Formalisation du concept d'assortiment idéal dans la grande distribution

Jocelyn Poncelet, Pierre-Antoine Jean, Michel Vasquez, Jacky Montmain

Formalisation du concept d'assortiment idéal dans la grande distribution

Formalization of the ideal assortment concept in retail

Jocelyn Poncelet [1,2] Pierre-Antoine Jean[2] Michel Vasquez[2] Jacky Montmain[2]

(1) TRF REtail, 116 allée Norbert Wiener, 30000 Nîmes, France
(2) EuroMov DHM, IMT Mines Ales, Univ Montpellier, Alès, France
6 Avenue de Clavières, 30319 Alès Cedex France
prénom.nom@mines-ales.fr

Résumé :
La survie d'une chaîne de supermarchés dépend fortement de sa capacité à fidéliser ses clients. La question de l'assortiment du magasin est donc cruciale. Avec des dizaines de milliers de produits sur les étagères, définir l'assortiment idéal est un problème d'optimisation combinatoire complexe. L'approche que nous proposons inclut des connaissances a priori sur l'organisation hiérarchique des produits par famille pour formaliser le problème de l'assortiment en un problème combinatoire de sac à dos. La principale difficulté du problème d'optimisation reste l'estimation des bénéfices attendus d'une évolution de gamme pour une famille de produits. Cette estimation est basée sur les résultats comptables de magasins similaires du réseau. La définition de la similarité entre deux magasins est alors centrale et repose sur la connaissance a priori de l'organisation hiérarchique des produits. Cette structuration permet une comparaison abstraite qui rend mieux compte des comportements d'achats.

Mots-clés :
Assortiment dans la grande distribution, Mesures de similarité sémantique, problème du sac à dos.

Abstract:
The survival of a supermarket chain is heavily dependent on its capacity to maintain the loyalty of its customers. Proposing adequate products to customers is the issue of the store's assortment. With tens thousands of products on shelves, designing the ideal assortment is a thorny combinatorial optimization problem. The approach we propose includes prior knowledge on the hierarchical organization of products by family to formalize the ideal assortment problem into a knapsack problem. The main difficulty of the optimization problem remains the estimation of the expected benefits associated to changes in the product range of products' families. This estimate is based on the accounting results of similar stores. The definition of the similarity between two stores is therefore central and is based on a priori knowledge of the hierarchical organization of the products, which allows an abstract comparison, which better accounts for purchasing behavior.

Keywords:
Optimal Assortment in Mass Distribution, Semantic Similarity Measures, Knapsack problem.

1 Introduction

La concurrence chez les grands distributeurs est toujours plus intense; par conséquent, afin de satisfaire la demande fluctuante et les attentes croissantes des clients, les grands distributeurs doivent se concentrer sur la recherche d'avantages durables. La survie d'une chaîne de supermarchés dépend fortement de sa capacité à fidéliser ses clients [14,15]. La question de l'assortiment le mieux adapté à un magasin est donc un problème central [13,16]. Par ailleurs, les enseignes doivent gérer des réseaux de magasins de grande taille. Elles mettent en place des assortiments communs partagés dans le réseau des magasins pour une gestion plus facile [17]. Par conséquent, les magasins partagent un assortiment commun et centralisé [20] avec quelques exceptions [18]. Pour améliorer leur performance globale, les gérants des enseignes essaient de tirer profit des résultats des différents magasins du réseau : ils essaient d'identifier les produits qui fonctionnent le mieux dans certains magasins du réseau pour les recommander à d'autres qui leur paraissent similaires. L'objectif est de définir l'assortiment optimal pour chaque magasin. Cependant, la définition de magasins

similaires n'est pas si évidente : elle peut être liée à la localisation des magasins, à la diversité des produits en rayon, à leurs chiffres d'affaires, à leur format (ex : hypermarché, supermarché, etc.) [14]. Ce concept de similarité joue un rôle central dans cette contribution.

Pour mieux comprendre la complexité de la tâche, il faut se rappeler que certains hypermarchés proposent jusqu'à 100 000 produits [18]. Définir l'assortiment idéal dans un grand magasin consiste à sélectionner cet ensemble de produits. Ce problème de gestion correspond à un problème d'optimisation combinatoire NP-complet. En pratique, les décisions sont prises localement par un responsable de catégorie, alors que le problème de l'assortiment idéal devrait correspondre à une décision globale au niveau du magasin. Pour résoudre ce problème, les grands distributeurs disposent néanmoins de connaissances a priori [19]. En effet, les magasins sont organisés en catégories, par exemple, la nourriture, les produits ménagers, les textiles, etc. Ces catégories sont elles-mêmes subdivisées en familles ou unités de besoin (par exemple, la catégorie textile se décline en sections : linge de maison, literie, vêtements, etc.). Cette organisation hiérarchique des produits va nous permettre de raisonner sur les familles de produits et non plus sur les produits eux-mêmes, d'abstraire le raisonnement catégoriel sur les produits et de structurer la décision pour éviter l'explosion combinatoire.

La plupart du temps, un hypermarché ne peut pas choisir un seul produit pour parfaire son offre. En effet, cet article appartient nécessairement à un niveau d'assortiment ou gamme de produits en adéquation avec la taille ou l'emplacement du magasin : le choix d'un produit nécessite de prendre tous les produits associés à ce niveau d'assortiment pour une famille de produits donnée [18, 20]. Par exemple, si un magasin propose une section de soda, il peut se satisfaire d'une offre minimale, par exemple, `Coca-Cola 1.5L`; mais il peut aussi revendiquer une offre supérieure : par exemple, il aimerait proposer du `Lipton 2L`. L'ajout à une offre n'est généralement pas autorisé produit par produit, mais par sous-ensemble de produits. Par exemple, pour pouvoir proposer du `Lipton 2L`. à sa clientèle, le gérant du magasin devra au final proposer la gamme : `Coca-Cola 1.5L + Lipton 2L + Orangina 1,5L + Schweppes 1,5L`.

2 Modélisation de l'assortiment idéal comme un problème d'optimisation combinatoire

Soit Ω un magasin. F_i est la $i^{ème}$ famille de produits, *i.e.* un ensemble de produits qui appartiennent à une même unité de besoin (*e.g.*, boissons non alcoolisées, électroménager, etc.). Récursivement, toute famille F_i est une spécialisation d'une super famille : *e.g.*, $Coca-cola \in F_{soda} \subset F_{boissons\ non\ alcoolisées} \subset F_{boissons}$. Les produits peuvent ainsi être organisés selon un ordre partiel taxonomique définissant une hiérarchie d'abstraction (cf. figure 1). Les produits sont les classes les plus spécifiques de cet ordre partiel, *i.e*, les feuilles de la taxonomie.

Distinguons le cas particulier des familles de produits, c'est-à-dire les classes les plus basses de la hiérarchie, les moins abstraites, puisque leurs descendants directs sont des produits. Pour chacune de ces familles de produits F_i, un niveau d'assortiment $s(F_i)$ est défini : pour chaque famille de produits, le magasin peut choisir son niveau d'assortiment $s(F_i)$ dans un ensemble fini de choix imposés par la direction de l'enseigne de magasins. Formellement, pour chaque famille, une suite ordonnée de sous-ensembles de produits $s^{k_i}(F_i), k_i = 1..n$ est définie au sens de la relation d'inclusion (*i.e.*, $s^{k_i}(F_i) \subset s^{k_i+1}(F_i)$) et le magasin peut choisir son assortiment uniquement parmi les sous-ensembles $s^{k_i}(F_i)$ (*e.g.* $s^1(F_{soda})$: `Coca-Cola 1,5L ;`

$s^2(F_{soda})$: `Coca-Cola 1,5L + Lipton 2L + Orangina 1,5L + Schweppes 1,5L`, etc.). La taille de $s(F_i)$ est ensuite le niveau d'assortiment k_i tel que $s(F_i) = s^{k_i}(F_i)$. En pratique, k_i est un naturel entre 1 et 9 ; $k_i = 1$ quand le niveau d'assortiment de la famille F_i est minimal et $k_i = 9$ quand il est maximal.

Ensuite, on peut écrire : $\Omega \triangleq \bigcup_{i=1}^{n} s^{k_i}(F_i)$. Un chiffre d'affaires $p\left(s^{k_i}(F_i)\right)$ et un coût de stockage (ou un prix de revient) $c\left(s^{k_i}(F_i)\right)$ sont associés à chaque $s^{k_i}(F_i)$. Pour toute super famille dans l'organisation hiérarchique des produits, son chiffre d'affaires et sa capacité de stockage sont simplement calculés récursivement comme la somme des chiffres d'affaires et des coûts de stockage des produits que la super famille recouvre.
La conception de l'assortiment d'un grand magasin consiste alors à choisir le rang k_i pour chaque famille de produits (cf. figure 1).

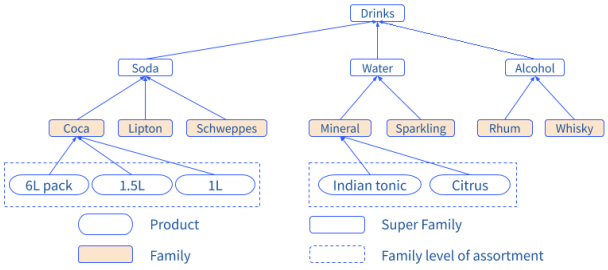

Figure 1: Organisation hiérarchique des produits et niveaux d'assortiment

Plus $p(\Omega) \triangleq \sum_{i=1}^{n} p\left(s^{k_i}(F_i)\right)$ est grand, meilleur est l'assortiment de Ω. Néanmoins, sans plus de contraintes, $p(\Omega)$ serait nécessairement maximal lorsque $k_i = 9$. En pratique, $\sum_{i=1}^{n} c\left(s^{k_i}(F_i)\right)$ doit rester bien en deçà de $\sum_{i=1}^{n} c(s^9(F_i))$ pour d'évidentes raisons budgétaires (C).

Soit I un sous-ensemble de familles. Il peut être nécessaire de modéliser des contraintes sur ces super familles. Par exemple, $c(s^{k_{boissons\ no\ naloccolisées}}(F_{boissons\ non\ aloccolisées}))$ $+c(s^{k_{bières}}(F_{bières}))$ $+c(s^{k_{eauxminérales}}(F_{eauxminérales}))$ $\leq C_{I=Boissons}$ signifie que le coût de stockage (ou le prix de revient) associé à `Boissons` (Super famille F_I) ne peut dépasser C_I. Une borne inférieure c_I peut aussi être introduite pour garantir une offre minimale par unité de besoin. Ce type de contraintes locales peut être ajouté pour chaque super famille.

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

Sous les contraintes :

$$\sum_{i=1}^{n} c\left(s^{k_i}(F_i)\right) \leq C \text{ - } \textit{contrainte globale}$$

Pour tout $I \subset 2^{\{1..n\}}$,

$$c_I \leq \sum_{i=1}^{|I|} c\left(s^{k_i}(F_i)\right) \leq C_I \text{ - } \textit{contrainte locale}$$

Ce problème d'optimisation combinatoire est connu sous le nom de problème du sac à dos à contraintes monodimensionnelles et variables naturelles bornées.

3 Estimation du chiffre d'affaires espéré

Considérons que l'un des assortiments à évaluer dans le problème d'optimisation comprenne l'augmentation de la gamme de produits de la famille de produits F_i : $s^{k_i}(F_i)$ est améliorée en $s^{k_i+1}(F_i)$. Le coût de stockage $c(s^{k_i+1}(F_i))$ peut être facilement renseigné par le magasin dans le problème d'optimisation. En revanche, il faut estimer $p\left(s^{k_i+1}(F_i)\right)$ car il s'agit d'un chiffre d'affaires espéré. Lorsque le niveau d'assortiment du magasin est k_i, $p\left(s^{k_i}(F_i)\right)$ est une mesure disponible du

magasin, mais $p\left(s^{k_i+1}(F_i)\right)$ ne peut être qu'une estimation.

$p\left(s^{k_i+1}(F_i)\right)$ peut seulement être estimé à partir des chiffres d'affaires observés dans d'autres magasins. L'idée de base est que plus ces magasins sont similaires au magasin concerné, plus l'estimation sera fiable. Le problème le plus difficile est de définir ce que signifie « similaire ». Intuitivement, ces magasins de référence doivent réaliser des chiffres d'affaires proches de ceux de Ω pour toutes les familles de produits $F_j \neq F_i$ et $s^{k_i+1}\left(F_i\right)$ pour la famille F_i. $p(s^{k_i^{\Omega}+1}(F_i))$ peut alors être calculé par exemple comme la somme pondérée des $p(s^{k_i^{\Omega'}+1}(F_i))$, où les Ω' sont les magasins de référence voisins de Ω et les poids seraient fonction de la similarité entre Ω et Ω'.

Ce concept de similarité entre deux grands magasins est la question cruciale. Ω devrait être similaire à Ω' lorsque les chiffres d'affaires de Ω et Ω'' sont répartis de la même manière sur l'organisation hiérarchique des produits. Cela implique qu'ils ont approximativement les mêmes types de clients.

Classiquement, la distance entre deux magasins pourrait être basée sur un espace métrique classique où les dimensions correspondraient à tous les produits proposés par les magasins de l'enseigne ; la valeur de chaque coordonnée serait le chiffre d'affaires du produit par exemple, et serait nulle si le grand magasin ne propose pas ce produit. Étant donné que certains hypermarchés proposent jusqu'à 100 000 produits, le processus de clustering souffrirait de la dimension de l'espace. De plus, une telle distance ne rendrait pas compte de l'organisation hiérarchique des produits dans le concept de similarité. On peut en effet noter qu'un grand magasin spécialiste de fruits et légumes est évidemment plus proche d'une grande épicerie que d'une quincaillerie car les deux premiers sont des magasins alimentaires alors que le dernier est un magasin spécialisé : les deux premiers proposent la même super famille $F_{nourriture}$. Cette similarité intuitive ne peut pas être évaluée avec des distances classiques. L'organisation hiérarchique des familles de produits est une connaissance a priori à prendre en compte lors de l'évaluation de la similarité de deux magasins. Il est nécessaire d'introduire des mesures de similarité adéquates. Cette notion de mesures de similarité est détaillée dans la section 3.

Notons que l'augmentation de $s^{k_i}\left(F_i\right)$ à $s^{k_i+1}\left(F_i\right)$ doit générer un chiffre d'affaires supérieur pour la famille F_i. Par contre, cet assortiment nécessite un coût $c(s^{k_i+1}\left(F_i\right))$ plus élevé que $c(s^{k_i}\left(F_i\right))$. Par conséquent, le coût de stockage d'au moins une famille de produits $F_{j,j\neq i}$ doit être réduit pour maintenir constant le coût global de stockage du magasin.

A ce stade, si l'on sait définir les magasins de référence pour tout magasin Ω, nous pouvons estimer $p\left(s^{k_i+1}(F_i)\right), \forall\left(k_1,k_2,..,k_n\right)\in\left[1..9\right]^n$ comme la somme pondérée par les similarités des chiffre d'affaires des magasins de référence associés à Ω pour une amélioration donnée.

Concevoir l'assortiment d'un grand magasin consiste à choisir le rang k_i pour chaque famille de produits, il faut énumérer et évaluer tout assortiment potentiel dans $\left[1..9\right]^n$ pour sélectionner le meilleur qui sera la solution du problème d'optimisation. La taxonomie de produits permet de réduire l'espace de recherche.

4 Taxonomie et similarités sémantiques

La mesure de similarité qui répond à nos attentes repose sur la structure taxonomique qui organise les produits et les familles de produits dans les magasins puisque Ω devrait être similaire à Ω' lorsque les chiffres d'affaires de Ω et Ω' sont répartis de la même manière sur cette structure hiérarchique. Dans la littérature de la recherche d'information et

de la gestion des connaissances, les éléments de la structure taxonomique sont nommés concepts (ou classes). Une structure taxonomique définit un ordre partiel sur les concepts clés d'un domaine en introduisant des relations de généralisation et de spécialisation entre les concepts (par exemple, le concept `boissons gazeuses` généralise le concept `soda` qui à son tour généralise les concepts `Coca` ou `Schweppes` ; inversement le concept `soda` spécialise le concept `boissons gazeuses`). Les taxonomies donnent accès à une abstraction consensuelle sur les concepts. Elles sont des outils centraux d'une grande variété d'applications qui s'appuient sur des connaissances expertes dans leur traitement informatique, *e.g.*, les systèmes d'information en biomédecine ou les systèmes d'aide à la décision clinique [1, 2]. Elles sont largement utilisées en Intelligence Artificielle, en Recherche d'Information, ou en Linguistique Computationnelle [3].

Dans notre étude, l'utilisation de la taxonomie des produits permet de synthétiser et de comparer la distribution des ventes des magasins. Dans la grande distribution, la taxonomie des produits est définie au niveau d'une enseigne car, d'une enseigne à l'autre, les strates intermédiaires diffèrent vite et seuls les niveaux les plus abstraits restent partagés (toutes les grandes surfaces auront au niveau le plus abstrait : `alimentaire` versus `non alimentaire`). Les éléments les plus tangibles des taxonomies correspondront aux unités de besoin. Le lecteur intéressé par des tentatives de leur déploiement dans la grande distribution pourra se référer à [4][5].

Plus formellement, nous considérons une taxinomie $T = (\leq, C)$ où C est l'ensemble des concepts (ou classes) et ($\leq$) l'ordre partiel. Nous désignons par $A(c) = \{x \in C / c \leq x\}$ et $D(c) = \{x \in C / x \leq c\}$ respectivement les ancêtres et les descendants du concept $c \in C$. La racine est l' unique concept sans ancêtre (sauf lui-même) ($A(root) = \{root\}$) et un concept sans descendant (sauf lui-même) est nommé feuille de la taxonomie (dans notre cas une feuille est donc un produit) et $D(leaf) = \{leaf\}$. Nous désignons également par *leaves-c* l'ensemble des feuilles (c'est-à-dire les produits dans notre étude) qui sont inclus dans le concept (la classe) c, c'est-à-dire *leaves-c* $= D(c) \cap leaves$.

4.1 Informativité basée sur la taxonomie

Un aspect intéressant des taxonomies est qu'elles donnent l'occasion d'analyser les propriétés intrinsèques et extrinsèques des concepts [8]. En effet, en analysant la topologie des taxonomies ou les statistiques sur l'usage des concepts, plusieurs travaux ont été consacrés à la définition de modèles pour le Contenu Informationnel (*IC*) de concepts [6]. Les modèles d'*IC* sont conçus pour imiter l'appréciation humaine, généralement consensuelle et intuitive, de l'informativité du concept. Par exemple, la plupart des gens conviendront que le concept `Concombre` est plus informatif que le concept `Légume` dans le sens où savoir qu'un client achète des concombres est plus informatif que de savoir qu'il achète des légumes. Diverses analyses basées sur des taxonomies, comme le calcul de la similarité entre concepts sur lequel nous reviendrons dans la section suivante, dépendent de modèles d'évaluation de l'*IC*.

Plus formellement, nous désignons par I l'ensemble des instances des concepts, et $I^*(c) \subset I$ les instances explicitement associées au concept c —ce qui exclut les instances des descendants. Autrement dit, nous considérons qu'aucune annotation associée à une instance ne peut être déduite, c'est-à-dire :

$$\forall c, c' \in C, avec\ c < c', I^*(c) \cap I^*(c') = \varnothing.$$

Nous désignons ensuite par $I(c) \subseteq I$ les instances qui peuvent être associées au concept c en considérant la transitivité de la relation taxonomique et l'ordre partiel $\leq$, *e.g.*, $I(Légume) \subseteq I(Aliment)$. On obtient donc $\forall c \in C, |I(c)| = \sum_{x \in D(c)} |I^*(x)|$.

Cette notion d'instance est généralement utilisée pour discuter de la spécificité d'un concept, c'est-à-dire à quel point un concept est restrictif au regard de I. Plus un concept est restrictif, plus il est considéré comme spécifique. Dans la littérature, la spécificité d'un concept est également considérée comme le Contenu Informationnel de celui-ci (*IC*). Dans cet article, l'*IC* est une fonction de C vers $\mathbb{R}^+$. Conformément aux contraintes de modélisation des connaissances, toute fonction *IC* doit décroître de façon monotone des feuilles à la racine de la taxonomie : $c \leq c' \Rightarrow IC(c) \geq IC(c')$.

Dans cet article, nous ne nous intéressons qu'aux propriétés extrinsèques des concepts pour estimer leur informativité (c'est-à-dire que l'on a recours à des informations sur les concepts qui ne sont pas portés par la topologie taxonomique). Le lecteur intéressé pourra se référer à [8] pour une comparaison des approches intrinsèques et extrinsèques de l'*IC*.

Dans notre approche, l'information exogène à la taxonomie est portée par le ticket de caisse. Celui-ci permet de « compter » les instances d'un concept (le nombre de fois que le produit est acheté, le chiffre d'affaires associé à ce produit) : en effet, seuls les produits apparaissent sur le ticket de caisse, et donc seules les instances de produits peuvent être observées en pratique. Ainsi, dans notre étude, les informations ne sont portées que par les feuilles de la taxonomie (produits dans notre cas) : $\forall c \notin leaves, \left|I^*(c)\right| = 0$.

L'approche extrinsèque de l'*IC* est basée sur la théorie de l'information de Shannon et propose d'évaluer l'informativité d'un concept en analysant ses occurrences dans une collection d'éléments. Initialement proposé par Resnik [6], l'*IC* d'un concept c est défini comme étant une fonction décroissante de $m(c)$, la « probabilité » que c se produise dans une collection, de documents en recherche d'information ou de produits dans le cas d'une enseigne.

La « probabilité » qu'une instance de I appartienne à $I(c)$ peut être définie ainsi :

$$m : 2^C \rightarrow [0;1] \text{ avec } m(c) = \frac{\left|I(c)\right|}{\left|I\right|} \quad (2)$$

L'informativité d'un concept est ensuite évaluée en définissant :

$$IC(c) = -\log m(c) \quad (3)$$

Notons que m n'est pas une distribution de probabilité puisque $m(root) = 1$ et que les éléments de C ne sont pas considérés comme disjoints pour m, mais une fonction de croyance contrairement à ce qui est parfois véhiculé en recherche d'information à partir des travaux de Resnik. Nous utilisons par la suite cette définition de l'*IC* extrinsèque dans notre proposition pour capturer la spécificité des concepts dans notre contexte d'application.

Revenons à la taxonomie de produits T et aux familles de produits F_i, les classes de cette taxonomie. Les produits en sont les feuilles. Les classes qui subsument directement les feuilles sont des familles de produits (par exemple `Soda`) auxquelles sont associés des niveaux d'assortiment ($(k_1, k_2, .., k_n) \in [1..9]^n$); les autres classes sont des super familles de produits (par exemple les `Boissons`). Déclinons les notions (2) et (3) dans notre modélisation en utilisant les chiffres d'affaires pour compter les instances :

$$\forall x \in leaves(T), \left|I^*(x)\right| = \left|I(x)\right| \triangleq \sum_\Omega p^\Omega(x)$$

$$m(x) = \sum_\Omega p^\Omega(x) / \sum_x \sum_\Omega p^\Omega(x)$$

$$IC(x) = -\log\left(m(x)\right) \quad (4)$$

$$\forall f \notin leaves, \left|I^*(f)\right| = 0, \left|I(f)\right| = \sum_{x \in leaves - f} \left|I(x)\right|$$

$$m(f) = \sum_{x \in leaves - f} \sum_\Omega p^\Omega(x) / \sum_x \sum_\Omega p^\Omega(x)$$

$$IC(f) = -\log\left(m(f)\right) \quad (5)$$

4.2 Mesures de similarité basées sur la taxonomie

Nous pouvons maintenant expliquer comment calculer la similarité de deux concepts sur la

base de l'informativité des concepts. Initialement, la mesure de similarité sémantique (SSM) a été conçue de manière « ad hoc » pour quelques domaines spécifiques et de nombreuses mesures ont été proposées dans la littérature [7, 8, 9]. Des recherches ont été effectuées afin d'obtenir un cadre unificateur théorique des SSM et pouvoir les comparer [10, 2]. Nous rappelons simplement ici l'une des mesures de similarité (SSM) basée sur le caractère informatif des concepts les plus utilisées en recherche d'informations. Elle est basée sur l'ancêtre commun le plus informatif (MICA). Par exemple, dans la figure 1, le MICA de `Coca-cola 1,5L` et `Schweppes 1,5L` est `Soda` tandis que le MICA de `Soda` et `Eaux` est `Boissons` (la racine de la figure 1). Resnik [6] est le premier à définir implicitement le MICA de deux concepts : c'est le concept qui subsume deux concepts c_1 et c_2 et qui a le plus grand IC (c'est-à-dire, l'ancêtre le plus spécifique):

$$sim_{\text{Resnik}}(c_1, c_2) = IC\left(MICA\left(c_1, c_2\right)\right) \quad (5)$$

Les SSM sont définies pour la comparaison de deux concepts. Cependant, la comparaison de deux magasins vus comme deux regroupements de classes de produits nécessite d'introduire une similarité de groupe pour comparer les deux sous-ensembles de concepts. Plusieurs mesures ont là encore été proposées [11, 12]. La *Best Match Average* (BMA) [11] est une des mesures les plus utilisées et est une moyenne composite entre deux ensembles de concepts A et B :

Si $sim_m(c, X) = \max_{c' \in X} sim\left(c, c'\right)$ et $sim\left(c, c'\right)$ une SSM alors,

$$sim_{BMA}(A, B) = \frac{1}{2|B|}\sum_{c \in B} sim_m(c, A) + \frac{1}{2|A|}\sum_{c \in A} sim_m(c, B) \quad (6)$$

Les SSM par paire et par groupe permettent de comparer deux sous-ensembles de concepts (ensembles de produits dans notre cas) lorsqu'une structure taxonomique a été définie au préalable.

De façon illustrative, si un magasin ne propose que du `Coca-Cola` dans son rayon `Boisson`, un autre uniquement du `Fanta` et un dernier simplement de la `Bière`, le MICA des deux premiers pourra être `Soda` alors que le MICA des deux premiers avec le dernier sera le concept plus abstrait de `Boisson`. La similarité entre les deux premiers sera alors perçue supérieure aux similarités avec le troisième. Dans notre étude, les SSM permettent de saisir l'idée que deux magasins Ω et Ω' sont similaires lorsque leurs chiffres d'affaires sont répartis de la même manière sur l'organisation hiérarchique des produits. Une première validation a été proposée dans [22].

5 Expérimentations

Pour garantir que ce modèle peut s'appliquer aux données massives de la grande distribution, nous avons construit trois benchmarks basés sur la taxonomie de Google. Des expériences ont été traitées sur 1 CPU IntelCore I7-2620M 2.7GHz et 8Go de RAM. Nous avons exploité la bibliothèque CPLEX (IBM CPLEX 1.25) et chaque benchmark requiert moins d'une seconde pour trouver la solution de l'équation (1). Ces benchmarks nous permettent d'affirmer que le traitement complet des données pour calculer l'assortiment idéal de magasins d'un réseau peut être réalisé à l'échelle de la grande distribution comme le montre le tableau 1. Les interprétations sémantiques de notre travail doivent encore être faites, mais nécessitent l'intervention d'experts de la grande distribution sur de longues périodes. Cette évaluation sort du cadre de cet article et sera réalisée dans le cadre de l'activité commerciale de TRF REtail.

Table 1 – Détails des Benchmarks

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

6 Conclusion

Le but de cet article est de proposer une méthodologie permettant d'améliorer l'assortiment de magasins de la grande distribution. L'objectif consiste à proposer aux magasins des produits adéquats en fonction de leurs contraintes spécifiques. L'approche que nous proposons inclut des connaissances a priori sur l'organisation hiérarchique des produits par famille pour formaliser le problème de l'assortiment en un problème combinatoire de sac à dos. La principale difficulté du problème d'optimisation reste l'estimation des bénéfices attendus d'une évolution de gamme pour une famille de produits. Cette estimation est basée sur les résultats comptables de magasins similaires du réseau. La définition de la similarité entre deux magasins est alors centrale et repose sur la connaissance a priori de l'organisation hiérarchique des produits. Cette structuration permet une comparaison abstraite qui rend mieux compte des comportements d'achats. Les SSM par paire et par groupe permettent de comparer deux sous-ensembles de concepts lorsqu'une structure taxonomique a été définie au préalable. Nous avons validé cette hypothèse dans le domaine du biomédical [21]. Enfin, la gestion de la taxonomie de produits réduit notablement l'espace de recherche du problème du sac à dos. Nos premières expérimentations montrent que le passage à l'échelle est tout à fait envisageable.

Références

[1] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain. The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies, *Bioinformatics*, **30**(5): 740-742, 2014.

[2] S. Harispe, D. Sanchez, S. Ranwez, S. Janaqi, J. Montmain. A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain, *J. of Biomedical Informatics*, **48**:38–53, 2014.

[3] S. Harispe, A. Imoussaten, F. Trousset, J. Montmain. On the consideration of a Bring-to-mind Model for Computing the Information Content of Concepts defined into Ontologies, *IEEE International Conference on Fuzzy Systems (FUZZ IEEE)*, Istanbul, Turkey, 2015.

[4] H. K. Kim, J. K. Kim, Q. Y. Chen. A product network analysis for extending the market basket analysis, *Expert Systems with Applications*, **39**(8):7403–7410, 2012.

[5] A. lbadvi, M. Shahbazi. A hybrid recommendation technique based on product category attributes. *Expert Systems with Applications*, **36**(9):11480–11488, 2009.

[6] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy, *in Proc. of IJCAI-95, Montréal, Canada,*1995.

[7] D. Sanchez and M. Batet. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of Biomedical Informatics*, **44**(5):749–759, 2011.

[8] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain. Semantic Similarity from Natural Language and Ontology Analysis, *Synthesis Lectures on Human Language Technologies, Morgan&Claypool Publishers*, 254 pages, 2015.

[9] S. Harispe, J. Montmain, M. Medjkoune. Summarizing conceptual descriptions using knowledge representations, *IEEE Symposium Series on Computational Intelligence (SSCI),* 2016.

[10] S. Janaqi, S. Harispe, J. Montmain, and S. Ranwez. Robust selection of domain-specific knowledge-based semantic similarity measures from uncertain expertise, *15th Int. Conference on Information processing and Management of uncertainty in Knowledge-Based Systems, IPMU, Montpellier, France,* 2014.

[11] A. Schlicker, F. S. Domingues, J. Rahnenfhrer, T. Lengauer. A new measure for functional similarity of gene products based on gene ontology, *BMC Bioinformatics*, **7**, 2006.

[12] C. Pesquita, D. Faria, H. Bastos, A. Falcao, F. Couto. Evaluating go-based semantic similarity measures, *In Proc. 10th Annual Bio-Ontologies Meeting, pp. 37-40*, 2007.

[13] E. Yücel, F. Karaesmen, F.S. Salman and M. Türkay. Optimizing product assortment under customer-driven demand substitution, *European Journal of Operational Research*, **199**(3):759-768, 2009.

[14] G. Kök, M.L. Fisher, R. Vaidyanathan. Assortment Planning: Review of Literature and Industry Practice, *Chapter in Retail Supply Chain Management, Eds. N. Agrawal and S.A. Smith, Kluwer Publishers,* 2006

[15] N. Agrawal, S.A. Smith. Optimal retail assortments for substitutable items purchased in sets, *Naval Research Logistics*, **50**(7):793-822, 2003.

[16] A. Alptekinoglu. Mass customization vs. mass production: variety and price competition, *Manufacturing & Service Operations Management*, **6**(1):98-103, 2004.

[17] P. Boatwright, J.C. Nunes. Reducing assortment: An attribute-based approach, *J. of Marketing*, **65**(3):50-63, 2001.

[18] C. Huffman C., B.E. Kahn. Variety for sale: Mass customization or mass confusion? *Journal of Retailing*, **74**:491-513, 1998.

[19] K. Pal. Ontology-Based Web Service Architecture for Retail Supply Chain Management. *European Journal of Operational Research,* **199**:759–768, 1998.

[20] S. Netessine, N. Rudi. Centralized and competitive inventory model with demand substitution, *Operational Research*, **51**:329–335, 2003.

[21] J. Poncelet, J., P.A. Jean, F. Trousset, S. Harispe, N. Pecheur, J. Montmain. Impact des mesures de similarité sémantique dans un algorithme de partitionnement : d'un cas biomédical à la détection de comportements de consommation, *26e Rencontres de la Société Francophone de Classification, Nancy, France*, 2019.

[22] Poncelet, J. Jean, P-A., Trousset, F., Montmain, J. Semantic Hierarchical Clustering: An Application in the Biomedical Domain, *19th International Conference on Artificial Intelligence and Soft Computing, (ICAISC 2020), Zakopane, Poland*.