V \subseteq X \subseteq W$). Ainsi, le plus petit profil assortiment est V, et le plus grand T. Un magasin associé au profil assortiment 2 (Figure 13) devra inclure les produits associés aux profils assortiment $V, X, 1$, et 2. Un autre magasin associé au profil assortiment A devra, lui, inclure les produits associés aux profils assortiment $V, 1$ et A.

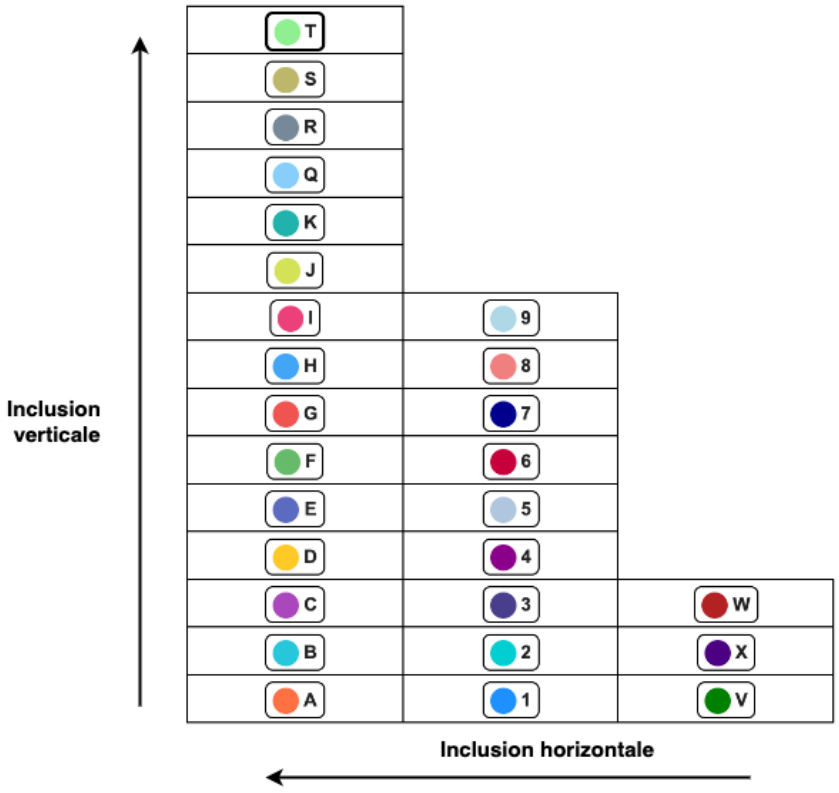


FIGURE 13 – Assortiment gigogne bidimensionnel

Concrètement, nous pouvons considérer que la structure marchandise de l'enseigne qui exploite cet assortiment gigogne s'étend alors comme le présente la Figure 14 où, chaque indicateur alphanumérique représente un sous-ensemble de produits propre à chaque catégorie. Cette vision bidimensionnelle de l'assortiment gigogne apporte plus de flexibilité quant aux produits à inclure dans les magasins.

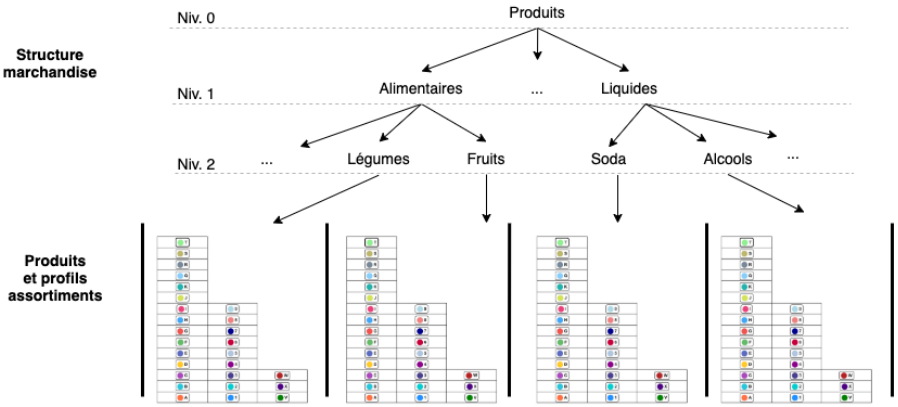


FIGURE 14 – Exemple illustratif d'une structure marchandise étendue avec les profils assortiment bidimensionnel

L'inclusion des sous-ensembles de produits pour un assortiment gigogne bidimensionnel est la même que pour un assortiment monodimensionnel. Cependant, il faut considérer les deux dimensions donc $S^{k_i}(F_i)$ devient $S^{k_i,k'_i}(F_i)$, où, k_i représente la vision horizontale et k'_i la vision verticale. Dans le cas bidimensionnel l'ordre n'est pas total, la résolution en est par conséquent modifiée.

Par souci de simplification, pour exprimer la problématique d'optimisation, nous nous focalisons uniquement sur l'aspect monodimensionnel. La généralisation à l'aspect bidimensionnel n'affecte pas la formulation du problème d'optimisation car il suffit simplement de prendre en considération la relation entre k_i et k'_i. C'est à dire, par exemple, dans la Figure 13, $k_i = 1..3$, respectivement $V, 1, A$ et :

— si $k_i = 1$ (V) alors $k'_i = 1..3$ (V, X, W)

— si $k_i = 2$ (1) alors $k'_i = 1..9$ $(1, 2, 3, 4, 5, 6, 7, 8, 9)$

— si $k_i = 3$ (A) alors $k'_i = 1..15$ $(A, B, C, D, E, F, G, H, I, J, K, Q, R, S, T)$

Ainsi, le petit profil assortiment est obtenu lorsque $k_i = 1$ et $k'_i = 1$ (V), et le plus grand lorsque $k_i = 3$ et $k'_i = 15$ (T).

Pour définir l'assortiment d'un magasin, sous l'hypothèse d'assortiment gigogne monodimensionnel, nous pouvons écrire :

$$\Omega \triangleq \bigcup_{i=1}^{n} S^{k_i}(F_i)$$

où, n correspond à l'ensemble des catégories de l'enseigne.
Autrement dit, l'assortiment d'un magasin correspond à l'union des profils assortiment associés aux différentes catégories F_i. Nous notons $P(S^{k_i}(F_i))$ l'utilité du sous ensemble de produits inclus dans $S^{k_i}(F_i)$. Cette utilité peut être, par exemple, le chiffre d'affaires, ou l'utilité définie dans la section 5.3.3. De la même façon, nous notons $C(S^{k_i}(F_i))$ les contraintes ou coûts associés aux sous-ensembles de produits $S^{k_i}(F_i)$: ce peut être un prix d'achat ou un volume de stockage. Dans le tableau suivant, par exemple, cette contrainte correspond au nombre de produits :

Catégorie	Profil Assortiment	Nombre de références
`Soda`	A	10
`Soda`	B	25
`Soda`	C	50
`Soda`	...	...
`Soda`	Z	X

TABLE 8 – Exemple de contraintes liées aux profils assortiment

Remarque : les contraintes (ou coûts) dont nous parlons dans ce chapitre sont liées aux caractéristiques des produits. Par exemple, il peut s'agir de l'encombrement des produits et par suite des mètres linéaires nécessaires pour pouvoir les mettre en rayon ou tout simplement le nombre maximal de produits d'une catégorie pouvant être disposés en magasin comme illustré dans le tableau 8. Elles sont différentes des contraintes utilisateurs proposées dans la contribution de la section 5.2. L'optimisation naïve de l'assortiment consisterait à essayer tous les sous-ensembles de produits possibles sans tenir compte des contraintes des magasins ou des profils assortiment. En d'autres termes, il faudrait tester toutes les combinaisons possibles d'assortiments gigognes pour toutes les catégories afin d'identifier celles qui maximisent l'utilité des produits d'un magasin. Toujours naïvement, proposer les assortiments gigognes les plus grands devrait systématiquement maximiser l'utilité de l'assortiment d'un magasin, cependant cela ne saurait être mis en place pour d'évidentes raisons de place ou de trésorerie pour ce qui concerne les plus petits magasins d'un réseau. L'utilité et les coûts peuvent être calculés pour les catégories parents des produits qui font l'objet de l'assortiment gigogne de manière récursive comme le montre la Figure 15.

Plus formellement, plus $P(\Omega) \triangleq \sum_{i=1}^{n} P(S^{k_i}(F_i))$ est grand, plus grande est l'utilité et donc meilleur est l'assortiment de Ω, sans autres contrain-

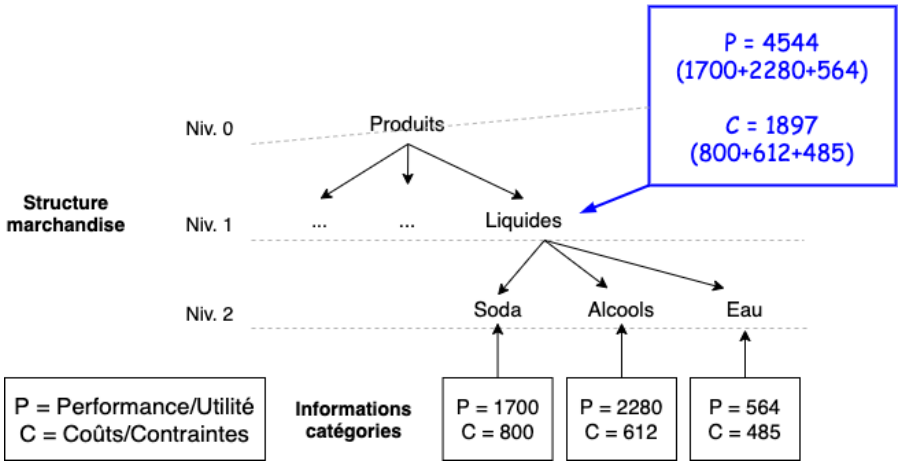


FIGURE 15 – Illustration de propagation d'utilité et de coûts

tes, $P(\Omega)$ devrait être nécessairement maximal lorsque $k_i = \mathcal{K} \; \forall \; i = 1..n$.

En pratique, $\sum_{i=1}^{n} C(S^{k_i}(F_i))$ est généralement inférieur ou égal à $\sum_{i=1}^{n} C(S^{\mathcal{K}}(F_i))$ pour des contraintes évidentes de stockage ou de coût ou de taille ... nous notons $\mathcal{C}$ cette borne supérieure. Considérons $\mathcal{I}$ un sous-ensemble de catégories de produits. Il peut être nécessaire de modéliser des contraintes liées aux catégories parents. Par exemple :

$$C(S^{k_{Soda}}(Soda)) + c(S^{k_{Alcools}}(Alcools)) + C(S^{k_{Eau}}(Eau)) \leq \mathcal{C}_{\mathcal{I}=Liquides}$$

Cela signifie que la capacité de la catégorie `Liquides` est bornée par $\mathcal{C}_{\mathcal{I}}$ et que le choix des assortiments gigognes des catégories enfants `Soda`, `Alcools` et `Eau` doit respecter cette contrainte globale. On peut également fixer une borne inférieure (c_I) sur chaque catégorie, *e.g.* il faut a minima proposer un échantillon de produits pour chaque catégorie I. La formalisation de la recherche de l'assortiment optimal en étudiant le bon profil assortiment peut alors être formalisé selon le problème d'optimisation combinatoire suivant :

$$\begin{aligned}
&\text{Arg} \max_{k_i, i=1..n} \sum_{i=1}^{n} P(S^{k_i}(F_i)) \\
&\text{Sous la contrainte :} \\
&\sum_{i=1}^{n} c(s^{k_i}(F_i)) \leq \mathcal{C} - \textit{ contrainte globale} \\
&\text{Pour chaque } \mathcal{I} \text{ dans } 2^{\{1..n\}}, c_{\mathcal{I}} \leq \sum_{i=1}^{|\mathcal{I}|} C(S^{k_i}(F_i)) \leq \mathcal{C}_{\mathcal{I}} - \textit{-contrainte locale}
\end{aligned}$$

Ce problème d'optimisation combinatoire est connu sous le nom du problème de sac à dos avec des contraintes mono-dimensionnelles et des variables numériques bornées [22].

Nous avons des contraintes pour chaque profil assortiment de chaque catégorie, tout comme nous avons des contraintes pour chaque magasin, par exemple, un nombre de références comme l'illustre le tableau suivant :

Magasin	Nombre de références
M1	55
M2	70
M3	40
...	...
MN	XX

TABLE 9 – Exemple des contraintes magasins liées aux profils assortiment

Nous recherchons donc les meilleurs profils assortiment pour chacune des catégories en fonction des magasins. Afin de conserver la stratégie véhiculée par l'enseigne, les magasins doivent avoir au moins le plus petit profil assortiment dans chacune des catégories. Cela permet de couvrir les unités de besoins identifiées par l'enseigne. Autrement dit, quel que soit le magasin et la catégorie analysés, nous retrouverons toujours les produits inclus dans le plus petit profil assortiment.

4.1.2 *Estimateur de gain d'assortiment supérieur*

A partir de ce problème d'optimisation, pour améliorer l'assortiment d'un magasin, deux actions sont envisageables. Elles consistent soit à réduire le profil assortiment d'une catégorie, soit à l'augmenter.

Le fait de réduire le profil assortiment d'une catégorie F_i correspond à passer de $S^{k_i}(F_i)$ à $S^{k_i-1}(F_i)$ dans l'assortiment gigogne. Le manque à gagner lié à cette restriction peut être estimé par $P(S^{k_i}(F_i) \backslash S^{k_i-1}(F_i))$. Cette perte est identique à celle calculée pour la rationalisation de l'assortiment (*cf.* Section 5.4).

Augmenter le profil assortiment d'une catégorie F_i correspond à une mise à niveau de F_i, de $S^{k_i}(F_i)$ à $S^{k_i+1}(F_i)$. L'estimation associée est plus complexe à calculer. Les contraintes $C(S^{k_i+1}(F_i))$ sont directement disponibles. En revanche, il est plus difficile d'estimer $P(S^{k_i+1}(F_i))$, information pourtant nécessaire pour évaluer l'utilité de l'assortiment du magasin testé à renseigner pour chaque alternative dans la fonction objectif du problème d'optimisation combinatoire.

Cette mise à niveau du profil assortiment ne peut être estimée qu'à partir d'autres mesures de référence rencontrées dans d'autres

magasins similaires. Nous nous plaçons ici dans un contexte ambigu entre exploitation et exploration (*cf.* Section 2.2). L'idée de base est que plus ces magasins de "référence" sont similaires au magasin concerné, plus les estimations seront fiables.

Le problème revient alors à définir ce que signifie "référence". Intuitivement, les magasins de "référence" sont ceux disposant de rayons similaires à celui du magasin étudié (Ω) et qui proposent de plus les produits inclus dans $S^{k_i+1}(F_i)$ pour la catégorie F_i si on s'intéresse à une augmentation du profil assortiment k_i pour cette catégorie de produits spécifiquement. A partir de ces magasins de "référence", $P(S^{k_i+1\Omega}(F_i))$ peut être calculé, par exemple, en prenant la moyenne pondérée (les poids sont les similarités), le maximum ou tout opérateur d'agrégation des $P(S^{k_i+1\Omega'}(F_i))$ où, Ω' est un des magasins de "référence" pour le magasin Ω. L'estimateur du gain associé à un assortiment qui va permettre de comparer les assortiments dans le problème d'optimisation repose donc sur la notion de voisinage : une amélioration du profil assortiment pour un magasin donné doit induire un gain additionnel (une augmentation de l'utilité plus généralement) dont la valeur est estimée à partir des valeurs références des magasins au profil voisin du magasin à l'étude. La similarité entre deux magasins devient alors une mesure cruciale pour obtenir des estimations fiables.

Nous pouvons considérer que Ω est similaire à Ω' si la répartition de leur chiffre d'affaire, par exemple, est la même pour les deux magasins sur la structure marchandise. Cela implique a minima que les consommateurs partagent approximativement les mêmes habitudes d'achats dans les différentes catégories de produits dans les deux magasins.

Intuitivement, la distance entre deux magasins quelconques pourrait être basée sur un espace métrique classique où les dimensions correspondraient à tous les produits proposés par l'ensemble des magasins d'une enseigne. Par exemple, la valeur de chaque coordonnée serait le chiffre d'affaires du produit et serait nulle si le magasin ne propose pas ce produit.

Étant donné que certains grands distributeurs offrent plus de 100 000 produits différents (*cf.* Tableau 2), le processus de clustering sur un tel espace serait basé sur une matrice creuse et souffrirait d'un trop grand nombre de dimensions. Qui plus est, une telle distance ne prendrait pas en considération les liens sémantiques donnés par la structure marchandise. Dans un tel processus les produits des catégories `Fruits` et `Légumes` seraient équidistants avec les produits de la catégorie `Bricolage`. Tout comme le seraient le `Coca-Cola 1L5`, le `Coca-Cola 1L25` et le `Stylo Bic bleu`. Ce qui en pratique est contre-intuitif.

Cette similarité intuitive ne peut pas être évaluée avec des distances classiques et nous semble pourtant indispensable pour proposer des solutions pertinentes au calcul de voisinage d'un magasin. Il est donc nécessaire d'introduire des mesures plus appropriées qui exploitent ces connaissances sémantiques pour évaluer la similarité entre les magasins.

Nous avons présenté les profils assortiment $S^{k_i}(F_i), k_i = 1..\mathcal{K}$ pour toute catégorie de produit F_i. Il est important de noter que l'action de passer le profil assortiment de k_i à $k_i + 1$ doit générer une augmentation de l'utilité pour la catégorie de produit concernée, tout comme l'action consistant à changer d'un profil assortiment k_i à $k_i - 1$ génère une perte d'utilité.

En revanche, proposer un assortiment gigogne plus conséquent (de k_i vers $k_i + 1$) a nécessairement des répercussions sur les contraintes (stockage, place, ...) ou coûts associés à la famille F_i puisque $C(S^{k_i}(F_i)) < C(S^{k_i+1}(F_i))$. Par conséquent, il est nécessaire de réduire le profil assortiment d'une autre catégorie de produit $F_{j,j\neq i}$ pour respecter le coût marchandise global ou la contenance globale du magasin. A ce stade, nous pouvons donc pour tout magasin Ω, et, pour tout profil assortiment k_i ($k_i \in [1..\mathcal{K}]^n$), estimer $P(S^{k_i+1}(F_i))$ grâce à l'utilité mesurée des magasins le plus similaires à Ω.

Finalement la méthode que nous proposons ici consiste à définir un profil assortiment k_i pour chacune des catégories et chacun des magasins. Il suffit alors d'énumérer et d'évaluer tout assortiment potentiel qui respecte les contraintes d'encombrement ou de coût pour sélectionner le meilleur qui sera considéré comme étant la solution du problème d'optimisation.

4.2 TAXONOMIE ET RAISONNEMENT D'ABSTRACTION

Comme nous venons de l'évoquer, la notion de similarité entre les magasins est complexe et nécessite de trouver une mesure adéquate qui tienne compte de la notion de structure marchandise. La mesure de similarité qui répond à nos attentes repose sur la structure taxonomique qui organise les produits et les catégories de produits : la structure marchandise. Le principe reste le même que dans l'analyse des clients que nous avons proposée au chapitre précédent. Rappelons que le magasin Ω devrait être similaire à Ω' si la répartition des chiffres d'affaires de Ω et Ω' sur la structure marchandise est semblable.

Dans notre étude, nous avons choisi d'utiliser uniquement la structure marchandise car elle se retrouve chez tous les distributeurs. Rappelons (*cf.* Chapitre 3) que les éléments de la structure marchandise

(appelée aussi taxonomie de produits) sont vus comme les concepts (ou classes) d'une taxonomie et qu'il est possible de définir un ordre partiel des concepts d'un domaine en généralisant et en spécialisant les relations entre les concepts (*e.g.* `Liquides` généralise `Soda` qui à son tour généralise `Coca` ou `Schweppes`).

Plus formellement, nous notons la taxonomie conceptuelle $T = (\preceq, C)$, où (C) représente l'ensemble des concepts (*i.e.* les catégories de produits dans notre cas) et ($\preceq$) l'ordre partiel. On note $A(c) = \{x \in C \mid c \preceq x\}$ et $D(c) = \{x \in C \mid x \preceq c\}$ respectivement les ancêtres et les descendants du concept $c \in C$. La racine est le concept unique sans ancêtres (sauf lui-même) ($A(root) = \{root\}$) et un concept sans descendant est noté une feuille (dans notre cas, une feuille est un produit) et $D(leave) = \{leave\}$. Nous notons également *leaves-c* l'ensemble des feuilles (*i.e.* les produits dans notre étude) qui sont incluses dans le concept (ou classe) c, *i.e.*, $leaves\text{-}c = D(c) \cap leaves$.

4.2.1 *Informativité basée sur la taxonomie*

Le Contenu Informationnel (IC) a été introduit dans le chapitre 3. Nous définissons I l'ensemble d'instances, et $I^*(c) \subseteq I$ les instances qui sont explicitement associées au concept c. Nous considérons qu'aucune annotation associée à une instance ne peut être déduite ou inférée, *i.e.*, $\forall\, c, c' \in C$, avec $c \preceq c'$, $I^*(c) \cap I^*(c') = \emptyset$. On note $I(c) = I$ les instances qui sont associées au concept c considérant la transitivité de la relation taxonomique et de l'ordre partiel du concept $\preceq$, *e.g.* $I(\texttt{Soda}) \subseteq I(\texttt{Liquides})$. On obtient alors :

$$\forall\, c \in C, |I(c)| = \sum_{x \in A(c)} |I^*(x)|$$

Dans cette approche, nous proposons d'utiliser les récapitulatifs de ventes (*cf.* Section 5.3.3) pour instancier les concepts de la taxonomie. En pratique seuls les produits sont vendus, l'information n'est donc initialement portée que par les feuilles de la structure marchandise (les produits), ce qui correspond à :

$$\forall c \notin leaves, |I^*(c)| = 0$$

Cependant, en raison de la transitivité de la relation taxonomique, les instances d'un concept $c \in C$ sont également des instances de tout concept subsumant c, *i.e.*, $\texttt{Soda} \preceq \texttt{Liquides} \Rightarrow I(\texttt{Soda}) \subseteq I(\texttt{Liquides})$. Cette notion centrale est généralement utilisée pour discuter de la spécificité d'un concept, *i.e.* à quel point un concept est restrictif par rapport à $I(root)$.

Plus un concept est restrictif, plus il est considéré comme spécifique. Dans la littérature, la spécificité d'un concept est également modélisée

par le Contenu Informationnel (IC) au sens de la théorie de Shannon. Nous faisons référence à la notion d'IC définie par une fonction $IC : C \longrightarrow \mathbb{R}^+$ telle que conformément aux contraintes de modélisation des connaissances, cette fonction IC doit diminuer de façon monotone, des feuilles à la racine de la taxonomie, de façon à respecter :

$$c \preceq c' \Rightarrow IC(c) \geq IC(c')$$

Dans ce manuscrit, des informations externes ont été utilisées pour estimer l'informativité du concept (*i.e.* qui ne sont pas portées intrinsèquement par la structure taxonomique). Il s'agit d'une approche extrinsèque, basée sur la théorie de l'information de Shannon qui propose d'évaluer le caractère informatif d'un concept en analysant les statistiques d'occurrences dans une collection d'items. Proposé à l'origine par Resnik dans le cadre de la recherche d'information dans une collection [134], le contenu informationnel d'un concept c est défini à partir de $pro(c)$, la probabilité d'occurrence de c dans la collection.

$I(c)$ peut être défini à l'aide de $pro : 2^c \rightarrow [0,1]$ avec $pro(c) = |I(c)|/|I|$. L'informativité d'un concept est ensuite définie par : $IC(c) = -log(pro(c))$.

On note T la structure marchandise (ou taxonomie de produits), F l'ensemble des catégories et $P^\Omega(x)$ (*i.e.* x est un produit) est le chiffre d'affaires lié au produit x dans le magasin Ω. Nous pouvons alors définir la probabilité pro comme :

$$\left|\begin{array}{l} \forall x \in \ leaves(T), |I^*(x)| \triangleq |I(x)| = \sum_\Omega P^\Omega(x) \text{ alors} \\ \\ pro(x) = \dfrac{\sum_\Omega P^\Omega(x)}{\sum_x \sum_\Omega P^\Omega(x)} \text{ et } IC(x) = -log(pro(x)) \\ \forall\ f \notin \ leaves, |I^*(f)| = 0, |I(f)| = \sum\limits_{x \in \{leaves\ -\ f\}} |I(x)|, \\ \\ pro(f) = \dfrac{\sum_{x \in \{leaves\ -\ f\}} \sum_\Omega P^\Omega(x)}{\sum_x \sum_\Omega P^\Omega(x)} \text{ et } IC(f) = -log(pro(f)) \end{array}\right.$$

Une fois le contenu informationnel (x) des concepts défini, et grâce à la structure marchandise, il est possible de mesurer la similarité entre concepts en ayant recours aux similarités sémantiques basées sur l'IC que nous avons introduites dans la Section 3.1.2). Ainsi, nous obtenons une famille de mesures de similarité qui permettent de modéliser l'idée initiale et intuitive que deux magasins sont semblables si leurs chiffres (par exemple) sont voisins sur toute la structure marchandise.

4.2.2 *Exemple illustratif de l'optimisation des magasins*

Cette section vise à illustrer la modélisation et la chaîne de traitement des données mise en œuvre pour optimiser les profils assor-

timent des magasins. Nous présentons, par la suite, les différentes étapes de la méthodologie.

Les paramètres et variables requis sont les suivants :

1. une structure marchandise commune à tous les points de vente ;
2. des profils assortiment ($S^{k_i}(F_i)$) définis pour chacune des catégories de produits F_i (ces catégories sont définies par l'enseigne) ;
3. un coût/contrainte associé à chaque profil assortiment pour chaque catégorie de produits $C(S^{k_i}(F_i))$. Ces contraintes peuvent être, par exemple, un nombre de références, un coût de stockage, un prix de revient. Dans cet exemple illustratif, nous utilisons le coût de stockage qui est plus représentatif des contraintes réelles mais en réalité plus complexe à capter ;
4. une utilité $P(S^{k_i}(F_i))$ (*e.g.* le chiffre d'affaire) pour chacun des magasins, associée aux différents profils assortiment S^{k_i} de F_i ;
5. Les seuils nécessaires pour exprimer des contraintes de bornes supérieures sur le coût/stockage de catégories (*e.g.* $\mathcal{C}_{\mathcal{I}=Liquides}$).

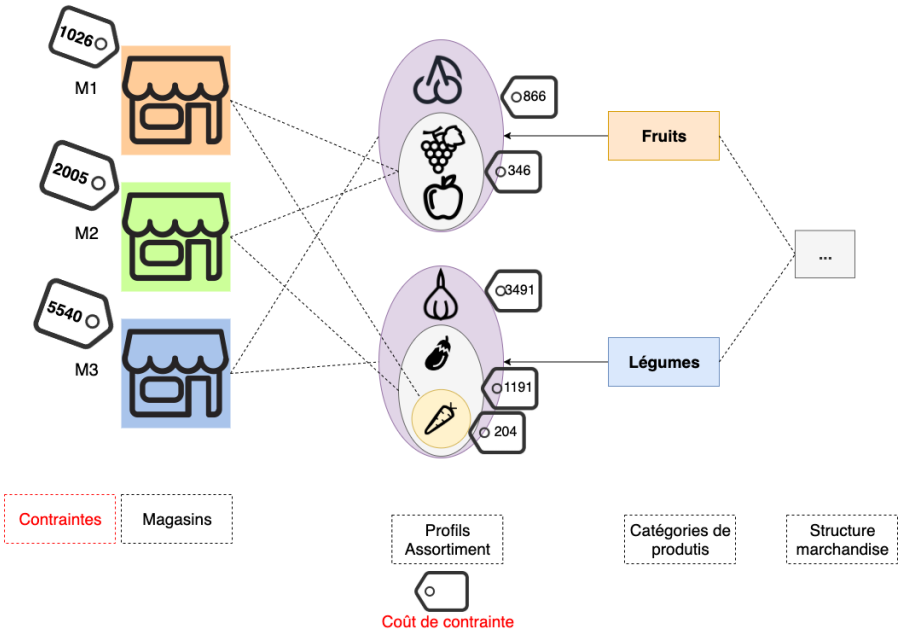


FIGURE 16 – Paramètres et variables requis

La Figure 16 illustre les différents paramètres et variables nécessaires. Dans cet exemple, nous considérons trois magasins (M1,M2,M3) :

1. Nous considérons seulement deux familles de produits notées F pour `Fruits` et L pour `Légumes`. Deux profils assortiment sont proposés pour la catégorie de produits `Fruits` (*i.e.*, $k_F \in \{1;2\}$) et trois pour la catégorie `Légumes` (*i.e.*, $k_L \in \{1;2;3\}$). Nous avons, les relations d'inclusion suivantes : $S^1(F) \subset S^2(F)$ et $S^1(L) \subset S^2(L) \subset S^3(L)$.

2. Chaque profil assortiment possède son propre coût, dans cet exemple un coût de stockage : $C(S^1(F)) = 346; C(S^2(F)) = 1191; C(S^1(L)) = 204; C(S^2(L)) = 866; C(S^3(L)) = 2400$.
3. A partir des profils assortiment associés à chaque magasin, dans cet exemple $(S^k_F(F), S^k_L(L))$, l'utilité totale peut être calculée $(P(S^{k_F}(F) + P(S^{k_L}(L)))$.
4. Chaque magasin a également une borne supérieure pour ses coûts de stockage notée en plus de son propre seuil noté respectivement $SC_1 = 1670; SC_2 = 2700; SC_3 = 5540$ ce qui implique que $C(S^{k_F}(F)) + C(S^{k_L}(L)) \leq SC$ pour chacun des magasins.

Tout changement de $S^{k_F}(F)$ ou $S^{k_L}(L)$ entraîne des variations de l'utilité de l'assortiment et la vérification des contraintes de stockage. L'objectif consiste donc à identifier le couple $(S^{k_F}(F), S^{k_L}(L))$ qui propose la plus grande utilité tout en respectant les contraintes de coûts. Ce résultat est obtenu en résolvant le problème d'optimisation que nous avons décrit précédemment. Comme nous l'avons souligné, la principale difficulté est l'évaluation de l'utilité lorsque $S^{k_F}(F)$ et $S^{k_L}(L)$ sont changés en $S^{k'_F}(F)$ et $S^{k'_L}(L)$ avec $k_F < k'_F$ et $k_L < k'_L$.

Une estimation de ces utilités doit être réalisée pour permettre l'évaluation de la performance des différents couples $(S^{k'_F}(F), S^{k'_L}(L))$ dans le problème d'optimisation. Comme expliqué ci-dessus, cette estimation est basée sur l'utilité de magasins similaires qui proposent $S^{k'_F}(F)$ et $S^{k'_L}(L)$ pour les familles `Fruits` et `Légumes`. À cette fin, nous pouvons appliquer des mesures de similarité sémantique sur la structure marchandise pour calculer la matrice de similarité entre les magasins (*cf.* Section 4.2.1).

Comme nous l'avons fait déjà souligné, à partir de cette matrice de similarité, des clusters sémantiques de magasins peuvent aussi être définis. Ils permettent d'apporter de nouvelles connaissances pour les enseignes notamment une vision très haut niveau des habitudes d'achats des consommateurs.

Une fois la matrice des similarités calculée à l'instar de la validation proposée au chapitre précédent, elle est utilisée pour estimer l'utilité des différents profils assortiment changés ($S^{k'_F}(F)$ et $S^{k'_L}(L)$) correspondant, dans cet exemple à l'utilité du magasin le plus similaire qui distribue $S^{k'_F}(F)$ et/ou $S^{k'_L}(L)$.

Un exemple d'estimation d'utilité des profils assortiment est proposé Figure 17. Nous considérons ici que les magasins M1 et M2 sont les plus proches, information obtenue grâce aux mesures de similarité sémantique. Comme le magasin de "référence" pour M1 est le magasin M2, l'estimation du passage de $S^1(L)$ vers $S^2(L)$ correspond à l'utilité du magasin M2. Dans cet exemple illustratif, nous proposons d'utiliser l'utilité maximale du magasin de "référence" pour estimer l'impact des

décisions pouvant être prises. En pratique n'importe quel agrégateur (*e.g.* moyenne pondérée) peut être mis en place. Comme M3 est le seul magasin à proposer $S^2(F)$, il est le seul à permettre d'estimer l'utilité de l'ensemble de produits associés à ce profil assortiment. Ainsi, les magasins M1 et M2 pourraient obtenir une utilité de 476 s'ils proposaient ce profil assortiment pour la catégorie `Fruits`. Tout comme pour $S^2(F)$, seul le magasin M3 propose le profil assortiment $S^3(L)$ raison pour laquelle son utilité sert d'estimation pour ce profil assortiment.

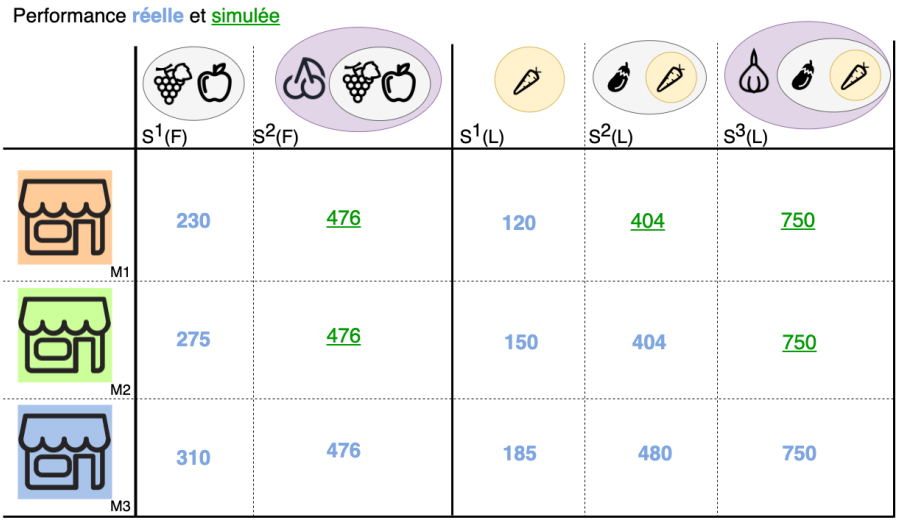


FIGURE 17 – Estimation de l'utilité

La dernière étape consiste à exploiter ces estimations dans le problème d'optimisation. Comme expliqué précédemment, nous cherchons à identifier le profil assortiment idéal pour chaque catégorie de produits pour trouver :

$$\text{Arg} \max_{k_i, i=1..n} \sum_{i=1}^{n} p(S^{k_i}(F_i))$$

tout en respectant la contrainte (dans cet exemple, nous introduisons seulement une contrainte globale) :

$$\sum_{i=1}^{n} c(S^{k_i}(F_i)) \leq \mathcal{C}$$

Cela revient à évaluer toutes les combinaisons de $S^{k_i}(F_i)$ pour toutes les catégories de produits F_i. Les contraintes sur les $C(S^{k_i}(F_i))$ sont appliquées et permettent ainsi de réduire l'espace des solutions du problème d'optimisation. Une illustration de la façon dont ces contraintes locales réduisent l'ensemble des solutions est proposée Figure 18.

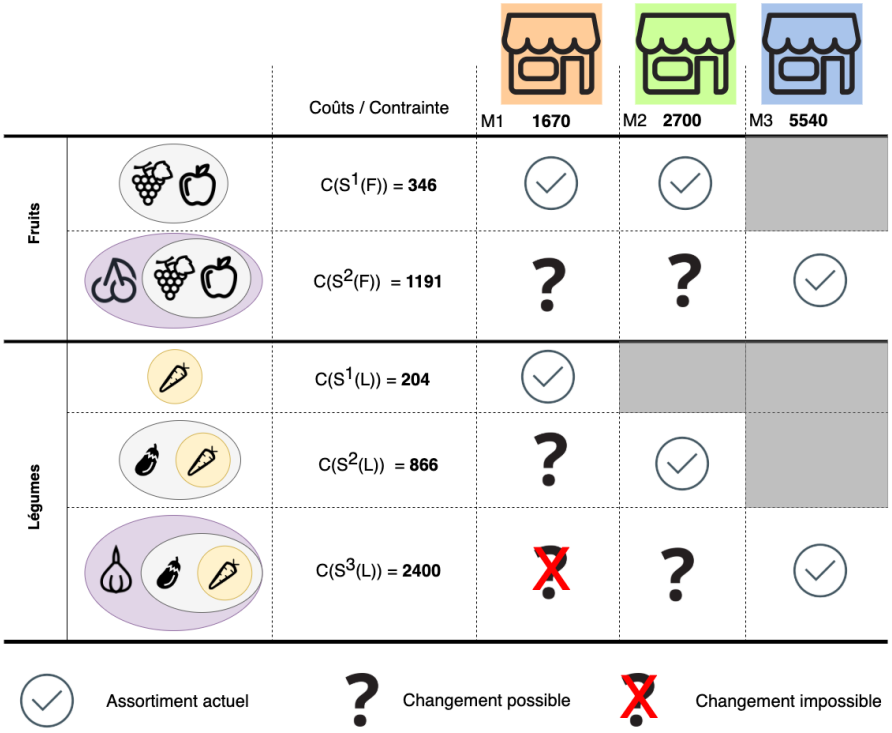


FIGURE 18 – Exemple de réduction des solutions grâce aux contraintes

Dans cet exemple, le profil assortiment le plus élevé pour les légumes $S^3(L)$ est supérieur à la capacité de stockage totale des magasins M1 et M2. Cette information permet d'éliminer la mise à niveau $S^3(L)$ pour la catégorie de produits `légumes` dans les magasins M1 et M2 en prenant en compte les assortiments de rang inférieur. Enfin, trois améliorations peuvent être envisagées :

1. Le magasin M1 peut améliorer son profil assortiment pour la catégorie de produits `Fruits` en passant de $S^1(F)$ à $S^2(F)$:
$$C(S^2(F)) + C(S^1(L)) \leq \text{SC}_1$$
2. Le magasin M1 peut améliorer son profil assortiment pour la catégorie de produits `Légumes` en passant de $S^1(L)$ à $S^2(L)$:
$$c(S^1(F)) + c(S^2(L)) \leq \text{SC}_1$$
3. Le magasin M2 peut améliorer son profil assortiment pour la catégorie de produits `Fruits` en passant de $S^1(F)$ à $S^2(F)$:
$$c(S^2(F)) + c(S^2(L)) \leq \text{SC}_2$$

Le magasin M3 possède déjà tous les produits, donc aucune amélioration n'est possible. En raison de sa capacité de stockage, le magasin M2 ne peut qu'améliorer son assortiment `Fruits`. Le magasin M1 peut améliorer ses assortiments `Fruits` ou `Légumes`. La Figure 17 fournit les estimations de l'utilité pour tous les changements de profils assortiment. Les améliorations optimales peuvent désormais en être déduites. Par conséquent, le magasin M1 devrait mettre à niveau son

profil assortiment de la catégorie `Légumes` pour améliorer au mieux son utilité totale. Cet exemple permet de souligner comment le problème d'optimisation peut être réduit grâce aux restrictions locales et au raisonnement taxonomique.

Pour mettre en place l'optimisation naïve dans cet exemple, nous devrions essentiellement raisonner sur l'ensemble des produits suivants :

— `pomme`
— `pamplemousse`
— `cerise`
— `carotte`
— `aubergine`
— `oignon`

Pour seulement 6 produits, nous avons 63 possibilités $[2^n - 1]$ qui doivent être testées pour chaque magasin. Grâce à notre approche, raisonner sur la taxonomie des produits et gérer les contraintes permet de réduire considérablement l'espace de recherche. Dans cet exemple, nous réduisons l'espace de recherche initial (les 63 possibilités) à seulement 5 possibilités comme l'illustre la Figure 18.

Si le nombre de magasins livrés diminue, alors les coûts qui ont été théoriquement minimisés risquent d'augmenter car ils dépendent de négociations entre une enseigne et un fournisseur. Cette négociation consiste à vendre certains produits dans un maximum de points de vente en échange de prix avantageux. Si le nombre de points de vente proposant l'article diminue, alors les prix seront mécaniquement moins avantageux et les coûts initialement pris en considération pour optimiser l'assortiment évolueront. Raison pour laquelle, les profils assortiment sont indispensables pour s'assurer de respecter au mieux ces ententes. Ce type de problème industriel met en avant l'intérêt des contraintes proposées et exploitées dans les différents scénario de la méthode Agile (*cf.* Chapitre 5).

Pour garantir que ce processus puisse être exploité sur les volumes de données que nous retrouvons dans la grande distribution, nous avons établi trois benchmark basés sur la taxonomie Google [1]. Les expériences ont été menées sur 1 processeur Intel Core I7-2620M 2,7GHz 8Go RAM. Nous avons exploité la bibliothèque CPLEX (IBM CPLEX 1.25) et chaque benchmark à requis moins d'une seconde. Ces benchmarks nous permettent d'affirmer que notre processus peut être appliqué sur un important volume de données comme indiqué dans le tableau 10.

1. https://www.google.com/basepages/producttype/taxonomy.fr-FR.txt

TABLE 10 – Détails des Benchmarks

	Benchmark 1	**Benchmark 2**	**Benchmark 3**
Nombre de magasins	15	30	50
Nombre de profils assortiment	4	16	20
Nombre de catégories de produits	12	80	200
Nombre de variables	180	2 400	10 000

La méthodologie formalisée ici permet d'améliorer les profils assortiment associés aux magasins. Elle s'intègre parfaitement dans un processus d'amélioration continue. En effet, proposer de changer le profil assortiment des magasins permet d'offrir de nouvelles solutions aux autres méthodes d'optimisation de l'assortiment. Par exemple, le nombre de références associées aux plus petits magasins restreint les profils assortiment (*cf.* Section 5.3.2). Changer le profil assortiment des plus petits magasins permet de changer cette contrainte et par conséquent, de nouveaux produits peuvent être inclus dans les différents profils assortiment.

Notre étude sémantique permet d'améliorer la précision des estimations d'utilité grâce à l'identification de magasins qui présentent des comportements d'achats similaires. Alors que généralement, les clusters utilisés sont ceux définis par une enseigne (*e.g.* Format, Région ...), nous pouvons maintenant, grâce aux mesures de similarité sémantique, exploiter des points communs plus précis et pertinents dans les tendances de vente.

4.3 CONCLUSION

Cette analyse sémantique des magasins permet d'identifier la typologie des produits achetés par les consommateurs. Cela donne une synthèse des habitudes d'achats des clients dans les différents magasins, qui nous semble être une information plus pertinente que la taille ou la localisation pour estimer les changements à apporter à l'assortiment d'un magasin.

L'exploitation des données des consommateurs, présentée dans le chapitre 3, permet d'aller plus loin dans l'analyse d'habitudes d'achats. Cette analyse peut en plus apporter de nouvelles contraintes pouvant-être appliquées sur les catégories les plus demandées/achetées. Ainsi, nous pouvons très bien imaginer rajouter des contraintes au problème d'optimisation issues de l'analyse de segmentation des clients : elles

pourront, par exemple, suggérer de modifier les valeurs seuils des contraintes de coûts sur les familles de produits. Enfin, la formalisation de l'amélioration de l'assortiment comme un problème d'optimisation nous a permis de développer des interfaces de recommandations permettant d'apporter des solutions aux grands distributeurs. Ces interfaces sont présentées dans le chapitre suivant.

5

DE LA STRATÉGIE DE L'ENSEIGNE À UN PROCESSUS D'AMÉLIORATION CONTINUE DE L'ASSORTIMENT

Dans ce manuscrit, nous nous intéressons à la problématique de l'assortiment des grands distributeurs. Nous avons souligné dans le chapitre 2 la divergence entre les travaux de recherche et les méthodes industrielles pour la planification d'assortiment. Tandis que des modèles théoriques complexes sont proposés pour une optimisation locale de l'assortiment, il est industriellement traité avec une vision globale qui vise, par la massification des produits, à baisser les coûts. Les derniers travaux de recherche (*e.g.* [51, 67, 81]) proposent des changements d'assortiments plusieurs fois dans une journée qui sont des solutions envisageables uniquement pour le e-commerce. La grande distribution quant à elle continue de gérer son assortiment en se focalisant sur l'optimisation de sa chaîne d'approvisionnement qui doit s'adapter à une croissance du nombre de magasins.

Aujourd'hui, seuls quelques acteurs sont en charge de gérer le volume gigantesque de produits (*cf.* Tableau 2). Appelés catégorie managers, ils sont, comme leur nom l'indique, responsables de l'ensemble des produits d'une unique catégorie de produits. Deux types de catégories managers peuvent être distingués, ceux qui considèrent le réseau de magasins complet que nous appellerons *catégorie managers centrale* et ceux qui ne s'occupent que d'un magasin qui seront appelés *catégories managers magasins*. Dans ce chapitre, nous exploitons la formalisation de la planification d'assortiment sous la forme d'un problème d'optimisation combinatoire faite dans le chapitre précédent pour proposer une méthode qui permettra d'aider ces différents acteurs dans leurs prises de décisions. Grâce aux structures de connaissances précédemment introduites, nous formalisons différents scénarios d'optimisation de l'assortiment. Ils se basent sur le problème d'optimisation présenté dans le chapitre 4 pour faire émerger un processus d'optimisation continue.

La section 5.1 rappelle les problématiques de la planification d'assortiment rencontrées aujourd'hui. La section 5.2 propose une syntaxe permettant de traduire la stratégie des distributeurs en contraintes. Dans la section 5.3, nous formalisons notre méthode d'optimisation Agile avec deux approches : l'optimisation globale et l'optimisation locale. La fin de cette section définit la performance d'un assortiment et le calcul des simulations mises en place dans nos systèmes de recommandations. La section 5.4 propose un nouveau scénario d'optimisation de l'assortiment qui permet de souligner l'aspect modulaire

de notre méthode Agile. Enfin, la section 5.5 souligne les avantages des contributions proposées dans ce chapitre.

5.1 LA PLANIFICATION D'ASSORTIMENT

Reprenons simplement la problématique de la planification d'assortiment. L'objectif est de définir un ensemble de produits qui maximise la performance d'un ou de plusieurs magasins. Tandis que les principaux travaux de recherche proposent des assortiments dynamiques appropriés pour le e-commerce, les grands distributeurs se focalisent sur une stratégie globale qui optimise la chaîne de distribution (approvisionnement et vente) dans de très nombreux points de vente. Intuitivement ces deux approches complémentaires définissent l'une l'assortiment optimal local, donc pour un magasin ; et l'autre l'assortiment optimal global qui, dans ce manuscrit, correspond à l'assortiment gigogne (partagé par l'ensemble des magasins d'une enseigne).

5.1.1 *Assortiment Local et Global*

Comme nous l'avons introduit, les travaux de recherche se sont focalisés sur la définition d'un sous-ensemble de produits pour une catégorie précise dans un magasin [83] en considérant qu'il suffit de reproduire cette optimisation sur chacune des catégories pour obtenir l'assortiment idéal d'un magasin (*cf.* Section 2.2) [96, 132, 155]. Pour aller plus loin, certains travaux proposent d'inclure des contraintes qui réduisent l'utilité des produits en fonction des coûts de distribution associés (*e.g.* complexité d'acheminement d'un produit dans un magasin). L'introduction de ces contraintes permet de prendre en considération la gestion multi-magasins (*cf.* Section 2.2.4). Nous appelons cette façon d'optimiser l'assortiment optimisation locale puisque chacun des magasins dispose de son meilleur assortiment incluant ses propres contraintes de chaîne de distribution. À leur échelle, les grands distributeurs demandent aux catégorie managers de chacun des magasins d'optimiser, grâce à leur expertise, les produits proposés. Cependant, la liberté de ces acteurs est fortement dépendante de l'assortiment gigogne défini par des catégorie managers qui gèrent eux l'assortiment gigogne distribué à travers tous les points de vente.

Les grands distributeurs, comme toutes les sociétés, essaient d'appliquer une stratégie globale qui a pour but d'être représentative de l'image véhiculée [29]. Par exemple, une enseigne de distribution alimentaire "Discount" (*e.g.* Lidl) et une enseigne alimentaire "Classique" (*e.g.* Carrefour) ne proposeront pas les mêmes assortiments de produits. Dans l'industrie, ce sont les catégorie managers centrale qui s'occupent de définir ces assortiments "globaux" partagés par tout le réseau de magasins de l'enseigne à l'aide notamment de l'assorti-

ment gigogne. Nous appelons cette approche optimisation globale, puisque l'assortiment sera répercuté dans tous les magasins. Dans cette approche, nous cherchons à définir les meilleurs sous-ensembles de produits. À notre connaissance, aucun travail de recherche ne s'est penché sur la traduction de la stratégie des enseignes pour l'intégrer au problème de planification d'assortiment. Des recherches portant sur l'optimisation de tournées de véhicules proposent néanmoins des sous-ensembles de produits optimisant la chaîne de distribution à travers un ensemble de magasins d'une enseigne pour minimiser les coûts. Cependant, les solutions proposées s'éloignent considérablement des solutions que nous retrouvons en optimisation locale.

Industriellement, ces deux approches coexistent : des acteurs locaux, dont le pouvoir de décision dépend d'acteurs plus globaux, définissent respectivement l'assortiment de leur magasin et l'assortiment gigogne.

5.1.2 *Les assortiments dans la grande distribution*

L'un des problèmes majeurs identifié aujourd'hui est lié à l'étanchéité qui existe dans l'environnement de la grande distribution. Le fonctionnement en silos, qui a permis d'accroître l'autonomie des acteurs, des services, des procédures etc. implique la disparition de liens transverses qui existaient entre ces éléments hétérogènes. Des notions, dont les connexions implicites sont définies directement par les distributeurs, sont traitées de manière indépendantes (*e.g.* les catégories de produits). Ce problème de transversalité est d'autant plus critique avec l'expansion permanente des distributeurs et devient maintenant un frein à l'amélioration de leur performance. L'exploitation de ces connexions que nous avons appelées sémantiques entre catégories de produits devient nécessaire pour les preneurs de décision.

Grâce aux structures de connaissances et au modèle de données de TRF, introduits dans le chapitre 3, nous sommes en mesure de retrouver ces liens propres à chaque enseigne. Afin d'aider les industriels dans l'amélioration de leurs assortiments, nous considérons qu'il est indispensable de permettre à certains experts d'exprimer leur stratégie. Pour cela, l'une de nos contributions repose sur la définition d'une syntaxe qui traduit la volonté d'une enseigne en contrainte(s). Cette syntaxe est formalisée dans la section suivante et servira dans notre proposition de méthodologie Agile permettant d'améliorer l'assortiment.

5.2 EXPRESSION DE CONTRAINTES

À partir des structures de connaissances (*e.g.* taxonomie de produits), nous avons mis en place une syntaxe d'expression de contraintes. L'objectif est de permettre aux utilisateurs de traduire la stratégie de l'enseigne à l'aide de contraintes qui seront appliquées lors de

l'optimisation de l'assortiment. Par exemple, une enseigne peut demander à ce que tous ses magasins possèdent un minimum de 10% de produits de la marque du distributeur. Cette contrainte peut se retrouver factuellement dans les assortiments. Néanmoins, le respect de cette stratégie est corrélé aux décisions prises par les acteurs locaux et globaux. Comme nous l'avons souligné, ces catégorie managers définissent l'assortiment d'une unique catégorie sans considération pour les autres. Le suivi de la "stratégie" proposée précédemment (10% de produits de marque du distributeur) implique que chaque catégorie manager devra respecter cette exigence. Cependant le respect de cette consigne peut être à l'origine d'une dégradation de l'assortiment général. Grâce aux structures de connaissances et au modèle de données de TRF, nous pouvons travailler uniquement sur le sous-ensemble de produits de marque distributeur pour ajouter jusqu'au seuil des 10% les produits les plus performants. Cela permet de respecter la stratégie de l'enseigne tout en considérant la transversalité entre les catégories. Le cas extrême peut proposer un assortiment dans lequel les produits de marque distributeur sont dans une unique catégorie.

5.2.1 *Les contraintes et les préférences*

Chaque enseigne dispose de sa propre stratégie. Tandis que certains veulent valoriser leurs propres produits, d'autres préfèrent mettre en avant des produits locaux. Cette stratégie est ensuite traduite par les catégorie managers magasin ou centrale. La notion d'assortiment idéal dépend donc de nombreuses contraintes. Ces *contraintes* sont issues des caractéristiques des produits (*e.g.* types, marques ...) mais également des profils d'assortiment, des exigences du distributeur, de la localisation des magasins (*e.g.* Région/Pays,) etc. Nous considérons qu'il existe deux types de *contraintes* :

— Les contraintes strictes : qui permettent de limiter l'espace de recherche pour optimiser l'assortiment. Par exemple, la taille d'un supermarché amène une contrainte sur le nombre de produits dont il peut disposer.

— Les contraintes préférentielles : qui expriment l'objectif de l'enseigne, par exemple maximiser le chiffre d'affaires. Ces préférences sont indispensables pour personnaliser la performance/l'utilité des produits. Cette notion est détaillée à la fin de la section suivante. Pour illustrer nos propos, un distributeur qui veut se concentrer sur le marché Bio voudra maximiser le nombre de produits Bio qu'il propose au détriment du profit maximal de chacun de ses magasins.

Le problème vient du fait que chaque acteur du monde de la grande distribution ajoute des contraintes. Par exemple, un fournisseur peut livrer un produit j à la condition que les magasins qu'il livre lui commande un autre produit i. Par conséquent, ces magasins devront pos-

séder dans leur assortiment les produits i et j. Les méthodes actuelles ne permettent pas d'identifier le lien transversal entre ces produits [84]. Pour gérer ces problématiques, les industriels définissent simplement les produits comme étant "Obligatoires". Nous proposons une syntaxe permettant aux utilisateurs d'exprimer des contraintes en exploitant les structures de connaissances et les liens transversaux entre les différentes entités sémantiques. Cette syntaxe ad hoc a été implémentée pour faciliter la connexion au modèle de données de TRF. Nous avons ainsi pu définir les contraintes les plus courantes de la grande distribution et une syntaxe générique d'expressions a pu voir le jour. Cette syntaxe se décompose en quatre éléments :

1. Un objectif.
   Noté $\ll OBJ \gg$, l'objectif est comme son nom l'indique la finalité recherchée. Il permet de définir, dès l'instanciation, la nature de la condition. Autrement dit, cela permet de savoir si la condition est une contrainte stricte ou préférentielle.
2. Un KPI (Key Performance Indicator).
   Noté $\ll KPI \gg$, il permet d'identifier l'indicateur, la caractéristique ou la notion métier sur lequel porte la condition.
3. Une valeur.
   Notée $\ll VAL \gg$, qui permet de pouvoir spécifier un seuil. Nous définissons cet élément optionnel comme un nombre entier.
4. Un périmètre.
   Noté $\ll PER \gg$, il définit l'ensemble des notions concernées dans les structures de connaissances. Cet élément est essentiel dans la définition de la contrainte.

À partir de cette syntaxe des contraintes simples et/ou plus complexes peuvent être exprimées. Nous proposons quelques exemples de contraintes avec la syntaxe associée :

"Maximiser le chiffre d'affaires"
$\ll OBJ \gg =$ "Maximiser" $\rightarrow$ Nature = "Préférence"
$\ll KPI \gg =$"chiffre d'affaires"
$\ll PER \gg =$"ALL"
$\ll VAL \gg = \emptyset$;

"10 produits pour les fournisseurs X & Y"
$\ll OBJ \gg =$ "VALEUR_EXACT" $\rightarrow$ Nature = "Contrainte stricte"
$\ll KPI \gg =$"Nombre de produits"
$\ll PER \gg =$ "fournisseurs X & Y"
$\ll VAL \gg =$"10";

"Au plus 10 marques par Unité de besoin"
$\ll OBJ \gg =$ "Au plus" $\rightarrow$ Nature = "Contrainte stricte"
$\ll KPI \gg =$"Nombre de marques"

$\ll PER \gg$="Unité de besoins"
$\ll VAL \gg$="10";

"Autant de produits BIO que de produits MDD"
$\ll OBJ \gg$= "Autant de" $\rightarrow$ Nature = "Contrainte stricte"
$\ll KPI \gg$="Nombre de produits"
$\ll PER \gg$="MDD/BIO"
$\ll VAL \gg$="AUTANT";

Remarque : la dernière contrainte proposée ("Autant de produits BIO que de produits MDD" (Marque Du Distributeur)) représente la stratégie de l'enseigne et peut être considérée comme préférentielle. Cependant, elle ne définit pas un objectif, mais une contrainte stricte entre le nombre de produits BIO et le nombre de produits MDD.

À partir du "dictionnaire", nous avons construit un arbre de décision qui représente toutes ces contraintes avec les règles de premier ordre associées. Un exemple illustratif de l'arbre de décision est présenté Figure 19.

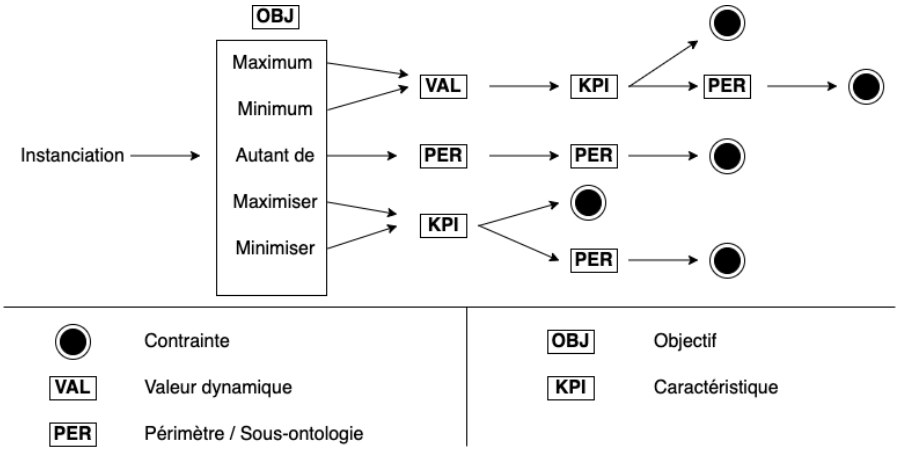


FIGURE 19 – Exemple illustratif de l'arbre de décision explicitant l'instanciation de contraintes avec les éléments nécessaires à chaque objectif

À partir des différentes combinaisons d'objectifs et KPI, des contraintes ont pu être définies. Par exemple, les contraintes qui définissent un maximum d'articles ($\ll OBJ \gg$= "Maximiser" & $\ll KPI \gg$= "Nombre de produits") sont traduites par $|S \cap \ll PER \gg| \ll VAL \gg$ où S représente l'ensemble des produits de l'enseigne, $\ll PER \gg$ le périmètre restreint des produits et $\ll VAL \gg$ le seuil à ne pas dépasser. $\ll PER \gg$ étant propre aux structures de connaissances spécifiques à chacune des enseignes, cela permet d'établir des conditions qui seront propres aux référentiels maîtrisés par les distributeurs.

Le processus d'expression des contraintes peut se décomposer en trois étapes distinctes :

1. Expression de la contrainte
2. Vérification de la validité de la contrainte
3. Traduction mathématique de la contrainte

La première étape consiste à permettre tout simplement à l'utilisateur d'exprimer sa contrainte. Ensuite, les éléments indispensables à la traduction mathématique sont vérifiés. Pour illustrer cette étape, nous considérons que l'utilisateur à entré la contrainte suivante : "Maximiser le chiffre d'affaires". Pour vérifier si la contrainte est acceptée, il est nécessaire de retrouver un "chemin" complet dans l'arbre de décision présenté dans la Figure 19. Pour la contrainte qui nous sert d'exemple, nous obtenons le chemin en vert dans la Figure 20.

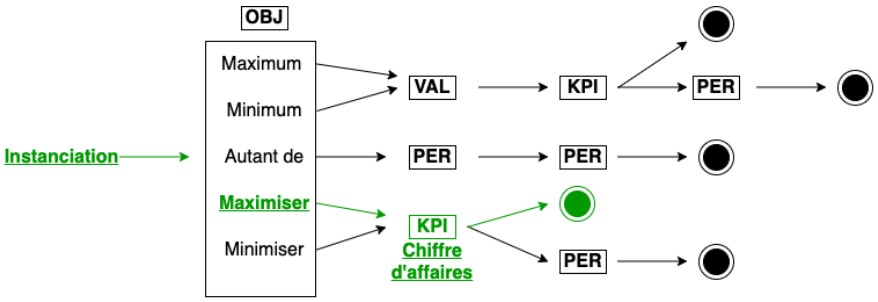


FIGURE 20 – Exemple illustratif de chemin dans l'arbre de décision

Dans la mesure où une contrainte exprimée aboutie à l'un des points final de l'arbre de décision, nous considérons que celle-ci est valide. Le chemin obtenu avec la contrainte précédente peut évoluer dans le cas où le périmètre est entré par l'utilisateur. Par exemple, si l'utilisateur souhaite "Maximiser le chiffre d'affaires des Fruits", le chemin sera modifié dans la mesure ou le périmètre est identifié comme illustré dans la Figure 21.

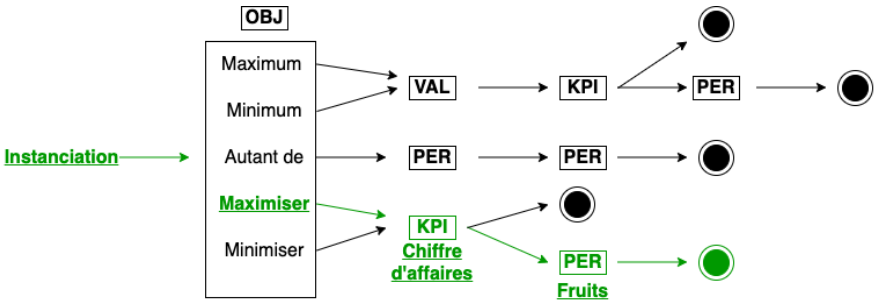


FIGURE 21 – Exemple illustratif de chemin dans l'arbre de décision

Dans cet exemple, le périmètre correspond à la catégorie de produits `Fruits`. Si le périmètre n'est pas automatiquement identifiable par l'algorithme, *e.g.* "Maximiser le chiffre d'affaires des *bons produits*",

alors soit un chemin plus court est possible et il est préféré (*cf.* Figure 20). Autrement, la contrainte est considérée comme invalide.

Une contrainte valide est par la suite étudié vis-à-vis des autres contraintes afin d'identifier d'éventuelles contraintes contradictoire. La contradiction de contraintes est une notion très intuitive comme par exemple les contraintes "Maximum 25 produits" et "Minimum 26 produits". Il existe aussi des contraintes liées au contexte de chaque enseigne comme par exemple "Minimum 250 produits Fruits" lorsque l'enseigne ne dispose que de 200 différents produits pour la catégorie de produits `Fruits`. Cependant, en fonction de chaque enseigne (*e.g.* le nombre de fournisseurs, le nombre de marques ...) l'identification de contradiction peut demander un temps de calcul relativement conséquent. Aujourd'hui, les contradictions évidentes sont identifiées et nous cherchons une méthode automatique permettant d'aller plus loin dans l'identification d'éventuelles contradiction.

Enfin, lorsque les contraintes sont validées, elles sont exploitées par les interfaces présentées dans la suite de ce chapitre. Il est important de souligner que les contraintes permettent de limiter les possibilités de changement dans l'assortiment tandis que les préférences viennent affiner la performance des produits. Ainsi, si une utilisateur souhaite "Maximiser le chiffre d'affaires", alors cet indicateurs se verra valoriser à l'aide d'un coefficient. Dans le cas ou une contrainte vise à limite le nombre de produits (*e.g.* "Maximum 50 produits Fruits"), alors l'algorithme qui permet de recommander des changements dans l'assortiment bloquera automatiquement le nombre de produits associés à la catégorie `Fruits` à 50. Cette contrainte permet de réduire drastiquement le nombre de possibilité envisageable.

Nous avons travaillé et travaillons encore sur la propagation de ces contraintes. Par exemple, définir un maximum de produits pour un ensemble de catégories de produits définit mécaniquement un maximum de produits pour le père et pour les fils. Certaines contraintes (notamment le nombre de produits) se propagent très facilement sur les structures de connaissances, néanmoins, d'autres impactent beaucoup plus d'éléments et leur propagation nécessite l'intervention d'experts. La Figure 22 clarifie la propagation de contraintes en considérant un nombre d'articles minimum pour les trois catégories de produits suivantes : `Soda`, `Alcools` et `Eau`. Nous constatons qu'à partir des contraintes définies sur les catégories `Soda` (min 15 produits), `Alcool` (min 8 produits) et `Eau` (min 5 produits) une nouvelle contrainte apparaît pour la catégorie plus abstraite que sont les `Liquides`. Ainsi, le nombre minimum de produits de type `Liquides` doit être de 28.

Cet exemple illustratif ne se base que sur la taxonomie de produits, en réalité, ces contraintes sont propagées sur les structures de connaissances dans lesquelles de nombreuses autres variables sont considérées (*e.g.* marques, fournisseurs ...). Ces contraintes, pour être exploitables, nécessitent l'expertise métier pour éviter de mettre

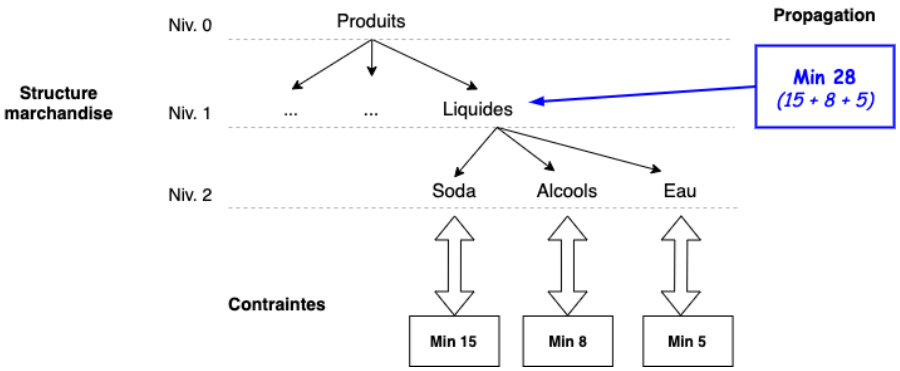


FIGURE 22 – Exemple illustratif de la propagation de contraintes

en place des contraintes contradictoires. Pour simplifier le travail demandé aux utilisateurs, nous avons développé une interface graphique qui traduit des spécifications en langage naturel vers des contraintes appartenant à l'arbre de décision (*cf.* Figure 19). Cette interface est introduite dans la section suivante. Les structures de connaissances sont exploitées dans notre méthode d'optimisation de l'assortiment formalisée dans la section 5.3.

5.2.2 *Expression des contraintes en langage naturel*

Une interface utilisateur a été développée pour faciliter l'expression des contraintes. La Figure 23 illustre l'interface disponible sur

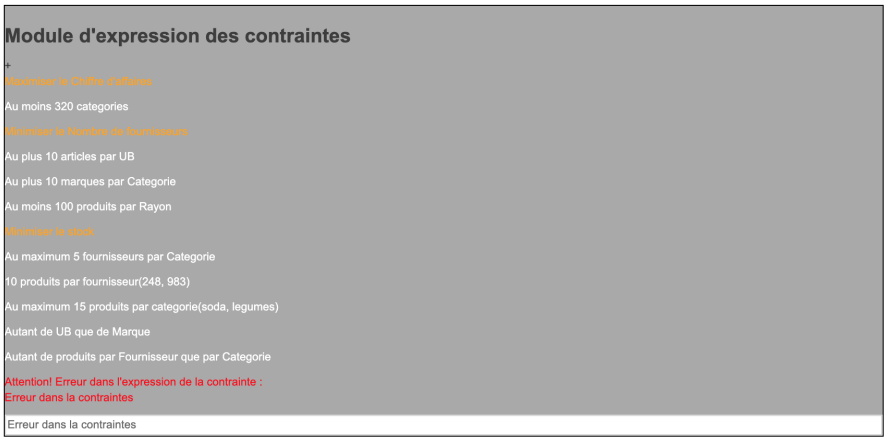


FIGURE 23 – Interface d'expression des contraintes

l'application de TRF Retail permettant aux utilisateurs d'exprimer leurs contraintes dans un vocabulaire contrôlé. Cette interface permet à tout utilisateur d'exprimer des contraintes plus ou moins complexes. Les contraintes en "Orange" représentent les contraintes préférentielles de l'enseigne. Nous retrouvons : "Maximiser le chiffre d'affaires", "Minimiser le nombre de fournisseurs" et "Minimiser le stock". Les contraintes en "Blanc" sont des contraintes strictes acceptées par notre

"traducteur" automatique qui a retrouvé l'ensemble des éléments nécessaires à notre syntaxe pour définir une règle de premier ordre sur le périmètre défini. Enfin, les contraintes en "Rouge" sont des expressions qui n'ont pu être traduites. Sur cette interface, nous pouvons voir de nombreuses contraintes pouvant être exprimées par les utilisateurs. La limite de cette interface vient de la syntaxe qui peut différer significativement d'une langue à l'autre.

Actuellement, nous avons développé le module en français et travaillons sur un module en anglais. Néanmoins, pour permettre l'expression de contraintes pour tous les clients de TRF Retail, nous avons développé un système simple de questions à choix multiples qui permet de parcourir naturellement l'arbre de décision. Cette interface est le résultat d'un long travail d'identification et de traduction des conditions. Il représente l'expertise métier de TRF Retail combinée à un formalisme mathématique complexe.
Grâce à ces deux interfaces, la stratégie des distributeurs peut être exprimée et exploitée dans la planification de l'assortiment. Les contraintes résultantes sont propres à chaque enseigne et peuvent être définies sur un sous-ensemble d'articles ou même un seul article. La section suivante formalise notre méthode Agile qui exploite ces contraintes.

5.3 FORMALISATION DE LA MÉTHODE AGILE

La problématique de ce manuscrit porte sur la planification d'assortiment. Autrement dit, nous cherchons à identifier le sous-ensemble de produits qui obtient l'utilité maximale (*cf.* Section 2.2.2). Le problème est complexe car il revêt à la fois l'aspect combinatoire du chapitre précédent, mais également la définition des assortiments gigogne de l'enseigne. Les grands distributeurs ne peuvent décemment pas changer l'intégralité des produits proposés aujourd'hui. Pour pallier ce problème, les structures de connaissances avec contraintes permettent d'observer l'environnement complet (notamment les assortiments d'une enseigne) avec un vocabulaire et des notions adaptées. Pour proposer une solution pragmatique optimisant l'assortiment, nous passons par deux approches complémentaires : l'optimisation globale et l'optimisation locale. De plus, nous considérons que l'expertise humaine est indispensable pour s'adapter aux tendances de ventes des produits qui évoluent chaque jour. Par exemple, le `handspinner` (toupie à main) a eu un cycle de vie relativement court mais des ventes exceptionnelles.

À présent, nous présentons la méthode mise en place. Généralement, les méthodes Agile sont utilisées pour la gestion et la réalisation de projets. Plus pragmatiques que des méthodes traditionnelles, elles nécessitent une très grande implication du "client" (dans notre cas, les preneurs de décisions/utilisateurs des enseignes) et se basent sur

des cycles itératifs. Ce que nous proposons c'est une méthode Agile d'amélioration de l'assortiment qui considère à la foi l'optimisation globale et locale. Le fait d'utiliser une méthode Agile permet de limiter les effets de bord en corrigeant une éventuelle erreur humaine ou l'émergence d'une toute nouvelle tendance. Cette approche permet de définir un cycle vertueux qui s'intègre parfaitement dans un processus d'amélioration continue de l'assortiment. Les structures de connaissances connectées à l'aide du modèle de données de TRF permettent aux utilisateurs d'appréhender le problème en basant leur raisonnement sur des ensembles de produits dont la sémantique associée est plus importante que les seules catégories de produits.

D'ailleurs, les liens transversaux permettent de travailler avec une vision multi-catégories. Pour s'affranchir des problèmes de complexité et pour être cohérent dans notre démarche pragmatique, nous proposons d'identifier des sous-ensembles de produits indépendants. Cette indépendance est liée à la fois aux contraintes exprimées par le(s) utilisateur(s) mais aussi à un ensemble de contraintes prédéfinies avec TRF Retail. L'expertise métier permet d'apporter des règles génériques associées aux liens présents dans les structures de connaissances. Par exemple, l'assortiment gigogne implique implicitement la présence de produits ("obligatoires") pour les magasins.

La suite de cette section formalise l'approche locale et l'approche globale.

5.3.1 *Optimisation Locale*

L'optimisation locale consiste à proposer les meilleurs produits pour un magasin donné. Dans l'industrie, c'est le catégorie manager d'un magasin qui choisit les produits qui viennent "personnaliser" l'assortiment pour une catégorie donnée. Par exemple, il peut choisir d'introduire des produits locaux pour satisfaire davantage les clients. Son pouvoir de décision est directement corrélé aux contraintes d'assortiment gigogne qui l'oblige à prendre un sous-ensemble initial de produits (*cf.* Section 2.1.3). Notre approche d'optimisation locale doit considérer ces contraintes pour identifier quels sont les produits qui peuvent être remplacés dans l'assortiment. La Figure 24 schématise l'identification des produits pouvant être retirés d'un magasin M_n. Les produits rouges sont ceux pour lesquels, grâce à la spécification des différentes contraintes, aucune action n'est envisageable. Les contraintes (dans cette Figure, l'assortiment gigogne) empêchent le retrait des produits pour garantir le respect de la stratégie définie par l'enseigne. Grâce à notre système d'expression de contraintes, n'importe quel utilisateur de l'enseigne peut en ajouter d'autres pour augmenter ou réduire le sous-ensemble de produits obligatoires. Finalement, la capacité d'action du catégorie manager est limitée à l'ensemble des produits verts qu'il peut remplacer. Pour s'assurer que les propositions

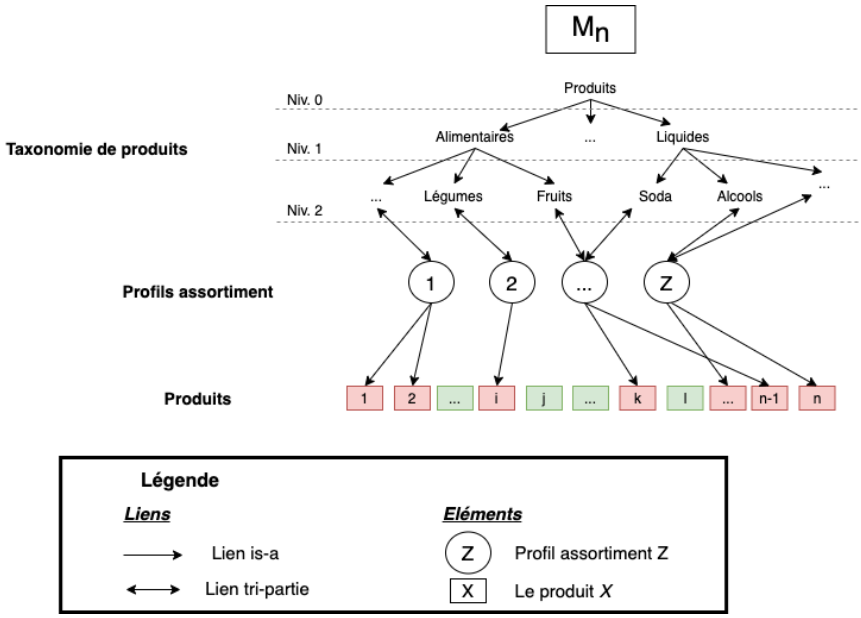


FIGURE 24 – Exemple de structure de connaissances

d'amélioration de l'assortiment sont cohérentes, nous contraignons l'algorithme à proposer un nombre de références équivalent. Autrement dit, si nous recommandons d'ajouter un article, alors un autre devra être déréférencé. Pour conserver l'image actuelle du magasin, seuls les produits d'une même catégorie peuvent être inter-changés.

Remarque : nous utilisons les catégories servant à définir l'assortiment gigogne pour nos recommandations (*cf.* Section 2.1.3).

Les nouveaux produits pouvant être ajoutés dans un magasin sont déduits des données périodiques de ventes provenant d'autres magasins de la même enseigne.
Finalement, nous avons deux ensembles de produits :

1. l'ensemble des produits de l'enseigne (tous magasins confondus) ;
2. les produits proposés dans un magasin

Pour reprendre les notations introduites dans la section 2.2, nous pouvons considérer que ces sous-ensembles de produits correspondent respectivement à :

— N l'ensemble des produits de l'enseigne, $N = \{1, 2, \ldots, n\}$,

— S le sous-ensemble de produits présents dans le magasin, $S \subset N$

La différence par rapport aux travaux de recherche introduit dans la section 2.2 est principalement liée au changement de référentiel. Comme nous l'avons souligné, ces travaux se focalisent sur une unique catégorie sélectionnée[84]. Grâce aux structures de connaissances, les

produits peuvent être traités en fonction de liens sémantiques supplémentaires. De plus, le volume de produits à analyser est relativement réduit grâce aux contraintes proposées dans la section 5.2. L'objectif dans l'optimisation locale est de définir le sous-ensemble de produits S' qui maximise l'utilité / la performance tel que :

$$\max \sum_{j \in S'} u_j(S') \wedge (|S'| = |S| + \epsilon)$$

où u_j représente la performance/l'utilité du produit j et ϵ, le rapport existant entre le volume des produits tel que :

$$\epsilon = \frac{volume(S) - volume(S')}{|S'|}$$

Cette notion subjective est formalisée dans la section 5.3.3.

En reprenant l'hypothèse proposée par [25] pour identifier l'assortiment optimal (*cf.* Section 2.2.2), nous considérons que le meilleur sous-ensemble de produits S' est celui qui contient les produits dont l'utilité est la plus grande. Le fait de contraindre le nombre de transpositions de produits ($|S'| = |S|$) simplifie l'identification du sous-ensemble de produits optimal. Si cet assortiment respecte les contraintes définies sur les structures de connaissances, alors il est recommandé dans l'interface proposée sur l'application de TRF Retail (*cf.* Image 25). Autrement, nous définissons une matrice binaire de tous les sous-ensembles possibles de produits $(A_1, A_2, \ldots, A_n)$ dans N avec la contrainte $|S'| = |S|$:

Solution	A_1	A_2	...	A_n	*Utilit*
1	0	1	...	1	7500
2	1	1	...	0	8992
3	0	0	...	0	1100
...			...		...
X	1	0	...	1	6300

Nous retrouvons dans cette matrice l'assortiment optimal et l'assortiment initial (S). Pour optimiser le temps de résolution, nous trions ces sous-ensembles de produits par utilité décroissante. Cela permet de ne se focaliser prioritairement sur les sous-ensembles de produits dont l'utilité totale est au moins supérieure à celle actuelle (S). L'algorithme analyse ensuite si l'un de ces assortiments respecte les contraintes. Le cas échéant, celui qui dispose de la plus grande utilité est retenu. Dans le cas où il n'existe pas de meilleur assortiment respectant les contraintes, nous indiquons sur l'interface quelles sont les contraintes

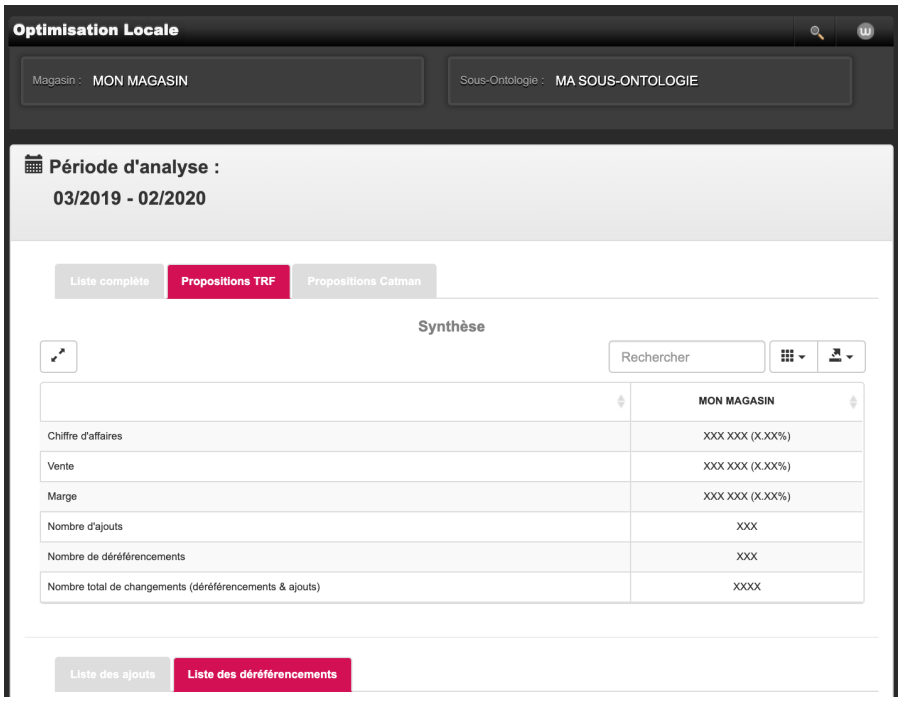


FIGURE 25 – Interface d'optimisation locale

qui ne permettent pas de recommander un meilleur sous-ensemble de produits. La Figure 25 illustre l'interface développée avec TRF Retail.

Cette interface permet aux utilisateurs d'optimiser manuellement l'assortiment local d'un magasin. Après avoir choisi un magasin et un sous-ensemble de produits, l'utilisateur accède à trois visions :

— L'ensemble des produits pouvant être utilisés N (appelé "Liste complète" dans l'interface).

— La proposition S' obtenue avec l'algorithme d'optimisation (appelé "Proposition TRF" dans l'interface).

— La simulation S'' que les utilisateurs peuvent modifier de façon dynamique à l'aide des deux premières visions (appelé "Proposition Catman" dans l'interface). C'est ce sous-ensemble de produits qui sera théoriquement mis en rayon.

Pour chacune de ces visions, un tableau de bord résume trois KPI comptables : chiffre d'affaires, vente et marge. La méthode utilisée pour estimer ces indicateurs est formalisée avec la performance/l'utilité des produits dans la section 5.3.3. En plus de ces informations, le nombre de produits devant être transposés (déréférencés et ajoutés) est indiqué pour apporter une vision détaillée de l'impact sur l'assortiment.

Nous proposons de travailler sur des sous-ensembles de produits en tenant compte des liens transversaux entre les catégories de produits. Ainsi, le raisonnement d'optimisation se base sur un très faible nombre d'articles. Pour réduire davantage la complexité ou pour correspondre davantage à la stratégie de l'enseigne, il est possible de

contraindre l'ensemble des produits à ajouter à l'ensemble des produits correspondant à l'assortiment supérieur de la catégorie. Cette nouvelle contrainte permet d'optimiser la chaîne d'approvisionnement grâce à une stabilisation du nombre de références à l'échelle de l'enseigne. Cependant, cela réduit les possibilités de personnalisation de l'assortiment des magasins.

La solution finalement proposée dépend alors de différentes notions propres à l'enseigne qui répondent à un ensemble de contraintes. Initialement, la simulation de l'utilisateur est identique à celle proposée par l'algorithme ($S' = S''$). À travers l'interface, l'utilisateur peut interagir avec les produits pour transformer l'assortiment S' en S''. L'assortiment résultant devrait être l'ensemble des produits qui sera réellement mis en place. En résumé, l'algorithme développé avec TRF Retail propose aux utilisateurs :

1. d'exprimer des contraintes pour respecter la stratégie définie par l'enseigne ;
2. d'identifier les sous-ensembles de produits pouvant être changés dans un magasin ;
3. le meilleur sous-ensemble de produits qui peut être mis en place dans le magasin ;
4. la possibilité de finaliser manuellement l'assortiment qui sera mis en place.

À l'aide de cette interface, les enseignes sont en mesure d'accorder plus d'autonomie aux magasins pour optimiser leurs assortiments tout en s'assurant de respecter la stratégie sous-jacente de l'enseigne. Pour que notre proposition d'optimisation de l'assortiment soit complète, nous devons nous intéresser maintenant à l'optimisation globale de l'assortiment.

5.3.2 *Optimisation Globale*

L'optimisation globale consiste à proposer les meilleurs produits pour l'ensemble des points de vente d'une enseigne. Dans l'industrie, ce sont les catégorie managers centrales qui définissent le profil assortiment des produits et des magasins pour représenter la stratégie d'une enseigne (*cf.* Section 2.1.3). L'objectif, contrairement à l'approche précédente, est d'améliorer les produits de chacun des profils assortiment. Autrement dit, en reprenant la Figure 24, nous cherchons ici à optimiser les produits (rouge) associés à chaque profil assortiment.

Dans cette approche il faut considérer des contraintes de plus haut niveau notamment celles liées à la détention (nombre de magasins possédant l'article). Le raisonnement proposé dans cette optimisation a pour objectif d'assurer le respect de la stratégie de l'enseigne à travers tous les magasins. Des contraintes fortes doivent être mises en avant. Par exemple, le plus petit magasin doit pouvoir recevoir

l'ensemble des produits des profils assortiment qui lui sont associés. Cette contrainte permet de simplifier drastiquement la complexité du problème. En effet, le nombre de références de chacun des profils assortiment est limité par les plus petits magasins. Nous pouvons donc définir la matrice suivante pour chacune des catégories :

Catégorie	Profil Assortiment	Nombre de références
Soda	1	10
Soda	2	25
Soda	3	50
Soda	...	...
Soda	Z	X

Un catégorie manager centrale voudra proposer les produits avec la plus grande utilité dans le plus petit profil assortiment afin que tous les magasins de l'enseigne puissent les distribuer. Cependant, il doit aussi respecter des contraintes comme, par exemple, posséder un certain pourcentage d'articles de la marque du distributeur dans tous les magasins. Pour les identifier, nous exploitons ces contraintes exprimées par les utilisateurs à l'aide de l'interface proposée dans la section 5.2. Deux ensembles d'articles peuvent de nouveau être discernés pour chacune des catégories :

— N l'ensemble des produits de la catégorie, $N = \{1, 2, \ldots, n\}$,

— S_k le sous-ensemble de produits de la catégorie associée au profil assortiment k, $S_k \subseteq N$ et $S_k \subset S_{k+1}$

Deux approches peuvent alors être envisagées :

— Remplacer les produits avec le moins d'utilité dans S_k par ceux avec le plus d'utilité dans S_{k+1} dans la mesure où ces produits ont une plus grande utilité en respectant la cardinalité de chacun des profils assortiment. Ensuite, il suffit de reproduire l'opération pour tous les profils assortiment.

— Proposer les produits avec le plus d'utilité dans le réseau de magasins pour S_k et réitérer l'opération pour S_{k+1} en gardant à l'esprit que $S_k \subset S_{k+1}$.

La seconde approche est celle qui permettrait, en principe, d'obtenir le meilleur assortiment gigogne. Néanmoins, les coûts logistiques associés ne sont pas du tout considérés car les produits du plus petit profil assortiment sont déjà proposés dans l'ensemble du réseau de magasins tandis que ceux inclus dans le profil assortiment le plus élevé ($\mathcal{K}$) ne sont inclus que dans quelques magasins. Autrement dit, le fait de changer un produit associé au profil assortiment S_{k+x} au profil assortiment S_k implique qu'il faudra modifier l'assortiment de tous les magasins associés aux profils assortiment inférieurs à $k + x$. Pour conserver une approche pragmatique industrialisable, nous avons

préféré la première approche qui, en réalité, minimise le nombre de changements de produits dans les différents points de vente. Si l'on souhaite changer le profil assortiment $k-1$ d'un produit vers le profil assortiment k, il suffira simplement d'ajouter ce produit dans les magasins associés au profil assortiment k. En effet, ce produit est déjà présent dans tous les magasins associés au profil assortiment $k-1$ ou inférieur. Il suffira donc de l'introduire dans l'assortiment des magasins associés aux profil assortiment k. Comme pour notre approche locale, nous considérons que les experts sont essentiels. Nous proposons donc l'interface présentée Figure 26 qui permet à un catégorie manager centrale d'optimiser un profil assortiment grâce aux sous-ensembles de produits identifiables à l'aide des structures de connaissances et des liens portés par le modèle de données de TRF. La vision principale de chacun des sous-ensembles de produits est un profil assortiment pour une catégorie donnée.

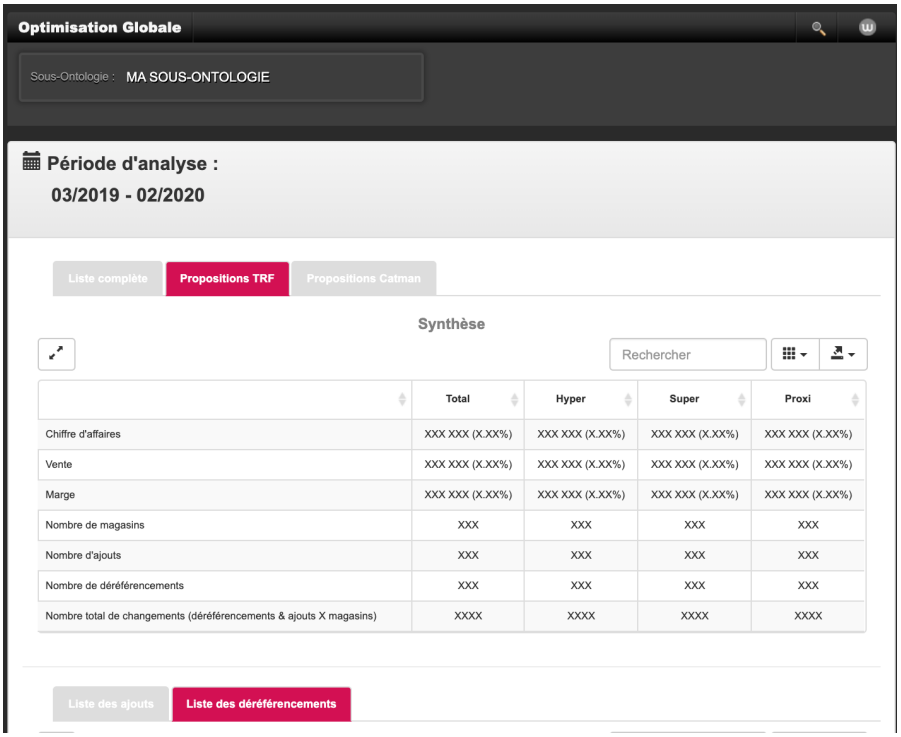


FIGURE 26 – Interface d'optimisation globale

Nous retrouvons dans cette interface des notions très similaires à celle utilisée pour l'optimisation locale. La différence notable provient du tableau de bord dans lequel les KPI sont répartis en fonction des clusters de magasins définis par l'enseigne. Dans la Figure 26, le cluster de format utilisé correspond à "Hyper" pour les hypermarchés, "Super" pour les supermarchés et "Proxy" pour les magasins de proximité ... Nous retrouvons aussi le nombre de magasins impactés ("Nombre de magasins") et le nombre de produits à échanger ("Nombre d'ajouts", "Nombre de déréférencements", "Nombre de changements (déréféren-

cements et ajouts X magasins)"). Afin d'apporter une vision globale de ces modifications, une vision consolidée est proposée dans la colonne "Total".

Grâce à ces deux interfaces, les grands distributeurs peuvent améliorer l'assortiment d'un magasin (local) ou l'assortiment gigogne distribué à travers le réseau complet de magasins (global). Les données utilisées sont mises à jour en fonction de la périodicité définie en amont par l'enseigne (*e.g.* mensuel, hebdomadaire ...). L'aspect temporel de ces recommandations permet de considérer le cycle de vie des produits et l'évolution des tendances. Ces interfaces servent finalement de système de recommandation pour les experts permettant ainsi de pouvoir améliorer petit à petit les assortiments tout en garantissant de la cohérence avec les attentes de l'enseigne. La section suivante formalise les méthodes de calcul d'utilité des produits.

5.3.3 *Performance des produits*

La performance ou l'utilité d'un produit, autrement dit, la métrique qui permet de définir leur rang est une notion très subjective. Deux personnes d'une même enseigne peuvent considérer des KPI très différents pour estimer la performance des produits. Dans l'industrie, les notions les plus courantes sont comptables avec principalement le chiffre d'affaires, la marge et les ventes. Certaines enseignes utilisent des notions plus complexes comme le "taux de rachat" ou encore les parts de marché. Dans la littérature, trois différents type de données peuvent être utilisées pour définir la performance d'un produit :

1. échantillonage de données ;
2. transaction de ventes ;
3. récapitulatif des ventes.

Le choix du type de données dépend généralement de la problématique étudiée.

Le premier correspond généralement aux modèles qui se basent sur l'approche Multinomial Logit (MNL) [53] utilisée surtout dans des articles marketing pour analyser l'impact des variables "mix-marketing" sur la demande [35, 84]. Les données échantillonnées permettent d'analyser le comportement d'achats au fil du temps. Pour garantir une certaine cohérence, les données liées aux cartes de fidélité sont utilisées. C'est aujourd'hui le seul moyen de garantir l'identification d'un ménage. Cependant, toutes les enseignes ne disposent pas de carte de fidélité et tous les clients ne s'en servent pas systématiquement.

Le modèle van Ryzin et Mahajan [104] correspond à la seconde source de données pour estimer de la performance des produits. Les transactions de vente peuvent être considérées comme des tickets de caisse. Cependant, dans ce modèle, les auteurs ne s'intéressent qu'à un instant de la journée [83]. Les données sont par conséquent

incomplètes dans le sens où seules les arrivées de clients ayant effectué un achat sont enregistrées. Pour corriger ce problème, l'algorithme Expectation-Maximization (EM) est généralement la méthode la plus utilisée en se basant sur la vraisemblance complète dans un algorithme itératif [39, 151].

Enfin, le récapitulatif des ventes permet de consolider l'ensemble des KPI sur une période pour un magasin et un produit. Cette approche permet de travailler sur un volume de données considérablement réduit et demeure le modèle le plus utilisé en industrie. D'autres travaux ont proposé des analyses en se basant sur ce type de données [83]. TRF Retail dispose a minima de ces données consolidées pour chacun des grands distributeurs avec lesquels il collabore. Évidemment, certaines enseignes peuvent les enrichir avec, par exemple, les tickets de caisse, les cartes de fidélité, les données concurrentielles ...

Pour proposer un système inter-opérant, nous définissons l'utilité d'un produit à partir des récapitulatifs des ventes. Les données supplémentaires permettent d'accroître la précision de l'utilité et par extension de perfectionner les recommandations.

La performance d'un produit dépend de l'objectif. Supposons simplement qu'une enseigne souhaite maximiser son chiffre d'affaires. Dans ce cas-là, l'utilité des produits sera simplement leur chiffre d'affaires : u_{mj} est le chiffre d'affaires du produit j dans le magasin m à optimiser. Si le produit j n'existe pas dans le magasin ($u_{mj} = 0$), alors nous devons estimer son utilité espérée u'_{mj} grâce au chiffre d'affaires qu'il obtient dans les autres points de vente de l'enseigne. Pour proposer une estimation réaliste de u'_{mj}, nous calculons tout d'abord la participation du produit dans sa catégorie, c'est à dire que nous identifions quelle est la part de chiffre d'affaires réalisée par le produit j par rapport au chiffre d'affaires total de sa catégorie (N produits) pour un magasin donné, et cela pour chacun des magasins possédant le produit. Nous recommandons d'utiliser les magasins similaires (*cf.* Chapitre 4) afin d'apporter des recommandations liées à la spécificité du magasin. Nous pouvons formuler l'utilité de ce produit, pour un magasin m, de la façon suivante :

$$u'_{mj} = \sum_{i=1, i\neq j}^{N} u_{mi} \times \left(\left[\sum_{m'=1, m'\neq m}^{M} \frac{u_{m'j}}{\sum_{i=1}^{N} u_{m'i}} \right] / M \right) - \sum_{i=1, i\neq j}^{N} u_{mi}$$

où, M représente l'ensemble des points de vente de l'enseigne qui possède le produit j et N le nombre de produits de la catégorie, u_{mi} correspond à l'utilité du produit i pour le magasin m. Grâce aux similarités entre les magasins, définies dans le chapitre 3, les recommandations peuvent considérer des éléments sémantiques comme les habitudes d'achats des consommateurs pour améliorer la précision de l'utilité potentielle u'_{mj} du produit j en intégrant un cœfficient de similarité (les magasins les plus semblables ont un poids plus conséquent).

Maintenant dans le cas de l'optimisation globale, nous conservons cette mécanique pour estimer l'utilité u'_{mj} d'un produit, néanmoins, nous la pondérons à l'aide d'un cœfficient. L'idée est de rectifier l'utilité précédemment calculée à l'aide d'un coefficient lors d'un retrait/ajout du produit en fonction de la distribution du produit dans le réseau de magasins. Ces cœfficients sont obtenus au regard :

— du nombre de magasins dans lesquels l'article sera ajouté (*i.e.* ceux associés au profil assortiment S_{k-1}), noté m^+ ;

— du nombre de magasins dans lesquels l'article sera retiré (*i.e.* ceux associés au profil assortiment S_k), noté m^- ;

— du nombre de magasins total, noté M.

Pour considérer l'impact du retrait d'un produit de certains magasins, nous formalisons le cœfficient pondérateur de u'_{mj} (> 1) suivant :

$$\frac{M + m^-}{M}$$

Nous considérons que plus un produit est présent dans le réseau de magasins, plus sont utilité doit être importante, car il est sans doute partie prenante de l'identité de l'enseigne. Son retrait doit impacté par conséquent davantage l'utilité globale de l'enseigne. Dans le même état d'esprit, pour considérer l'impact d'ajout d'un produit dans certains magasins, nous formalisons le cœfficient pondérateur de u'_{mj} (< 1) suivant :

$$\frac{M - m^+}{M}$$

Ainsi, plus le nombre de magasins impactés par l'ajout de l'article est élevé, plus les coûts d'approvisionnement seront importants. Ces deux cœfficients permettent d'améliorer ou de réduire l'utilité des produits en fonction des actions pouvant être mises en place. Cela apporte une utilité plus réaliste concernant l'impact de telles actions pour l'enseigne. En plus de cela, il devient possible d'identifier des changements inexécutables à cause d'une trop forte densité de magasins associés à certains profils assortiment.

L'utilité d'un article est illustrée ici par le chiffre d'affaires, néanmoins, l'algorithme développé avec les experts de TRF Retail agrège de nombreux KPI et peut considérer des notions beaucoup plus avancées en fonction des données à disposition (*cf.* 5.4).

Les contraintes préférentielles viennent évidemment affiner cette utilité. Par exemple, la contrainte :
"Maximiser la marge"
$\ll OBJ \gg =$ "Maximiser" $\rightarrow$ Nature = "Préférence"
$\ll KPI \gg =$"marge"
$\ll PER \gg =$"ALL"
$\ll VAL \gg = \emptyset$;

permet de valoriser ce KPI au détriment des autres. De la même façon, différents cœfficients peuvent être utilisés en fonction de la stratégie des enseignes. Une enseigne qui souhaite privilégier les produits d'origine BIO verra leurs utilités être significativement augmentées à l'aide d'un cœfficient valorisant (> 1). À l'inverse, si elle souhaite par exemple réduire les produits d'un certain fournisseur, alors un cœfficient dévalorisant (> 0 et < 1) réduira l'utilité des produits liés au fournisseur. Dans ces cas là, les coefficients pondérateurs associés aux nombres de magasins impactés sont supprimés pour valoriser davantage les préférences de l'enseigne.

TRF Retail a en plus défini une métrique permettant d'évaluer la performance des produits à l'aide d'index allant de 0 à 20. Comme pour un étudiant ayant différentes notes pour différentes matières, les produits ont différents index pour différentes notions clés de la grande distribution. Nous retrouvons :

— l'index assortiment ;

— l'index vente ;

— l'index marge ;

— l'index linéaire ;

— l'index stock.

À l'aide de ces différents indicateurs un index général, appelé Index TRF, est calculé et correspond à la "moyenne générale" de l'article, il est le résultat d'une somme pondérée des critères précédents. Comparer les ventes de produits `High-tech` avec celles de produits `Alimentaire` n'a pas réellement de sens, c'est pourquoi ces index sont calculés par magasin et par catégorie de produits. Le niveau de catégorie de produits est défini conjointement avec l'enseigne et TRF Retail. Chaque produit est ainsi évalué en référence à sa famille. Cet index permet de rendre comparables les produits `High-tech` et les produits `Alimentaire` à l'aide d'un indicateur facilement compréhensible (note de 0 à 20).

Comme certains produits peuvent avoir des performances très différentes d'un magasin à un autre, ces index sont aussi calculés au niveau du cluster format (*e.g.* supermarché, hypermarché ...). Cela permet d'avoir une vision beaucoup plus générale de la performance des produits. Ces différents index sont utilisés pour affiner l'utilité et les recommandations.

L'utilité d'un produit est une agrégation de KPI pondérés par des notions métiers et des contraintes préférentielles exprimées par une enseigne. Plus les données sont riches, plus l'utilité est précise et plus les recommandations seront cohérentes et qualitatives.

Le nombre de produits gérés aujourd'hui par les grands distributeurs est si conséquent qu'il finit par restreindre la capacité d'optimisation, ne serait-ce que d'un point de vue combinatoire. Pour les aider encore davantage, nous avons travaillé sur une approche permettant

de rationaliser le nombre de produits. Cette approche est présentée dans la section suivante.

5.4 RATIONALISATION DE L'ASSORTIMENT

Cette section formalise une méthode complémentaire au processus d'amélioration continue d'optimisation de l'assortiment. Actuellement, l'optimisation globale permet d'améliorer l'assortiment gigogne et l'optimisation locale permet de personnaliser et de perfectionner l'assortiment d'un magasin. Cependant, les contraintes apportées par l'assortiment gigogne réduisent considérablement la liberté des magasins dans l'optimisation locale. De plus, comme nous l'avons souligné dans le chapitre 2, l'ajout permanent de nouveaux produits réduit petit à petit les "degrés de libertés" des magasins. Aujourd'hui, les travaux de recherche et les industriels s'accordent pour souligner l'importance d'une rationalisation drastique de l'assortiment. La démesure ayant été atteinte, aujourd'hui les enseignes ont plus à gagner en réduisant la variété de produits de leurs assortiments. Ils ont élargi leurs assortiments à tel point que des études démontrent que réduire l'assortiment ne diminue pas les ventes [20, 23, 45].

Avec TRF Retail, nous avons développé un algorithme qui permet d'aider à rationaliser l'assortiment. Autrement dit, l'objectif est de réduire le nombre de produits distribués dans les points de vente.

5.4.1 *Méthode de rationalisation*

Pour rester cohérent avec notre processus d'amélioration continue introduit précédemment, nous avons commencé par définir, avec TRF Retail, la structure de connaissances optimale nécessaire à une rationalisation pragmatique et efficace. Cette structure est illustrée Figure 27 dans laquelle nous retrouvons :

- la structure marchandise (*cf.* Section 2.1.2) ;
- les rôles des catégories (*cf.* Section 2.3) ;
- le cluster de magasins "Format" (*e.g.* supermarchés, hypermarchés ...) ;
- le cluster de magasins "Région" (*e.g.* Nord, Sud, Est, Ouest) ;
- l'assortiment gigogne (*cf.* Section 2.1.3) ;
- l'unité de besoin à laquelle un ou plusieurs produits répondent (*cf.* Section 2.1.2).

L'objectif de la rationalisation, contrairement à l'optimisation globale (*cf.* Section 5.3.2) ou locale(*cf.* Section 5.3.1), est d'identifier les produits ayant la plus faible utilité afin de les retirer de l'assortiment. Ceci doit être fait en s'assurant :

- de respecter au mieux la stratégie véhiculée par l'enseigne ;

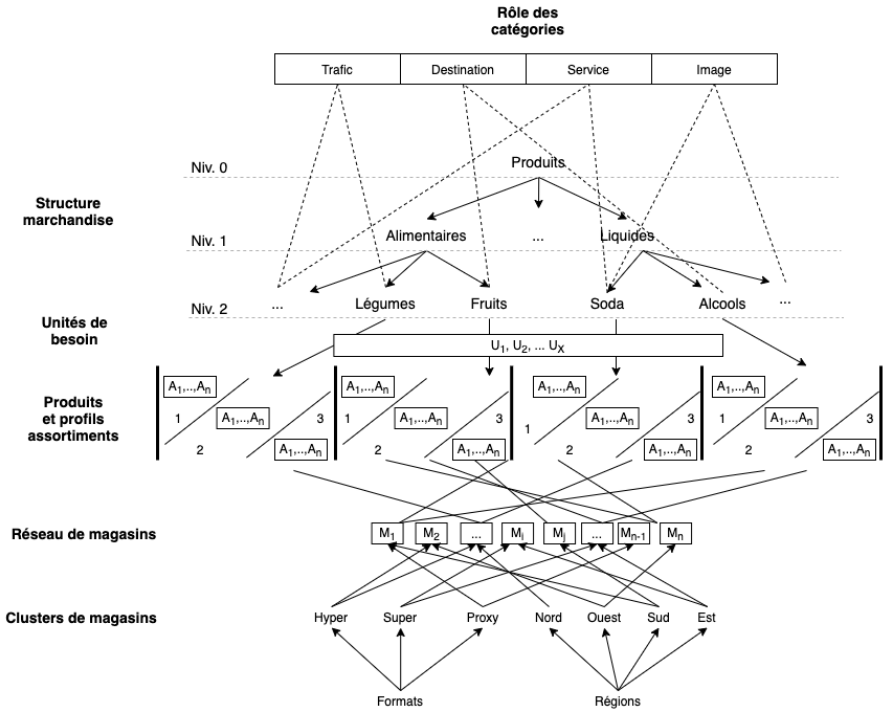


FIGURE 27 – Schéma de la structure de connaissances servant à la rationalisation de l'assortiment

— de minimiser les pertes liées aux retraits des produits ;

— de conserver au moins un produit pour chacune des unités de besoin.

Les gains liés à la rationalisation de l'assortiment sont principalement associés au stockage des produits ainsi qu'à la gestion de l'assortiment. De plus, comme nous l'avons introduit, les produits de substitution devraient garantir une compensation des ventes [20, 23, 45].

Ces différentes contraintes ont été traduites en une heuristique pour définir une fonction objectif valorisant une diminution de l'utilité de l'assortiment initial. Intuitivement, pour rationaliser un assortiment, il suffit de retirer les produits qui ne se vendent jamais. Néanmoins, très peu sont dans ce cas-là lorsque l'on considère l'ensemble du réseau de magasins avec les spécificités locales mais aussi, à cause d'éventuelles ruptures dans certains points de vente (*cf.* la substitution, Section 2.2). Pour obtenir les informations essentielles sur les produits, nous avons à disposition les données suivantes :

1. récapitulatif des ventes périodiques ;
2. tickets de caisse ;
3. récapitulatif des ventes concurrentielles.

La première source de données permet d'analyser la performance intrinsèque des produits dans les points de vente de l'enseigne. Les tickets de caisse permettent d'ajouter l'intérêt que peuvent avoir les consommateurs pour les produits notamment à travers la notion de

taux de rachats (repurchase rate). Enfin, les données concurrentielles déterminent la performance des produits vis-à-vis des concurrents. Cette dernière source peut amener également des éléments clés comme, par exemple, les parts de marché. Elle permet, en plus, de mettre en lumière d'éventuels problèmes propres à l'enseigne (*e.g.* prix de vente trop élevé, trop de produits de substitution, etc.).

À partir de cette structure, des contraintes et des différentes sources de données, nous avons développé un système de recommandation permettant de proposer les produits à retirer de l'assortiment.

5.4.2 *Interface de rationalisation*

Avec TRF Retail, un système dynamique de rationalisation a été mis en place. L'interface de rationalisation a été développée conjointement avec TRF Retail et une enseigne de la grande distribution alimentaire.

Il propose, à partir d'une catégorie sélectionnée par l'utilisateur (permettant d'accéder aux sous-ensembles de produits) et du pourcentage de produits à déréférencer ("Taux de rationalisation"), de choisir les articles à retirer. L'interface finale proposée aux utilisateurs comporte trois onglets : "Matrice des rôles catégories", "Scorecard Catégorie" et "Simulation". Le fait de rationaliser un assortiment nécessite d'avoir un certain nombre d'informations à disposition.

Nous retrouvons tout d'abord la matrice des rôles des catégories illustrée Figure 28.

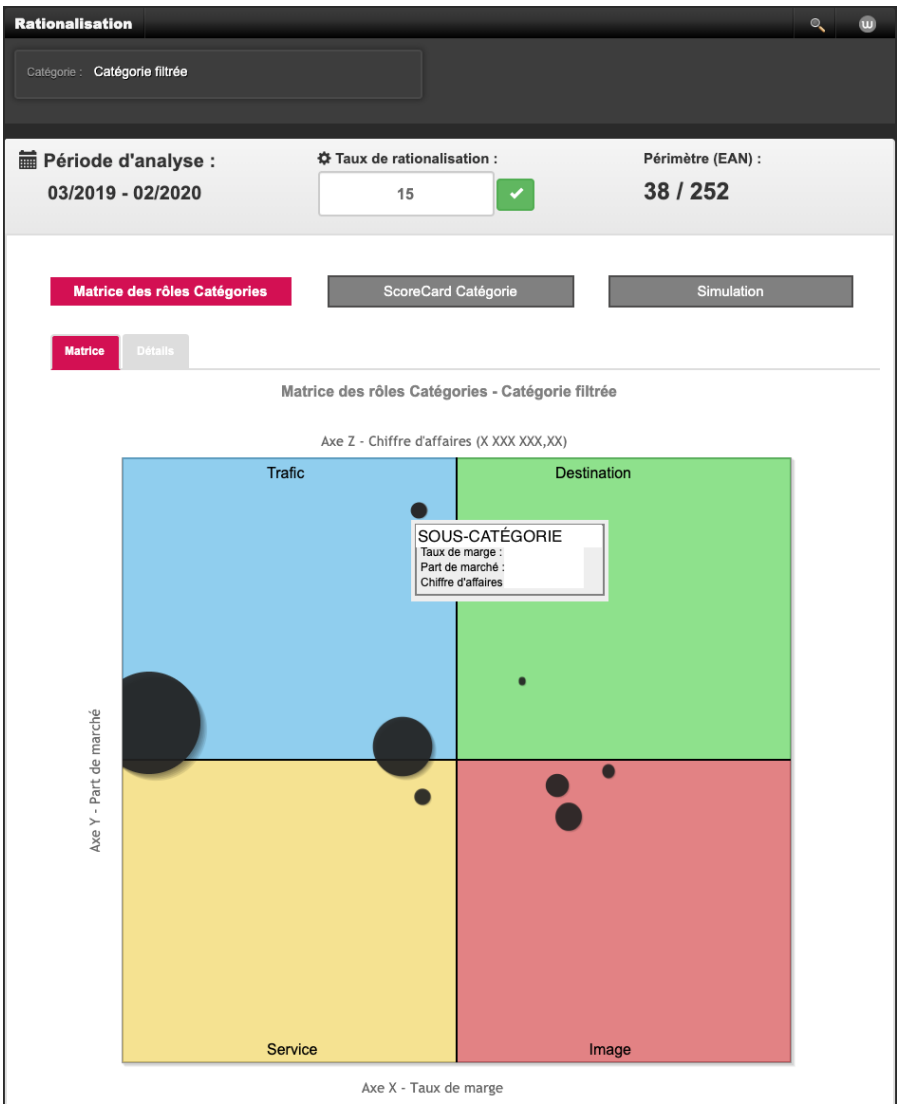


FIGURE 28 – Interface rationalisation - Matrice des rôles Catégories

Cette matrice situe chacune des sous-catégories (catégories filles) de la catégorie sélectionnée par rapport aux deux axes que sont les parts de marché et le taux de marge. En fonction du groupe auquel sont associées les sous-catégories, l'utilisateur aura tendance à privilégier le retrait de produits. Par exemple, si les parts de marché et le taux de marge sont élevés, il n'y a aucune raison de retirer des produits de cette sous-catégorie. À l'inverse, si les parts de marchés et/ou les taux de marge sont faibles alors la décision de retirer des produits sera plus évidente. Ainsi, plus une sous-catégorie est située en bas à gauche de la matrice, plus les produits inclus sont susceptibles d'être retiré du magasin. A l'inverse, plus une catégorie est située en haut à gauche de

la matrice, plus les produits inclus dans celle-ci sont indispensables au bon fonctionnement du magasin.

Le second onglet "Scorecard Catégorie" récapitule les KPI nécessaires à l'analyse des sous-catégories. Nous y retrouvons l'ensemble des indicateurs comptables déjà introduits (*e.g.* le chiffre d'affaires, les ventes ...) en plus de notions plus stratégiques.

Ce second onglet donne aux utilisateurs les connaissances nécessaires à la prise de décision pour la rationalisation d'un assortiment.

Enfin, dans l'onglet "Simulation" présenté en Figure 29, nous retrouvons une interface similaire à celles utilisées pour l'optimisation locale et globale de l'assortiment. La différence majeure provient du

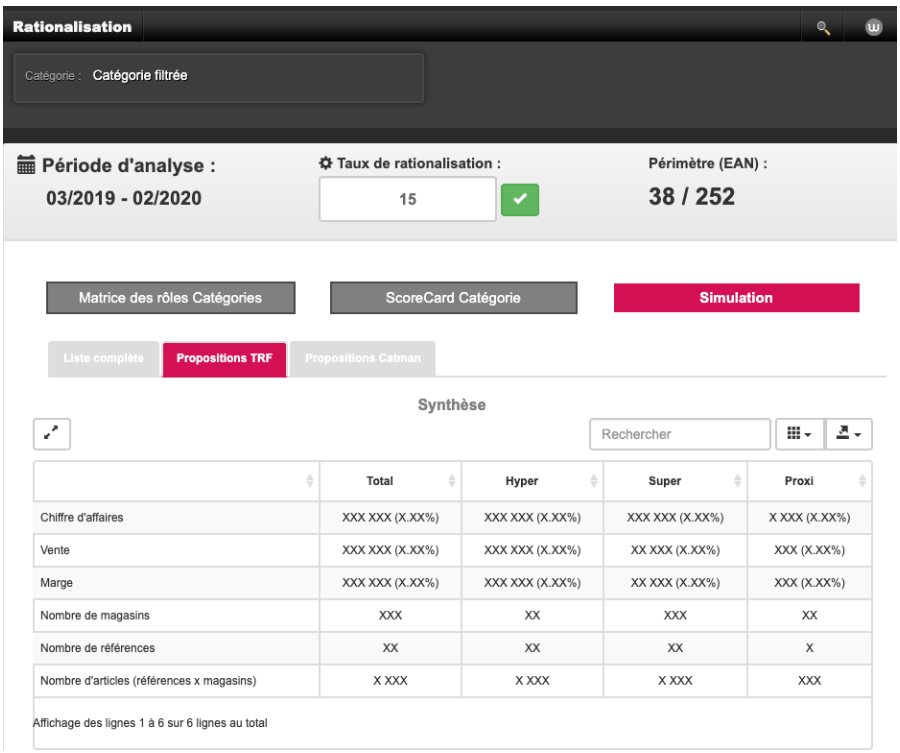


FIGURE 29 – Interface rationalisation - Recommandations

fait qu'ici les KPI analysés correspondent à la perte associée au déréférencement des produits. Comme nous l'avons dit précédemment, très peu de produits ne se vendent pas du tout dans l'ensemble des points de vente. La particularité de l'algorithme de rationalisation est qu'il identifiera toujours un sous-ensemble de produits à retirer avec une perte factuelle obtenue à partir des récapitulatifs de ventes. C'est-à-dire que si l'on recommande de retirer un produit, alors ses KPI factuels (*e.g.* chiffre d'affaires, ventes, marges ...) sont considérés comme de la perte pour l'enseigne. Néanmoins, cette perte est seulement théorique car la contrainte imposant au moins un produit pour une unité de besoin assure un report des ventes vers un autre produit de substitution. De plus, les produits que l'on recommande de retirer

sont ceux ayant l'utilité la plus basse par rapport aux préférences de l'enseigne.

Aujourd'hui, cette interface est dédiée uniquement à une enseigne mais nous travaillons sur sa généralisation. La principale difficulté est liée à l'identification des unités de besoins qui est indispensable pour optimiser le report des ventes et, par extension, minimiser les pertes.

5.5 CONCLUSION

Dans cette section, nous illustrons un processus d'amélioration continue permettant d'optimiser l'assortiment développé avec TRF Retail. Les structures de connaissances formalisées dans le chapitre 3 permettent l'expression de contraintes introduisant la stratégie des enseignes dans l'optimisation de l'assortiment. Ces contraintes apportent des sous-ensembles de produits restreints, nous affranchissant des problèmes de complexité liés aux volumes de données [51]. De plus, les liens transversaux introduits entre les éléments propres à chaque enseigne fournissent une structure de connaissances sémantiques jusqu'alors inexploitée.

Nous proposons une méthode Agile pour les raisons suivantes :

— l'expertise humaine est indispensable qu'elle soit apportée par un responsable magasin ou centrale ;

— les assortiments sont, comme nous l'avons indiqué, dynamique et l'évolution des tendances de vente est permanente nécessitant de nombreuses itérations.

Les interfaces de recommandation développées permettent de prendre en considération ces deux aspects et apportent une solution industrialisée proposant l'amélioration de l'assortiment. L'adaptabilité de notre méthode repose sur l'identification de "scénarios" complémentaires. Dans ce chapitre, nous avons distingué l'optimisation locale (d'un magasin) et globale (de tous les magasins) faisant émerger un cycle vertueux. Ces deux scénarios se basent sur les structures de connaissances propres à chaque enseigne. Cela permet de souligner l'interopérabilité de notre méthode. Non seulement, par la capacité à s'adapter à différentes sources de données mais, aussi grâce au calcul de l'utilité de chaque produit qui est lui aussi spécifique à chaque enseigne.

L'intérêt de notre proposition, en plus d'être une méthode applicable/appliquée, repose sur sa capacité à évoluer. Ceci est très largement souligné avec la rationalisation d'assortiment qui s'intègre tout à fait dans le cycle vertueux d'optimisation d'assortiment.

Rationaliser l'assortiment permet d'offrir de nouvelles possibilités aux algorithmes d'optimisation locale et globale. Reproduire fréquemment ces scénarios permet d'améliorer petit à petit l'assortiment s'inscrivant parfaitement dans un processus d'amélioration continue.

N'importe quel autre scénario peut être intégré à partir du moment où il répond à la procédure illustrée dans la Figure suivante.

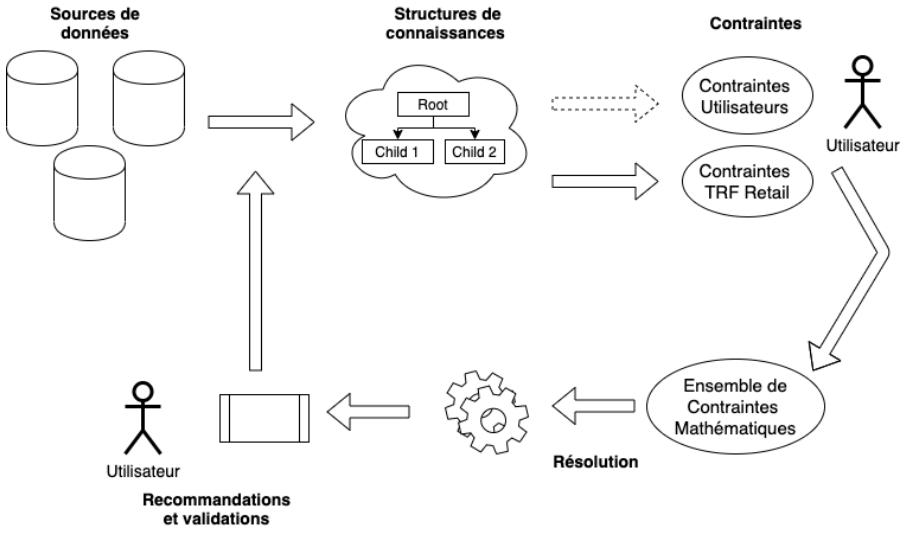


FIGURE 30 – Méthode Agile, procédure

Dans cette procédure, nous retrouvons les utilisateurs, les structures de connaissances, les contraintes et les recommandations. Les scénarios introduits dans ce chapitre n'utilisent les structures de connaissances que pour exprimer des contraintes et apporter des informations complémentaires aux utilisateurs. Cependant, comme nous l'avons souligné dans le chapitre précédent, elles peuvent servir à apporter des nouvelles informations indispensables aux grands distributeurs.

6

CONCLUSION ET PERSPECTIVES

6.1 CONCLUSION

Dans cette thèse, nous nous sommes intéressés à la problématique de l'assortiment pour les grands distributeurs. Cette problématique est liée à un éco-système complexe dont les éléments sont inter-connectés. Nous avons défini des structures de connaissances permettant d'établir ces liens indispensables à l'identification de l'assortiment optimal.

Notre **première contribution** repose sur la formalisation de l'assortiment optimal sous la forme d'un problème d'optimisation combinatoire. Nous exploitons les structures de connaissances de la grande distribution dans la formalisation et la résolution de ce problème. Nous avons également illustré nos approches sémantiques utilisant la structure marchandise qui sert de référentiel à chacune des enseignes. Une analyse portant sur les magasins permet d'identifier ceux qui présentent des ventes similaires sur la structure marchandise. Cette vision améliore la précision des recommandations. Pour aller plus loin, nous avons proposé d'analyser les habitudes d'achats des clients grâce aux cartes de fidélité. Cette analyse apporte de nouvelles connaissances, notamment les habitudes d'achats des consommateurs. Les clients sont la cible finale de l'assortiment et une meilleure compréhension de leur comportement permettra une identification plus fine de l'assortiment optimal. Par extension, cela permettra de les fidéliser et donc de garantir une performance durable pour chacun des magasins.

A partir de cette première contribution, nous proposons une **deuxième contribution** qui permet, à l'aide de contraintes, l'expression d'une stratégie sous-jacente propre à chaque grand distributeur. Pour exprimer ces contraintes, nous proposons une syntaxe intuitive qui permet aux enseignes d'introduire un cadre représentatif d'éléments pouvant provenir de la chaîne d'approvisionnement (*e.g.* contraintes de fournisseurs) ou encore des éléments représentatifs de la stratégie véhiculée (*e.g.* valoriser les produits "Bio"). Ces contraintes ont la particularité d'être exprimées sur un référentiel qui leur est propre grâce au format de données de TRF Retail. De plus, nous avons souligné la possibilité d'identifier des sous-ensembles de produits indépendants, ce qui permet de réduire drastiquement la complexité du problème d'optimisation.

Enfin, notre **troisième contribution** repose sur la méthode Agile d'optimisation de l'assortiment. A partir des contraintes, différents scénarios d'optimisation de l'assortiment sont proposés. Dans ce manuscrit, nous avons explicité quatre scénarios avec des objectifs différents. Cependant, comme nous l'avons souligné, ces approches sont complé-

mentaires et s'inscrivent dans un processus d'amélioration continue de l'assortiment. Nous avons d'abord formalisé le problème de changement des profils assortiment de chacun des magasins comme un problème d'optimisation. Ce scénario ne considère pas directement les produits mais des ensembles de produits définis pas les enseignes. Actuellement, nous développons l'interface qui pourra être utilisée par les utilisateurs pour simuler et appliquer les changements sur leurs assortiments. Ensuite, nous avons présenté trois scénarios qui se focalisent uniquement sur le niveau le plus fin de l'assortiment : les produits. Le premier scénario propose pour chacun des magasins des changements permettant de venir combler l'assortiment. Ce scénario est appelé l'**optimisation locale**. Le second scénario propose de challenger les produits inclus dans les différents profils assortiment pour améliorer la performance de l'ensemble du réseau de magasins. Ce scénario est appelé l'**optimisation globale**. Enfin, comme nous l'avons souligné, la gestion d'un nombre gigantesque de produits est aujourd'hui un véritable problème pour les quelques personnes dont la mission est la gestion de l'assortiment. C'est pourquoi notre troisième scénario propose de rationaliser l'assortiment en considérant des gains de stock et des reports de ventes grâce aux produits substituables. Pour chacun des trois derniers scénarios, une interface d'aide à la décision est proposée.

Notons que l'utilisation d'une méthode Agile permet d'apporter de nouvelles recommandations au fur et à mesure que les enseignes mettent leurs données à disposition. Ainsi, nous travaillons sur un assortiment dynamique qui évolue au même rythme que les nouvelles tendances.

6.2 PERSPECTIVES

Les perspectives d'amélioration et d'extension des travaux proposés sont nombreuses. Dans ce manuscrit, nous apportons une réponse pragmatique à la vaste problématique de l'assortiment en adressant chacun des problèmes majeurs aujourd'hui identifiés.

Comme nous l'avons souligné, l'expression des contraintes reste complexe. Nous aimerions permettre aux enseignes d'exprimer en langage naturel leurs souhaits et la stratégie qu'ils veulent mettre en place. Ainsi, les simulations pourraient être plus adaptées à leurs attentes. Néanmoins, la transformation mathématique de ces contraintes reste un problème épineux. Pour atteindre cet objectif, il est indispensable de ramener ces phrases en langage naturel aux différents éléments de l'arbre de décision permettant de définir formellement une contrainte.

Nous travaillons sur l'intégration de contraintes pouvant découler des différentes analyses sémantiques proposées. En effet, à partir de l'analyse sémantique des consommateurs, de nouvelles orientations peuvent émerger pour définir l'assortiment idéal. Idéalement, nous

aimerions être en mesure d'intégrer ces contraintes caractéristiques d'habitudes d'achats afin de préserver les produits qui fidélisent aujourd'hui les clients. De plus, nous pensons que les enseignes et chacun de leurs magasins devraient se focaliser sur les catégories peu sollicitées par la clientèle. Ce sont ces catégories de produits qui devraient, selon nous, apporter une amélioration significative de la performance car elles ne semblent pas répondre aujourd'hui à la demande des clients.

Grâce à l'identification de magasins dont le comportement d'achats des clients est similaire, nous pouvons dès à présent affiner nos simulations. Cependant, nous pouvons aller plus loin en exploitant, par exemple, les analyses clients ou encore en croisant différentes approches sémantiques. De plus, un apprentissage pourra être mis en place pour identifier des facteurs qui influent sur le calcul de l'utilité. Ainsi, des situations exceptionnelles peuvent venir affiner nos simulations, par exemple, la coupe du monde de football qui entraîne une augmentation significative des ventes de bière et de chips. L'identification et la formalisation d'éléments externes impactant les ventes permettraient d'améliorer la précision des simulations.

Les trois scénarios qui proposent d'améliorer l'assortiment en changeant les produits exploitent des sous-ensembles indépendants. Actuellement, nous proposons une méthode pragmatique, basée sur les concepts du modèle Multinomial Logit (MNL), qui consiste à sélectionner un à un les produits par ordre d'utilité décroissant. Cependant, rien n'empêche de mettre en place un système d'élicitation des produits plus complexe comme par exemple le modèle Multinomial Logit Imbriqué (*Nested Multinomial Logit*), ou encore, le modèle de demandes exogènes.

Les grands distributeurs ont réussi à créer des avantages qui, en plus d'être écologiques, leur assurent une certaine stabilité sur le marché. En proposant des produits de marques distributeurs, manufacturés à proximité associés souvent aux appellations Bio, Régional ... les enseignes s'assurent une production locale qui, en plus de répondre à la demande écologique croissante sur le marché, leur offre par la même occasion une image éco-responsable. Ces avantages durables sont encore difficiles à valoriser. Néanmoins, leur plus-value est plus que notable, car en satisfaisant une clientèle de plus en plus impliquée, ils ont l'opportunité d'optimiser localement leur chaîne de distribution de façon à réduire leur Co2, améliorant ainsi leur image ... Ce cycle vertueux souligne parfaitement l'intérêt pour les grands distributeurs d'améliorer le critère de durabilité.

Enfin, l'une des perspectives majeures de notre problématique consiste à évaluer nos recommandations. La meilleure évaluation qui puisse être faite repose sur une étude empirique des résultats.

Autrement dit, cela passe par la mise en place (ou le retrait) effectif des produits en magasin. De telles actions impliquent des coûts conséquents pour les grands distributeurs. Il semble inévitable d'atteindre une période suffisamment longue pour mesurer les résultats obtenus. Les éléments environnementaux, comme par exemple la météo, impactent significativement la performance des produits. Pour améliorer la précision de l'évaluation des recommandations, nous considérons qu'une période de minimum trois mois est requise pour des produits dont les ventes ne sont pas saisonnières.

Pour pallier ce problème de temps et améliorer nos recommandations efficacement, nous travaillons sur leur évolution. Dans cette optique, nous identifions les produits que nous recommandons d'ajouter ou de retirer. Dans le cas où ces produits sont récurrents (souvent recommandés), nous considérons que leur plus value est certaine. Il nous parait donc indispensable d'accorder une importance toute particulière à ces produits qui selon nous, sont ceux qui auront l'impact le plus positif sur l'assortiment des grands distributeurs.

7

APPENDIX : EXEMPLE ILLUSTRATIF DE L'ASSORTIMENT GIGOGNE

Nous allons, à l'aide d'un exemple illustratif, présenter simplement le fonctionnement de l'assortiment gigogne. La mise en place de l'assortiment gigogne nécessite deux étapes :

1. définir un profil assortiment pour chacun des produits
2. définir un profil assortiment pour chacun des magasins

Ces deux étapes sont faites par les catégories managers centrale qui l'appliquent à la catégorie dont ils sont responsables. Pour illustrer ces étapes, nous utiliserons la structure marchandise et les produits présentés dans la figure 31. Dans cette exemple nous considérons

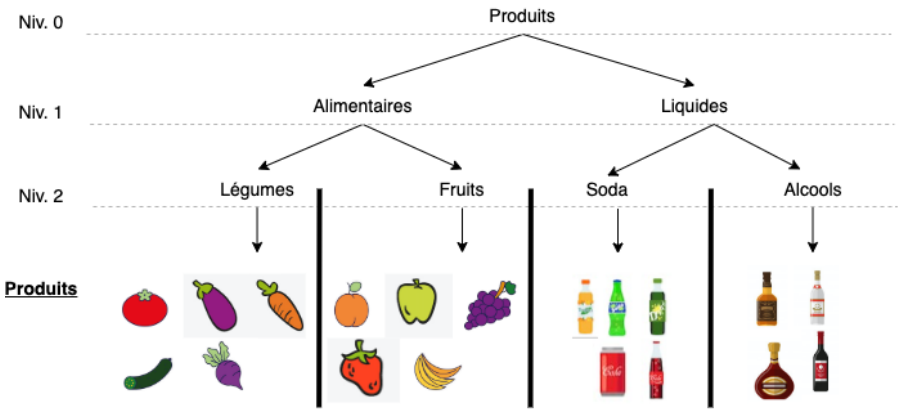


FIGURE 31 – Exemple de structure marchandise

quatre catégories de produits (`Légumes`, `Fruits`, `Soda` et `Alcools`) qui implique implicitement quatre différents catégories managers. Ces catégories managers définissent les ensembles de produits de chacun des profils assortiment comme illustré dans la figure 32. Dans cet exemple illustratif nous considérons quatre profils assortiment. Le plus petit profil assortiment est noté "1" et le plus grand "4". Avec la règle d'inclusion formalisé dans le manuscrit, les produits associés au profil assortiment "1" sont inclus dans le profil assortiment et ainsi de suite.

Profil Assortiment / Catégorie de produits	1	2	3	4
Légumes				
Fruits				
Soda				
Alcools				

FIGURE 32 – Profil assortiment des produits

Ensuite, chacun des catégories managers définit le profil assortiment des magasins comme présenté dans la figure 33. Dans cet exemple, nous considérons trois magasins.

Magasin / Catégorie de produits			
Légumes	4	2	1
Fruits	4	2	1
Soda	3	2	2
Alcools	3	2	3

FIGURE 33 – Profil assortiment des magasins

Magasin / Produit			
	1	1	1
	1	1	1
	1	1	0
	1	0	0
	1	0	0
	1	1	1
	1	1	0
	1	1	0
	1	0	0
	1	0	0
	1	1	1
	1	1	1
	1	1	1
	1	0	0
	0	1	0
	1	1	1
	1	1	1
	1	0	1
	0	0	1

FIGURE 34 – Exemple de vecteur binaire représentant l'assortiment des magasins

Enfin, grâce aux profils assortiments des articles et des magasins, nous pouvons définir l'assortiment final de chacun des magasins. Il est représenté dans la figure 34 par un vecteur binaire. Si la valeur du vecteur est égal à 1 alors le magasin doit posséder l'article, autrement, le magasin ne doit pas l'avoir.

BIBLIOGRAPHIE

[1] Ouvrage Collectif AFGI. « Evaluer pour évoluer, les indicateurs de performance au service du pilotage ». In : *Association Française de Gestion Industrielle* (1992).

[2] H. Aeron, A. Kumar et J. Moorthy. « Data mining framework for customer lifetime value-based segmentation. » In : *Expert Systems With Applications* 19 (2012), p. 17-30.

[3] N. Agrawal et S.A. Smith. « Estimating negative binomial demand for retail inventory management with unobservable lost sales ». In : *Naval Research Logistics (NRL)* 43 (1996), p. 839-861.

[4] N. Agrawal et S.A. Smith. « Management of multi-item retail inventory systems with demand substitution. » In : *Operations Research* 48 (2000), p. 50-64.

[5] N. Agrawal et S.A. Smith. « Optimal retail assortments for substitutable items purchased in sets. » In : *Naval Research Logistics* 50 (2003), p. 793-822.

[6] S.P. Anderson, A. de Palma. et J.F. Thisse. « Discrete Choice Theory of Product Differentiation. » In : *The MIT Press, Cambridge* (1992).

[7] Kurt Salmon Associates. « Efficient consumer response : Enhancing consumer value in the grocery industry. » In : *Food Marketing Institute* (1993), p. 9-526.

[8] J.L. Balintfy. « On a Basic Class of Multi-Item Inventory Problems ». In : *Management Science* 10 (1964), p. 287-297.

[9] Y. Bassok, R. Anupindi et R. Akella. « Single-period multiproduct inventory models with substitution. » In : *Operations Research* 47 (1999), p. 632-642.

[10] S. Basuroy et D. Nguyen. « Multinomial logit market share models : Equilibrium characteristics and strategic implications. » In : *Management Science.* 44 (1998), p. 1396-1408.

[11] M. Batet, D. Sanchez, A. Valls et K. Gibert. « Exploiting taxonomical knowledge to compute semantic similarity : an evaluation in the biomedical domain. » In : *In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (2010), p. 274-283.

[12] M. Ben-Akiva et S.R. Lerman. « Discrete choice analysis : Theory and application to travel demand. » In : *Cambridge, The MIT Press.* (1985).

[13] L. Berrah. « Une approche d'évaluation de la performance industrielle : Modèle d'indicateur et techniques floues pour un pilotage réactif. » Thèse de doct. 1997.

[14] L. Berrah. « Les indicateurs de performance : concepts et applications ». In : *Cépaduès Editions* (2002).

[15] L. Berrah, G. Mauris et J. Montmain. « Monitoring the improvement of an overall industrial performance based on a Choquet integral aggregation ». In : *The International Journal of Management Science OMEGA* 36 (2008), p. 340-351.

[16] L. Berrah, G. Mauris, J. Montmain et V. Clivillé. « Efficacy and efficiency indexes for a multi-criteria industrial performance synthesized by Choquet integral aggregation ». In : *International Journal of Data Analysis Techniques and Strategies* 21 (2008), p. 415-425.

[17] P. Besson, J.B. Cavaillé, A. Charles, P. Lorino, C. Pourcel, F. Roubellat, T. Sybord, A. Vergnenègre et M.P. Wurtz. « Aide à la conception des systèmes de conduite des systèmes de production ». In : *Actes du 3ème Congrès International de Génie Industriel* (1991).

[18] U.S. Bititci. « Modelling of performance measurement systems in manufacturing enterprises ». In : *International Journal of Production Economics* 42 (1995), p. 137-147.

[19] M. Bitton. « Ecograi : Méthode de conception et d'implantation de systèmes de mesure de performances pour organisations industrielles ». Thèse de doct. 1990.

[20] P. Boatwright et J.C. Nunes. « Reducing assortment : An attribute-based approach. » In : *Journal of Marketing* 65 (2001), p. 50-63.

[21] N. Borin et P. Farris. « A sensitivity analysis of retailer shelf management models. » In : *Journal of Retailing* 71 (1995), p. 153-171.

[22] S. Boussier, M. Vasquez, Y. Vimont, S. Hanafi et P.A. Michelon. « Multi-level search strategy for the 0-1 Multidimensional Knapsack Problem ». In : *Discrete Applied Mathematics* (2010), p. 97-109.

[23] S.M Broniarczyk, W.D. Hoyer et L. McAlister. « Consumers' perception of the assortment offered in a grocery category : The impact of item reduction ». In : *Journal of Research in Marketing* 35 (1998), p. 166-176.

[24] G.P. Cachon et A.G. Kok. « Category management and coordination of categories in retail assortment planning in the presence of basket shoppers. » In : *Management Science* 53 (2007), p. 934-951.

[25] G.P. Cachon, C. Terwiesch et Y. Xu. « Retail assortment planning in the presence of consumer search. » In : *Manufacturing & Service Operations Management.* 7 (2005), p. 330-346.

[26] G.P. Cachon, C. Terwiesch et Y. Xu. « On the Effects of Consumer Search and Firm Entry in a Multiproduct Competitive Market. » In : *Marketing Science. Forthcoming.* (2006).

[27] F. Caro et J. Gallien. « Dynamic assortment with demand learning for seasonal consumer goods. » In : *Working paper. Sloan School of Management.* (2005).

[28] E.J. Chang et K. Sycara. « Advances in Web Semantics : Ontologies, Web Services and Applied Semantic Web ». In : *Information Systems and Applications* (2009), p. 357-381.

[29] J.M. Charpentier. « Communication d'entreprise, de l'"image" au social. » In : *Communication et langages* (2006), p. 113-121.

[30] F. Chen, J. Eliashberg et P. Zipkin. « Customer preferences, supply-chain costs, and product line design. » In : *Product Variety Management : Research Advances, Kluwer Academic Publishers.* (1998).

[31] R.R. Chen, T.C.E. Cheng, T.M. Choi et Y. Wang. « Novel Advances in Applications of the Newsvendor Model ». In : *Decision Sciences* 48 (2016), p. 8-10.

[32] Y.L. Chen, M.H. Kuo, S.Y. Wu et K. Tang. « Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. » In : *Electronic Commerce Research and Applications* 8 (2009), p. 241-251.

[33] C. Ching-Hsue et C. You-Shyang. « Classifying the segmentation of customer value via RFM model and RS theory. » In : *Expert Systems With Applications* 36 (2009), p. 4176-4184.

[34] Y.H. Cho et J.K. Kim. « Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. » In : *Expert Systems with Applications* 23 (2004), p. 233-246.

[35] J-K. Chong, T-H. Ho et C.S. Tang. « A modeling framework for category assortment planning. » In : *Manufacturing & Service Operations Mgmt* 3 (2001), p. 191-210.

[36] M. Corstjens et P. Doyle. « A model for optimizing retail space allocations. » In : *Management Science* 27 (1981), p. 822-833.

[37] V. Cross et X. Yu. « Investigating ontological similarity theoretically with fuzzy set theory, information content, and Tversky similarity and empirically with the gene ontology. » In : *Scalable Uncertainty Management* (2011), p. 387-400.

[38] D.Gourc. « Contribution à la réingénierie des systèmes de production. » Thèse de doct. 1997.

[39] A.P. Dempster, N.M. Laird et D.B. Rubin. « Maximum likelihood from incomplete data via the EM algorithm. » In : *Journal of the Royal Statistical Society* 39 (1977), p. 1-38.

[40] P. Desai, S. Radhakrishnan et K. Srinivasan. « Product differentiation and commonality in design : balancing revenue and cost drivers. » In : *Management Science* 47 (2001), p. 37-51.

[41] S.K. Dhar, S.J. Hoch et N. Kumar. « Effective category management depends on the role of the category. » In : *Journal of Retailing* 77 (2001), p. 165-184.

[42] G. Dobson et S. Kalish. « Heuristics for pricing and positioning a product line. » In : *Management Science* 39 (1993), p. 160-175.

[43] G. Doumeingts et Y. Ducq. « Enterprise modelling techniques to improve efficiency of enterprises ». In : *Production Planning & Control* 12 (2001), p. 146-163.

[44] B. Downs, R. Metters et J. Semple. « Managing inventory with multiple products, lags in delivery, resource constraints, and lost sales : A mathematical programming approach ». In : *Management Science* 47 (2002), p. 464-479.

[45] X. Dreze, S.J. Hoch et M.E. Purk. « Shelf management and space elasticity. » In : *Journal of Retailing* 70 (1994), p. 301-326.

[46] T. Dromard. « Pourquoi Carrefour Planet n'a pas fonctionné ». In : *Challenges Entreprises* (2010). URL : https://www.challenges.fr/entreprise/pourquoi-carrefour-planet-n-a-pas-fonctionne_354870.

[47] H. EL-Ghalayini, M. Odeh et R. McClatchey. « Deriving Conceptual Data Models from Domain Ontologies for Bioinformatics. » In : *ICTTA 2006* 2 (2006), 3562—3567.

[48] V. Gaur et D. Honhon. « Assortment planning and inventory decisions under a locational choice model. » In : *Management Science* 52 (2006), p. 1528-1543.

[49] R. Gentleman. « Visualizing and distances using GO ». In : 38 (2005), p. 164-179. URL : http://master.bioconductor.org/help/package-vignettes/.

[50] R. Gentleman et Z. Jiang. « Extensions to gene set enrichment ». In : *Bioinformatics* 23 (2007), p. 306-313.

[51] V. Goyal, R. Levi et D. Segev. « Near-Optimal Algorithms for the Assortment Planning Problem Under Dynamic Substitution and Stochastic Demand. » In : *Operation Research* 64 (2016).

[52] A. Griva, C. Bardaki, K. Pramatari et D. Papakiriakopoulos. « Retail business analytics : Customer visit segmentation using market basket data. » In : *Expert Systems With Applications* 16 (2018), p. 1-16.

[53] P.M. Guadagni et J.D.C. Little. « A logit model of brand choice calibrated on scanner data. » In : *Marketing Science* 2 (1983), p. 203-238.

[54] N. Guarino. « Understanding, building and using ontologies. » In : *International Journal of Human-Computer Studies* 46 (1997), p. 293-310.

[55] N. Guarino. « Formal Ontology and Information Systems. » In : *Proceedings of FOIS'98* (1998), 3—15.

[56] G. Hadley et T.M. Whitin. « Analysis of Inventory Systems ». In : *Analysis of Inventory Systems. Prentice Hall* (1963).

[57] M. Hammer et J. Champy. « Le Reengineering : Réinventer l'entreprise pour une amélioration de ses performances ». In : *Dunod, Paris* (1993).

[58] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi et J. Montmain. « A Framework for Unifying Ontology-based Semantic Similarity Measures : a Study in the Biomedical Domain. » In : *Journal of Biomedical Informatics* 48 (2014), p. 38-53.

[59] S. Harispe, S. Ranwez, S. Janaqi et J. Montmain. « The Semantic Measures Library and Toolkit : fast computation of semantic similarity and relatedness using biomedical ontologies. » In : *Bioinformatics* (2014), 740—742.

[60] S. Harispe, S. Ranwez, S. Janaqi et J. Montmain. « Semantic Similarity from Natural Language and Ontology Analysis. » In : *Synthesis Lectures on Human Language Technologies* (2015).

[61] S. Harispe, S. Ranwez, S. Janaqi et J. Montmain. « Semantic similarity from natural language and ontology analysis. » In : *Synthesis Lectures on Human Language Technologies* 8 (2015), p. 1-254.

[62] M. Harzallah. « Modélisation des aspects organisationnels et des compétences pour la réorganisation d'entreprises industrielles. » Thèse de doct. 2000.

[63] M. Harzallah et F. Vernadat. « Méthodologie de réorganisation d'entreprise industrielle ». In : *Actes 3ème Congrès Franco-Québecois de Génie Industriel, Montréal, Canada* (1999).

[64] W.H. Hausman et G. Aydin. « Supply chain coordination and assortment planning. » In : *Working paper. University of Michigan.* (2003).

[65] T. Hong et E. Kim. « Segmenting customers in online stores based on factors that affect the customer's intention to purchase. » In : *Expert Systems With Applications* 39 (2012), p. 2127-2131.

[66] H. Hotelling. « Stability in Competition ». In : *Economic Journal* (1929), p. 50-63.

[67] A. Hubner et K. Schaa. « A shelf-space optimization model when demand is stochastic and space-elastic ». In : *Science Direct, Omega* 68 (2017), p. 139-154.

[68] L.P. Hung. « A personalized recommendation system based on product taxonomy for one-to-one marketing online. » In : *Expert Systems with Applications* 29 (2005), p. 383-392.

[69] J. Irion, F. Al-Khayyal et J. Lu. « A Piecewise Linearization Framework for Retail Shelf Space Management Models. » In : *Working paper. Georgia Institute of Technology.* (2004).

[70] G. Jacob. « Le Reengineering de l'Entreprise ». In : *éditions Hermès, Paris* (1994).

[71] G. Jacob. « La refonte des systèmes d'information ». In : *éditions Hermès, Paris* (1995).

[72] J.H. Jacot et J.P. Micaelli. « La performance économique en entreprise ». In : *éditions Hermès, Paris* (1996).

[73] M. Jarrar, J. Demey et R. Meersman. « On Using Conceptual Data Modeling for Ontology Engineering. » In : *Journal on Data Semantics* 2800 (2003), 185—207.

[74] T.Q. Jia. « Vers une meilleure gestion des ressources d'un groupe autonome de fabrication ». Thèse de doct. 1998.

[75] J.J. Jiang et D.W. Conrath. « Combining local context and WordNet similarity for word sense identification. » In : *ArXiv Preprint Cmp-lg/9709008* (1997).

[76] K. Sparck Jones. « A statistical interpretation of term specificity and its application in retrieval. » In : *Journal of documentation* (1972), p. 11-21.

[77] I.M. Kaizen. « La clé de la compétitivité japonaise ». In : *Eyrolles, Paris* (1989).

[78] R.M. Karp. « Reducibility Among Combinatorial Problems ». In : *Complexity of Computer Computations* (1972), p. 85-103.

[79] M. Khajvand, K. Zolfaghar, S. Ashoori et S. Alizadeh. « Estimating customer lifetime value based on RFM analysis of customer purchase behavior : Case study. » In : *Procedia Computer Science* 3 (2011), p. 57-63.

[80] M. Khajvand, K. Zolfaghar, S. Ashoori et S. Alizadeh. « A multi-category customer base analysis. » In : *International Journal of Research in Marketing* 31 (2014), p. 266-279.

[81] H. K. Kim, J. K. Kim et Q. Y. Chen. « A product network analysis for extending the market basket analysis. » In : *Expert Systems with Applications* 39 (2012), p. 7403-7410.

[82] A.G. Kok. « Management of product variety in retail operations. » In : *Ph.D. Dissertation. The Wharton School, University of Pennsylvania.* (2003).

[83] A.G. Kok et M.L. Fisher. « Demand estimation and assortment optimization under substitution : methodology and application. » In : *Operations Research* 55 (2007), p. 1001-1021.

[84] A.G. Kok, M.L. Fisher et R. Vaidyanathan. « Assortment Planning : Review of Literature and Industry Practice. » In : *Retail Supply Chain Management* (2006), p. 1-46.

[85] P. Kueng et A.J.W Krahn. « Building a process performance measurement system : Some early experiences. » In : *Journal of Scientific & Industrial Research* 58 (1999), p. 149-159.

[86] H. Kuhn et MG Sternbeck. « Integrative retail logistics : an exploratory study. » In : *Operations Management Research* 6 (2013), p. 2-18.

[87] M. Kurtulus. « Supply chain collaboration practices in consumer goods industry. » In : *Ph. D. INSEAD.* (2005).

[88] M. Kurtulus et B. Toktay. « Category captainship : Outsourcing retail category management. » In : *Working paper. INSEAD.* (2005).

[89] L.W. Lacy. « Owl : Representing Information Using the Web Ontology Language ». In : (2005), p. 259.

[90] K. Lancaster. « The economics of product variety : A survey. » In : *Marketing Science* 9 (1990), p. 189-210.

[91] C. Leacock et M. Chodorow. « Combining local context and WordNet similarity for word sense identification. » In : *WordNet : An Electronic Lexical Database* 49 (1998), p. 265-283.

[92] C.W. Leung, S.C. Chan et F.L. Chung. « A collaborative filtering framework based on fuzzy association rules and multiple-level similarity. » In : *Knowledge and Information Systems* 10 (2006), p. 357-381.

[93] M. Levy et B.A. Weitz. « Retailing Management ». In : *McGraw-Hill/Irwin, NY* (2004), p. 398-400.

[94] B. Li, J.Z. Wang, F.A. Feltus, J. Zhou et F. Luo. « Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins. » In : (2010).

[95] J. Li, A. Sato, R. Huang et D. Chen. « A Rule-Based Knowledge Discovery Engine Embedded Semantic Graph Knowledge Repository for Retail Business. » In : *2016 International Conference on Advanced Cloud and Big Data (CBD)* (2016), p. 81-86.

[96] Z. Li. « A Single-Period Assortment Optimization Model ». In : *Production and Operations Demand* 16 (2007), p. 369-380.

[97] D. Lin. « An information-theoretic definition of similarity ». In : *Proceedings of the Fifteenth International Conference on Machine Learning* 98 (1998), p. 296-304.

[98] P. Lingras, A. Elagamy, A. Ammar et Z. Elouedi. « Iterative meta-clustering through granular hierarchy of supermarket customers and products. » In : *Information Sciences* 257 (2014), p. 14-31.

[99] C. Lopez, F. Nooralahzadeh, E. Cabrio, F. Segond et F. Gandon. « ProVoc : une ontologie pour décrire des produits sur le Web ». In : *IC2016 : 27es Journées francophones d'Ingénierie des Connaissances* (2016), p. 1-12.

[100] P. Lorino. « Méthodes et Pratiques de la Performance ». In : *Les Editions d'Organisation* (1996).

[101] B. Maddah et E.K. Bish. « Joint pricing, assortment, and inventory decisions for a retailer's product line. » In : *Working Paper, Virginia Polytechnic Institute and State University.* (2004).

[102] K. Madhavi. « The impact of product cannibalization on consumer purchasing decision - an attitudinal conflict paradigm ». In : *IMPACT : International Journal of Research in Business Management (IMPACT : IJRBM)* (2012).

[103] A. Maedche et S. Staab. « Comparing ontologies-similarity measures and a comparison study. » In : *Institute of Applied Informatics and Formal Description Methods (AIFB)* (2001).

[104] S. Mahajan et G. van Ryzin. « On the relationship between inventory costs and variety benefits in retail assortments. » In : *Management Science.* 45 (1999), p. 1496-1509.

[105] S. Mahajan et G. van Ryzin. « Inventory competition under dynamic consumer choice ». In : *Operations Research* 49(5) (2001), p. 646-657.

[106] S. Mahajan et G. van Ryzin. « Stocking retail assortments under dynamic consumer substitution. » In : *Operations Research* 49(3) (2001), p. 334-351.

[107] P. Manchanda, A. Ansari et S. Gupta. « The "shopping basket" : A model for multicategory purchase incidence decisions. » In : *Marketing Science* 18 (1999), p. 95-114.

[108] C. Manning, P. Raghavan et Hinrich Schütze. « Introduction to Information Retrieval. » In : *Cambridge University Press* (2009).

[109] H. Martin. « GoodRelations : An Ontology for Describing Products and Services Offers on the Web ». In : *Knowledge Engineering : Practice and Patterns* (2008), p. 329-346.

[110] G.G. Mazandu et N.J. Mulder. « IT-GOM : An Integrative Tool for IC-based GO Semantic Similarity Measures. » In : *Technical report, University of Cape Town (South Africa)* (2011).

[111] G.G. Mazandu et N.J. Mulder. « Information content-based gene ontology semantic similarity approaches : toward a unified framework theory. » In : *BioMed Research International, 2013* (2013).

[112] D. McFadden. « Conditional logit analysis of qualitative choice behavior. » In : *Frontiers in Econometrics. Academic Press, NY.* (1974).

[113] A.R. McGillivray et E.A. Silver. « Some concepts for inventory control under substitutable demand. » In : *Information Systems and Operational Research* 16 (1978), p. 47-63.

[114] J. Montmain. « Monitoring and control of an efficient industrial performance improvement within a MAUT assessment framework ». In : *Supply Chain Forum : an International Journal* 12 (2011).

[115] S. Moorthy. « arket segmentation, self-selection, and product line design ». In : *Marketing Science* 3 (1984), p. 288-307.

[116] F. Murtagh. « Ward's Hierarchical Agglomerative Clustering Method : Which Algorithms Implement Ward's Criterion ? » In : *Journal of Classification* 31 (2014), p. 274-295.

[117] M. Mussa et S. Rosen. « Monopoly and product quality. Journal of Economic Theory. » In : *Journal of Economic Theory* 18 (1978), p. 301-317.

[118] A. Neely. « The performance measurement revolution : why now and what next ? » In : *International Journal of Operations and Production Management* 19 (1999), p. 205-64.

[119] S. Netessine et N. Rudi. « Centralized and competitive inventory model with demand substitution ». In : *Operational Research* 51 (2003), p. 329-335.

[120] S. Netessine et T.A. Taylor. « Product line design and production technology. » In : *Working paper, Wharton School* (2005).

[121] E.W.T. Ngai, L. Xiu et D.C.K. Chau. « Application of data mining techniques in customer relationship management : A literature review and classification. » In : *Expert Systems with Applications* 36 (2009), p. 2592-2602.

[122] AC Nielsen. « Eighth Annual Survey of Trade Promotion Practices ». In : *Chicago, IL. ACNielsen* (1998).

[123] P.S. Noonan. « When consumers choose : a multi-product, multi-location newsboy model with substitution. » In : *Working paper. Emory University.* (1995).

[124] K. Pal. « Ontology-Based Web Service Architecture for Retail Supply Chain Management ». In : *European Journal of Operational Research* 199 (2009), p. 759-768.

[125] K. Pal. « Ontology-Based Web Service Architecture for Retail Supply Chain Management. » In : *Sustainable Energy Information Technology, (SEIT-2018) , 2018* 130 (2018), p. 985-990.

[126] M. Parlar et S.K. Goyal. « Optimal ordering policies for two substitutable products with stochastic demand. » In : *Information Systems and Operational Research* 21 (1984), p. 1-15.

[127] R. Paul, T. Groza, A. Zankl et J. Hunter. « Semantic similarity-driven decision support in the skeletal dysplasia domain ». In : *International Semantic Web Conference. Springer, Berlin Heidelberg* (2012), p. 164-179.

[128] C. Pesquita, D. Faria, H. Bastos, A. Falcao et F. Couto. « Evaluating go-based semantic similarity measures ». In : *Proc 10th Annual Bio-Ontologies Meeting* 37 (2007).

[129] G. Pirró. « A semantic similarity metric combining features and intrinsic information content. » In : *Data Knowledge Engineering* 68 (2009), p. 1289-1308.

[130] G. Pirró et J. Euzenat. « A semantic similarity framework exploiting multiple parts-of speech. » In : *OTM Confederated International Conferences On the Move to Meaningful Internet Systems. Springer Berlin Heidelberg.* (2010), p. 1118-1125.

[131] R. Rada, H. Mili, E. Bicknell et M. Blettner. « Development and application of a metric on semantic nets. » In : *IEEE Transactions on Systems, Man, and Cybernetics* 19 (1989), p. 17-30.

[132] K. Rajaram. « Assortment planning in fashion retailing : methodology, application and analysis. » In : *European Journal of Operational Research* 129 (2001), p. 186-208.

[133] K. Rajaram et C.S. Tang. « The impact of product substitution on retail merchandising. » In : *European Journal of Operational Research* 135 (2001), p. 582-601.

[134] P. Resnik. « Using information content to evaluate semantic similarity in a taxonomy. » In : *Proceedings of IJCAI-95* (1995), 448—453.

[135] P. Resnik. « Semantic Similarity in a taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. » In : *Journal of Artificial Intelligence Research* 11 (1999), p. 95-130.

[136] M.A. Rodriguez et M.J. Egenhofer. « Determining semantic similarity among entity classes from different ontologies. » In : *IEEE Transactions on Knowledge and Data Engineering* 15 (2003), p. 442-456.

[137] E.J. Ruiz et B.C. Grau. « LogMap : Logic-based and Scalable Ontology Matching ». In : *The 10 th International Semantic Web Conference* (2011).

[138] S. SAHRAOUI, L. BERRAH et J. MONTMAIN. « Techniques d'optimisation pour la définition d'une démarche d'amélioration industrielle : une approche par analyse et agrégation des performances. » In : *e-revue des Sciences et Technologies de l'Automatique* 5 (2008).

[139] D. SÀNCHEZ et M. BATET. « Semantic similarity estimation in the biomedical domain : An ontology-based information-theoretic perspective. » In : *Journal of Biomedical Informatics* 44 (2011), p. 749-759.

[140] T. SCHAEL. « Théorie et pratique du worflow. Des processus métier renouvelés. » In : *Springer-Verlag* 5 (1997).

[141] V. SCHICKEL-ZUBER et B. FALTINGS. « OSS : A Semantic Similarity Function based on Hierarchical Ontologies. » In : *International Joint Conference on Artificial Intelligence (IJCAI)* 7 (2007), p. 551-556.

[142] A. SCHLICKER, F.S. DOMINGUES, J. RAHNENFÜHRER et T. LENGAUER. « A new measure for functional similarity of gene products based on Gene Ontology. » In : *BMC Bioinformatics* 7 (2006).

[143] N. SECO, T. VEALE et J. HAYES. « An Intrinsic Information Content Metric for Semantic Similarity in WordNet. » In : *16th European Conference on Artificial Intelligence* (2004), p. 1-5.

[144] A. SEGEV et Q.Z. SHENG. « Bootstrapping Ontologies for Web Services ». In : *IEEE Transactions of Service Computin* 5 (2012), p. 33-44.

[145] C.E. SHANNON. « A mathematical theory of communication. » In : *The Bell System Technical Journal* 27 (1984), p. 379-423.

[146] A.J. SHENHAR, D. DVIR et Y. SHULMAN. « A two-dimensional taxonomy of products and innovations. » In : *Journal of Engineering and Technology Management* 12 (1995), p. 175-200.

[147] I. SIMONSON. « The effect of product assortment on buyer preferences ». In : *Journal of Retailing* 75 (1999), p. 347-370.

[148] P. SINGH, H. GROENEVELT et N. RUDI. « Product variety and supply chain structures. » In : *Working paper. University of Rochester.* (2005).

[149] R. SRIKANT et R. AGRAWAL. « Mining Generalized Association Rules. » In : *Future Generation Computer Systems* 13 (1997), p. 161-180.

[150] M. STAMMINGER. « Key-Driven Business Reengineering-How to get Reengineering more efficient and effective ». In : *Information Infrastructure Systems for Manufacturing* (1996), p. 265-274.

[151] K. Talluri et G. van Ryzin. « Revenue management under a general discrete choice model of consumer behavior. » In : *Management Science* 50 (2004), p. 15-33.

[152] A. Tversky et G. Itamar. « Studies of Similarity ». In : *Cognition and categorization* 9 (1978), p. 79-98.

[153] M. Uschold et M. Gruninger. « Ontologies and Semantics for Seamless Connectivity ». In : *SIGMOD Rec.* 33 (2004), 58—64.

[154] Z. Wu et M. Palmer. « Verb semantics and lexical selection. » In : *Annual Meeting of the Association for Computational Linguistics* (1994), p. 133-138.

[155] E. Yucel, F. Karaesmen, F.S. Salman et M. Turkay. « Optimizing product assortment under customer-driven demand substitution ». In : *European Journal of Operational Research* 199 (2009), p. 759-768.

[156] M.F. Zarandi. « A Retail Ontology : Formal Semantics and Efficient Implementation. » In : *Ph.D. University of Toronto* (2007).

[157] X. Zhou, J. Menche, A.L. Barabási et A. Sharma. « Human symptoms-disease network. » In : *Nature communications* (2014), p. 1-10.

[158] Z. Zhou, Y. Wang et J. Gu. « A new model of information content for semantic similarity in WordNet. » In : *FGCNS 08. Second International Conference* 3 (2008), p. 85-89.

[159] A. lbadvi et M. Shahbazi. « A hybrid recommendation technique based on product category attributes. » In : *Expert Systems with Applications* 36 (2009), p. 11480-11488.