Structuration semi-supervisée des données complexes*

Marian-Andrei Rizoiu
Laboratoire ERIC
Université Lumière Lyon 2
Lyon, France
Marian-Andrei.Rizoiu@univ-lyon2.fr

Résumé

L'objectif de la thèse est d'explorer la façon dont les données complexes peuvent être analysées en utilisant des techniques d'apprentissage non-supervisé, dans lesquelles de l'information supplémentaire est injectée pour guider le processus exploratoire. A partir de problèmes spécifiques, nos contributions prennent en compte les différentes dimensions des données complexes : leur nature (par exemple, image, texte), l'information additionnelle attachée aux données (par exemple, étiquettes, structure, ontologie de concepts) et la dimension temporelle. Le travail de recherche réalisé dans le cadre de cette thèse porte sur deux grandes problématiques : introduire des connaissances sémantiques dans la représentation des données et prendre en compte le temps dans le processus d'apprentissage.

1 Cadre général

Le Web initial (*i.e.*, le *Web 1.0*) était composé de pages statiques, que les utilisateurs pouvaient seulement consulter, et non pas modifier. L'arrivée du *Web 2.0* a donné la possibilité aux utilisateurs d'interagir avec les pages internet, de générer dynamiquement le contenu et de travailler d'une façon collaborative. Ce nouveau paradigme a contribué au changement de la façon dont l'information est produite, partagée et consommée. Les utilisateurs lissent, regardent, écoutent le contenu existant, puis ils réagissent, ils postent, ils décrivent et ils mettent des annotations, en enrichissant ainsi les informations disponibles Toute cette information disponible gratuitement est une source non-exhaustive de données. Les données originaires de l'internet sont juste un exemple d'une classe plus large de données, appelés **données complexes**. Les données complexes sont des données hétérogènes (par exemple, du texte, des images, des vidéos, de l'audio *etc.*), qui sont inter-reliés par la structure du document complexe dans lequel elle résident (par exemple la page web, dans le cas de données issues de l'internet). Ces données complexes ont une très grande dimensionnalité et, très souvent, ont une dimension temporelle. L'aspect temporel est particulièrement important dans le cas des articles de presse ou pour les réseaux sociaux en ligne.

Les difficultés rencontrées dans l'analyse de données complexes issues du *Web 2.0*, notamment les immenses quantités de données hétérogènes non-structurées et semi-structurées, sont les points centraux des principales applications liées à l'Internet. Des exemples de telles applications sont *la recherche d'information* (en anglais *Information Search and Retrieval*) : la recherche des information utiles dans la quantité énorme des données disponibles est encore aujourd'hui, pour un utilisateur moyen, la tâche la plus répandue sur internet. D'autres exemples sont la *catégorisation*, qui est l'effort collectif d'organiser l'information disponible pour obtenir, par exemple, des *folksonomies* comme Delicious [1], ou bien les *systèmes de recommandation*, qui ont pour but de recommander

*Thèse préparée sous la direction de Stéphane Lallich et Julien Velcin, professeurs à l'Université Lumière Lyon 2.

1. `https://delicious.com/`

de nouveaux contenus, en fonction des habitudes de l'utilisateur déduites du contenu qu'il regarde actuellement.

Les difficultés introduites par le *Web 2.0* ont conduit à l'émergence du *Web sémantique* [2], qui a pour mission de convertir la collection des documents non-structurés et semi-structurés qui composent aujourd'hui l'internet en une structure de documents inter-reliés, en incluant dans le contenu des documents web de la sémantique, dans un format lisible par la machine. Le but du *Web sémantique* est de fournir un cadre commun qui permet à l'information d'être partagée et réutilisée à travers les frontières des applications, des entreprises et des communautés. Ce processus implique la publication de l'information dans des langages créés spécialement pour les données sur l'internet, comme le RDF [3], le OWL [4] et le XML [5]. L'addition des descripteurs lisibles par la machine permet aux gestionnaires de contenu d'ajouter du sens à leur contenu et aux machines de traiter l'information au niveau sémantique, plutôt qu'au niveau du texte, obtenant ainsi des résultats plus significatifs. Cette information sémantique est recueillie dans des bases de connaissances, comme des ontologies librement accessibles (par exemple, DBpedia [6] [7] ou Freebase [7]). L'un des principaux défis du *Web sémantique* est d'obtenir une représentation sémantique des données. Le principal problème de la représentation des données de différentes natures (par exemple, image, texte) est que les caractéristiques de bas niveau utilisées pour représenter numériquement les données sont trop éloignées de la sémantique de leur contenu et n'arrivent pas à capturer leur signification.

Notre travail : enjeux de recherche et méthodes privilégiées. Le principal enjeu du travail de recherche de cette thèse est d'**utiliser de la sémantique lors de l'analyse de données complexes**. Dans les sections 4, 5 et 6, nous abordons les problématiques de l'introduction de connaissance humaine (par exemple, des annotation, des bases de connaissances) dans le processus d'apprentissage, ainsi que la reconstruction sémantique de l'espace de représentation des données. On distingue deux sous-problèmes : (i) *plonger les données dans un espace de représentation capable de capturer la sémantique sous-jacente aux données*, ce qui se traduit par la construction d'un espace de représentation qui intègre mieux la sémantique et qui peut être, ensuite, utilisé directement avec des algorithmes d'apprentissage machine classiques, et (ii) *injecter des connaissances externes dans les algorithmes d'apprentissage automatique*, ce qui implique la modification des algorithmes d'apprentissage machine afin de prendre en compte la sémantique dans le processus d'extraction des connaissances.

Un deuxième enjeu de cette thèse est de **prendre en compte la dimension temporelle des données complexes dans le processus d'analyse**. La dimension temporelle est plus que juste une autre dimension descriptive des données, car elle modifie profondément la définition du problème d'apprentissage. La description des données devient contextualisée (*i.e.*, la description peut n'être vraie que pendant un laps de temps donné) et de nouveaux problèmes se posent : suivre l'évolution temporelle des individus, détecter les tendances, le « burstiness » des thématiques dans les réseaux de microblogging (comme Twitter), le suivi des événements populaires à la mode, *etc.* La dimension temporelle est intimement liée à l'aspect interactif du Web 2.0. Nous décrivons les travaux que nous avons effectués concernant la dimension temporelle dans la section 3, dans laquelle nous développons un algorithme de clustering qui prend en compte le temps afin de construire des clusters cohérents du point de vue descriptif et temporel. Dans le travail présenté dans cette thèse, nous traitons chacun de ces enjeux de recherche de façon individuelle. Nous avons actuellement des travaux en cours (mentionnés dans la section 8), qui permettront de faire le lien entre les deux enjeux de recherche.

2. Le *Web sémantique* et le *Web 3.0* sont souvent utilisés comme synonymes, leur définition n'étant pas encore standardisée dans la littérature.
3. `http://www.w3.org/RDF/`
4. `http://www.w3.org/OWL/`
5. `http://www.w3.org/XML/`
6. `http://www.dbpedia.org`
7. `http://www.freebase.com/`

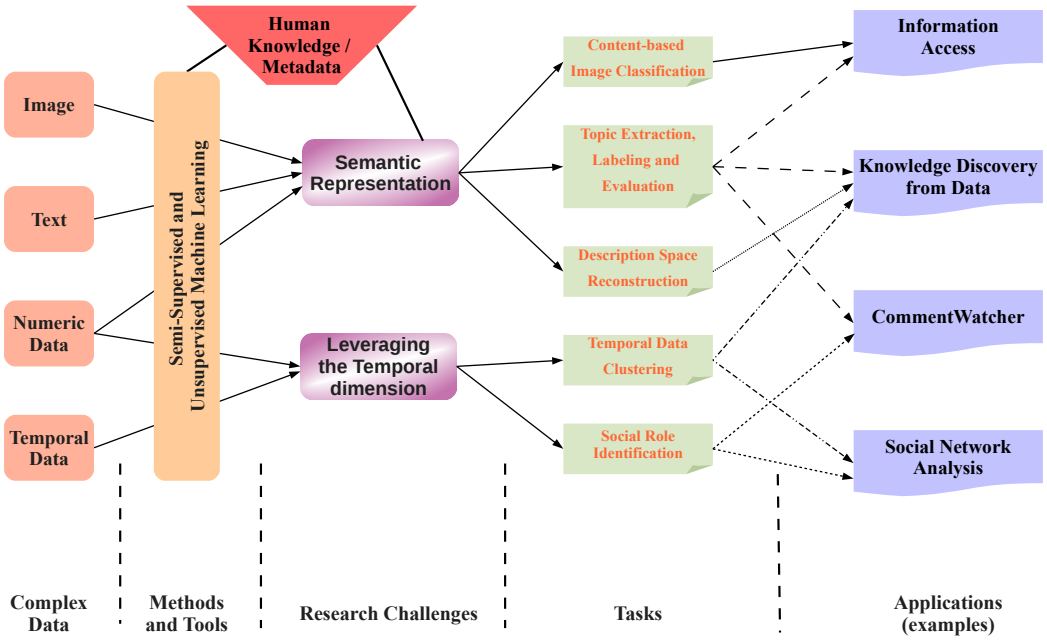

FIGURE 1 – Le schéma qui organise les travaux effectues dans le cadre de cette thèse : à partir des données complexes prises en entrée et se terminant par des exemples d'applications.

Les méthodes que nous privilégions en abordant les deux principales problématiques à la base de notre recherche sont le clustering non-supervisé et, notamment, le **clustering semi-supervisé**. Le clustering semi-supervisé [16] est une méthode essentiellement non-supervisée, dans laquelle des connaissances partielles sont injectées afin de guider le processus d'apprentissage. Contrairement à la classification semi-supervisée [11], dans laquelle l'accent est mis sur l'application des algorithmes supervisés en présence de petits corpora d'apprentissage, le clustering semi-supervisé est utilisé lorsque les informations additionnelles sont incomplètes ou en quantité trop faible pour que l'on puisse appliquer des techniques supervisées. Nous utilisons des techniques issues du clustering semi-supervisé pour modéliser et utiliser à la fois les connaissances additionnelles attachées aux données complexes et la dimension temporelle des données. Par conséquent, le contexte général de la recherche effectuée dans cette thèse se situe au croisement des domaines de l'**analyse de données complexes** et du **clustering semi-supervisé**.

Le reste de ce document est structuré de façon suivante. La fin de la présente section décrit le projet de recherche qui sous-tend cette thèse et résume les différentes contributions que nous avons apportées. Dans la section qui suit 2, nous donnons un aperçu des deux domaines autour desquels s'articule notre travail, l'analyse de données complexes et le clustering semi-supervisé. Dans les autres sections, nous présentons brièvement nos travaux concernant la dimension temporelle (section 3), la reconstruction sémantique de l'espace de représentation de données (section 4), la construction d'une représentation des images sémantiquement enrichi (section 5) et l'analyse de texte, notamment l'extraction, le nommage et l'évaluation de thématiques (section 6). Nous terminons par la section 8, où nous présentons quelques conclusions, des travaux en cours et les aspects pratiques de notre travail, notamment le logiciel `CommentWatcher`. Les travaux de recherche présentés dans cette thèse ont été effectués au sein du Laboratoire ERIC (Entrepôts, Représentation et Ingénierie des Connaissances) à Lyon, France, dans l'équipe Data Mining et Décision.

1.1 Projet de recherche

L'analyse de données complexes est un très vaste domaine, qui touche au traitement des images, au traitement du langage naturel, à l'intelligence artificielle et même à la sociologie (par exemple, la construction et l'analyse des réseaux sociaux en ligne). Nous avons donc dérivé les deux grandes problématiques de recherche et nous nous sommes concentré sur des tâches plus spécifiques, qui varient de *la détection des évolutions typiques* à *la reconstruction sémantique de l'espace de représentation des données* ou à *l'intégration des informations externes dans la description numérique des images*. Le projet de recherche sous-jacent à cette thèse a été construit progressivement, au travers d'une relation dialectique entre la théorie et la pratique. Les projets ou contrats de recherche dans lesquels j'ai été impliqué au fil de cette thèse ont soulevé plusieurs problèmes spécifiques (comme ceux mentionnées auparavant), qui ont souvent nécessité de traiter des données complexes (données hétérogènes de différentes natures, par exemple, texte, image) et l'intégration d'informations supplémentaires dans le processus d'apprentissage.

Les différents travaux de cette thèse sont illustrés et structurés de manière conceptuelle par la figure 1. Nous commençons, sur la partie gauche du schéma, avec un sous-ensemble de types de données complexes (*i.e.*, nous nous sommes intéressé aux données de type texte, image et numérique, ainsi qu'aux données ayant une dimension temporelle). Dans nos travaux, nous analysons chaque type de données indépendamment. Sur la partie droite du schéma sont illustrées les applications finales et abstraites de notre travail, par exemple, l'*accès à l'information*, l'*extraction des connaissances à partir des données*, l'*analyse des réseaux sociaux* ou `CommentWatcher`, un outil open-source pour l'analyse de média en ligne, qui est le résultat de nos travaux appliqués. Au milieu, nous présentons, de droite à gauche (plus précisément depuis l'objet de notre travail, *i.e.*, de la sortie vers l'entrée), (a) les tâches plus spécifiques sur lesquelles nous avons axé notre travail, (b) le enjeux de recherche qui en découlent dans les tâches spécifiques et (c) les méthodes et les outils que nous avons privilégiés afin d'atteindre les buts de notre recherche.

1.2 Parties constitutives de cette thèse

Compte tenu de la grande diversité des sujets abordés, nous présentons les travaux en quatre parties distinctes, pourtant complémentaires. Les quatre parties s'intéressent, respectivement, (a) à la dimension temporelle, (b) à la représentation sémantique des données, (c) aux données de type image et (d) aux données de type texte. Les différentes partie peut être considérées comme autonomes, tout en restant connectées entre eux par les **lignes directrices**, les **liens transversaux** et l'**articulation conceptuelle**. Par la suite, nous détaillons chacun de ces concepts.

Lignes directrices Les principaux enjeux de notre recherche sont traduits en lignes directrices, qui se retrouvent tout au long de notre travail : (i) compréhension des résultats par un utilisateur, (ii) plongement des données de natures différentes dans un espace de description sémantique et (iii) conception des algorithmes et des méthodes qui incorporent la sémantique et la dimension temporelle.

Dans chacune de nos propositions, nous considérons comme crucial de générer des **résultats compréhensible par un utilisateur humain**. Pour de nombreux problèmes , il existe des approches de type boîte noire (par exemple, l'analyse de composantes principales est une solution pour réorganiser l'espace de description), mais la signification sémantique de leur résultat est rarement évidente et, par conséquent, ces derniers sont difficiles à interpréter. Nos propositions sont développées pour être intelligibles par un esprit humain.

Une autre ligne directrice de nos travaux est de **plonger des données de natures différentes dans un espace de description sémantique**. La construction d'un tel espace de description généralement consiste à (a) ramener les données dans un format numérique commun, après avoir capturé les informations présentes dans le format natif, et à (b) utiliser efficacement des informations externes pour améliorer cette représentation numérique.

Enfin, un axe central de notre recherche est de **concevoir des algorithmes et des méthodes qui incorporent la sémantique et la dimension temporelle**, en se basant sur des techniques non-supervisées et semi-supervisées. Souvent, de l'information supplémentaire est jointe aux données, sous la forme de (a) annotations des experts/utilisateurs, (b) structure de documents inter-reliés ou (c) bases de connaissances externes (par exemple, des ontologies). Nous utilisons ces connaissances supplémentaires à plusieurs reprises, généralement en nous servant de techniques de clustering semi-supervisé. Nous utilisons également des contraintes semi-supervisées pour modéliser les dépendances temporelles dans les données.

Liens transversaux Il existe plusieurs **liens transversaux** entre les différentes parties de notre recherche. Notre travail sur les données textuelles est intimement lié avec le logiciel `CommentWatcher`. A partir des textes des forums de discussion en ligne que nous récupérons, nous extrayons les thématiques et nous construisons un réseau social en utilisant la relation de réponse spécifique aux forums. Le réseau social est modélisé et visualisé comme un multidigraphe, dans lequel les liens entre les noeuds sont associés aux thématiques. En plus, l'algorithme de clustering temporel que nous proposons est appliqué à la détection des rôles sociaux dans le réseau social que nous avons construit. Un autre lien transversal concerne notre algorithme de construction des attributs, qui a été initialement motivé par la nécessité de réorganiser une collection d'étiquettes définies par les utilisateurs et que nous utilisons afin de créer une représentation sémantique des images. Nous avons aussi des travaux en cours qui portent sur l' intégration de la dimension temporelle dans cet algorithme de construction des attributs. L'idée est de détecter si les attributs qui définissent les données sont corrélés avec un certain décalage dans le temps.

1.3 Contributions de la thèse

Les principales contribution de cette thèse sont articulées entre elles autour des deux principales problématiques à la base de notre recherche, construire un espace de représentation capable de capturer la sémantique sous-jacente aux données et prendre en compte l'aspect temporel des données. Ces deux problématiques de recherche se traduisent par des lignes directrices que l'on retrouve au travers notre travail : (a) obtenir des résultats faciles à interpréter par un esprit humain (b) plonger les données de différentes natures dans des espaces numériques capables de capturer la sémantique et (c) construire des algorithmes et méthodes prenant en compte des informations sémantiques en même temps que la dimension temporelle.

Par la suite, nous détaillons les contributions les plus importantes de notre travail de recherche.

Détecter les évolutions typiques Une de nos principales problématiques de recherche est de tirer parti de l'information temporelle dans le processus de regroupement non-supervisé. Dans [43], nous détectons les évolutions typiques des entités en proposant une nouvelle mesure de dissimilitude temporelle et une fonction de pénalité inspirée de la loi normale. La fonction de pénalité est utilisée avec des techniques de clustering semi-supervisé et a pour but d'encourager la segmentation contiguë des observations correspondant à une entité. Nous proposons un nouvel algorithme de clustering temporel, appelé TDCK-Means, qui crée une partition de clusters cohérents à la fois dans l'espace multidimensionnel et dans l'espace temporel.

Utiliser la sémantique des données pour améliorer l'espace de représentation Comme présenté dans la section 2.1, le traitement des données complexes de natures différentes (par exemple, image, texte) se résume habituellement à représenter les données dans un espace numérique et à appliquer des algorithmes classiques d'apprentissage. Nous jugeons crucial d'améliorer cet espace de représentation des données pour prendre en compte les relations sémantiques issues du jeu de données. Nous construisons, en utilisant des algorithmes non-supervisés, de nouveaux attributs qui

sont plus adaptés pour décrire l'ensemble des données tout en étant compréhensibles par un utilisateur. Nous proposons dans [44] deux algorithmes pour construire de nouveaux attributs sous la forme de conjonctions ou de négations d'attributs initiaux. Les attributs ainsi générés sont moins corrélés entre eux et mettent en valeur les relations sémantiques cachées entre les individus.

Améliorer la représentation des images en utilisant une construction semi-supervisée du vocabulaire visuel Une des façons les plus souvent utilisées pour traduire les images de leur format natif vers un format numérique est la représentation « sac-de-mots-visuels » (en anglais « bag-of-features »). Dans notre travail concernant les images, nous utilisons des connaissances expertes, sous la forme d'annotations, dans le processus de construction de la représentation numérique, à l'aide de techniques issues du clustering semi-supervisé. Nous proposons deux approches : dans la première nous construisons un vocabulaire visuel adapté pour décrire chaque objet qui apparaît dans la collection d'images, tandis que le second porte sur le filtrage des points d'intérêt qui ne concernent pas l'objet en cause.

Analyser des données textuelles : extraire et évaluer les thématiques Les données textuelles peuvent être transformées en format numérique en utilisant une représentation « sac-de-mots ». Une fois cette représentation mise en place, les thématiques extraites des textes peuvent être utilisées, par exemple, pour la construction automatique d'ontologies de concepts [42]. L'extraction des thématiques peut aussi être améliorée en utilisant des informations supplémentaires. Dans [36], nous montrons comment une hiérarchie de concepts peut être utilisée pour évaluer les thématiques extraites en utilisant des approches statistiques (par exemple, LDA [8]).

2 Présentation du domaine

2.1 L'analyse des données complexes

Il est difficile de traiter efficacement les données complexes [55], celles-ci étant de natures variées (*i.e.*, texte, image ou audio/vidéo) et étant profondément hétérogènes puisque pouvant provenir de sources multiples. Les données complexes sont souvent temporelles, puisque l'évolution des entités dans le temps peut être enregistrée. De plus, des informations additionnelles sont souvent attachées aux données, sous la forme d'annotations d'experts, de structure des documents interconnectés ou de bases de connaissances librement accessibles (par exemple, des ontologies comme DBpedia [7]).

Une méthode simple pour traiter les données complexes de différentes natures est de les représenter dans un espace numérique et d'appliquer des algorithmes classiques d'apprentissage automatique. La plupart des algorithmes d'analyse de données ont été développés pour utiliser des données décrites dans cet espace de représentation. Chaque document est représenté par un vecteur multidimensionnel, où chaque dimension correspond à une variable prédéfinie. Le défi actuel consiste à représenter les données de natures différentes dans un espace numérique capable de capturer l'information sémantique présente dans le format natif tout en utilisant efficacement de l'information externe pour enrichir la sémantique de cet espace.

Par la suite, nous synthétisons les spécificités des données complexes.

Différentes natures des données. (texte, image ou audio/vidéo) Analyser des données non-numériques pose des problèmes, parmi lesquels nous citons le fait que (a) elles ne sont pas directement « compréhensibles » par un ordinateur (*i.e.*, elles doivent être traduites d'abord dans un espace numérique) et (b) l'espace numérique dans lequel les données sont converties arrive à capter très peu d'information sémantique et, par conséquent, les algorithmes d'apprentissage automatique obtiennent de faibles performances. Certaines méthodes [27, 35] approchent cette problématique en utilisant des

données de différentes natures (par exemple, images et textes) afin de mieux guider le processus d'apprentissage.

L'information additionnelle (par exemple, des bases des connaissances externes) Des informations ou des ressources externes pourraient être disponibles pour compléter les informations sémantiques déjà présentes dans les données. Ces informations supplémentaires peuvent être sous la forme de (a) annotations fournies par des experts pour étiqueter les données ou (b) bases des connaissances inter-reliées (*i.e.*, des ontologies).

La dimension temporelle/dynamique. Il arrive souvent que la même entité est décrite selon les mêmes caractéristiques à différents moments du temps ou à des endroits différents (par exemple, un patient peut consulter plusieurs médecins, à différents moments du temps). Ces différentes données sont associées à la même entité et, du coup, les données complexes décrivent l'évolution de l'entité dans l'espace de description. Un type particulier de données temporelles est constitué par les données dynamiques, qui sont disponibles sous forme de flux (ces données ne peuvent être stockées et elles doivent être analysées en ligne).

La grande dimensionnalité. Prendre en compte des données de natures multiples et les bases externes de connaissances posent des problèmes de dimensionnalité. Ces problèmes de dimensionnalité peuvent soit porter sur les volumes de données qui doivent être traités (le problème de la « *scalabilité* »), ou, plus souvent, la grande dimensionnalité de l'espace de description (le « *malédiction de la dimension* »).

Les sources diverses et distribuées. Les données complexes peuvent provenir de sources différentes, qui, par ailleurs, n'ont pas besoin d'être co-localisées. Ceci n'est pas un problème nouveau (par exemple, depuis des temps anciens, la même information peut être trouvée dans des livres différents, dans des bibliothèques différentes), mais elle a été exacerbée avec l'arrivée du Web 2.0. L'information est aujourd'hui essentiellement distribuée dans de nombreuses sources, au lieu d'être centralisée dans les bibliothèques. Le paradigme de la récupération de l'information a également changé, passant de la classification (par exemple, trier des livres dans une bibliothèque basée sur un ensemble de critères) à la recherche (les moteurs modernes de recherche sur internet interrogent plusieurs bases des connaissances distribuées afin de compiler une réponse à une requête).

Parmi les méthodes permettant de prendre en compte l'information externe, nous avons privilégié les approches issues du clustering semi-supervisé que nous détaillons dans la section suivante.

2.2 Clustering semi-supervisé

Introduire des connaissances expertes partielles dans un algorithme d'apprentissage automatique relève du domaine de l'apprentissage semi-supervisé. Les connaissances partielles sont soit incomplètes (par exemple, le cas où des relations entre certains individus sont connues) ou tout simplement les exemples étiquetés ne sont pas en quantité suffisante pour que l'on puisse appliquer des algorithmes supervisés. Les domaines d'apprentissage semi-supervisé peuvent être globalement divisés en deux catégories :

– **la classification semi-supervisée** [11, 54] est une tâche essentiellement supervisée, qui permet d'apprendre à partir de données étiquetées et non étiquetées. Bien que ces approches puissent apprendre à partir d'un faible nombre d'exemples pour chaque catégorie, elles posent toujours un certain nombre de restrictions, parmi lesquelles (a) le nombre de catégories doit être fixé et connu à l'avance et (b) des exemples étiquetés doivent être présents pour chaque catégorie. Ces restrictions peuvent se révéler trop sévères pour les données complexes, pour lesquelles la liste des étiquettes pourrait ne pas être connue à l'avance (ou même le nombre

de classes pourrait être inconnu). En outre, les informations supervisées pourraient n'être disponible que sous la forme de quelques connexions entre des paires d'exemples (par exemple, on connaît le fait que deux personnes doivent être classées ensemble, mais aucune autre information n'est connue concernant la catégorie dans laquelle elles devraient être classées).

– **le clustering semi-supervisé** est une tâche d'apprentissage essentiellement non-supervisée, qui cherche notamment à guider le processus de clustering à l'aide des informations externes. Le clustering semi-supervisé est utile lorsque : (i) les classes ne sont pas connues à l'avance, (ii) il n'y a pas assez d'exemples disponibles pour chaque classe ou (ii) les connaissances experts ne sont pas représentatives pour les données. Nous considérons que cette approche est plus appropriée pour prendre en compte l'information supplémentaire intégrée dans des données complexes et, par conséquent, nous la présentons plus en détail dans la suite de cette section.

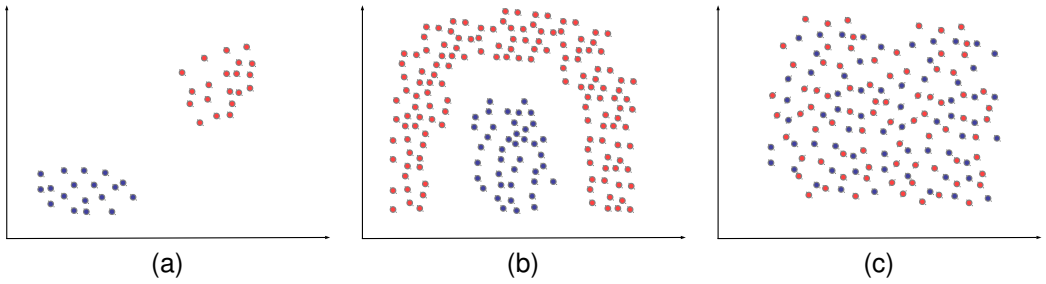

FIGURE 2 – Trois problèmes de clustering : (a) un problème facile, (b) un problème difficile et (c) un problème impossible.

Pourquoi guider le processus de clustering ? Les algorithmes traditionnels de clustering sont adaptés pour structurer les données des domaines jusque-là inconnus (par exemple, le problème de Yahoo ! [13]), mais ils ne parviennent pas à satisfaire nos attentes lorsque certaines connaissances existent sur le domaine. Par exemple, si les données forment naturellement des clusters compacts qui sont bien séparés (comme dans la figure 2a), il n'y a pas besoin de connaissances de base externes, la plupart des algorithmes de clustering peuvent détecter la partition souhaitée. De même, si aucune distinction ne peut être faite entre les classes dans l'espace de description (comme dans la figure 2c), alors peu d'information utile peut être trouvé dans les données elles-mêmes, et la supervision sera, de nouveau, de peu d'utilité. Les connaissances externes seront donc utiles lorsque des motifs sont présents dans les données, au moins partiellement, mais l'algorithme de clustering ne les détecte pas correctement sans assistance (comme on le voit dans la figure 2b). L'idée est d'utiliser les connaissances externes pour guider l'algorithme de clustering pour trouver la partition « correcte ».

Comment modéliser la supervision ? Les connaissances experts peut être modélisée soit par l'utilisation d'étiquettes de classe (*i.e.*, comme dans l'apprentissage supervisé), soit en utilisant des contraintes. Les contraintes peuvent être définies, par exemple, sur un sous-ensemble d'individus ou sur les clusters (utiles pour créer des clusters qui respectent certaines conditions, à part la cohésion implicite). Les contraintes les plus utilisés dans la littérature [16] sont les contraintes qui établissent des relations entre paires d'individus (en anglais « pairwise constraints »).

Dans [49], deux types de contraintes entre paires d'individus sont introduites. Une contrainte « must-link » entre les individus x et y signifie que, dans la partition créée, x et y doivent être placés dans le même cluster. De même, une contrainte « cannot-link » entre x et y signifie que les deux individus ne peuvent pas être placés dans le même cluster. Les contraintes must-link sont transitives

(*i.e.*, $(x, y) \in \mathcal{M}$ and $(y, z) \in \mathcal{M} \Rightarrow (x, z) \in \mathcal{M}$, où $\mathcal{M}$ est l'ensemble des contraintes must-link), ce qui signifie que l'ensemble des contraintes entre paires d'individus peut être enrichi avec de nouvelles contraintes en calculant la fermeture transitive de l'ensemble must-link.

Taxonomie Les algorithmes traditionnels de clustering utilisent une *mesure de similarité* donnée et une *stratégie de recherche* dans l'espace des solutions, afin de construire une partition formée de clusters cohérents. Les connaissances supervisés disponibles sont utilisées par les algorithmes de clustering semi-supervisé pour modifier la *mesure de similarité* ou la *stratégie de recherche* (ou les deux). Par conséquent, les méthodes de clustering semi-supervisée peuvent être divisées [4, 24] en deux classes : (a) les approches à base de similarité [2, 6, 30, 51], qui cherchent à apprendre de nouvelles mesures de similarité afin de satisfaire les contraintes et (b) les approches à base de recherche [3, 17, 23, 31, 50], dans lesquelles l'algorithme de clustering est lui-même modifié.

3 Détecter des évolutions typiques

Les travaux présentés dans cette section abordent une de nos principales problématiques de recherche, *i.e.*, utiliser la dimension temporelle dans l'analyse de données complexes, en employant des techniques de clustering semi-supervisé. Les solutions et les algorithmes présentés par la suite ont été élaborés pour résoudre une tâche d'apprentissage spécifique : *détecter les évolutions typiques dans l'ensemble des entités*. Ce problème spécifique a été motivé par les besoins des chercheurs en Sciences Politiques, impliqués dans le projet IMAGIWEB [8]. Nous utilisons également nos propositions dans une tâche d'apprentissage qui relève du domaine de l'*Analyse de Réseaux Sociaux* : détecter les rôles sociaux des utilisateurs dans un réseau social en ligne, construit à partir de forums de discussion en ligne.

Les chercheurs en Sciences Politiques ont toujours compilé des bases de données, dans lesquelles les informations ont souvent une composante temporelle : l'évolution d'un certain nombre d'entités est enregistrée sur une période de temps donnée. L'idée est d'injecter la dimension temporelle des données complexes dans un algorithme d'apprentissage automatique pour **détecter les évolutions typiques**. Cette idée est particulièrement intéressante pour les chercheurs en sciences politiques, car elle devrait permettre la détection de liens cachés entre l'évolution des entités et certains événements qui ont eu lieu plus tard dans le temps (par exemple, l'arrivé au pouvoir des leaders politiques extrémistes et les guerres ultérieures). Cette recherche n'est pas limitée aux sciences politiques, mais elle peut être généralisée à d'autres domaines des sciences sociales et humaines. En psychologie, il est bien connu que l'état mental d'un patient est très influencé par des événements traumatisants de son passé.

3.1 Formalisation

Les données que l'on analyse décrivent un ensemble d'entités $\phi_j \in \Phi$, où chaque entité est décrite pour chaque moment de temps considéré $t_m \in \mathcal{T}$ dans un espace de description multidimensionnel $\mathcal{D}$. Par conséquent, une entrée dans une telle base de données serait une observation, un triplet $(entité, moment\_temps, description)$. Par exemple, une base de données qui étudie l'évolution des états démocratiques [1] enregistre, pour chaque pays et chaque année, les valeurs pour des multiples indicateurs économiques, sociaux, politiques et financiers. Les pays concernés sont les entités et les années sont les périodes de temps.

Un des intérêts des chercheurs en sciences politiques est de détecter les évolutions typiques des individus dans une base de données créée comme mentionné ci-dessus. Il existe un double intérêt : a) obtenir une compréhension approfondie des phases que l'ensemble des entités traversent dans la

8. `http://eric.univ-lyon2.fr/~jvelcin/imagiweb`

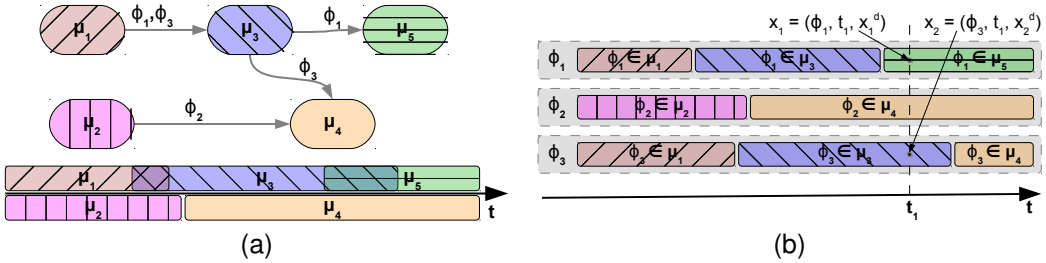

FIGURE 3 – Desired output : (a) the evolution phases and the entity trajectories, (b) the observations of 3 entities contiguously partitioned into 5 clusters.

période de temps (par exemple détecter les périodes d'instabilité politique mondiale, les crises économiques *etc.*) ; b) construire la trajectoire d'une entité à travers les différentes phases (par exemple, un pays aurait pu avoir traversé une période de dictature militaire, suivie d'une période de démocratie). Les critères qui décrivent chaque phase ne sont pas connus à l'avance (quels indicateurs annoncent une crise économique mondiale ?) et ils peuvent varier d'une phase à l'autre.

Nous abordons ces problématiques en proposant un algorithme de clustering temporel avec des contraintes, qui partitionnent les observations en clusters $\mu_j \in \mathcal{M}$, qui sont cohérents à la fois dans l'espace multidimensionnel de description et dans l'espace temporel. Les clusters obtenus servent à représenter les phases typiques de l'évolution des entités à travers le temps. Figure 3 montre le résultat souhaité de notre algorithme de clustering. Chacune des trois entités représentées (ϕ_1, ϕ_2 et ϕ_3) est décrite à 10 moments de temps ($t_m, m = 1, 2, ..., 10$) et les 30 observations de l'ensemble de données sont réparties en 5 clusters ($\mu_j, j = 1, 2, ..., 5$). Dans la figure 3a nous observons comment les clusters μ_j sont organisés dans le temps et la figure 3b montre comment la série d'observations appartenant à chaque entité sont attribuées à des clusters, formant ainsi des segments continus. Cette succession de segments est interprétée comme la succession des phases par lesquelles passe l'entité.

Nous identifions les condition que la partitions résultante doit remplir :

- **regrouper les observations ayant des descriptions similaires dans le même cluster** (tout comme le regroupement traditionnel le fait). Les clusters représentent un certain type d'évolution ;
- **créer des clusters temporellement cohérents, avec une étendue limitée dans le temps.** Pour qu'un cluster soit valable, il faut regrouper les observations qui sont temporellement proches. S'il existe deux périodes différentes avec des évolutions similaires, il est préférable de les avoir regroupées séparément, car elles représentent deux phases distinctes. En outre, même s'il est acceptable que certaines évolutions existent pendant toute la période étudiée, habituellement les clusters doivent avoir une étendue temporelle limitée ;
- **segmenter, le plus contigument possible, la série d'observations associées à chaque entité.** La séquence de segments sera interprétée comme la séquence des phases par lesquelles passe l'entité.

3.2 TDCK-Means, l'algorithme de clustering temporel

Mesure de dissimilarité temporelle Pour construire une partition qui respecte les conditions indiquées ci-dessus, nous proposons une nouvelle mesure de dissimilarité temporelle qui prend en compte tant la dimension temporelle que la dimension descriptive. Les observations qui sont proches dans l'espace de description, mais éloignées dans le temps sont considérées comme dissemblables. La mesure de dissimilarité temporelle que nous proposons $||x_i - x_j||_{TA}$ combine la distance Euclidienne dans l'espace multidimensionnel de description $\mathcal{D}$ et la distance temporelle

entre les dates. Nous proposons la définition suivante :

$$||x_i - x_j||_{TA} = 1 - \left(1 - \gamma_d \frac{||x_i^d - x_j^d||^2}{\Delta x_{max}^2}\right)\left(1 - \gamma_t \frac{||x_i^t - x_j^t||^2}{\Delta t_{max}^2}\right) \quad (1)$$

où $||\bullet||$ est la norme classique L^2 et Δx_{max} et Δt_{max} sont, respectivement, les diamètres de $\mathcal{D}$ et $\mathcal{T}$ (la plus grande distance entre deux observations dans l'espace multidimensionnel ou temporel).

γ_d et γ_t sont des paramètres qui permettent de régler le rapport entre la composante descriptive multidimensionnelle et la composante temporelle. γ_d et γ_t ne sont pas indépendants un de l'autre, leurs valeurs étant fixées en utilisant un paramètre unique α :

$$\gamma_d = \begin{cases} 1+\alpha, & \text{si } \alpha \leq 0 \\ 1, & \text{si } \alpha > 0 \end{cases} \; ; \; \gamma_t = \begin{cases} 1, & \text{si } \alpha \leq 0 \\ 1-\alpha, & \text{si } \alpha > 0 \end{cases} \quad (2)$$

α prend des valeurs entre -1 (prise en compte de la seule composante temporelle) et 1 (prise en compte de la seule composante descriptive).

Contraintes de contiguïté Nous proposons également une méthode pour renforcer la contiguïté de la segmentation des observations associées à une entité, en introduisant un terme de pénalisation inspiré de la densité de la distribution normale. La fonction de pénalisation encourage les observations qui sont proches dans le temps à être affectées au même cluster. Nous utilisons des *contraintes entre paires d'observations*, provenant de la littérature de clustering semi-supervisé, et nous ajoutons des contraintes « must-link » entre toutes les paires d'observations appartenant à la même entité. L'algorithme est autorisé à briser les contraintes, mais une pénalité est infligée pour chaque transgression. La pénalité est d'autant plus sévère que les observations sont proches dans le temps. Nous proposons la fonction de pénalité suivante :

$$w(x_i, x_j) = \beta * e^{-\frac{1}{2}\left(\frac{||x_i^t - x_j^t||}{\delta}\right)^2} \mathbb{1}\left[x_i^\phi = x_j^\phi\right] \quad (3)$$

où β est un paramètre ayant pour objet de contrôler l'impact de la fonction de contiguïté, et la valeur maximale que prend la fonction, alors que δ est un paramètre qui contrôle la largeur de la fonction. β est dépendant du jeu de données et il peut être fixé comme un pourcentage de la distance moyenne entre les observations. $\mathbb{1}\left[statement\right]$ est une fonction qui renvoie 1 si $statement$ est vrai et 0 dans le cas contraire.

TDCK-Means Nous combinons les deux propositions dans un algorithme de clustering temporel avec contraintes, TDCK-Means, qui crée une partition de clusters cohérents, à la fois dans l'espace multidimensionnel et dans l'espace temporel. Les clusters que l'algorithme construit servent de phases d'évolution. L'algorithme TDCK-Means suit un structure similaire à l'algorithme classique K-Means. L'algorithme cherche à minimiser la fonction objectif $\mathcal{J}$, définie dans l'équation 4, en effectuant de multiples itérations dans lesquels deux phases sont alternées jusqu'au point où la partition ne change plus entre deux itérations successives. Dans une première phase, les observations sont affectées aux clusters et dans une deuxième phase, les centroïdes des clusters sont recalculés, en se fondant sur l'affectation précédente. La mesure objective utilisée est :

$$\begin{aligned} \mathcal{J} = |\mathcal{X}| - \sum_{j=1}^{k} \sum_{x_i \in \mathcal{C}_j} \left[\left(1 - \gamma_d \frac{||x_i^d - \mu_j^d||^2}{\Delta x_{max}^2}\right)\left(1 - \gamma_t \frac{||x_i^t - \mu_j^t||^2}{\Delta t_{max}^2}\right)\right] \\ + \sum_{x_i \in \mathcal{X}} \sum_{x_k \notin \mathcal{C}_j} \beta * e^{-\frac{1}{2}\left(\frac{||x_i^t - x_j^t||}{\delta}\right)^2} \mathbb{1}\left[x_i^\phi = x_j^\phi\right] \end{aligned} \quad (4)$$

où $\mathcal{X}$ est l'ensemble de toutes les observations dans le jeu de données. Visiblement, elle inclut la mesure de dissimilarité temporelle définie dans l'équation 1 et la fonction de pénalisation définie dans l'équation 3.

3.3 Évaluation

Jeu de données Les expérimentation avec TDCK-Means ont été effectuées sur un jeu de données issues du domaine des Sciences Politiques *Comparative Political Data Set I* [1]. Ce jeu de données est une collection de données politiques et institutionnelles, qui consiste en des données annuelles pour 23 pays démocratiques, durant la période 1960 à 2009, et pour chaque année, les valeurs de 207 indicateurs politiques, démographiques, sociales et économiques sont enregistrées.

Le jeu de données a été nettoyé en enlevant les variables redondantes et, ensuite, le corpus a été prétraité en éliminent certaines caractéristiques prédominantes des entités. Par exemple, il est difficile de comparer, sur les données brutes, l'évolution de population entre des pays très peuplés (par exemple, la Chine) et des pays avec moins d'habitants (par exemple, l'Andorre), puisque toute évolution sur la période donnée de 50 années de l'ensemble de données sera éclipsée par la différence initiale. Inspiré des méthodes de l'économétrie de panels [18], nous supprimons les effets propres à l'entité, invariables dans le temps. Nous soustrayons de chaque valeur de la moyenne sur chaque attribut et sur chaque entité. Nous conservons la composante variable dans le temps, qui est à son tour normalisée, afin d'éviter de donner trop d'importance à certaines variables. L'ensemble de données obtenu se présente sous la forme de triplets $(pays, année, description)$.

Mesures d'évaluation Le jeu de données choisi ne contient pas de données étiquetées. Par conséquent, nous utilisons des mesures classiques de la théorie de l'information afin d'évaluer numériquement l'algorithme proposé. Nous évaluons séparément chacun des trois objectifs que nous proposons dans la section 3.1

Pour évaluer **la cohérence des clusters dans l'espace de définition multidimensionnel**, nous utilisons la variance moyenne intra-cluster, défini comme :

$$MDvar = \frac{1}{|\mathcal{X}|} \times \sum_{j=1}^{k} \sum_{x_i \in \mathcal{C}_j} ||x_i^d - \mu_j^d||^2$$

Nous utilisons également la variance intra-cluster pour évaluer **la cohérence temporelle des clusters** :

$$Tvar = \frac{1}{|\mathcal{X}|} \times \sum_{j=1}^{k} \sum_{x_i \in \mathcal{C}_j} ||x_i^t - \mu_j^t||^2$$

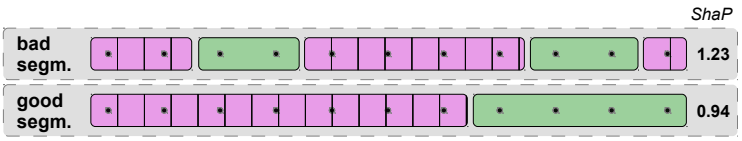

FIGURE 4 – Examples of a good and a bad segmentation in contiguous chunks and their related *ShaP* score.

Pour évaluer la **contiguïté de la segmentation des observations appartenant à un individu**, nous proposons une mesure basée sur l'entropie de Shannon, adaptée pour prendre en compte l'alternance entre deux phases. Un exemple d'une bonne segmentation et d'une mauvaise segmentation est donné dans la figure 4, avec leurs scores respectifs *ShaP* (Shannon Pénalisé). Ce score est défini

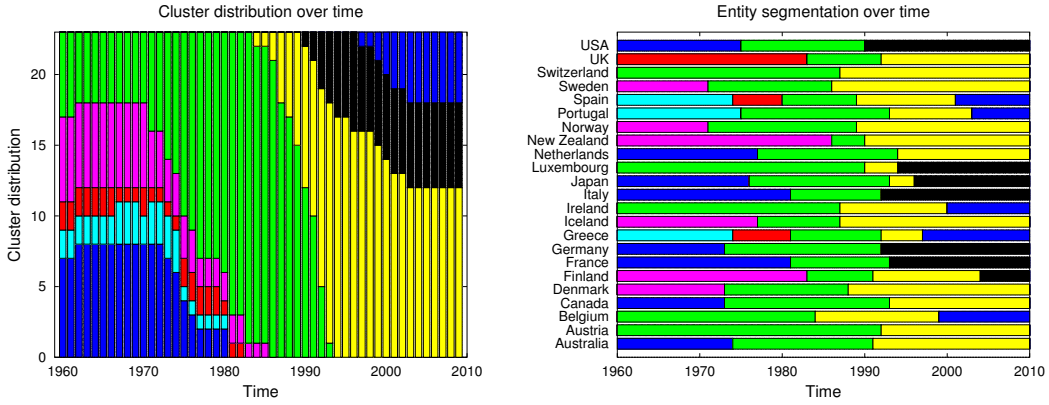

FIGURE 5 – Les évolutions typiques construites par TDCK-Means sur le jeu de données *Comparative Political Data Set I* avec 8 phases. (a) Combien d'entités sont associées à un cluster donné pour chaque année et (b) la segmentation des observations associées à chaque entité.

par :

$$ShaP = \frac{1}{|\mathcal{X}|} \times \sum_{x_i \in \mathcal{X}} \sum_{j=1}^{k} \left(-p(\mu_j) \times \log_2(p(\mu_j)) \times \left(1 + \frac{n_{ch} - n_{min}}{n_{obs} - 1} \right) \right)$$

où n_{ch} est le nombre de changements dans la série d'affectations aux clusters des observations appartenant à un individu, n_{min} est le nombre minimal de changements requis et n_{obs} est le nombre d'observations associées à une entité.

Évaluation qualitative Dans la figure 5, nous montrons les évolutions typiques construites par TDCK-Means (avec les paramètres $\beta = 0.003$ et $\delta = 3$), avec 8 phases. La figure 5a montre combien d'entités sont associées à un cluster donné pour chaque année. Les clusters μ_5 et μ_6 contiennent le plus d'observations, ce qui suggère que $\mu_5 \longrightarrow \mu_6$ est une évolution typique.

La signification de chaque cluster ne commence à se démêler que lorsqu'on étudie la segmentation des observations associées aux pays par rapport aux clusters, montrée dans la figure 5b. Par exemple, le cluster μ_2 regroupe les observations appartenant à l'Espagne, le Portugal et la Grèce, de 1960 jusque vers 1975. Historiquement, cela coïncide avec des régimes non démocratiques dans ces pays (la dictature de Franco en Espagne, le « régime des colonels en Grèce »). De même, le cluster μ_4 contient des observations de pays comme le Danemark, la Finlande, l'Islande, la Norvège, la Suède et la Nouvelle-Zélande. Ce groupe peut être interprété comme le « modèle socio-économique suédois ». Dans la seconde partie de la période étudiée, le cluster groupe μ_8 regroupe les observations de la Grèce, l'Irlande, l'Espagne, le Portugal et la Belgique, les pays qui semblaient le plus fragiles pendent la crise économique de 2008.

Évaluation quantitative Pour l'évaluation quantitative, nous avons comparé 5 algorithmes : (1) **Simple K-Means** ; (2) **Temporal-Driven K-Means** - combine K-Means avec la mesure de dissimilarité temporelle ; (3) **Constrained K-Means** - ajoute les contraintes de contiguïté aux K-Means ; (4) **TDCK-Means** et (5) **tcK-Means** - un algorithme de clustering avec contraintes proposé dans [31].

Nous présentons dans le tableau 1, dans le cas de chacun des 5 algorithmes, la moyenne et l'écart-type (présenté en italique) des différents indicateurs pour plusieurs valeurs de k, Les algorithmes Simple K-Means, Temporal-Driven K-Means et Constrained K-Means étant conçus pour

TABLE 1 – Les valeurs moyennes pour les indicateurs et leur écart-type.

Algorithm	*MDvar*		*Tvar*		*ShaP*	
Simple K-Means	**120.59**	*2.97*	48.01	*8.87*	2.15	*0.23*
Temp-Driven K-Means	122.98	*2.85*	**19.97**	*5.39*	2.58	*0.18*
Constrained K-Means	132.69	*8.07*	103.15	*42.98*	**1.24**	*0.5*
TDCK-Means	127.81	*3.96*	27.54	*5.81*	2.06	*0.2*
tcK-Means	123,04	*3.8*	62.44	*24.16*	1.79	*0.32*

optimiser principalement un seul critère, il n'est pas étonnant qu'ils montrent les meilleurs résultats pour, respectivement, la variance multidimensionnelle, la variance temporelle et l'entropie (meilleurs résultats en gras). TDCK-Means cherche à offrir un compromis et il obtient, dans deux cas sur trois, le deuxième meilleur score. Il est à noter que la mesure de dissimilarité temporelle, utilisé dans Temporal-Driven K-Means, offre la meilleure stabilité (le plus faible écart-type moyen) pour tous les indicateurs. Cependant, les algorithmes avec contraintes (Constrained K-Means et TCK-Means) montrent une forte instabilité, en particulier pour *Tvar*. Globalement, TDCK-Means montre une très bonne stabilité.

4 Utiliser la sémantique des données pour améliorer leur représentation

Une des principales problématiques de notre recherche, énoncées dans la section 1, est d'utiliser la sémantique dans l'analyse des données complexes. Cette section porte sur une tâche d'apprentissage cruciale : **la construction d'un espace de représentation des données, avec une sémantique enrichie**. Alors que dans les autres parties de notre travail (*i.e.*, dans les sections 3, 5 et 6), nous utilisons soit des connaissances expert externes, soit la dimension temporelle des données afin de déduire une connaissance plus complète, dans le travail présenté dans la suite de cette section, nous nous concentrons sur la sémantique déjà disponible dans les données elles-mêmes.

Motivations Dans le contexte de l'apprentissage automatique, pour être utile, un attribut doit apporter des nouvelles informations par rapport aux autres attributs. Les attributs corrélés n'apportent pas d'information l'un par rapport à l'autre, mais leur cooccurrence est souvent le résultat d'une relation sémantique entre les deux. Par conséquent, notre travail concernant la reconstruction de la représentation des données a deux missions : (a) améliorer l'espace de représentation en enlevant les corrélations entre les attributs et (b) découvrir des liens sémantiques entre les attributs en analysant les cooccurrences dans les données. Pour faire face à ces défis, nous proposons un nouvel algorithme non-supervisé, **uFC**, qui améliore l'espace de représentation en réduisant la corrélation totale entre les attributs, tout en découvrant les liens sémantiques entre les attributs grâce à la construction de nouveaux attributs. Les paires d'attributs initiaux avec une forte corrélation sont remplacées par des conjonctions booléennes et les cooccurrences sémantiquement induites dans le jeu de données initial sont mises en évidence.

4.1 Pourquoi construire un nouveau ensemble des attributs ?

Un attribut p_j qui est fortement corrélé avec un autre attribut p_i n'apporte aucune information nouvelle, puisque la valeur de p_j peut être déduite de celle de p_i. Par conséquent, nous pouvons filtrer les attributs « non pertinents » avant d'appliquer l'algorithme d'apprentissage. Mais si on supprime simplement certains attributs, on court le risque de perdre de l'information importantes

FIGURE 6 – Exemples des images annotées avec $\{groupes, route, bâtiment, intérieur\}$.

des **liens sémantiques entre les attributs**, et c'est la raison pour laquelle nous avons choisi de **construire de nouveau attributs**, au lieu de simplement filtrer les attributs corrélés.

Notre travail a pour objet les jeux de données dans lesquels les données sont décrites en utilisant des attributs booléens, parce que dans la plupart des jeux de données réels les attributs binaires ont des significations particulières. Par exemple, une collection d'images est annotée avec un ensemble d'étiquettes et chaque attribut booléen indique la présence (valeur **vrais**) ou l'absence (valeur **faux**) d'un certain objet dans l'image. Dans une telle liste d'objets (par exemple, eau, $cascade$, $manifestation$, $urbain$, $groupes$ et $intérieur$) une partie de la structure sémantique de l'ensemble des attributs peut être devinée facilement, parce que des relations de type « c'est-un » ou « partie-de » sont assez intuitives : une $cascade$ c'est un type d'eau, une $patte$ est une partie d'un $animal$, *etc.* Mais d'autres relations peuvent être induites par la sémantique des données (*i.e.*, les images dans notre exemple), par exemple il y a une cooccurrence entre $manifestation$ et $urbain$, car les manifestations ont généralement lieu dans la ville. Figure 6 représente un jeu de données d'images, qui est décrit en utilisant les attributs $\{groupes, route, bâtiment, intérieur\}$. L'ensemble des attributs est assez redondant et certaines ne sont informatives (par exemple, l'attribut $groupes$ est présent pour tous les individus). Compte tenu de cooccurrences entre les attributs, nous pourrions créer les nouveaux attributs plus éloquents *personnes à l'intérieur et non sur la route* ($groupes \wedge \neg route \wedge intérieur$, décrivant la rangée du haut) et *personnes sur la route avec des bâtiments sur le fond* ($groupes \wedge route \wedge bâtiment$, décrivant la rangée du bas). L'idée sous-jacente est de créer un ensemble d'attributs, dépendant des données, de telle sorte que les nouveaux attributs ont une faible cooccurrence entre eux.

4.2 Nos propositions

Pour améliorer les résultat des algorithmes d'apprentissage automatique, des approches existent dans la littérature (par exemple, SVM [14], PCA [19] *etc.*) qui traitent le problème des attributs non adéquats en modifiant l'espace de description. Le principal inconvénient de ces approches est qu'elles fonctionnent comme une boîte noire, où le nouvel espace de représentation est soit caché (pour SVM) soit totalement synthétique et non directement compréhensible pour un esprit humain (PCA). La littérature propose également des algorithmes qui construisent de nouveaux attributs fondés sur les attributs primitifs, mais, à notre connaissance, tous ces algorithmes construisent un nouvel espace de représentation de manière supervisée, à partir de l'information de classe, fournie

a priori avec les données.

La nouveauté de nos propositions Par rapport aux solutions existantes dans la littérature, nos solutions novatrices ont deux grands avantages : en plus de construire un espace de représentation dans lequel les variables cooccurrent moins, elles (a) produisent des nouveau attributs compréhensible pour l'utilisateur humain et (b) fonctionnent sans exemples pré-étiquetés, d'une manière non supervisée. Le premier algorithme que nous proposons est une adaptation d'un algorithme supervisé, en le transformant en non-supervisé. Pour le deuxième algorithme, nous avons développé une nouvelle heuristique qui choisit, à chaque itération, des paires d'attributs fortement corrélés et qui les remplace par des conjonctions de littéraux. Comme résultat, la redondance globale de l'ensemble des attributs est réduite. Les itérations ultérieures créent des formules booléennes plus complexes, qui peuvent contenir des négations. Nous utilisons des considérations statistiques (tests d'hypothèses) pour déterminer automatiquement la valeur des paramètres en fonction de l'ensemble de données, et nous évaluons nos propositions à l'aide d'une méthode inspirée du *front du Pareto* [45].

uFRINGE Nous proposons **uFRINGE**, une version non-supervisée de FRINGE [38], un algorithme qui construit de nouveaux attributs en utilisant la sortie d'un algorithme d'induction d'arbres de décision, comme ID3 [40] ou C4.5 [41]. Pour rendre FRINGE non-supervisé, nous remplaçons l'algorithme d'induction des arbres de décision par un algorithme non-supervisé qui construit des arbres de clustering [9].

Limitations Cependant, uFRINGE a quelques inconvénients. L'ensemble des attributs construits a tendance à être redondant et à contenir des doublons. Les attributs nouvellement construits sont ajoutés à l'ensemble des attributs et ils sont utilisés, à côté des anciennes variables, dans les itérations ultérieures. Les anciennes variables ne sont jamais retirés de l'ensemble des attributs et elles peuvent être combinées à plusieurs reprises, ce qui entraîne des doublons dans l'ensemble des attributs construit.

uFC - une heuristique gloutonne Nous dépassons les limites d'**uFRINGE** en proposant un deuxième algorithme innovant appelé **uFC**. Notre approche réduit la corrélation globale entre les attributs en remplaçant, d'une façon itérative, les paires d'attributs fortement corrélés par des conjonctions de littéraux. Nous utilisons une stratégie de recherche gloutonne afin d'identifier les variables qui sont fortement corrélées entre elles et, par la suite, nous utilisons un opérateur de construction pour créer de nouveaux attributs. A partir de deux attributs corrélés f_i et f_j, nous créons trois nouvelles variables : $f_i \wedge f_j$, $f_i \wedge \overline{f_j}$ et $\overline{f_i} \wedge f_j$. Ensuite, f_i et f_j sont tous deux retirés de l'ensemble des attributs. L'algorithme s'arrête lorsqu'il n'y a plus de nouvelles variables créés ou lorsqu'il a effectué un nombre maximum d'itérations fixé à l'avance.

La figure 7 illustre visuellement, à l'aide de diagrammes de Venn, comment l'algorithme remplace les anciennes variables par de nouvelles variables. Les attributs sont représentés par des rectangles, chaque rectangle contenant les individus ayant un certain nombre d'attributs avec la valeur **vrai**. Naturellement, les individus situés à l'intersection de deux rectangles ont les deux attributs fixés à **vrai**. La figure 7a montre la configuration initiale de l'ensemble des attributs : f_1 et f_2 ont une grande intersection, ce qui signifie qu'ils apparaissent souvent ensemble. Au contraire, f_2 et f_5 ont une petite intersection, ce qui suggère que leur co-occurrence est inférieure à celle due au hasard (corrélation négative). f_3 est inclus dans l'intersection de f_1 et f_2, tandis que f_4 n'a pas d'éléments en commun avec les autres (f_4 est incompatible avec tous les autres attributs). Le but de l'algorithme est de construire un nouvel ensemble d'attributs, dans lequel il n'y a pas d'intersection entre les diagrammes de Venn correspondants. A la première itération (c.f. figure 7b), f_1 et f_2 sont combinés et 3 nouvelles variables sont créées : $f_1 \wedge f_2$, $f_1 \wedge \overline{f_2}$ et $\overline{f_1} \wedge f_2$. Ces nouvelles variables vont remplacer les attributs originaux f_1 et f_2. Lors de la deuxième itération (c.f. figure 7c), $f_1 \wedge f_2$ est combiné avec f_3. Comme f_3 est contenu dans $f_1 \wedge f_2$, la variable $\overline{f_1 \wedge f_2} \wedge f_3$ aura un support égal à zéro et

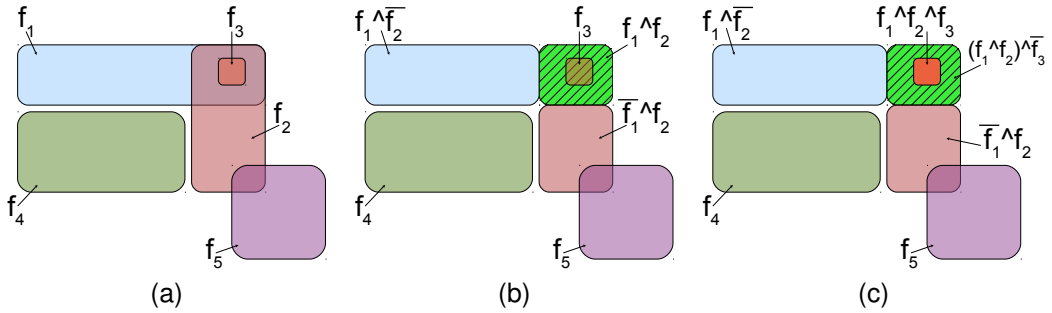

FIGURE 7 – Représentation graphique du processus de construction des nouveaux attributs en utilisant des diagrammes de Venn. (a) Iter. 0 : variables initiales (les primitives), (b) Iter. 1 : combinaisons de f_1 et f_2 and (c) Iter. 2 : Combinaisons de $f_1 \wedge f_2$ et f_3.

elles sera supprimée. Notez que f_2 et f_5 ne sont jamais combinés car ils sont considérés comme non corrélés. Le nouvel ensemble d'attributs sera donc $\{f_1 \wedge \overline{f_2}, f_1 \wedge f_2 \wedge f_3, f_1 \wedge f_2 \wedge \overline{f_3}, \overline{f_1} \wedge f_2, f_4, f_5\}$.

Chercher les paires d'attributs corrélés Pour trouver les paires d'attributs corrélés, l'algorithme effectue une énumération intelligente de toutes les paires d'attributs $(f_i, f_j) \in F \times F$. Afin de mesurer la corrélation entre deux variables, nous utilisons le coefficient de corrélation de Pearson (**r**), mesure classique pour évaluer la dépendance linéaire entre deux attributs. Cette mesure prend ses valeurs entre -1 et 1 ; une valeur de 0 indique une absence de corrélation linéaire entre les deux variables. Quand la corrélation d'une paire de variables est supérieure à un certain seuil λ, les deux attributs sont jugés comme corrélés et ils sont ajoutés à l'ensemble des paires corrélées O. Formellement, nous avons :

$$O = \{(f_i, f_j) \in F \times F,\, i \neq j \,|\, \mathbf{r}((f_i, f_j)) > \lambda\} \tag{5}$$

où le paramètre λ sert à régler le nombre des paires sélectionnées. Nous proposons, dans le paragraphe suivant, une méthode basée sur un test d'hypothèse statistique afin de déterminer automatiquement la meilleure valeur pour λ.

Construction et filtrage des attributs Après avoir construit l'ensemble O de paires corrélées, on parcourt toutes les paires en suivant l'ordre décroissant du score de corrélation. A partir d'une paire (f_i, f_j), nous construisons trois nouveaux attributs : $f_i \wedge f_j$, $\overline{f_i} \wedge f_j$ et $f_i \wedge \overline{f_j}$. Ces nouveaux attributs sont garantis, par construction, d'être corrélés négativement. f_i et f_j peuvent être soit des attributs initiaux, soit des attributs plus complexes construits lors des itérations précédentes. Chaque itération construit des attributs à l'aide d'opérateurs très simples (conjonction de deux littéraux). Cependant, des attributs complexes et plus riches du point de vue sémantique apparaissent au fil des itérations.

Après avoir construit les nouveau attributs, nous enlevons de l'ensemble O la paire (f_i, f_j) et toutes autres paires contenant f_i ou f_j. A la fin de chaque itération, nous filtrons l'ensemble des attributs construits pour enlever : (a) les attributs qui ont un support de zéro (les attributs qui prennent la valeur « faux » pour tous les exemples), et (b) les attributs qui ont participé à la construction des nouveau attributs (les nouveaux attributs remplacent les anciens). Autrement dit :

$$\{f_i, f_j \in F | (f_i, f_j) \in O\} \xrightarrow{\text{remplacé par}} \{f_i \wedge f_j, \overline{f_i} \wedge f_j, f_i \wedge \overline{f_j}\}$$

Choix automatique du paramètre λ Le paramètre λ, introduit dans l'équation 5, est très dépendent du jeu de données considéré et difficile à déterminer de manière générale. Nous proposons

de le supprimer en introduisant une technique qui choisit seulement les paires d'attributs pour lesquelles la corrélation est jugée significative du point de vue statistique. Nous utilisons pour chaque paire d'attributs candidate une méthode statistique, le test d'hypothèse, où nous confrontons l'hypothèse d'indépendance H_0 à l'hypothèse de corrélation positive H_1. Pour effectuer le test statistique, nous choisissons d'utiliser le coefficient de corrélation de Pearson. Formellement, nous testons les hypothèses $H_0 : \rho = 0$ et $H_1 : \rho > 0$, où ρ est le coefficient théorique de corrélation entre les deux attributs candidats. On peut montrer que, dans le cas d'attributs booléens, la valeur observée du χ^2 d'indépendance est $\chi^2_{obs} = n \times r^2$ (n est la taille du jeu de données). Par conséquent, en considérant comme vraie l'hypothèse H_0, $n \times r^2$ suit approximativement une distribution du χ^2 avec un degré de liberté ($n \times r^2 \sim \chi^2_1$). Comme résultat $r\sqrt{n}$ suit une distribution normale ($r\sqrt{n} \sim N(0,1)$).

En conséquence, nous rejetons l'hypothèse H_0 en faveur de l'hypothèse H_1 si et seulement si $r\sqrt{n} \geq u_{1-\alpha}$, où $u_{1-\alpha}$ est la valeur critique à droite de la distribution normale. Les deux attributs candidats sont considérés comme significativement corrélés quand $r(f_i, f_j) \geq \frac{u_{1-\alpha}}{\sqrt{n}}$. Le niveau de signification α représente le risque de rejeter l'hypothèse d'indépendance H_0 alors qu'elle était vraie en réalité.

4.3 Mesures d'évaluation et quelque résultats

Corrélation totale d'un ensemble d'attributs Afin d'évaluer la corrélation totale d'un ensemble d'attributs, nous proposons une mesure inspirée de la formule de Poincaré [22]. Dans sa forme booléenne, cette formule est utilisée pour calculer la cardinalité d'une réunion finie d'ensembles finis, et cela en fonction du nombre d'éléments de ces ensembles et de leurs intersections. Étant donné un ensemble d'attributs $F = \{f_1, f_2, ..., f_m\}$, sa formulation généralisée est comme suit :

$$p(f_1 \vee f_2 \vee ... \vee f_m) = \sum_{k=1}^{m} \left((-1)^{k-1} \sum_{1 \leq i_1 < ... < i_k \leq m} p(f_{i_1} \wedge f_{i_2} \wedge ... \wedge f_{i_k}) \right)$$

En nous basant sur cette formule, nous proposons l'**indice de recouvrement** (OI pour *Overlapping Index*) comme suit :

$$OI(F) = \frac{\sum_{i=1}^{m} p(f_i) - 1}{m - 1}$$

où $OI(F) \in [0, 1]$ et doit être minimisé.

Complexité d'un ensemble d'attributs Souvent, réduire la corrélation totale d'un ensemble d'attributs revient à augmenter la taille de cet ensemble. Si l'on considère que la paire (f_i, f_j) est corrélée, à l'exception du cas où $f_i \supseteq f_j$ ou $f_i \subseteq f_j$, alors l'algorithme remplacera $\{f_i, f_j\}$ par $\{f_i \wedge f_j, \overline{f_i} \wedge f_j, f_i \wedge \overline{f_j}\}$ en augmentant le nombre total des attributs. Comme le nombre maximal des attributs pouvant être construits est limité par le nombre des individus qui composent le jeu de données ($|F| \leq unique(I) \leq |I|$), nous proposons la mesure suivante pour évaluer la complexité d'un ensemble d'attributs :

$$C_0(F) = \frac{|F| - |P|}{unique(I) - |P|}$$

où P est l'ensemble des attributs dits primitifs (les attributs initiaux). $C_0(F) \in [0, 1]$ et doit être minimisé.

Recherche d'un compromis entre deux critères opposés La corrélation totale d'un ensemble de données (mesurée par OI) et sa complexité (mesurée par C_0) sont associées à des critères opposés qu'il n'est pas possible d'optimiser simultanément. Obtenir un compromis entre des critères opposés est un problème classique dans le domaine de l'optimisation multicritère. Nous choisissons d'utiliser le concept d'**optimalité de Pareto** [45] afin de déterminer notre solution. Une solution

TABLE 2 – Ensemble des attributs construits par **uFC** avec les heuristiques « point le plus proche » et « basée sur le risque ».

primitives	**uFC*(0.194, 2)**	**uFC$_\alpha$(0.001)**
$person$	$groups \wedge \overline{road} \wedge interior$	$groups \wedge \overline{road} \wedge interior$
$groups$	$\overline{groups \wedge road \wedge interior}$	$\overline{groups \wedge road \wedge interior}$
$water$	$\overline{groups \wedge road} \wedge interior$	$\overline{groups \wedge road} \wedge interior$
$cascade$	$water \wedge cascade \wedge tree \wedge forest$	$water \wedge cascade \wedge tree \wedge forest$
sky	$\overline{water \wedge cascade} \wedge \overline{tree \wedge forest}$	$\overline{water \wedge cascade} \wedge \overline{tree \wedge forest}$
$tree$	$\overline{water \wedge cascade} \wedge tree \wedge forest$	$\overline{water \wedge cascade} \wedge tree \wedge forest$
$grass$	$sky \wedge building \wedge \overline{tree \wedge forest}$	$sky \wedge building \wedge \overline{tree \wedge forest}$
$forest$	$sky \wedge building \wedge tree \wedge \overline{forest}$	$sky \wedge building \wedge tree \wedge \overline{forest}$
$statue$	$\overline{sky \wedge building} \wedge tree \wedge forest$	$\overline{sky \wedge building} \wedge tree \wedge forest$
$building$	$sky \wedge \overline{building} \wedge panorama$	$sky \wedge \overline{building} \wedge panorama$
$road$	$sky \wedge \overline{building} \wedge \overline{panorama}$	$sky \wedge \overline{building} \wedge \overline{panorama}$
$interior$	$\overline{sky \wedge building} \wedge panorama$	$\overline{sky \wedge building} \wedge panorama$
$panorama$	$\overline{groups} \wedge road \wedge person$	$\overline{groups} \wedge road \wedge person$
	$\overline{groups} \wedge road \wedge \overline{person}$	$\overline{groups} \wedge road \wedge \overline{person}$
	$\overline{groups} \wedge road \wedge person$	$\overline{groups} \wedge road \wedge person$
	$water \wedge \overline{cascade}$	$\overline{\mathbf{sky}} \wedge \mathbf{building} \wedge \mathbf{groups} \wedge \mathbf{road}$
	$\overline{\mathbf{sky}} \wedge \mathbf{building}$	$\overline{\mathbf{sky} \wedge \mathbf{building}} \wedge \overline{\mathbf{groups} \wedge \mathbf{road}}$
	$\overline{tree} \wedge forest$	$\overline{\overline{\mathbf{sky}} \wedge \mathbf{building}} \wedge \mathbf{groups} \wedge \mathbf{road}$
	$\mathbf{groups} \wedge \mathbf{road}$	$water \wedge \overline{cascade}$
	$grass$	$\overline{tree} \wedge forest$
	$statue$	$grass$
		$statue$

est considérée comme Pareto optimale si et seulement s'il n'existe pas une autre solution avec à la fois un meilleure score de corrélation *et* un meilleur score de complexité. L'ensemble des solutions Pareto optimales forment le front de Pareto.

En pratique, nous faisons varier les paramètres du système et nous plongeons les solutions obtenues dans le plan défini par OI et C_0. Ensuite, nous utilisons deux heuristiques pour choisir la « solution optimale » : (a) heuristique dite du « point le plus proche », où nous choisissons sur le front de Pareto la solution la plus proche du point idéale déterminé par les coordonnées $(0; 0)$; (b) heuristique « basée sur le risque », où nous combinons la méthode pour choisir la valeur du paramètre λ avec une condition d'arrêt fixée sur le nombre d'itérations : les mesures OI et C_0 sont combinées dans un seul indicateur en utilisant la moyenne quadratique, avant d'itérer tant que la valeur de cette moyenne quadratique continue à diminuer.

Évaluation qualitative Dans le tableau 2, nous montrons l'ensemble des attributs qui peuvent être construits avec notre proposition à partir d'un ensemble de primitives. **uFC*(0.194, 2)** représente l'exécution de notre algorithme avec les paramètres choisis en utilisant l'heuristique « point le plus proche » et **uFC$_\alpha$(0.001)** est l'exécution avec l'heuristique « basée sur le risque » et un risque de 0,001 . Nous avons mis en gras les attributs qui différent entre les deux exécutions.

Évaluation quantitative Le tableau 3 montre une comparaison quantitative entre les deux heuristiques proposées ci-dessus. Nous utilisons un risque de 0.001 pour le jeux de données `hungarian` et 0.0001 pour `spect` et `street` (à cause de la dimension de jeux de données). Les ensemble

TABLE 3 – Heuristiques dites « point le plus proche » et « basée sur le risque » .

	Strategy	λ	$limit_{iter}$	$\#feat$	$\#common$	$length$	OI	C_0
hung.	**Primitives**	-	-	13	-	1.00	0.235	0.000
	uFC*(0.194, 2)	0.194	2	21	19	2.95	0.076	0.069
	uFC$_\alpha$(0.001)	0.190	2	22		3.18	0.071	0.078
street	**Primitives**	-	-	66	-	1.00	0.121	0.000
	uFC*(0.446, 3)	0.446	3	87	33	2.14	0.062	0.038
	uFC$_\alpha$(0.0001)	0.150	1	90		1.84	0.060	0.060
spect	**Primitives**	-	-	22	-	1.00	0.279	0.000
	uFC*(0.432, 3)	0.432	3	36	19	2.83	0.086	0.071
	uFC$_\alpha$(0.0001)	0.228	2	39		2.97	0.078	0.086

des attributs construits par les deux heuristiques sont très similaires, pas seulement les différences pour l'OI, le C_0, la longueur moyenne des attributs et la dimension du set des attributs sont négligeables, mais la plupart des attributs créés sont identiques. Pour `hungarian`, 19 attributs à partir de 21 créés par les heuristiques sont identiques. Cela montre que la heuristique « basée sur le risque » arrive à des résultats très similaires à ceux créés par la heuristique dite « point le plus proche », sans avoir besoin de varier les paramètres et de ré-exécuter l'algorithme **uFC** un grand nombre de fois.

5 Analyser des données issues d'images : construction sémantique du vocabulaire visuel

Cette section aborde l'un des problème centraux de nos travaux : comment utiliser des connaissances sémantiques lors de l'analyse de données complexes. Nous nous intéressons en particulier aux données de type image et, pour être plus précis, nous cherchons à construire une représentation numérique des images avec une sémantique enrichie. L'une des tâches d'apprentissage les plus importantes lorsque l'on traite des images est *la classification supervisée basée sur le contenu*. C'est une tâche particulièrement difficile, surtout parce que les caractéristiques de bas niveau utilisées pour décrire numériquement les images captent en général peu d'information sur leur sémantique. Dans notre travail, nous choisissons d'aborder ce problème en enrichissant la sémantique associée à la représentation des images en utilisant des connaissances externes. L'hypothèse sous-jacente est que la création d'une représentation basées sur une sémantique enrichie permet d'obtenir des performances en apprentissage plus élevées, et ce sans qu'il soit nécessaire de modifier les algorithmes d'apprentissage eux-mêmes. Pour tester notre hypothèse, nous appliquons notre proposition à la tâche de classification supervisée basée sur le contenu, et nous montrons que l'enrichissement sémantique de la représentation des images améliore les performances en classification.

Le format habituel pour stocker des images sur un support informatique est une matrice composée de pixels. Or, ce genre de caractéristiques « bas niveau » apporte très peu d'information concernant le contenu sémantique de l'image. L'une des représentations qui présente des résultats très prometteurs est la représentation en « sac-de-caractéristiques » (en anglais *bag-of-features* ou *BoF*), inspirée de la représentation textuelle « sac-de-mots » (an anglais *bag-of-words*) ; cette représentation utilisée pour les textes est présentée dans la section 6.1. Notre proposition consiste justement à utiliser de l'information experte, sous la forme d'annotations non positionnelles, afin d'améliorer la sémantique d'une représentation de type *BoF*. Nous introduisons cette information additionnelle au niveau de la construction du vocabulaire visuel. Pour cela, nous proposons deux nouvelles contributions qui s'appuient sur des informations sémantiques externes et qui permettent au vocabulaire

visuel de cerner plus précisément la sémantique qui peut être associée à une collection d'images. La première proposition porte sur l'introduction de l'information supplémentaire tôt dans la création du vocabulaire visuel, ce qui permet de construire un vocabulaire visuel dédié aux images annotées avec un tag donné. Dans la deuxième proposition, nous ajoutons une phase de filtrage comme prétraitement dans la construction du vocabulaire visuel. L'idée est d'éliminer les points d'intérêt qui ont de faibles chances d'appartenir à un objet donné et d'augmenter ainsi la précision du processus de classification qui suit.

5.1 Construire une représentation numérique de type « sac-de-caractéristiques »

Typiquement, la construction d'une représentation *BoF* est un processus composé de quatre phases, comme le montre la figure 8. A partir d'une collection $\mathcal{P}$ contenant n images, le but est de plonger ces images dans un espace numérique dans lequel les algorithmes sont les plus efficaces. Dans *la phase 1*, chaque image $p_i \in \mathcal{P}$ est échantillonnée et l_i caractéristiques [9] sont extraites. Les techniques d'échantillonnage les plus populaires sont celles basées sur une grille dense [21, 48] et des détecteurs de points d'intérêt [15, 21, 46]. Dans *la phase 2*, en utilisant un descripteur local, comme le SIFT [32] ou le SURF [5], chaque caractéristique est décrite à l'aide d'un vecteur à h dimensions [10] Par conséquent, après cette phase, chaque image p_i est décrite numériquement par $V_i \subset \mathbb{R}^h$, l'ensemble des vecteurs à h dimensions décrivant les caractéristiques échantillonnées à partir de p_i.

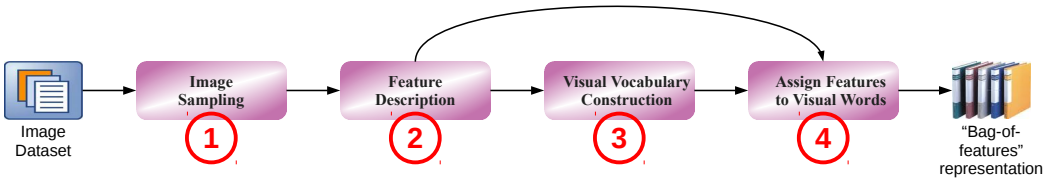

FIGURE 8 – Schéma de construction d'une représentation numérique de type « sac-de-caractéristiques » pour décrire des images.

Ensuite, sur la base des caractéristiques extraites et décrites dans les phases antérieures, *la phase 3* consiste à construire le vocabulaire visuel. La technique employée est généralement un algorithme de type *clustering*. Le vocabulaire visuel est une collection de M mots visuels ; ces mots sont décrits dans le même espace numérique que celui des caractéristiques visuelles et ils servent de bases à l'espace numérique dans lequel les images sont traduites. Plus précisément, les centroïdes des clusters créés par l'algorithme de clustering servent de mots visuels. Dans *la phase 4*, chaque caractéristique échantillonnée est affectée à l'un de ces mots visuels. Par conséquent, chaque image est décrite comme une distribution sur les mots visuels en utilisant l'un des systèmes de pondération de termes inspirés de la fouille de textes (par exemple, tf, $tf\text{x}idf$, etc.). La description numérique qui en résulte peut ensuite être utilisée pour des tâches de classification, d'extraction d'information ou d'indexation.

5.2 Enrichir la sémantique d'une représentation numérique de type *BoF* en utilisant des connaissances externes

Nous présentons à présent deux nouvelles méthodes qui s'appuient sur des informations sémantiques externes, fournies sous la forme d'*étiquettes d'objet* non positionnelles, afin d'enrichir la sémantique du vocabulaire visuel. Notre travail se situe dans un contexte de faible supervision,

9. l_i dépend du contenu de l'image (nombre d'objets, forme, *etc.*) et de l'algorithme d'extraction utilisés. l_i peut varier de quelques centaines de caractéristique allant jusqu'à plusieurs dizaines de milliers.
10. par exemple, pour le descripteur SIFT $h = 128$.

similaire à celui définit par [52], où chaque étiquette signale la présence d'un objet donné dans une image, mais pas sa position, ni sa forme ou sa taille. De plus, seule une fraction de l'ensemble des images est étiquetée et nous utilisons, tout à la fois, les images étiquetées et non étiquetées pour construire la représentation comportant une sémantique enrichie. Par conséquent, notre travail se positionne clairement dans un contexte semi supervisé.

Pour chacune des étiquettes, nous construisons un *vocabulaire visuel dédié* qui se base uniquement sur les images associées à une étiquette en particulier. Ce type d'approches a montré [28, 39] qu'il améliore la précision par rapport à un vocabulaire généraliste, grâce au fait que les vocabulaires spécialisés contiennent des mots visuels capables de décrire de manière plus appropriée les objets qui apparaissent dans la collection d'images. Pour notre deuxième approche, nous améliorons encore la précision en proposant une phase de prétraitement qui filtre les caractéristiques visuelles susceptibles de ne pas être associées à un objet donné. Cette proposition suit l'idée de l'algorithme de reconnaissance d'objets proposé dans [32]. On utilise en effet un ensemble d'exemples positifs et un ensemble d'exemples négatifs, construits chacun sur la base des informations d'étiquette. Le prétraitement de filtrage est finalement combiné avec la construction des *vocabulaires visuels dédiés*. Nos expérimentations montrent que cette approche permet d'obtenir systématiquement un gain de précision, à la fois avec un vocabulaire spécialisé (sans filtrage) et avec un vocabulaire généraliste.

Construire un vocabulaire visuel dédié L'idée qui se trouve derrière l'utilisation d'une représentation de type *BoF* est que les mots visuels ont un pouvoir prédictif pour certains objets. La qualité des mots visuels (et leur pouvoir prédictif) peut être améliorée s'ils sont construits seulement à partir des caractéristiques extraites de leurs objets respectifs. Cela permet d'éliminer les éléments caractéristiques du fond de l'image ou d'autres objets. Dans un contexte de faible supervision, le contours des objets est inconnu ; mais pouvoir sélectionner uniquement les images qui contiennent un certain objet (information connue grâce à l'étiquette) augmente le rapport entre le nombre de caractéristiques pertinentes et le bruit. Par conséquent, les mots visuels construits de cette manière permettent d'obtenir des descriptions plus précises pour les objets désignés par les étiquettes. C'est pourquoi nous proposons de construire un *vocabulaire visuel dédié* pour chaque étiquette $t_i \in \mathcal{T}$, c'est-à-dire généré à partir des caractéristiques extraites des images marquées avec l'étiquette t_i.

Nous distribuons les mots visuels de manière uniforme entre les étiquettes et nous construisons k vocabulaires spécialisés, chacun contenant m/k mots visuels. Chaque vocabulaire dédié est créé en utilisant une approche *BoF* standard, comme expliqué dans la section 5.1. Pour une étiquette donnée t_i, nous créons $\mathcal{C}_i$, l'ensemble de toutes les caractéristiques extraites à partir des images marquées avec t_i. Formellement, cela donne :

$$\mathcal{C}_i = \bigcup_{\substack{j=1 \\ y_{j,i}=1}}^{n_1} V_j$$

où V_j est l'ensemble des caractéristiques extraites à partir de l'image p_j. L'ensemble des mots visuels qui en résulte est plus représentatif pour décrire l'objet désigné par l'étiquette t_i. A la fin de l'algorithme, nous fusionnons les *vocabulaires spécialisés* pour obtenir un vocabulaire visuel général unique. Cette constructions garantit que le vocabulaire visuel généré contient les mots visuels représentatifs pour tous les objets annotés avec les étiquettes de $\mathcal{T}$.

Filtrage des points d'intérêt Dans cette partie, nous détaillons un mécanisme de filtrage pour augmenter encore davantage le ratio entre les caractéristiques pertinentes et le bruit. Nous l'appliquons comme une phase de prétraitement à la technique de construction des vocabulaires dédiés, présentée précédemment : nous filtrons ainsi les caractéristiques visuelles qui sont susceptibles de ne pas être reliées à l'objet en question. Étant donnée une image $p_i \in \mathcal{P}_1$, nous construisons deux collections d'images auxiliaires : l'*ensemble des exemples positifs*, qui contient uniquement les

(a) (b) (c)

FIGURE 9 – (a) Image avec l'annotation « moto », (b) image appartenant à l'*ensemble des exemples positifs* et (c) image appartenant à l'*ensemble des exemples négatifs*

images étiquetées de manière identique à p_i, et l'*ensemble des exemples négatifs*, qui contient les images qui ne partagent aucune annotation commune avec p_i. Nous définissons alors KP_{p_i} comme l'ensemble des caractéristiques échantillonnées à partir des images qui se trouvent dans l'ensemble des exemples positives, et KN_{p_i} comme l'ensemble des caractéristiques extraites de l'ensemble des exemples négatifs :

$$KP_{p_i} = \{f^+ \in V_j \,|\, \forall\, t_l \in \mathcal{T} \text{ pour laquelle } y_{i,l} = 1 \implies y_{j,l} = 1\}$$
$$KN_{p_i} = \{f^- \in V_j \,|\, \forall\, t_l \in \mathcal{T} \text{ pour laquelle } y_{i,l} = 1 \implies y_{j,l} = 0\}$$

Prenons le cas d'une caractéristique visuelle extraite à partir de l'image p_i ($f \in V_i$) qui est davantage similaire aux caractéristiques de l'ensemble des exemple négatifs ($f^- \in KN_{p_i}$) qu'aux caractéristiques de l'ensemble des exemples positifs ($f^+ \in KP_{p_i}$). Cette caractéristique a plus de chances d'appartenir à l'arrière plan de l'image p_i qu'à l'un des objets annotés dans l'image ; elle peut, par conséquent, être filtrée. Nous utilisons la *distance euclidienne* pour mesurer la similarité entre deux caractéristiques visuelles, décrites numériquement à l'aide d'un descripteur SIFT : $||f_1 - f_2|| = \sqrt{\Sigma_{i=1}^{h} (f_{1,i} - f_{2,i})^2}$. Formellement, pour une caractéristique f extraite à partir de l'image p_i, nous avons :

$$f \in V_i \text{ est filtrée } \Leftrightarrow \nexists f^+ \in KP_{p_i} \text{ tel que } ||f - f^+|| \leq \delta$$
$$\text{avec } \delta = \alpha \times \min_{f \in KN_{p_i}} ||f - f^-|| \tag{6}$$

où δ est un seuil de filtrage et $\alpha \in \mathbb{R}^+$ est un paramètre qui permet de régler ce seuil de filtrage. Ces valeurs correspondent à la distance entre la caractéristique f et la caractéristique la plus proche provenant de l'ensemble des exemples négatifs. La caractéristique f est considérée comme similaire à une caractéristique $f^+ \in KP_{p_i}$ de l'ensemble des exemples positifs si et seulement si la distance $||f - f^+||$ est plus petite que le seuil de filtrage. Par conséquent, une caractéristique f est filtrée quand elle n'a pas de caractéristique similaire dans l'ensemble des exemples positifs.

Prenons l'exemple d'une collection d'images représentées dans la figure 9. Les images 9a et 9b sont étiquetées avec « moto », tandis que l'image 9c est annotée avec « ville ». L'image cible 9a possède des bâtiments en arrière plan, et toute caractéristique échantillonnée de cette région ne sera pas pertinente pour l'objet moto. L'image 9b sert d'*ensemble des exemples positifs*, tandis que 9c sert d'*ensemble des exemples négatifs*. Prenons l'exemple de deux caractéristiques f_1, échantillonnée à partir de la roue de la moto (en vert), et f_2, prélevée à partir des bâtiments situés en arrière-plan (en rouge), de l'image cible. Pour f_1, il existe au moins une caractéristique similaire dans l'ensemble des exemples positifs, tandis que f_2 n'a pas de caractéristique similaire dans cet ensemble. En conséquence de quoi f_2 est éliminée, car elle n'est pas jugée pertinente pour l'objet « moto ».

5.3 Évaluation

Protocole expérimental Le but du protocole d'évaluation est de quantifier le gain de performance obtenu en enrichissant la sémantique d'une représentation des images de type « *BoF* baseline » (comme celui présenté dans la section 5.1) en utilisant des connaissances expertes. L'évaluation est réalisée dans le contexte d'une tâche d'apprentissage de classification supervisée des images basée sur le contenu. Pour chaque image, nous construisons plusieurs représentation numériques, correspondant chacune à une technique particulière. Ensuite, le même algorithme de classification, utilisant les mêmes paramètres, est appliqué sur chacune de ces représentations. La différence constatée dans les performances est alors imputée à la pertinence des représentations.

Construction de représentations numériques Nous construisons chaque représentation numérique comme présenté dans la figure 8. Les phases 1, 2 et 4 sont identiques pour chaque représentation. Dans *la phase 1*, nous extrayons les caractéristiques visuelles en utilisant un détecteur de régions Hessian-Affines et chaque caractéristique est décrite dans *la phase 2* à l'aide de descripteurs SIFT. *La phase 3* diffère suivant la représentation employée : le vocabulaire visuel est construit soit (a) en utilisant des caractéristiques tirées au hasard (cas appelé **random**), (b) à l'aide d'un algorithme simple de clustering (méthode « baseline » **random+km**), (c) avec notre proposition de construction des vocabulaires visuels dédiés (appelé **model**), ou (d) avec nos propositions de filtrage et de construction du vocabulaire visuel dédié (appelé **filt+model**). Dans *la phase 4*, chaque caractéristique visuelle est associée à un mot visuel ; la représentation *BoF* est construite et nous appliquons ensuite un classifieur SVM [14] ou un classifieur à base de clustering pour classer les images sur la base de leur contenu.

Jeux de données Nous évaluons nos proposition sur trois bases d'images : `Caltech101` [20], `RandCaltech101` [29] et `Caltech101-3`. `RandCaltech101` est une version de `Caltech101` dans laquelle l'arrière plan de chaque image et l'orientation des objets plan ont été modifiés au hasard afin de rendre la classification plus difficile. Nous avons créé `Caltech101-3` en conservant uniquement les 3 classes les plus importantes de `Caltech101`, et ce afin d'éliminer le déséquilibre présent dans le jeu de données initial.

FIGURE 10 – Exemples d'images appartenant aux classes « faciles à apprendre » (rangée du haut) et des classes « difficiles à apprendre » (rangée du bas)

Évaluation qualitative Dans une tâche de classification, certaines classes sont naturellement plus faciles a apprendre que d'autres. Dans la figure 10, nous présentons des exemples d'images appartenant aux classes « faciles à apprendre » (une bonne précision est obtenue en classification) et de classes « difficiles à apprendre » (on obtient une précision plus basse). Les objets appartiennent à la classe facile soit parce qu'ils apparaissent toujours dans la même posture (par exemple, *airplanes*,

TABLE 4 – Des classes « faciles à apprendre » et des classes « difficiles à apprendre » sur `Caltech101` et `RandCaltech101`

CLASSES « FACILES »		CLASSES « DIFFICILES »	
Caltech101	*RandCaltech101*	*Caltech101*	*RandCaltech101*
airplanes	**accordion**	**beaver**	**bass**
car_side	*airplanes*	*buddha*	**binocular**
dalmatian	*car_side*	*butterfly*	**brontosaurus**
dollar_bill	*dalmatian*	**ceiling_fan**	*buddha*
Faces_easy	*dollar_bill*	**cougar_body**	*butterfly*
garfield	*Faces_easy*	*crab*	*crab*
grand_piano	*garfield*	*crayfish*	*crayfish*
Leopards	**laptop**	*cup*	**crocodile**
metronome	*Motorbikes*	*dragonfly*	*cup*
Motorbikes	*panda*	*ewer*	*dragonfly*
panda	*snoopy*	**ferry**	*ewer*
scissors	*soccer_ball*	*flamingo*	*flamingo*
snoopy	*stop_sign*	*flamingo_head*	*flamingo_head*
soccer_ball	*watch*	*ibis*	**gerenuk**
stop_sign	*windsor_chair*	*kangaroo*	**helicopter**
tick	*yin_yang*	*lamp*	*ibis*
watch		*lobster*	*kangaroo*
windsor_chair		*mandolin*	*lamp*
yin_yang		*mayfly*	*lobster*
		minaret	*mandolin*
		pigeon	*mayfly*
		platypus	**metronome**
		pyramid	*minaret*
		rhino	**okapi**
		saxophone	*pigeon*
		schooner	*platypus*
		sea_horse	*saxophone*
		stapler	*sea_horse*
		strawberry	*stapler*
		wild_cat	*wrench*
		wrench	

garfield), soit parce qu'ils ont un motif de couleurs facile à reconnaître (*e.g., yin_yang*, *soccer_ball* ou *dalmatian*).

Le tableau 4 montre des classes faciles et difficiles à apprendre pour `Caltech101` et `RandCaltech101`, en soulignant les classes différentes en gras. Nous observons que la plupart des classes n'ont pas changé de difficulté malgré les modifications réalisées dans `RandCaltech101`. Cela montre que, tout en rendant les images plus difficiles à discriminer, `RandCaltech101` ne change pas fondamentalement la difficulté relative entre les classes.

Évaluation quantitative Du point de vue quantitatif, nous avons fait varier les différents paramètres de nos algorithmes et nous avons comparé les résultats en termes de précision, F_{score} et *True Positive Rate*. La figure 11 présente les résultats que nous avons obtenus dans nos expérimentations sur `Caltech101` (figure 11a) et sur `RandCaltech101` (figure 11b). Nous observons que nos propositions obtiennent constamment des meilleures résultats en termes de F_{score} que l'approche « baseline ». Cela montre que le fait d'introduire un peu de sémantique dans la représentation des images rend la représentation plus adaptée pour décrire les images, et permet donc d'améliorer les résultats d'un algorithme de classification sans avoir besoin de changer l'algorithme lui-même.

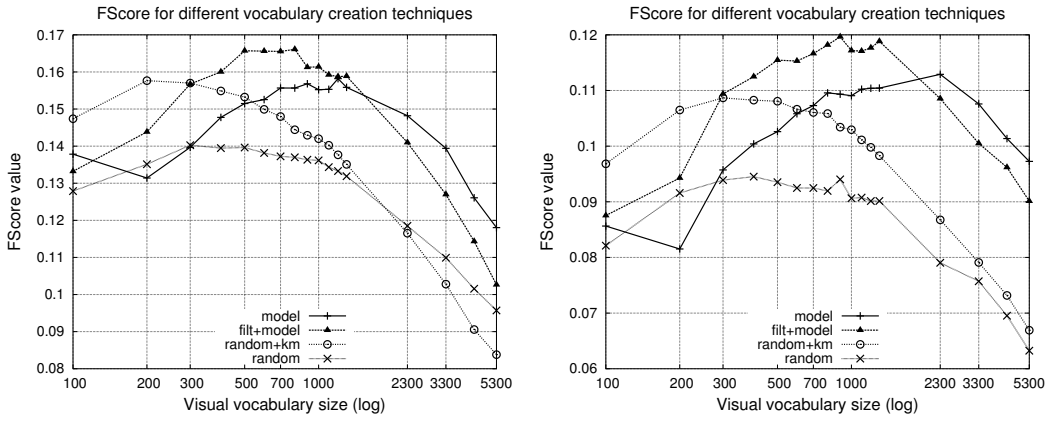

FIGURE 11 – F_{score}obtenu par le classifieur à base de clustering sur `Caltech101` (a) et sur `RandCaltech101` (b).

6 Analyser des données textuelles : extraction, nommage et évaluation de thématiques

Dans cette section, nous présentons les travaux que nous avons réalisés concernant l'un des types les plus importants et les plus abondants parmi les données complexes : le texte. Plus précisément, nous nous concentrons sur les *thématiques textuelles* qui fournissent un moyen rapide pour résumer les « idées » principales abordées dans une collection de textes. Les thématiques sont généralement définies comme des distributions statistiques de probabilités sur les mots et, par conséquent, ils sont souvent difficiles à interpréter pour les êtres humains. Nos travaux cherchent ici à **utiliser des connaissances sémantiques dans l'analyse des données textuelles**.

Dans le cadre de notre travail, cet enjeu de recherche se matérialise dans trois tâches associées à deux des lignes directrices de cette thèse : (i) l'extraction des thématiques, (ii) l'étiquetage des thématiques avec des noms compréhensibles pour un être humain, et (iii) l'utilisation ds connaissances sémantiques, sous la forme d'une hiérarchie de concepts, dans l'évaluation de ces thématiques. Pour la première tâche, nous proposons une solution pour l'extraction des thématiques à base de clustering avec recouvrement, qui autorise les documents à appartenir à plusieurs clusters. Pour la deuxième tâche, nous attribuons aux thématiques des noms que nous postulons comme humainement compréhensibles en utilisant une approche basée sur les « tableaux de suffixes » [33] (an anglais *suffix arrays*). Pour la dernière tâche, nous proposons d'associer les thématiques à des sous-arbres de concepts dans une base de connaissances donnée à l'avance (en l'occurrence, WordNet [34]). La cohésion sémantique des thématiques est évaluée en fonction de la hauteur et de la profondeur des sous-arbres thématiques correspondants dans la hiérarchie des concepts.

Les travaux concernant les données textuelles sont intimement liés au différents projets dans lesquels j'ai été impliqué tout au long de ma thèse, à savoir CONVERSESSION [11], ERIC-ELICO [12], CRTT-ERIC [13] et IMAGIWEB. Ils sont également liés à des travaux de développement, en particulier à `CommentWatcher` une plateforme open-source pour analyser les discussion sur des forums en ligne.

11. Projet lié à la création d'une entreprise startup `http://www.conversationnel.fr/`
12. Collaboration avec le laboratoire ELICO `http://www.elico-recherche.eu`
13. Collaboration avec le laboratoire CRTT `http://recherche.univ-lyon2.fr/crtt/`

6.1 Construire une représentation numérique de type « sac-de-mots »

L'une des représentations numériques les plus utilisées pour traiter des données textuelles est la représentation sous forme de « sac-de-mots » (en anglais *Bag of Words* ou *BoW*) [25, 26]. Cette représentation est largement utilisée dans le traitement du langage naturel et pour les tâches de recherche d'information. Récemment, cette représentation a été également utilisée dans la représentation des images, comme nous l'avons vu dans la section 5.1 ; on parle alors de représentation de type « sac-de-caractéristiques ».

Dans le modèle *BoW*, un texte est représenté comme une collection non ordonnée de mots, dans laquelle la grammaire et même l'ordre des mots sont ignorés. On attribue un score à chaque mot ou terme en utilisant un système de pondération.

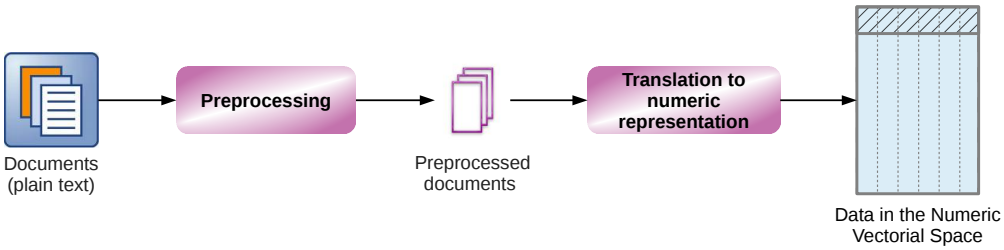

FIGURE 12 – Schéma pour transformer les données textuelles en une représentation numérique de type « sac-de-mots ».

La figure 12 montre la chaîne de traitement typique qui permet de mettre des données textuelles sous une forme numérique. Dans la phase de *prétraitement*, chaque document subit un certain nombre de transformations, parmi lesquelles : (i) élimination des mots-outils (le, la, les, de…), ainsi que des mots fréquents qui n'apportent aucune information sur la thématique du texte ; (ii) ramener les mots fléchis à leur racine (à l'aide du *stemming* ou de la *lemmatisation*) afin d'augmenter leur valeur descriptive. Dans la phase de *traduction dans une représentation numérique*, les documents prétraités sont convertis dans un espace numérique vectoriel en utilisant l'un des systèmes de pondération des termes (par exemple présence/absence, fréquence des termes, TFxIDF *etc.*). A la fin de cette phase, la collection textuelle peut être représentée comme une **matrice documents / termes**, dans laquelle chaque colonne correspond à un mot du vocabulaire et chaque ligne correspond à un document textuel.

6.2 Extraire et nommer les thématiques

Le système Le système que nous proposons aborde le problème d'extraction de thématiques à l'aide d'une approche de clustering textuel avec recouvrement. Une telle approche présente l'avantage d'autoriser les documents textuels à faire partie de plusieurs thématiques, en fonction des sujets abordés par le texte. Contrairement aux travaux précédents qui prennent en compte le chevauchement [12], notre approche aborde également le problème de la compréhension humaine des thématiques en attribuant à chaque groupe un nom « humainement compréhensible ».

Pour cela, un titre est choisi à partir d'une liste d'expressions complètes fréquentes, et l'utilisateur peut utiliser cette étiquette pour se faire une idée du contenu des textes au lieu d'une distribution de fréquences sur une (longue) liste de mots. Le système d'extraction de thématiques que nous présentons dans cette section a été implémenté dans `CKP`, un logiciel open-source d'extraction de thématiques qui a été intégré à la plateforme `CommentWatcher`.

Nous présentons dans la figure 13 le schéma du système proposé. A partir d'une collection de documents textuels (par exemple des discussions en ligne, des forums, des chats, des articles de journaux *etc.*), l'algorithme extrait les thématiques de la discussion et présente en sortie (i) les

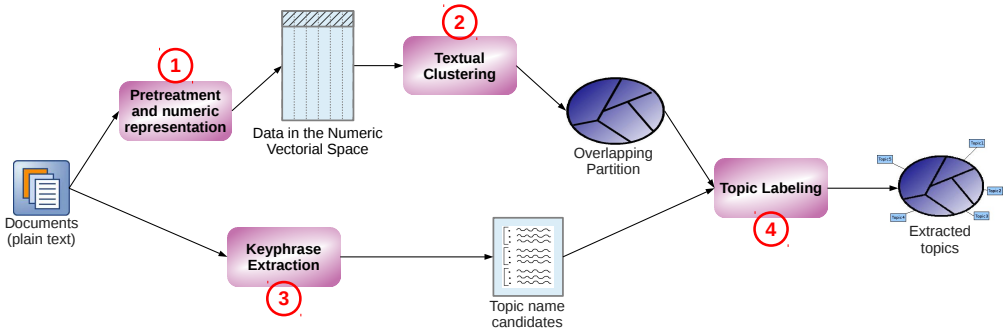

FIGURE 13 – Schéma d'un algorithme d'extraction de thématiques utilisant le clustering textuel avec recouvrement.

étiquettes des thématiques sous la forme d'expressions lisibles destinées à un utilisateur humain et (ii) la partition des textes associée à ces thématiques. Dans *la phase 1*, chaque document de la collection est prétraité comme indiqué dans la section 6.1 : les mots-outils sont supprimés et les mots sont ramenés à leur racine. Après le prétraitement, les documents sont transformés dans un espace de représentation vectoriel en utilisant l'un des systèmes de pondération de termes. Dans *la phase 2*, les documents sont regroupés en utilisant l'algorithme OKM [12] et une partition avec recouvrement est créée. OKM suit le schéma général des K-Means. Il optimise une fonction objectif de manière itérative en alternant deux étapes : les centres de gravité des clusters (les centroïdes) sont recalculés et les documents sont ensuite ré-attribués aux clusters les plus pertinents. Dans les K-Means, chaque document est associée à un seul centre de gravité et c'est le centre le plus proche en termes de la distance qui est utilisé. La différence est qu'OKM assigne chaque document à un *ou plusieurs* clusters. Pour cela, il associe un document à une image constituée du centre de gravité de l'ensemble des centroïdes auxquels ce document est associé. Par conséquent, chaque document peut faire partie d'une ou de plusieurs catégories. Dans *la phase 3*, des expressions complètes fréquentes (suite fréquente de mots) sont extraites à partir du texte original des documents en utilisant un algorithme basé sur des tableaux de suffixes. Les expressions extraites servent d'étiquettes candidates qui sont, dans *la phase 4*, réintroduites comme des pseudo-documents. Ainsi, les expressions sont décrites dans le même espace multidimensionnel dans lequel sont déjà décrits les documents. Il ne reste plus qu'à employer une distance classique, comme le cosinus dans notre cas, afin de choisir le meilleur nom pour chaque thématique.

Expérimentations et évaluation Nous avons effectué nos expérimentations sur des corpus écrits en anglais et en français, avec des textes de styles d'écriture différents. Pour l'anglais, nous utilisons un sous-ensemble du corpus Reuters [14]. Nous avons choisi les documents qui ont été annotés avec au moins une étiquette. En ce qui concerne le français, nous avons utilisé une discussion sur un forum en ligne, intitulée « Y a t-il trop commémorations en France ? », utilisé initialement dans [47].

Évaluation du regroupement Pour le regroupement issu de la classification non supervisée, nous avons utilisé une méthode d'évaluation à base d'experts : la sortie de l'algorithme est comparée au regroupement proposé par des experts. Nous considérons deux documents comme étant associés s'ils appartiennent à une même classe dans le processus de regroupement. De plus, nous considérons que cette association est correcte si les deux documents ont une étiquette Reuters commune. Cela permet ensuite de calculer les mesures classiques employées en recherche d'information de précision, rappel et F_{score}.

Nos expérimentations ont confirmé l'hypothèse initiale selon laquelle une approche par chevau-

14. `http://mlr.cs.umass.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection`

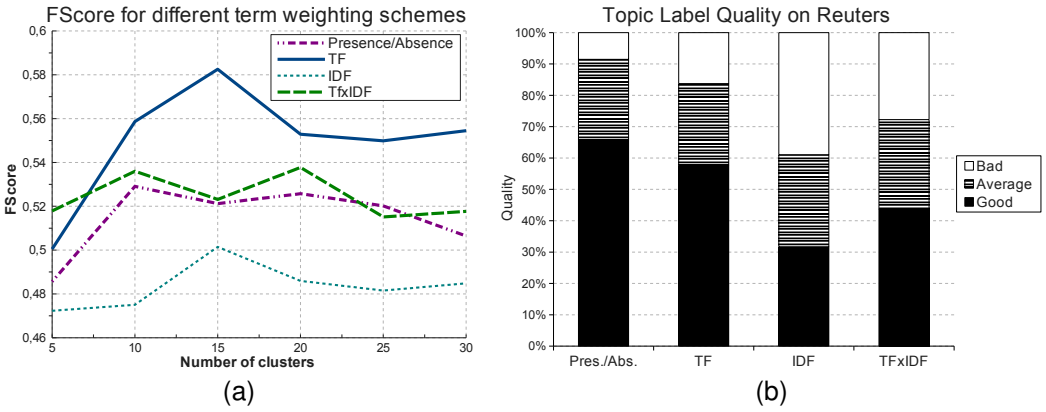

FIGURE 14 – Résultats obtenus sur `Reuters`, pour les différents systèmes de pondération de terme : (a) F_{score} obtenue par OKM et (b) qualité des noms obtenus.

chement est plus adaptée à la fouille de textes. Nous avons aussi fait varier le système de pondération des termes ; les résultats de cette comparaison se trouvent dans la figure 14a. Il est intéressant d'observer que la mesure qui obtient les meilleurs résultats ici est le TF (« Term Frequency »).

Évaluation du nom des catégories Les experts humains sont souvent utilisés dans la littérature (voir par exemple [10, 37]) pour quantifier la lisibilité des thématiques obtenues. Nous avons choisi une approche similaire pour évaluer le nom attribué aux thématiques : on a demandé à 5 experts de noter la qualité de ces noms en mettant 0 pour un nom qui n'apporte pas d'information, 1 pour un nom de qualité moyenne et 2 pour un nom qui exprime une idée réellement compréhensible. Les résultats obtenus sur `Reuters` sont présentés dans la figure 14b. Ils montrent une bonne acceptation par les utilisateurs des étiquettes construites par notre approche pour nommer les thématiques.

6.3 Évaluation des thématiques à l'aide d'une hiérarchie de concepts

Pour la tâche d'évaluation des thématiques, l'hypothèse sous-jacente est que les mesures statistiques ne parviennent pas totalement à émuler le jugement humain. Par conséquent, nous proposons une approche qui utilise une ressource sémantique existante, comme une hiérarchie de concepts (en l'occurrence, WordNet [34]). En utilisant les termes les plus pertinents attachés à la thématiques (par exemple, les mots de plus forte probabilité), nous faisons correspondre la thématique à une sous-arborescence dans la hiérarchie des concepts. De cette manière, nous relions une distribution statistique de fréquences à une structure sémantique. Nous définissons la spécificité et la couverture d'une sous-arborescence, en fonction de sa hauteur et de sa profondeur dans la hiérarchie, puis nous utilisons ces nouvelles mesures pour évaluer la cohésion sémantique des thématiques. D'un point de vue applicatif, nous travaillons actuellement à l'intégration de cette évaluation sémantique dans `CommentWatcher`.

Sous-arborescences thématiques Le point central du système original que nous proposons est d'aligner les thématiques à des sous-arborescences de la base de concepts. L'hypothèse sous-jacente consiste à passer par le sens des mots les plus représentatifs de ces thématiques, comme montré schématiquement dans la figure 15. Nous recherchons un concept ou un ensemble de concepts qui sont sémantiquement liés à au moins l'un des sens de ces mots.

Chaque concept de la hiérarchie des concepts est associé à un ensemble de mots synonymes (appelé *synset* dans WordNet). Etant donnée la propriété de polysémie des mots, chaque mot peut être associé à plusieurs concepts dans la hiérarchie comme autant de sens différents du mot. La sous-

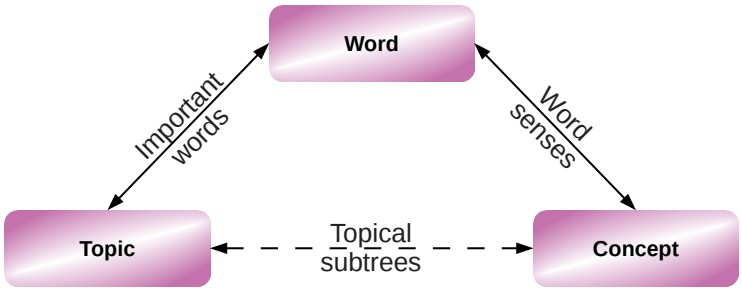

FIGURE 15 – Alignement des thématiques à des concepts, en passant par les différents sens des mots.

arborescence d'un mot est un sous-arbre qui possède comme feuilles tous les concepts qui sont des sens associés à un mot donné. Nous définissons alors la sous-arborescence d'une thématique comme étant un sous-arbre qui contient au moins un sens pour chacun des mots les plus représentatifs de la thématique.

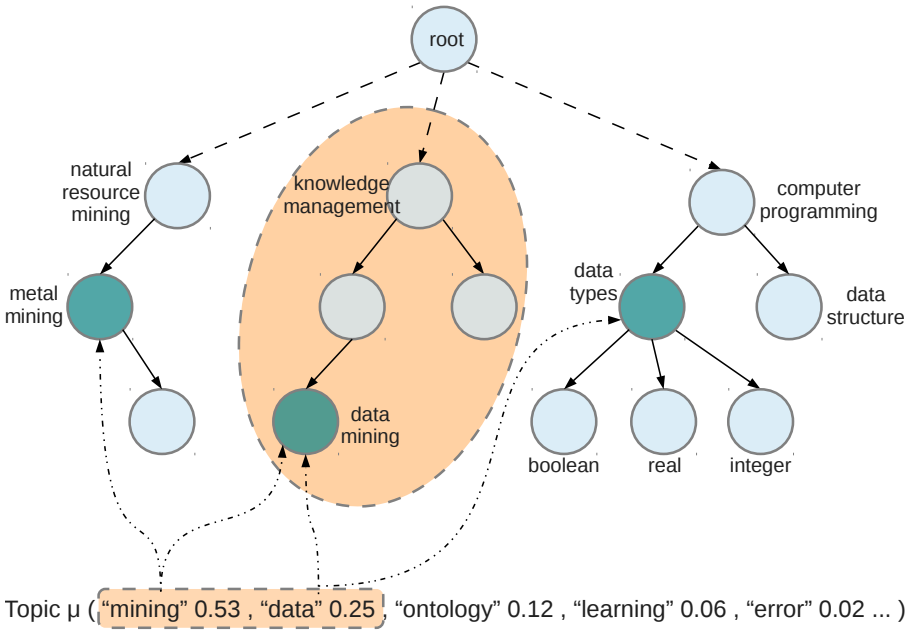

FIGURE 16 – La sous-arborescence d'une thématique ayant comme mots les mieux noté « data » et « mining », la partie commune des sous-arborescences de chaque mot étant soulignée.

Nous donnons dans la figure 16 l'exemple d'une sous-arborescence associée à une thématique ayant comme mots les mieux notés « data » et « mining ». Les sous-arbres de gauche et du centre sont associés avec le mot « data », alors que les sous-arbres de droite et du centre sont associés avec « mining » ; nous avons souligné la partie partagée par ces deux sous-arbres. Cet alignement est similaire à une procédure de désambiguïsation du sens des mots. Nous pouvons tout de suite remarquer que le sens de l'arbre gauche attaché à « mining » est relié à l'extraction de ressources naturelles, tandis que l'arbre central est relié à l'extraction de connaissances à partir des données. De même, l'arbre central peut être associé à « data », ce qui donne un sens global à la thématique.

En conclusion, nous associons à chaque thématique un concept « le plus lié », qui est la racine de l'arborescence la plus spécifique englobant au moins un sens pour chacun des mots importants. Pour

la thématique présentée dans la figure 16, ce concept est celui de « knowledge management ». La force de la relation entre la thématique et le concept est calculée en redéfinissant *la couverture* de la sous-arborescence associée à une thématique et ayant comme racine le concept donné, ainsi que *la spécificité* de son concept racine. Nous combinons ces deux mesures de couverture et de spécificité pour former une seule fonction de pertinence. C'est cette fonction que nous utilisons pour choisir le concept qui sera associé à chaque thématique. Le meilleur score de pertinence obtenu par un concept est alors utilisé pour quantifier la pertinence d'une thématique.

Expérimentations et évaluation Nous avons évalué notre proposition sur deux jeux de données : `Suall11` [15] et un jeu de données économiques [16]. Nous avons utilisé le même protocole expérimental que celui proposé dans [10], en injectant dans chaque thématique un mot « étranger » ou intrus (*spurious word* en anglais) et en demandant à 37 experts de l'identifier. L'hypothèse sous-jacente est que plus la thématique est bien formée (cohérente), plus le mot intrus est simple à identifier.

Évaluation par détection de mots intrus Nous avons divisé les thématiques en deux catégories : celles avec un score élevé de pertinence (+) et celles avec un score faible de pertinence (-). Le but est de détecter une amélioration dans le taux de détection des mots intrus entre les thématiques pertinentes et celles non pertinentes. Basé sur la réponse des évaluateurs, nous calculons $\overline{hit}_+$, le pourcentage moyen de détection des mots intrus pour les thématiques pertinentes, et $\overline{hit}_-$, le pourcentage moyen de détection des mots intrus pour les thématiques non pertinentes. Les résultats obtenus se trouvent dans le tableau 5, ainsi que le gain relatif entre $\overline{hit}_+$ et $\overline{hit}_-$.

TABLE 5 – Taux de détection des mots intrus.

Dataset	$\overline{hit}_+$	$\overline{hit}_-$	Gain $\overline{hit}$
AP	**0.69**	0.65	6.93%
Suall	**0.75**	0.59	28.55%

Les résultats obtenus montrent que le taux de détection moyen augmente entre les thématiques jugées comme pertinentes par notre algorithme et les thématiques non pertinentes, avec une amélioration du gain relatif de 6,93% pour le corpus économique et de 28,55% pour `Suall`. Le fait que le taux de détection est constamment meilleur pour les thématiques pertinentes est la preuve que l'alignement et la technique proposés nous permettent, d'une certaine manière, d'émuler le jugement humain qui a été réalisé sur la cohérence des thématiques.

7 Travaux appliqués et logiciels

Les recherches théoriques de cette thèse ont été menées conjointement à la réalisation de prototypes logiciels. L'aboutissement de ces prototypes est `CommentWatcher` [17], une plateforme libre dédiée à l'analyse des discussions en ligne, et en particulier l'analyse des forums. Construite comme une plateforme web, `CommentWatcher` dispose d'un module de récupération automatique des forums, d'un module d'extraction des thématiques à partir d'une sélection de textes, d'un module de visualisation des thématiques extraites ainsi que du réseau social sous-jacent des internautes qui participent à la discussion. La plateforme est utile tant pour la veille de presse (en permettant l'identification rapide des sujets importants dans les forums) que pour la recherche sur les médias sociaux de manière générale (en permettant aux chercheurs de constituer des corpus diachroniques de données textuelles).

15. Télécharger : `http://www.gutenberg.org/dirs/etext04/suall11.txt`
16. Télécharger : `http://eric.univ-lyon2.fr/~arizoiu/files/economic_corpus_AP.tar.bz2`
17. Site de présentation et courte vidéo : `http://mediamining.univ-lyon2.fr/commentwatcher`

8 Conclusion, travaux en cours et perspectives.

Les travaux effectués dans le contexte de cette thèse se situent au croisement de l'**analyse de données complexes** et du **clustering semi-supervisé**. Nous étudions comment les données de différentes natures peuvent être traitées tout en tenant compte de leur dimension temporelle et de l'information complémentaire qui peut y être attachée. Les problématiques abordées dans cette thèse sont très vastes et soulèvent de nombreuses perspectives de travail.

Les travaux futurs incluent notamment des améliorations pour les différentes propositions, présentées dans ce document. Une de directions de travail que nous avons envisagé est de **déterminer automatiquement les valeurs de paramètres du TDCK-Means** (α, β, δ et γ), en utilisant une approche inspiré de l'optimisation multi-critère à l'aide des algorithmes génétiques [53]. Une extension prévu est d'**adapter l'algorithme de construction de représentation des images à l'annotation incomplète**. Ce développement revient à adapter l'algorithme de construction des attributs aux données issues de l'internet (par exemple, des annotations sur des plateformes de partage des images). Pour la partie appliqué de notre travail, nos intentions sont d'**implémenter notre proposition d'évaluation sémantique de thématiques dans `CommentWatcher`**. Nous envisageons aussi d'intégrer des algorithmes supplémentaires d'extraction de thématiques, ce qui permettra de faire un comparaison plus en détails de différentes modèles de thématiques, dans le contexte de textes extraites à partir de l'internet. Le but au long terme, pour les parties théorique et pratique, est une meilleure intégration de toutes l'information issue de données complexes : du texte, de l'image, du vidéo, de l'audio, mesures numériques, la dimension temporelle, des annotation et des ontologies.

Travaux en cours Actuellement, nous travaillons à améliorer et étendre nos contributions. Par exemple, nous développons une extension de notre algorithme de clustering temporel pour introduire une construction simultanée d'une structure de graphe entre les clusters obtenus. Nous définissons une mesure de similarité d'intersection entre deux clusters, basé sur le nombre des entités qui présentent une transition entre les deux clusters. Nous calculons, après la construction des clusters, la matrice d'adjacence, basé sur cette mesure de similarité. Une autre vois sur laquelle nous travaillons en parallèle est d'appliquer l'algorithme TDCK-Means à un autre type de données et une autre problématique d'apprentissage : la détection de rôles sociaux dans les communautés en ligne. Nous détectons des rôles comportementaux, en étant similaires aux phases d'évolution, et nous définissons un rôle social comme une suite de rôles comportementaux.

Nous sommes aussi intéressés par améliorer notre algorithme de construction d'attributs afin de prendre en compte la dimension temporelle en plus de la sémantique du jeu de données. L'idée est de détecter des corrélation avec un certain délais de temps δ. Le problème fondamental qui se pose est « qu'est-ce que la corrélation dans un contexte temporel ? » Nous proposons une extension de la mesure de corrélation afin de prendre en compte le temps et, pour une paire de attributs, nous pouvons calculer δ « optimal », qui maximise la corrélation temporelle. À présent nous travaillons sur une méthodologie pour construire des attributs temporels, comme des extensions des conjonctions, des chaînes temporelles $f_i \xrightarrow{\delta_1} f_j \xrightarrow{\delta_2} f_k$, avec la signification f_i précèdes f_j avec un délais de temps δ_1, et f_j précèdes f_k avec un délais de temps δ_2.

Références

[1] Klaus Armingeon, David Weisstanner, Sarah Engler, Panajotis Potolidis, Marlène Gerber, and Philipp Leimgruber. Comparative political data set 1960-2009. Institute of Political Science, University of Berne., 2011.

[2] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning distance functions using equivalence relations. In *Machine Learning, International Workshop then conference*, volume 20, page 11, 2003.

[3] Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Semi-supervised clustering by seeding. In *International Conference on Machine Learning*, pages 19–26, 2002.

[4] Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney. Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering. In *Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, Collocated with ICML '03*, ICML '03, pages 42–49, 2003.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf : Speeded up robust features. *Computer Vision–ECCV 2006*, pages 404–417, 2006.

[6] Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Knowledge Discovery and Data Mining, Proceedings of the ninth ACM SIGKDD international conference on*, pages 39–48. ACM New York, NY, USA, 2003.

[7] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören. Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia-a crystallization point for the web of data. *Web Semantics : Science, Services and Agents on the World Wide Web*, 7(3) :154–165, 2009.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3 :993–1022, 2003.

[9] Hendrik Blockeel, Luc De Raedt, and Jan Ramon. Top-down induction of clustering trees. In *Proceedings of the 15th International Conference on Machine Learning*, pages 55–63, 1998.

[10] Jonathan Chang, Jonathan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves : How humans interpret topic models. In *Advances in Neural Information Processing Systems, Proceedings of the 23rd Annual Conference on*, volume 31 of *NIPS 2009*, 2009.

[11] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*, volume 2 of *Adaptive Computation and Machine Learning*. The MIT Press, September 2006.

[12] Guillaume Cleuziou. An extended version of the k-means method for overlapping clustering. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[13] David Cohn, Rich Caruana, and Andrew McCallum. Semi-supervised clustering with user feedback. In *Constrained Clustering : Advances in Algorithms, Theory, and Applications*, volume 4, pages 17–32. Cornell University, 2003.

[14] Corrina Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.

[15] Gabriela Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–22, 2004.

[16] Ian Davidson and Sugato Basu. A survey of clustering with instance level constraints. *ACM Transactions on Knowledge Discovery from Data*, pages 1–41, 2007.

[17] Ayhan Demiriz, Kristin Bennett, and Mark J. Embrechts. Semi-supervised clustering using genetic algorithms. In *Artificial Neural Networks in Engineering*, pages 809–814. ASME Press, 1999.

[18] Brigitte Dormont. Petite apologie des données de panel. *Économie & prévision*, 87(1) :19–32, 1989.

[19] G. H. Dunteman. *Principal components analysis*, volume 69. SAGE publications, Inc, 1989.

[20] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples : An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1) :59–70, 2007.

[21] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 2 of *CVPR 2005*, pages 524–531. IEEE, 2005.

[22] W. Feller. *An introduction to probability theory and its applications. Vol. I.* Wiley, 1950.

[23] Jing Gao, Pang-Ning Tan, and Haibin Cheng. Semi-supervised clustering with partial background information. In *In Proceedings of the Sixth SIAM International Conference on Data Mining*, 2006.

[24] Nizar Grira, Michel Crucianu, and Nozha Boujemaa. Unsupervised and semi-supervised clustering : a brief survey. Technical report, A Review of Machine Learning Techniques for Processing Multimedia Content, Report of the MUSCLE European Network of Excellence (FP6), 2005.

[25] Zellig S. Harris. Distributional structure. *Word*, 10 :146–162, 1954.

[26] Zellig S. Harris. *Mathematical structures of language*. Wiley, 1968.

[27] Rongrong Ji, Hongxun Yao, Xiaoshuai Sun, Bineng Zhong, and Wen Gao. Towards semantic embedding in visual vocabulary. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 918–925, 2010.

[28] Zhang Jianjia and Luo Limin. Combined category visual vocabulary : A new approach to visual vocabulary construction. In *Image and Signal Processing, 4th International Congress on*, volume 3 of *CISP 2011*, pages 1409–1415, October 2011.

[29] Teemu Kinnunen, Joni Kristian Kamarainen, Lasse Lensu, Jukka Lankinen, and Heikki Kälviäinen. Making visual object categorization more challenging : Randomized caltech-101 data set. In *2010 International Conference on Pattern Recognition*, pages 476–479. IEEE, 2010.

[30] Dan Klein, Sepandar D. Kamvar, and Christopher D. Manning. From instance-level constraints to space-level constraints : Making the most of prior knowledge in data clustering. In *International Conference on Machine Learning*, pages 307–314, 2002.

[31] Wei-Hao Lin and Er Hauptmann. Structuring continuous video recordings of everyday life using time-constrained clustering. In *IS&T/SPIE Symposium on Electronic Imaging*, 2006.

[32] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[33] Udi Manber and Gene Myers. Suffix arrays : A new method for on-line string searches. *SIAM Journal on Computing*, 22(5) :935–948, 1993.

[34] George A. Miller. Wordnet : a lexical database for english. *Communications of the ACM*, 38(11) :39–41, 1995.

[35] Raymond J. Mooney, Sonal Gupta, Joohyun Kim, and Kristen Grauman. Watch, listen & learn : Co-training on captioned images and videos. *Machine Learning and Knowledge Discovery in Databases*, pages 457–472, September 2008.

[36] Claudiu Musat, Julien Velcin, Stefan Trausan-Matu, and Marian-Andrei Rizoiu. Improving topic evaluation using conceptual knowledge. In *International Joint Conference on Artificial Intelligence, Proceedings of the Twenty-Second*, volume 3 of *IJCAI 2011*, pages 1866–1871. AAAI Press, 2011.

[37] Stanislaw Osinski. An algorithm for clustering of web search results. Master's thesis, Poznań University of Technology, Poland, June 2003.

[38] Giulia Pagallo and David Haussler. Boolean feature discovery in empirical learning. *Machine learning*, 5(1) :71–99, 1990.

[39] Florent Perronnin, Christopher R. Dance, Gabriela Csurka, and Marco Bressan. Adapted vocabularies for generic visual categorization. *Computer Vision–ECCV 2006*, pages 464–475, 2006.

[40] John R. Quinlan. Induction of decision trees. *Machine learning*, 1(1) :81–106, 1986.

[41] John R. Quinlan. *C4.5 : programs for machine learning*. Morgan Kaufmann, 1993.

[42] Marian-Andrei Rizoiu and Julien Velcin. Topic extraction for ontology learning. In Wilson Wong, Wei Liu, and Mohammed Bennamoun, editors, *Ontology Learning and Knowledge Discovery Using the Web : Challenges and Recent Advances*, chapter 3, pages 38–61. Hershey, PA : Information Science Reference, 2011.

[43] Marian-Andrei Rizoiu, Julien Velcin, and Stéphane Lallich. Structuring typical evolutions using temporal-driven constrained clustering. In *International Conference on Tools with Artificial Intelligence, Proceedings of the Twenty-Forth*, ICTAI 2012, pages 610–617. IEEE, November 2012.

[44] Marian-Andrei Rizoiu, Julien Velcin, and Stéphane Lallich. Unsupervised feature construction for improving data representation and semantics. *Journal of Intelligent Information Systems*, 40(3) :501–527, 2013.

[45] Y. Sawaragi, H. Nakayama, and T. Tanino. *Theory of multiobjective optimization*, volume 176. Academic Press New York, 1985.

[46] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering objects and their location in images. In *Computer Vision, Tenth IEEE International Conference on*, volume 1 of *ICCV 2005*, pages 370–377. IEEE, 2005.

[47] Anna Stavrianou. *Modeling and mining of web discussions*. PhD thesis, Université Lumière Lyon 2, 2010.

[48] Julia Vogel and Bernt Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2) :133–157, 2007.

[49] Kiri Wagstaff and Claire Cardie. Clustering with instance-level constraints. In *International Conference on Machine Learning, Proceedings of the Seventeenth*, pages 1103–1110, 2000.

[50] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schroedl. Constrained k-means clustering with background knowledge. In *International Conference on Machine Learning, Proceedings of the Eighteenth*, pages 577–584. Morgan Kaufmann, 2001.

[51] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. *Advances in Neural Information Processing Systems*, 15 :505–512, 2002.

[52] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories : A comprehensive study. *International Journal of Computer Vision*, 73(2) :213–238, 2007.

[53] Qingfu Zhang and Hui Li. Moea/d : A multiobjective evolutionary algorithm based on decomposition. *Evolutionary Computation, IEEE Transactions on*, 11(6) :712–731, 2007.

[54] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[55] Djamel A. Zighed, Shusaku Tsumoto, Zbigniew W. Ras, and Hakim Hacid, editors. *Mining Complex Data*, volume 165 of *Studies in Computational Intelligence*. Springer, 2009.