CREA : méthode d'analyse, d'adaptation et de réutilisation des processus à forte intensité de connaissance : cas d'utilisation dans l'enseignement supérieur en informatique

Fabrice Boissier

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PARIS 1 PANTHÉON - SORBONNE

Préparée au sein du Centre de Recherche en Informatique (EA 1445)
à L'École Doctorale de Management Panthéon - Sorbonne (ED 559)

Présentée par :

FABRICE BOISSIER

Pour l'obtention du grade de Docteur de L'Université Paris 1 Panthéon - Sorbonne

CREA : Méthode d'analyse, d'adaptation et de réutilisation des processus à forte intensité de connaissances

cas d'utilisation dans l'enseignement supérieur en informatique

Spécialité : Informatique

Thèse présentée et soutenue publiquement à Paris, le 17 janvier 2022
Composition du jury :

M. Guillaume CLEUZIOU	Professeur, Université d'Orléans	Rapporteur
M. Alain WEGMANN	Professeur, École Polytechnique Fédérale de Lausanne	Rapporteur
Mme. Daniela GRIGORI	Professeur, Université Paris Dauphine - PSL	Examinatrice
M. Camille SALINESI	Professeur, Université Paris 1 Panthéon - Sorbonne	Examinateur
Mme. Bénédicte LE GRAND	Professeur, Université Paris 1 Panthéon - Sorbonne	Directrice de Thèse
Mme. Irina RYCHKOVA	Maître de Conférence, Université Paris 1 Panthéon - Sorbonne	Co-Encadrante de Thèse

Remerciements

Je remercie tout d'abord mes encadrantes, Bénédicte Le Grand et Irina Rychkova de m'avoir encadré et accompagné pour cette thèse pendant laquelle beaucoup trop d'évènements imprévus se sont produits (mais au moins, on peut littéralement appeler ça l'aventure de la thèse !). Beaucoup de chemins, détours, impasses, raccourcis, ... ont été explorés, pour finalement construire cette carte. Ce fut complexe, long, parfois très laborieux, mais nous y sommes arrivés. Lorsque j'ai lu pour la première fois l'intitulé du sujet (« *Exploration des higraphs et de l'analyse formelle de concepts pour la modélisation des processus métier à forte intensité de connaissance et pour l'évaluation des modèles* »), je ne voyais absolument pas où commençait et finissait chaque terme... mais je suis aujourd'hui ravi de pouvoir expliquer plusieurs d'entre eux ! Même chose sur les méthodologies de recherche : design science, action research, behavior, et bien d'autres qu'il me reste à découvrir. Je ne pensais absolument pas toucher à autant de domaines en démarrant une thèse en informatique, mais je suis comblé par toutes ces connaissances (et je souhaite réellement continuer à apprendre et comprendre). Merci encore Bénédicte et Irina !

Je tiens également à remercier les rapporteurs qui ont accepté de relire mon manuscrit : Alain Wegmann et Guillaume Cleuziou. Leurs précieux conseils et remarques m'ont effectivement aidé à développer ce qu'il manquait, et j'espère pouvoir continuer à creuser certaines de leurs pistes dans de futurs travaux. Je remercie aussi Daniela Grigori et Camille Salinesi d'avoir accepté les rôles d'examinateurs.

Merci énormément au CRI et à l'ensemble de ses membres pour l'ensemble de vos avis critiques et les interminables discussions. Une pensée émue et pleine de souvenirs à Ali Jaffal, Elena Viorica Epure, Elena Kushnareva, Nourhène Ben Rabah, Afef Awadid, Danillo Sprovieri, Asmaa Achtaich : merci pour vos soutiens. Une pensée moins émue et beaucoup plus actuelle à David Beserra, Floriane Owczarek, Angela Patricia Villota Gomez, Luisa Fernanda Rincon, Sabrine Edded, Houssem Chemingui, et aux tous derniers (qui vont y arriver) : Nicolas Six, Claudia Negri, Ramona Elally, Camilo Correa. Sans oublier les collègues : Stéphane, Emmanuelle, Corinne, Astrid, Diem, Gabriel, Sarah, Julien, et Brigitte. Et les anciens camarades de classe : Gözde Kiraç, Nassima Mazouz, Wassila Bey Zekkoub.

Je n'oublie pas les membres du LSE à EPITA pour m'avoir accueilli sur la fin de cette thèse : Marc Espie (les maths, OpenBSD, ... je continue d'apprendre grâce à toi !), Robert Erra (les métriques : nous allons les traiter), Mark Angousturès (on ne lâche pas le data science, ni la recherche), Reda Dehak, Alizée Pénel, Laurent & Loïca (bip !), et Alexandre Letois.

Plus largement, je remercie *infiniment* l'ensemble des enseignantes et enseignants qui m'ont fait adorer les études et les diverses matières qui existent. C'est-à-dire, les professeurs des écoles en maternelle et en primaire, les professeurs des collèges et lycées, les professeurs et enseignants[-chercheurs] titulaires ou non à l'EPITA, à l'UQAC, et évidemment à l'Université Paris 1 Panthéon - Sorbonne. Je garderai éternellement d'excellents souvenirs de vous toutes et tous : Laurence, Cécile, André Ménini, Maria Raison, Élisabeth Aubin, M^{me} Croué (même si c'était difficile de retenir toutes les définitions par cœur), M^{me} Oeuvrard, M^{me} Corbic, M. Brassard, Bernard Roquejoffre, M^{me} Gautry, Gaston Époté Myke, M. Huchon, Peggy Sultan, M^{me} Mouchez, M^{me} Lorgeril, M^{me} Le Texier, M^{me} Luccioni, M. Azam, M. Tobaty, M. Nefati, Michel Volkovitch (et ses fabuleuses phrases à traduire concernant les flûtes et les synthétiseurs), M. Garcia, M. Oucif (Pierre Loti, « Vers Ispahan », « Aziyadé »), M. Begis (pour sa patience infinie), M^{me} Staszak (surtout le tout dernier cours de philosophie le 6

juin 2006), M^me Ehanno, Nathalie "Junior" Bouquet, Christophe "Krisboul" Boullay, Christelle Trémoulet, Anne-Sophie Dujardin, Olivier Rodot, Marwan Burelle, mes ACD Nicolas Ballas (ballas_n) et Mathieu Sabarly (sabarl_m), Didier Verna, Akim Demaille, Thierry Géraud, Sébastien Bombal, Éric Gaillard, Aurélien Borde, Julien Sterckeman, M. Bouchard, Samira Si-said Cherfi, Saïd Assar, Fayçal Hamdi, Camille Salinesi, Rébecca Deneckère, Irina Rychkova, Bénédicte Le Grand.

Cette thèse a beau être la somme des travaux d'une équipe de chercheurs, sans vous toutes et tous pour m'enseigner autant de choses, je n'aurais jamais acquis assez de connaissances pour mener à bien ce projet de recherche. Une pensée également pour André Rossano qui m'a particulièrement touché lorsqu'il nous expliquait le mainframe, et ses anecdotes de travail d'une autre ère sur S/370 et l'ASR 33.

Je pense également à vous toutes et tous... Jason Brillante en premier pour absolument tout, y compris les moments au Liberty Rock Studio gravés à jamais dans nos cœurs et mémoires. Éva Debray pour ses conseils, relectures, encouragements, appels, jus de citrons, explications, remises en question de mes certitudes, ... Si j'ai appris les méthodes de recherche en informatique au labo, j'ai appris beaucoup de choses sur la philosophie, le contrôle social, Spinoza, les sciences de l'éducation, l'histoire de l'école en France, ... grâce à toi, tes explications, et tes corrections. Merci Éva. Soutiens indéfectibles depuis plus de 10 ans maintenant (merci de me remonter le moral) : Geoffrey Tan, Laurent Garcin, Shanand Seeram, Jérémy Meng, Riyad Yakine. Merci à Kathleen Ripert pour le soutien un peu trop lointain depuis l'Allemagne... mais le soutien est là, et c'est le plus important ! (Tu reviens quand à Paris ?) Une pensée particulière à Ghislain Guillot (oui, tu comptes beaucoup quand même).

Je remercie toutes et tous les amis et potes du monde entier (GameSeries, Red-Network, Dimension MMX, D.Folk, ...) : Merlin, Vadim, Diana, Renan, Rémi, Frédéric, Justin, Christian, Marion, Jean-Baptiste, Lisa, Guillaume, Sophie, Éléonore, Samantha, Angélina, Johanna, Clara, Marine, Sarah, Elisa, Axel, Stacy, Alex(andra), Axel, Julie, Lionel, Patricia, ... sans oublier Robert Rafie et les épi-potes ! (Raph', Cédric, Pierre, Justin, Fabien, Cyril, Alexandre, Aurélie, Gabriel, Lucas, François, Alban, Florent, Robin, Joe, Nils, ...)

Alexandre Abraham, Lydie Gustafsson, et Egle Tomasi m'ont permis de sauter le pas et réellement faire une thèse : ce manuscrit n'existerait pas sans vous, ni mon changement radical de carrière.

Grâce aux doctorantes et doctorants de Paris 1 (parfois avec le doctorat) et d'ailleurs, j'ai découvert les SHS. Et sans vous, je serais encore un nigaud qui dirait des bêtises sur vos domaines : Hélène Bénistand, Léon Guillot, Armand Desprairies, Guillaume Noblet, Orianne Tercerie, Juliette Fontaine, Bastien Rueff, Cécile Bourgade, Maëlle de Seze, Milan Bonté, Hugo Vidon, Justine Audebrand, Anaïs Bonano, Sahra Rausch, Karim Abou-Merhi, Evélia Mayenga, Marine Lassery, Matthieu Febvre-Issaly, Typhaine Rahault, Guillemette Prevot, Michaël Pierrelée, Nicolas Jouvin, Clément, ... Et merci aux étudiantes et étudiants qui m'ont beaucoup appris ou avec qui j'ai bien ri : Joachim "Jojo" Loysel, Sami, Émilie, Thib', Hasna, Lucas, Maximilien, Trystan, Emilio, Jo, Marie, Elsa, Lorenz, Jaspal, Ulysse, Adèle, Diu, Nurlan, ...

Plus généralement, je remercie tou · te · s les camarades de Mobdoc et autres organisations. Vivement le 27 Nivôse (Zinc) de l'an CCXXX (230) !

Je n'oublie pas les plus importants : mes parents (Robert Boissier et Patricia Kissling), mes grands-parents (Willy, Augusta, Rosa), tous les autres membres (Yvette, Frédérique, Bertrand, Isabelle, Tiffany, Ornella, Robin, ...). Sans ma mère, je ne sais pas si j'aurais eu le bac (et je rigolerais probablement moins). Sans mon père, je ne sais pas si j'aurais fait de l'informatique (et je m'intéresserais probablement à beaucoup moins de choses).

Résumé

CREA : Méthode d'analyse, d'adaptation et de réutilisation des processus à forte intensité de connaissances
cas d'utilisation dans l'enseignement supérieur en informatique

La crise sanitaire du COVID-19 a particulièrement accéléré le mouvement de numérisation pourtant déjà initié depuis quelques décennies dans l'enseignement supérieur. De nombreuses activités ont dû être adaptées dans l'urgence, tout particulièrement les réunions entre enseignants, l'évaluation des étudiants et les enseignements. Ces activités sont des exemples de processus « *à forte intensité de connaissances* » (ou « *knowledge intensive processes* » en anglais) qui partagent des caractéristiques rendant difficile l'intégration du numérique, telles que :

— l'abondance de connaissances mobilisables, autant de la part des étudiants lors de leurs travaux que de la part des enseignants évaluant ou adaptant leurs cours,

— la collaboration entre toutes les parties prenantes du monde de l'enseignement supérieur,

— la créativité requise pour s'adapter au contexte incertain.

Ce besoin rapide de déployer de nouveaux processus à forte intensité de connaissances, ou d'adapter ceux qui existent, se confronte à de nombreux défis connus de ce domaine de recherche spécifique. La question est de savoir comment réutiliser des connaissances existantes, par exemple des connaissances entreposées en ligne dont l'abondance rend difficile la sélection des plus adaptées aux besoins des enseignants.

Dans cette thèse, nous proposons la méthode CREA réutilisant des cas passés dans le domaine de l'enseignement supérieur, en particulier pour la construction de cours. La méthode CREA permet de réutiliser des supports de cours existants pour tout d'abord représenter visuellement l'écart entre eux, mais également de proposer des séances de cours présentées sous forme de regroupements de sujets majeurs à aborder. D'autres types de documents peuvent également être intégrés parmi les supports de cours (des pages webs, ou des articles de recherche), afin de proposer des regroupements adaptés à un public particulier, voire de proposer des regroupements à l'état de l'art de la recherche. Cette méthode s'appuie sur des outils de traitement automatique de la langue pour extraire les termes employés indépendamment de la langue d'origine, puis sur l'analyse de concepts formels pour calculer des métriques permettant de construire des regroupements de termes et évaluer la similarité des cours fournis en entrée. Nous proposons également des résultats préliminaires d'une méthode d'ordonnancement des séances.

Mots clés : Processus à forte intensité de connaissances, Adaptive case management, Gestion de cas, Réutilisation de connaissances, Extraction de connaissances, Analyse de concepts formels, Traitement automatique du langage, Numérisation de l'enseignement, Enseignement

Abstract

CREA: Method for knowledge intensive process analysis, adaptation and reuse

use case in postgraduate computer science studies

Similarly to business domains, digitalisation and virtualisation of processes in higher education started a few decades ago. During COVID-19 pandemics, the interest in solutions for developing and delivering classes on-line grew substantially. Nevertheless, solutions allowing for efficient adaptation and reuse of the existing courses and teaching materials in the new circumstances are still lagging. Transformation of a traditional course to a virtual one, developing a new course adapted for the audience are some examples of " *knowledge intensive processes* ". These processes share common properties making digitalization hard, to cite some:

— they depend on the extensive knowledge and experience of teachers analysing the context and adapting class accordingly,

— they involve an intense collaboration between all stakeholders in the higher education environment,

— they require creativity for adapting to uncertain context.

Deploying new knowledge intensive processes, or adapting existing ones in this unexpected situation, faced multiple challenges already known from this specific research domain. The main issue is: how to best reuse existing knowledge, including online stored knowledge, where the abundance of sources and data makes it difficult to select the most suited to teachers' requirements.

In this thesis we propose the CREA method that enables a reuse of past cases in the higher education domain, particularly in courses preparation. The CREA method supports the reuse of existing courses materials: (i) it presents graphically the gap (i.e. semantic difference) between the courses and (ii) it summarises the class sessions in a form of clusters of main terms to discuss. Other types of documents can also be integrated within the input courses materials (including web pages, or research articles) in order to propose adapted clusters for specific audiences, or even state-of-the-art materials. This method relies on natural language processing tools in order to extract the terms regardless of the input language, then it uses formal concept analysis for computing metrics to build clusters of terms and assess the similarity of input materials. We also propose some preliminary results of a session scheduling method.

Keywords: Knowledge intensive process, Adaptive case management, Case Management, Knowledge Reuse, Knowledge Extraction, Formal concept analysis, Natural language processing, Digitalisation of education, Education

Table des matières

Table des figures

Liste des tableaux

1. INTRODUCTION

Dans ce chapitre, nous introduisons le sujet de cette thèse en expliquant succinctement les processus à forte intensité de connaissances, puis leurs concrétisations dans le contexte du domaine de l'enseignement supérieur, en particulier dans le cas de la construction d'un cours et de la réutilisation des connaissances. La problématique de recherche est ensuite détaillée ainsi que les hypothèses et stratégies de validation.

Sommaire

1.1 Contexte sociétal : l'enseignement supérieur et la réutilisation des connaissances

Le numérique un peu plus présent chaque jour, nous montre simultanément les nombreuses opportunités pour accroître notre confort et notre productivité, mais également les limites rencontrées face à certains traits humains encore difficiles à mesurer par des machines et leurs capteurs (les subtilités des jeux de mots et de l'humour, par exemple). Ces limites se retrouvent dans de nombreux aspects des métiers, y compris dans la gestion des processus qui a pourtant su profiter de l'automatisation croissante [124]. La gestion des processus métier modélise traditionnellement les tâches et activités, ainsi que leur ordonnancement, afin de décrire *comment* réaliser un processus [65]. Cependant, des caractéristiques liées au contexte ou à l'expertise des différents participants sont difficilement capturées et représentées dans les modèles existants, rendant certains processus peu répétables en l'état. Ces processus en particulier ont été étudiés et font l'objet d'un champ de recherche supplémentaire dans la gestion et les systèmes d'informations : les *processus à forte intensité de connaissances*, ou *knowledge intensive processes* (KIP) en anglais. Les processus à forte intensité de connaissances sont caractérisés [21] par : la manipulation de connaissances (implicites ou explicites), la collaboration entre participants au processus, l'imprédictibilité du contexte et de l'ordre des activités (voire l'apparition d'activités inconnues lors de la conception du processus), l'émergence d'informations au fur et à mesure qui influent sur la décision des activités à exécuter, l'apparition de buts intermédiaires à atteindre pour mener à bien le processus, la prise en charge d'évènements parfois imprévisibles, le respect de règles et contraintes limitant les activités possibles, et enfin, la non-répétabilité due aux situations uniques.

Les processus à forte intensité de connaissances se retrouvent dans de nombreuses activités où l'usage de connaissances dans un contexte collaboratif prédomine : une réunion, par exemple, est annoncée avec un ordre du jour, mais celui-ci ne sera pas toujours respecté, voire, des points seront ajoutés à cause d'évènements inattendus se produisant entre temps ou parce que des participants auront partagé des informations importantes. Les processus à forte intensité de connaissances peuvent parfois se rapprocher des processus classiques (prévisibles, détaillés, et dont les états sont connus à l'avance), par exemple lors de la résolution de problèmes connus et récurrents, ou au contraire s'en éloigner totalement lorsqu'une très forte exigence de créativité est nécessaire, comme dans les productions artistiques par exemple. Il existe plusieurs points de vue permettant de mieux apprécier le parcours entre plusieurs activités. En utilisant deux points de vue sur le suivi d'un patient dans un parcours médical, on se rend compte qu'un hôpital est bel et bien géré par des processus parfaitement prévisibles, mais le patient venant en consultation ne sait pas du tout à l'avance s'il sera hospitalisé ni quels tests il devra effectuer. Par exemple, effectuer une radiographie s'exécute avec des techniciens de santé précis en respectant une procédure bien définie, mais le personnel de santé ne sait pas à l'avance quels tests seront nécessaires pour chaque patient. Cette distinction entre les points de vue *processus* (la structure, l'ordre, et la connaissance en amont des activités à effectuer) et *cas* (chaque instance de processus évolue différemment des autres, et dépend des personnes impliquées) illustre

comment le contexte et les participants influencent le déroulement des activités. Là où l'hôpital gère des équipes et des machines selon des procédures précises, tout en devant s'adapter aux évènements survenant dans la société (en 2020 et 2021, il est plus probable de voir arriver des patients en détresse respiratoire, par exemple), le patient sera au contraire vu comme un cas évoluant au fur et à mesure des tests et résultats afin de l'orienter vers les spécialistes adaptés et lui fournir le traitement spécifique à sa situation.

Les caractéristiques des processus à forte intensité de connaissances soulèvent plusieurs défis concernant la gestion du flux de connaissances manipulées, l'aspect collaboratif et les décisions dépendant du contexte, l'intégration du contexte à la conception, la conformité des processus et de leurs instances, la flexibilité requise par les utilisateurs, et enfin, la réutilisation de fragments de processus [8]. Ce dernier défi est au croisement des deux points de vue *cas* et *processus* car il vise à réutiliser les résultats des instances passées de processus (assimilables à des cas terminés et validés) pour aider les utilisateurs à exécuter de nouvelles instances et construire de nouveaux processus (tous les deux assimilables à de nouveaux cas).

Dans le contexte de l'enseignement supérieur, l'intégration du numérique s'effectue depuis une quinzaine d'années grâce aux *innovations pédagogiques numériques* (IPN) [26]. Les IPN ont permis de transformer en outils en ligne de nombreux processus administratifs utilisant auparavant des données sur support papier, personnaliser les parcours de formation en les rendant plus flexibles, mais aussi soutenir le développement de nouvelles solutions pédagogiques s'appuyant sur le numérique [27]. Comme dans toutes les organisations, intégrer complètement ou partiellement les processus à forte intensité de connaissances aux systèmes informatiques est beaucoup plus difficile que pour les processus plus classiques : la crise du COVID-19 a montré que le déploiement massif de solutions de visio-conférences est certes un soutien technique supplémentaire, mais pas une prise en charge complète de l'activité d'enseignement. La construction d'un cours, ou sa reconstruction lors du changement de l'enseignant responsable, est typiquement une activité impliquant d'analyser précisément le contexte dans lequel le cours doit être donné (niveau de la filière/du diplôme, autres cours dépendants des connaissances transmises aux étudiants, cours précédemment suivis par les étudiants, ...) afin de sélectionner les informations précises à transmettre parmi celles existantes dans le domaine. Un enseignant développant un nouveau cours, ou devant reprendre un cours existant en se l'appropriant et en le mettant à jour, doit donc s'assurer de la pertinence des sources utilisées pour le contexte qu'il vise, et sélectionner des parties réutilisables pour former un support de cours à jour, compréhensible, et utile aux étudiants.

Cette thèse présente nos travaux visant à extraire des connaissances issues de cas passés, évaluer leur pertinence, et proposer à un utilisateur de les réutiliser dans un format organisé. Pour cela, nous avons choisi le domaine de l'enseignement supérieur et de la recherche, et plus spécifiquement le cas d'un enseignant cherchant à construire un cours sur plusieurs séances à partir de supports de cours existants, et éventuellement en y ajoutant des articles de recherche traitant du même sujet. L'étude d'un domaine

de recherche nécessite de maintenir une connaissance des questions et défis actuellement posés, auxquelles les contributions tentent de répondre, tout en collaborant avec l'ensemble de la communauté. Une grande créativité est donc requise pour poser des hypothèses et concevoir des contributions permettant de les confirmer ou infirmer. Lorsqu'un domaine devient suffisamment mature grâce à une somme conséquente de contributions, une diffusion de ces connaissances au-delà des experts peut se réaliser au travers de monographies ou de cours de spécialisation pour former de nouveaux experts. La préparation d'un cours constitue donc une synthèse des connaissances connues et vérifiées, afin de la transmettre à des personnes potentiellement peu voire non-initiées. Nous n'aborderons pas la façon de transmettre ces connaissances (représentant la gestion du flux de connaissances), ceci appartenant au domaine de la science de l'éducation, mais nous nous pencherons sur la façon d'extraire ces connaissances des documents existants afin d'évaluer leur pertinence vis-à-vis d'un sujet en particulier (prise en compte du contexte) et proposer une réorganisation de ces connaissances sous forme de séances de cours (réutilisation de fragments d'instances passées).

1.2 Problématique de recherche

Les diverses crises récemment rencontrées ont mis en évidence les limites des outils mis à disposition, et plus largement les difficultés à intégrer les processus à forte intensité de connaissances aux systèmes informatiques. D'après la littérature [8] plusieurs défis restent à relever, en particulier celui de la réutilisation de fragments de processus. Plusieurs travaux ont étudié ces défis et ont proposé des contributions s'adressant à plusieurs domaines d'application. Les fragments de processus [29][28] sont étudiés dans quelques travaux soit en s'intéressant au *raisonnement à base de cas* [104], ou *Case-Based Reasoning* (CBR) en anglais, qui réutilise explicitement les cas passés [14], soit en reconstituant un modèle de processus avec des contraintes d'exécution locales à chaque activité [22]. Enfin, certains travaux [132] génèrent des *motifs* (ou *patterns* en anglais) à partir de textes légaux afin de s'assurer du respect de la réglementation dans l'industrie alimentaire. Cette méthode s'appuie sur la rigidité des règles de rédaction des documents réglementaires pour en extraire des motifs, et exploite les expressions régulières pour s'assurer de leur respect.

Dans le cadre de l'enseignement supérieur et de la recherche, nous nous intéressons en particulier au processus de création de cours, et particulièrement à l'aide à la construction de cours à partir de supports existants. La préparation d'un cours exige beaucoup de temps, y compris lors de sa mise à jour pour inclure des nouveautés ou avec les dernières avancées de la recherche académique. Les récents évènements nous montrant également qu'il est parfois nécessaire de modifier rapidement les contenus pour s'adapter, nous proposons une méthode permettant d'utiliser des documents texte afin de mesurer leur cohérence vis-à-vis d'un sujet donné et d'en extraire une structure de séances réutilisable. La proposition de cette thèse vise donc à aider les enseignants à mieux gérer certains de leurs processus à forte intensité de connaissances, en particulier ceux impliqués dans le cadre de la construction d'un cours. Nos travaux visent donc à répondre à la question de recherche suivante :

Comment aider un enseignant à construire un cours à partir de documents existants, et en s'appuyant sur la réutilisation de connaissances ?

Cette question vise donc deux défis posés par les processus à forte intensité de connaissances : la réutilisation de fragments pour former des nouveaux cours, et la prise en compte du contexte au travers de la vérification de la pertinence des documents. Pour répondre à cette question et aux défis associés, nous posons plusieurs sous-questions de recherche et hypothèses :

1.a Comment définir le contexte d'un cours ?

H1	Le contexte d'un cours est défini par les documents sélectionnés	Comparer plusieurs corpus documentaires : certains homogènes sur un seul sujet, d'autres contenant plusieurs sujets *[Voir scénarios n°1-2-3-4-5]*

1.b Comment extraire les termes pertinents par rapport au contexte ?

H2	Des techniques de TAL permettent d'analyser les documents et d'extraire les termes pertinents	Vérifier les termes retenus par les outils de TAL une fois leurs analyses terminées, le tout sur plusieurs sujets distincts *[Voir scénarios n°1-2-3-5]*

2 Comment s'assurer de la pertinence des documents sélectionnés dans le contexte visé et des termes extraits ?

H3	L'analyse du graphe d'impact mutuel permet d'évaluer la pertinence des sous-ensembles de documents et d'identifier les écarts entre les documents	Comparer plusieurs corpus documentaires : des corpus homogènes sur un seul sujet doivent montrer un ensemble de documents très rapprochés ou tous aussi éloignés les uns des autres, d'autres corpus contenant plusieurs sujets doivent montrer que des documents sont plus éloignés du sujet central traité par la majorité des documents *[Voir scénarios n°1-2-3-4-5]*

3.a Quelles connaissances tacites sont présentes dans les documents étudiés ?

H4	Étant donné que la rédaction de supports de cours est une activité humaine avec un objectif de transmission de connaissances sur un sujet, considérons que les connaissances tacites suivantes peuvent être extraites d'un tel support :	Expliciter les liens en regroupant les termes liés et évaluer la qualité de ces regroupements, expliciter l'ordre des regroupements et évaluer la qualité de ces regroupements *[Voir scénarios n°1-3-5]*
H4a	Sélection de termes particuliers pour un contexte particulier visé (et non l'ensemble du champs lexical possible)	
H4b	Organisation de termes sous forme de groupes logiques (séquence, section, chapitre, ...) pour traiter un aspect particulier à la fois	
H4c	Présentation de thèmes au fur et à mesure selon un ordre précis (permettant d'assurer que les pré-requis soient abordés en premier)	

3.b Comment extraire les connaissances tacites présentes dans les documents étudiés ?

H5	La similarité conceptuelle permet l'extraction de connaissances tacites, en particulier les regroupements logiques de termes traitant d'un ou quelques aspects particuliers d'un sujet	Construire des regroupements de termes en utilisant la similarité conceptuelle, et évaluer ces regroupements *[Voir scénarios n°1-3-5]*

4 Comment la sélection initiale des documents peut-elle impacter les résultats d'application de la méthode CREA ?

H6	Les résultats d'application de la méthode CREA sont exploitables par les enseignants et ne sont pas impactés par :	Tester plusieurs scénarios dont les sujets, langues, et nombre de documents varient *[Voir scénarios n°1-3-5]*
H6a	Le choix initial du sujet de cours	
H6b	Le nombre de document initialement sélectionnés	
H6c	La langue des documents sélectionnés	
H7	Les résultats d'application de la méthode CREA peuvent être impactés par :	Tester la méthode avec des documents hétérogènes et des contenus diversifiés afin d'évaluer la qualité des résultats *[Voir scénarios n°2-4-5]*
H7a	La présence de documents hétérogènes dans la sélection initiale (articles de recherches, supports de cours, livres,...)	
H7b	La présence de parties (section, chapitre, ...) hors sujet au sein d'un document	
H7c	La nature du contenu des documents (images, listings du code, ...)	

5 Comment présenter les résultats de la méthode CREA afin d'améliorer l'exploitabilité pour un enseignant ?

H8	Le graphe d'impact mutuel permet d'aider un enseignant en rendant certaines informations plus visuelles :	Comparer plusieurs corpus documentaires : des corpus homogènes sur un seul sujet doivent montrer un ensemble de documents très rapprochés ou tous aussi éloignés les uns des autres, d'autres corpus contenant plusieurs sujets doivent montrer que des documents sont plus éloignés du sujet central traité par la majorité des documents *[Voir scénarios n°1-2-3-4-5]*
H8a	Le graphe d'impact mutuel permet de visualiser le contexte traité par le corpus documentaire	
H8b	Le graphe d'impact mutuel permet de visualiser l'écart de chaque document par rapport au contexte (cf H3)	
H8c	Le graphe d'impact mutuel n'est pas une représentation suffisante pour visualiser avec le maximum de précision les écarts entre documents	
H9	La présentation de clusters de termes sous la forme de tableaux impose implicitement l'ordre de lecture	Utiliser des méthodes de visualisation de données (nuage de mots, ...) *[Non traité/Voir discussions et conclusion]*

Les travaux précédemment cités se concentrent soit sur des domaines trop éloignés (industrie alimentaire), soit sur des processus trop structurés où les tâches sont au cœur d'un modèle, soit n'exploitent pas assez la sémantique des textes et des connaissances contenues. Des travaux existants [110] se sont intéressés à l'extraction de connaissances et leur analyse avec des techniques mathématiques d'analyse de données, en particulier avec l'*analyse de concepts formels*, ou *formal concept analysis* (FCA) en anglais. Notre contribution, la méthode CREA (*Case REuse and Adaptation*), y adjoint des métriques et visualisations issues de l'analyse de concepts formels étudiées dans d'autres travaux [58] afin de pouvoir valider la réutilisation de fragments de processus et la vérification de la pertinence. Plusieurs scénarios d'utilisation sont possibles : construire un tout nouveau cours à partir de supports existants (des documents sont rassemblés et une structure de cours est proposée), mettre à jour son propre cours en le comparant avec les supports d'autres cours (une carte indique l'éloignement de son cours par rapport aux autres), voire en l'étoffant avec des notions issues d'articles de recherche du monde académique (des articles de recherche sont ajoutés à la base de documents insérés afin de proposer une structure plus étendue encore).

Nous avons également commencé à étudier l'aspect temporel en proposant succinctement parmi les perspectives un ordonnancement pour les séances générées. Les documents dont l'organisation est chronologique permettent d'en extraire non seulement des connaissances, mais également l'ordre de présentation de ces connaissances. L'ordonnancement temporel a déjà été étudié dans certains travaux [22]. Nous présentons cependant des résultats préliminaires d'une version adaptée à l'organisation temporelle de clusters de notions. Une extension à la contribution principale et une expérience sont présentées en conclusion pour ordonnancer partiellement les clusters précédemment générés.

1.3 Plan du manuscrit

La méthode CREA proposée dans cette thèse vise à répondre au problème de recherche grâce à deux de ses productions en sortie :

— une visualisation graphique permettant de déterminer la pertinence des documents en entrée les uns par rapport aux autres et selon le(s) sujet(s) traité(s),

— des regroupements de termes issus des documents représentant les notions à aborder dans le nouveau cours en construction.

Le manuscrit de thèse est organisé comme suit :

— le chapitre 1 a introduit le contexte sociétal et la problématique de recherche visée ;

— le chapitre 2 présente une revue de la littérature concernant les domaines de la gestion des connaissances et des processus à forte intensité de connaissances, puis les techniques d'analyse de données utilisées pour la méthode CREA, et enfin les travaux connexes et similaires ;

— le chapitre 3 expose la méthode CREA en présentant tout d'abord le fonctionnement général et le cadre de travail, puis les deux phases de pré-traitement sémantique et d'analyse structurelle sont détaillées ;

— le chapitre 4 détaille la méthodologie d'évaluation, l'ensemble des expérimentations et leurs résultats en utilisant la méthode CREA dans plusieurs scénarios, puis discute ces résultats ;

— le chapitre 5 effectue une synthèse des contributions, usages possibles de la méthode CREA, et menaces de validité, puis nous proposons plusieurs perspectives pour chaque phase, et enfin nous présentons des résultats préliminaires concernant l'ordonnancement temporel des regroupements de termes afin de proposer un syllabus précis à l'enseignant.

2. CONTEXTE

Dans ce chapitre, nous introduisons le contexte par la gestion des connaissances, en particulier la description des connaissances tacites, des connaissances explicites, et du cycle SECI. Nous détaillons ensuite les processus à forte intensité de connaissances et les défis rencontrés dans ce domaine de recherche. Nous analysons particulièrement le défi de la réutilisation des point de vues de la gestion des connaissances et des processus à forte intensité de connaissances. Quelques techniques de traitement automatique de la langue, d'analyse de données, et de clustering utiles pour exploiter les connaissances issues de ces processus sont ensuite détaillées. Enfin, nous présentons plusieurs travaux traitant de la réutilisation afin de conclure sur le positionnement de la méthode CREA et en quoi l'usage des techniques précédemment citées permet de répondre aux besoins de la gestion des connaissances et des processus à forte intensité de connaissances.

Sommaire

2.1 Processus à forte intensité de connaissances : revue de la littérature

Durant les dernières décennies, l'essor du numérique a permis de passer d'objectifs liés à l'accroissement de la productivité à l'exploitation de connaissances [84][108]. Cette réorientation depuis les ressources physiques vers les connaissances a entraîné de nombreux changements dans les organisations. Plusieurs exemples [45] concernent par exemple Samsung et l'envoi de ses meilleurs employés à l'étranger pendant un an ou plus pour nouer des contacts en s'appuyant sur l'aspect relationnel, 3M laissant 15% de temps de travail à ses employés de la branche recherche et développement pour travailler sur les projets de leur choix et laisser libre cours à leur créativité afin de continuellement disposer de nouveaux produits, ou encore HP et sa « *HP Way* » mettant en avant l'individu et ses spécificités plutôt que de les formater à l'entreprise (ce qui inspira Google et d'autres). Plus généralement, on peut également observer le développement de méthodes orientées collaboratif (méthodes agiles), ou encore l'intérêt croissant pour l'innovation en exploitant les idées et la créativité.

Afin de mieux comprendre ces implications pour notre problématique, nous nous intéressons tout d'abord aux spécificités des connaissances tacites (ou implicites) par opposition aux connaissances explicites. Puis nous détaillons comment les processus à forte intensité de connaissances travaillent avec ces connaissances et quels défis restent à relever. Enfin, nous expliquons en particulier le problème de la réutilisation dans le cadre de la gestion des connaissances, puis des processus à forte intensité de connaissances.

2.1.1 La connaissance du point de vue de la gestion des connaissances

Les connaissances sont de plus en plus reconnues comme un avantage compétitif au sein des entreprises [1][37][46]. Dans les premiers niveaux de *l'échelle de la connaissance* [84] (« *knowledge ladder* » en anglais), illustrée par la figure 2.1, plusieurs concepts sont présentés pour obtenir des connaissances. Les *symboles* sont organisés grâce à des règles de syntaxes afin d'obtenir des *données* (des textes ou des nombres, par exemple). Ces *données*, une fois enrichies d'une signification, permettent d'établir des *informations* (une valeur devient une température lorsqu'on y adjoint une unité de mesure appropriée). Enfin, les *informations* interprétées selon le contexte, l'expérience, et les attentes de l'individu les manipulant deviennent des *connaissances*. Les connaissances sont avant tout un processus et non pas un objet manipulable [84]. Lorsqu'un individu agit en exécutant une tâche (*actions*), une partie de ses connaissances se matérialisent dans ses gestes, habitudes, et réflexes (*savoir faire*), mais également dans l'intention qu'il s'est fixé (*pourquoi faire*, avec quel but). Les connaissances correspondent à la capacité des individus à comprendre les relations entre les phénomènes. Les connaissances et leur transmission sont des sujets d'intérêts centraux du domaine de la *gestion des connaissances* [108][84], ou *knowledge management* (KM) en anglais. Deux types de connaissances ont été identifiés : les connaissances dites « *explicites* » et les connaissances dites « *tacites* » (ou *implicites*) dont la transmission est difficile de par leur nature [82].

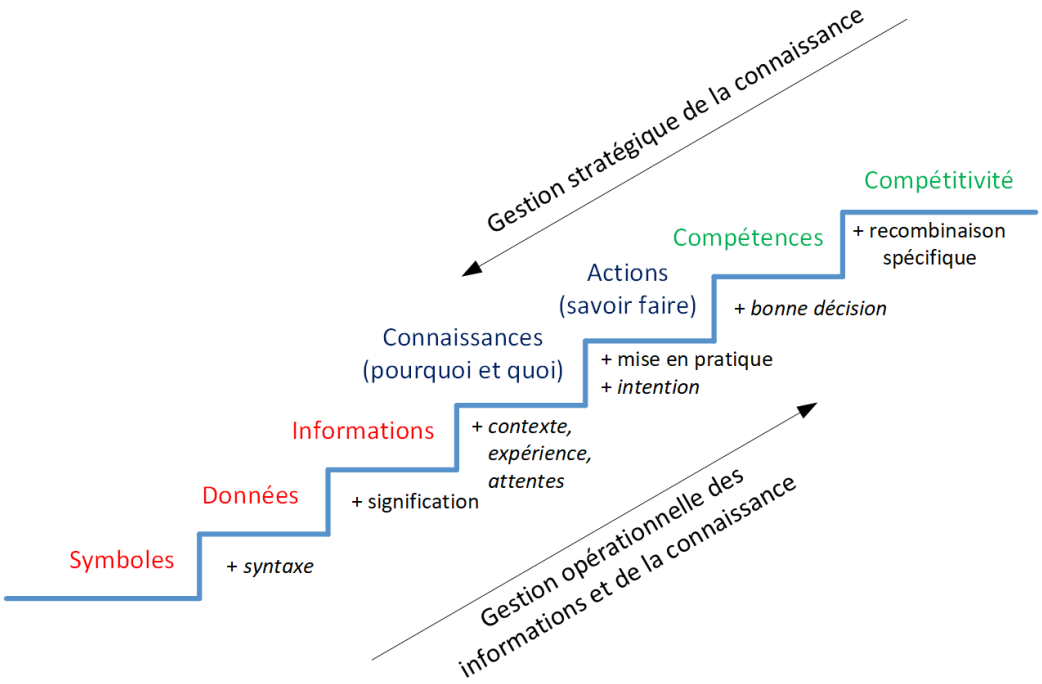


FIG. 2.1 – Échelle des connaissances traduite de [84]

Connaissances tacites et explicites

Les *connaissances explicites* sont formalisées, structurées, systématiques, codifiées et peuvent donc aisément être manipulées, conservées, et partagées [84][108][82][83]. Les écrits et schémas sont les exemples typiques de connaissances explicites permettant de les transférer facilement d'un individu à l'autre. Les aspects formel et structuré sont particulièrement visibles dans les formules mathématiques et les spécifications techniques. Un exemple concret abordé dans [82] concerne la conception d'une machine à pain dont le pétrissage est insatisfaisant. Afin d'améliorer le pétrissage, un boulanger expert est observé afin de recopier sa technique et permettre à la machine de générer des pâtes d'une qualité similaire aux siennes. Les connaissances explicites pouvant être extraites concernent les spécifications de la pâte en entrée puis en sortie, c'est-à-dire ses qualités mesurables (dureté, humidité, ...).

Les *connaissances tacites* concernent au contraire toute l'expertise acquise par un individu au travers de ses sens et de ses facultés. Les gestes d'un artisan, ou l'intuition issue d'expériences personnelle sont des exemples typiques de connaissances tacites [84][108][82]. Ces connaissances sont particulièrement difficiles à transmettre car elles dépendent complètement de l'individu et ne peuvent pas être formalisées ou exprimées aussi facilement que les connaissances explicites. La transmission de certaines connaissances tacites peut se faire au travers de l'apprentissage : un *maître* exécute des gestes, et son *apprenti* les répète jusqu'à pouvoir les maîtriser, ou au moins les répéter seul pour s'améliorer de lui-même (voire pour développer sa propre technique). Dans l'exemple de la machine à pain et du boulanger observé [82], les connaissances tacites concernent la technique du pétrissage manuel. Le boulanger peut expliquer

succinctement quelques étapes, mais l'expertise précise transformant la pâte implique de répéter le geste et ressentir ses effets. Une ingénieure a donc pratiqué la technique de pétrissage avec le boulanger pour comprendre le geste et ses effets, afin de pouvoir en déduire les parties manquantes dans la machine à pain conçue.

Appliqué à l'enseignement, on peut retrouver les connaissances qu'un enseignant souhaite transmettre à ses étudiants en s'appuyant sur des connaissances explicites (le syllabus, les notions ciblées, et le support de cours associé) et les connaissances tacites de l'enseignant (expérience personnelle en tant qu'ancien étudiant, cours déjà enseignés dans le passé, par exemple). Les connaissances tacites constituent précisément dans ce cas une valeur ajoutée à l'enseignant : choix dans l'ordre des notions présentées, accent mis sur certaines d'entre elles selon le profil des étudiants, et explications préférées par chaque groupe d'étudiants.

Cycle SECI

Ces deux types de connaissances participent à un cycle dénommé *SECI* [82] (Socialisation, Extériorisation, Combinaison, Intériorisation) illustré par la figure 2.2. Ce cycle s'appuie sur les quatre façons de créer et transformer des connaissances : de tacite à tacite (socialisation), de tacite à explicite (extériorisation), d'explicite à explicite (combinaison), d'explicite à tacite (intériorisation). Chacune de ces façons correspond à des situations particulières qui se répètent :

— Socialisation (de tacite à tacite) : Deux individus partagent des connaissances tacites par la socialisation. Il s'agit précisément de la relation où un apprenti observe, imite, et pratique ce que le maître fait, c'est-à-dire un transfert d'expérience d'une personne à l'autre. La socialisation repose sur l'accumulation et le transfert de connaissances tacites d'un individu à l'autre.

— Extériorisation (de tacite à explicite) : L'extériorisation de connaissances tacites permet de les expliciter, c'est-à-dire les formaliser et en tirer une expérience utile à l'avenir. Modéliser un processus est une forme d'extériorisation, tout comme décrire le déroulement d'un projet pour pouvoir apprendre des points positifs et éviter certaines erreurs. L'extériorisation repose sur l'articulation de connaissances tacites des membres d'un groupe pour pouvoir les traduire en concepts explicites.

— Combinaison (d'explicite à explicite) : Plusieurs sources de connaissances explicites, telles que des documents ou des informations, peuvent être combinées pour générer de nouvelles connaissances explicites. Lire plusieurs rapports ou articles pour en déduire une décision stratégique ou une nouvelle piste de recherche et présenter ces résultats aux collaborateurs sont des exemples. La combinaison repose sur la recherche et l'intégration de connaissances explicites existantes, pour pouvoir en générer de nouvelles et les diffuser aux groupes d'une organisation.

— Intériorisation (d'explicite à tacite) : Une fois de nouvelles connaissances explicites reçues, un individu les intériorise en les utilisant dans ses activités. Un ingénieur en lisant une documentation formalisée découvrira des nouveautés qu'il pourra mettre en application par la suite et se les approprier, créant des connaissances tacites. L'intériorisation concerne un individu qui apprend de nouvelles connaissances explicites au sein d'un groupe ou d'une organisation et se les approprie par la pratique, créant ainsi des connaissances tacites.

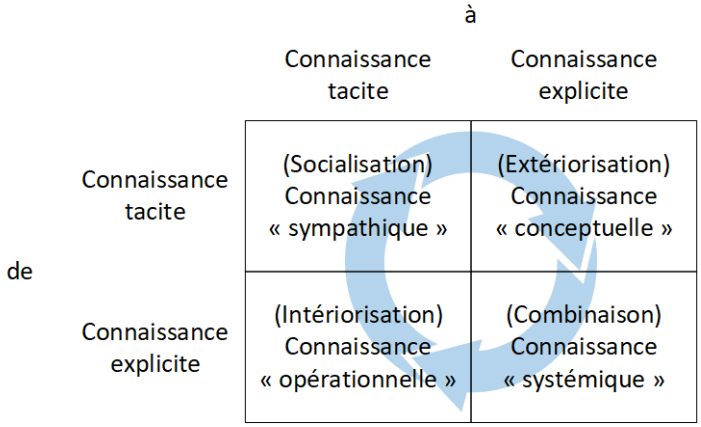


Fig. 2.2 – Modèle SECI de [82] (traduction française de [111])

Dans l'enseignement supérieur et la recherche, ce cycle se retrouve à différentes échelles. Lors d'une conférence, des chercheurs vont présenter leurs contributions (extériorisation) et discuter les uns avec les autres (socialisation), ce qui peut engendrer la création d'une nouvelle communauté de recherche partageant des objectifs communs (combinaison) et dont chaque membre développera de nouveaux travaux selon sa propre expérience et son contexte personnel (intériorisation).

Pour un cours, les supports de cours représentent les connaissances explicites, et les échanges entre enseignant et étudiants représentent les connaissances tacites. Un enseignant s'informera tout d'abord sur le sujet à enseigner en recherchant plusieurs sources (combinaison). Ensuite, l'enseignant lira les sources pour s'imprégner du sujet et sélectionner les parties à présenter aux étudiants selon ses exigences (intériorisation). Enfin, l'enseignant délivrera son cours lors de séances en travaux dirigés ou éventuellement en cours magistraux en s'appuyant sur un support de cours formel (extériorisation). Les travaux dirigés étant particulièrement adaptés aux questions individuelles, et donc à l'apprentissage pas à pas (socialisation), cette étape est répétée en parallèle du cours magistral. À la fin de chaque séance, l'enseignant vérifie l'avancement par rapport aux objectifs fixés (intériorisation) pour éventuellement modifier son support de cours en conséquences par rapport à ses sources (combinaison).

Inversement, du côté des étudiants, ceux-ci reçoivent un support de cours qu'ils doivent lire (intériorisation) et mettre en pratique lors des travaux dirigés (socialisation). Les notes de cours prises (extériorisation) sont relues (combinaison) pour pouvoir se préparer aux séances suivantes. Les notes de cours, devoirs, examens, et

présentations sont des exemples de mise en forme explicite de connaissances tacites acquises.

L'exemple de l'enseignement montre que l'ordre des étapes n'est pas toujours rigoureusement le même, mais respecte constamment la transition entre chaque type de connaissances : les étudiants peuvent éventuellement travailler en groupe après une séance (socialisation) avant un examen (extériorisation). Les connaissances tacites sont échangées entre camarades avant d'être explicitées dans un examen ou dans une présentation dont les membres recombineront les informations au fur et à mesure.

Ces activités centrées sur les connaissances, qu'elles soient tacites ou explicites, sont au cœur de processus actuellement difficiles à gérer pour de nombreuses raisons. Les individus et leurs connaissances tacites sont un exemple de difficulté : il est impossible de savoir à l'avance quelles connaissances tacites un individus possèdera en lui, ni les décisions qu'il prendra. Un domaine de recherche s'intéresse en particulier aux défis posés par les connaissances dans les processus au sein des organisations, c'est-à-dire aux *processus à forte intensité de connaissances*, que nous présentons dans les sous-sections suivantes.

2.1.2 Les processus à forte intensité de connaissances

Observer des phénomènes se produire, c'est observer des processus s'exécuter. Les processus tirent leurs origines linguistiques du latin *processus* ou *processioat* qui signifient *action exécutée* ou *quelque chose de fait*, et la *façon dont elle a été faite* [121]. Depuis Adam Smith et Frederick Taylor, le point de vue général adopté sur les processus métier est centré sur les activités, en particulier sur des tâches et leur ordonnancement [8]. Au sein des organisations actuelles, de nombreux processus métier sont modélisés et pris en charge par la *gestion de processus métier*, ou *business process management* [124] (BPM) en anglais. Certains processus centrés sur les connaissances sont cependant encore difficiles à gérer au sein des systèmes d'informations [8], il s'agit des *processus à forte intensité de connaissances*, ou *knowledge intensive processes* (KIP) en anglais. Des synonymes peuvent être retrouvés dans la littérature tels que *knowledge intensive business processes* [55][67] ou *artful processes* [50], mais certains autre termes comme *case* [18][116] ont une signification un peu plus particulière que nous présentons par la suite.

Comme présenté dans la sous-section précédente, les connaissances tacites et explicites impliquent certaines spécificités lors de leur intégration dans les systèmes d'informations. Les connaissances explicites sont par définition formalisées et visent à être transmises, celles-ci sont donc plus facilement intégrées aux méthodes classiques de gestion des processus. En effet, les activités de modélisation de processus, visant à formaliser un processus en le représentant sous forme de schéma, sont déjà intégrées dans le cycle de vie du BPM et sont un exemple d'extériorisation. Un autre exemple illustrant cette fois la combinaison concerne la rédaction des rapports annuels : les informations agrégées sont directement issues du système d'informations, parfois même

de façon automatique. À l'inverse, les connaissances tacites étant par définition informelles et uniques à chaque individu, celles-ci sont beaucoup moins aisées à gérer. La socialisation peut éventuellement être facilitée par l'organisation d'évènements ou d'ateliers, mais l'intériorisation est au contraire une étape totalement individuelle et ne peut donc pas simplement être provoquée par une volonté extérieure.

L'étude des processus à forte intensité de connaissances s'intéresse particulièrement, mais pas exclusivement, aux connaissances tacites afin de mieux les prendre en charge et exploiter leur potentiel. L'importance des connaissances dans ces processus leur confère plusieurs caractéristiques et exigences telles que la flexibilité dans l'ordre des activités, des évènements inattendus impliquant des tâches imprévisibles (donc absentes des modèles de processus conçus à priori), de la créativité, et bien d'autres.

Caractéristiques des KIP

Plusieurs études ont cherché à comprendre plus en profondeur les KIP et leurs caractéristiques afin de pouvoir, si possible, les intégrer aux outils et méthodes BPM déjà en place. Les travaux présentés dans [55] ont cherché dans la littérature les différences entre les KIP et les processus métier plus traditionnels. La plupart des caractéristiques n'opposent pas directement les deux types de processus, mais les KIP en accentuent certaines. Typiquement, « *les degrés exigés de complexité, répétabilité, et créativité ne sont pas les mêmes, mais ils ne sont pas moins éligibles à l'automatisation ou l'organisation structurée* » [55] (« *different levels of complexity, repeatability and creativity required for these processes, but are not necessarily less eligible for automation or structured* » [55]). Cependant, l'article conclut tout de même que les KIP sont « *plus complexes, moins répétables, et exigent beaucoup de créativité* » [55].

Les travaux dans [21] insistent tout d'abord sur les liens entre le BPM et le domaine du KM qu'il faut développer. En effet, la connaissance est un élément clé jusque là peu exploité dans le BPM, alors qu'elle contient d'après [16] « *l'expérience, le contexte, l'interprétation et la réflexion, et implique plus de coopération humaine que d'informations* » [21] (« *experience, context, interpretation and reflection and involves more human participation than information* » [21]). Cette vision met l'accent sur les connaissances tacites, étant donné que les connaissances explicites s'appuient sur des informations et des données déjà connues des systèmes d'informations, et donc des méthodes BPM mises en place. Plusieurs autres définitions sont également étudiées, mais une en particulier [115] insiste sur le rôle des *travailleurs du savoir* (*knowledge workers* en anglais) qui effectuent des tâches impliquant des prises de décisions à forte intensité de connaissances, c'est-à-dire en rassemblant de nombreuses informations pour en générer de nouvelles, voire pour produire des artefacts (par exemple des plans, des recommandations, ou encore des exigences liés aux décisions prises). Une définition des *artful processes* proposée dans [50] insiste également sur l'importance des individus exécutant le processus et surtout leurs connaissances tacites (« *compétences, expérience, jugement* » [50]) : le processus, voire les tâches, sont définissables d'un point de vue haut niveau, mais il est impossible de fixer à l'avance les détails. [21] souligne également le fait qu'un *artful process* est en fait défini par les personnes (et

leurs connaissances) qui l'exécutent : il n'est pas possible de définir un processus à proprement parler, mais il faut plutôt parler de l'instance en elle-même.

Toujours dans [21], huit caractéristiques des KIPs ont été retenues et sont présentées afin d'en déduire vingt cinq exigences. Les KIPs y sont considérés comme :

- Dirigés par les connaissances (« *knowledge-driven* ») : les données et connaissances disponibles servent à prendre les décisions et donc à diriger le processus.
- Orientés collaboratif (« *collaboration-oriented* ») : le contexte d'exécution du processus est collaboratif et multi-utilisateurs, et les connaissances et artefacts générés (autant par les humains que par l'instance de processus) sont partagés.
- Imprédictibles (« *unpredictable* ») : chaque instance dispose de son propre contexte et d'éléments imprévisibles qui font varier les connaissances exploitées, le flot d'exécution, et l'ordre des tâches d'une instance à l'autre.
- Émergents (« *emergent* ») : l'exécution du processus fait émerger au fur et à mesure des informations qui permettent de choisir les actions et décisions à prendre.
- Orientés buts (« *goal-oriented* ») : des buts intermédiaires servent à avancer dans l'exécution du processus.
- Dirigés par les évènements (« *event-driven* ») : les décisions prises par les travailleurs du savoir pour avancer dans l'exécution du processus dépendent des évènements rencontrés.
- Dirigés par les contraintes et les règles (« *constraint- and rule-driven* ») : les actions et décisions prises par les utilisateurs sont guidées par des règles ou contraintes à respecter.
- Non-répétables (« *non-repeatable* ») : les situations variant à chaque instance de processus, il est quasiment impossible de parfaitement répéter une précédente exécution.

Ces caractéristiques ne sont pas exhaustives, étant donné que le domaine est toujours actif et que les tâches non-liées à un système ou environnement d'exécution BPM n'ont pas été complètement dénombrées, mais elles sont particulièrement pertinentes. Une réunion, par exemple, implique par définition de la collaboration entre toutes et tous les participants, chaque point de l'ordre du jour doit être abordé (buts intermédiaires), elle est dirigée par les connaissances partagées au fur et à mesure des discussions et des points de vues de chacun, certains sujets seront peut être plus développés que d'autres (émergence et imprédictibilité), voire, la réunion peut être interrompue par un évènement inattendu. Dans tous les cas, il sera impossible de répéter scrupuleusement la réunion dans sa séquence et son contenu.

Il faut tout de même retenir qu'un KIP ne répond pas nécessairement à l'ensemble de ces caractéristiques, mais au moins à certaines d'entre elles. Le développement d'un logiciel peut tout à fait se réaliser seul (et sans utiliser de forum d'entre-aide) : il suffit d'avoir des connaissances en développement logiciel, et de respecter les limitations imposées par la machine et système d'exploitation ciblés. Cependant, ce logiciel sera peu réutilisable en l'absence de documentation explicitant le contexte précis d'utilisation.

De plus, l'absence de prise en compte des retours utilisateurs (d'ailleurs souvent issus de *l'expérience utilisateur*) limitera les évolutions. Les différentes méthodes agiles de développement d'applications [35] visent justement à essayer d'intégrer au mieux toutes ces caractéristiques : fixer des buts intermédiaires régulièrement, prévoir de la collaboration entre développeurs et utilisateurs, s'adapter au changement et limiter l'impact des évènements imprévus, ... tout en essayant de produire les applications les plus réutilisables possibles mais en s'adaptant au contexte de chaque cas.

Difficultés pour gérer les KIP

Les participants et leurs connaissances sont donc au cœur de ces processus dont les caractéristiques précédemment relevées sont difficilement intégrables telles quelles dans l'environnement BPM classique [96][79]. Nous avons identifié six défis concernant les KIPs et les avons décrits dans une publication [8].

1. « *Comment gérer de façon efficiente et efficace les connaissances et informations manipulées (c'est-à-dire créées, utilisées, mises à jours) par les KIPs ?* » L'absence de gestion des connaissances et des informations impacte négativement les KIPs. Les travaux de [51] exposent justement les difficultés auxquelles font face dans la pratique les organisations dans la gestion des grandes quantités d'informations. Bien que le domaine des *Mégadonnées* (*Big Data* en anglais) et ses 3V visent les problèmes liés aux *Volume, Vélocité et Variété*, il s'agit de manipuler des *données* et non pas des *informations* ni des *connaissances* [20].

2. « *Comment supporter l'aspect collaboratif et les prises de décisions spécifiques au contexte dans les KIPs ?* » Les organisations ont plutôt tendance à standardiser et contrôler les processus [96], or, limiter la collaboration et la créativité limite les capacités des utilisateurs [43] et contrevient aux caractéristiques mêmes des KIPs. Bien que des solutions tentent d'améliorer les communications informelles [24], l'aspect collaboratif entre les participants aux KIPs reste un défi à relever [79].

3. « *Comment intégrer les informations de contexte lors de la conception d'un KIP ?* » Le contexte d'exécution d'un KIP sous-entend la situation ou l'environnement d'exécution (l'état de l'organisation, des équipes impliquées, ou de la société en général) ainsi que les connaissances tacites des participants au KIP. Prévoir lors de la conception d'un processus dans quelle situation sera l'organisation et quelles seront les connaissances tacites des utilisateurs est quasiment impossible : chaque instance de KIP étant unique, des activités ou évènements inattendus peuvent se produire et apporter de nouvelles expériences aux utilisateurs qui ne répèteront pas nécessairement le même cheminement y compris si un cas similaire se reproduit. Le paradigme déclaratif [114] permet à minima de proposer des règles et un cadre à respecter, mais aussi dans lequel évoluer. Le contexte n'est donc pas complètement prévisible ni intégré, mais certaines conditions d'exécutions peuvent indiquer quelles activités faire ou préférer.

4. « *Comment supporter la conformité dans les KIPs ?* » Par définition, le paradigme impératif du BPM simplifie la mise en place de règles de conformité. Le domaine de la *gestion de règles métier* [98], ou *Business Rules Management* (BRM) en anglais, apporte une réponse particulièrement adaptée en séparant les règles des processus [134]. Comme indiqué dans le défi précédent : le paradigme déclaratif offre des pistes pour cadrer les KIPs, mais assurer une conformité stricte de façon automatisée est encore difficile. Le déroulement des activités dépendant surtout des décisions prises par les participants [66], ces derniers ont donc majoritairement un contrôle sur le processus et donc la responsabilité concernant le respect de la réglementation. Une autre contrainte réside aussi dans le fait que la règlementation est exprimée dans le langage naturel, voire avec des spécificités du droit, et qu'il est parfois long et difficile de les transformer en règles métier accessibles aux utilisateurs des KIPs [134]. Quelques solutions sont néanmoins étudiées pour simplifier cette transformation [132].

5. « *Comment supporter la flexibilité dans les KIPs ?* » L'imprédictibilité et la créativité, mais aussi l'émergence et les évènements se produisant au fil du temps, impliquent que les utilisateurs de KIPs ont besoin d'une grande liberté d'action. Comme nous l'avons vu, il s'agit cependant aussi de restreindre cette liberté selon les réglementations du domaine, voire, les contraintes liées au contexte particulier du cas étudié. La flexibilité nécessaire pour entretenir une certaine liberté d'action, tout en respectant les contraintes, est grandement étudiée dans la littérature [96][47][52][114][19][43][133], car il s'agit d'un problème encore ouvert.

6. « *Comment explorer et réutiliser des fragments de processus dans les KIPs ?* » Le domaine du *raisonnement à base de cas* [104], ou *Case-Based Reasoning* (CBR) en anglais, vise à ré-« *utiliser les expériences passées pour comprendre et résoudre de nouveaux problèmes* » [62] (« *using old experiences to understand and solve new problems* » [62]). Cette idée a été étudiée dans quelques travaux [21][14] en proposant aux utilisateurs des KIPs des *motifs* (ou *patterns* en anglais) issus d'instances précédemment exécutées. La validation des motifs par des experts du domaine peut permettre leur réutilisation en accord avec les règles en place. Ces motifs, selon leur nature, peuvent également embarquer une partie du contexte des exécutions précédentes sans intervention manuelle, et donc répondre à plusieurs défis précédemment énoncés.

Comme nous venons de le voir, les KIPs font encore face à de nombreux défis partiellement interdépendants. Plusieurs domaines de recherche connexes contribuent à affiner les réponses possibles, mais ne permettent pas en l'état d'intégrer les KIPs aux systèmes d'informations s'appuyant uniquement sur le BPM. La littérature propose plusieurs points de vues sur les KIPs permettant de mieux se concentrer sur certains défis.

2.1.3 Le défi de la réutilisation des fragments de processus et des connaissances

Le raisonnement à base de cas [104][62] s'appuie énormément sur la réutilisation des cas passés afin d'en résoudre de nouveaux. Ce point de vue des cas est issu de la *gestion de cas* [18][116][106], ou *case management* en anglais, qui s'appuie sur des pratiques où l'ensemble des informations nécessaires pour réaliser des activités sont rassemblées dans ce qui s'appelle un *cas* [106]. L'expérience acquise lors d'une activité se répercute dans les connaissances tacites des individus impliqués, et éventuellement sous forme explicite si des traces ou des artefacts ont été générés. Il s'agit de l'intériorisation dans le cycle SECI présenté précédemment : chaque individu a pratiqué une activité en s'appuyant sur des connaissances explicites qui sont devenues progressivement tacites. Ces connaissances tacites sont cependant uniques à la situation dans laquelle s'est réalisée l'activité. Réutiliser ces connaissances implique que l'individu se retrouve dans une situation similaire, c'est-à-dire que le contexte soit suffisamment proche pour qu'il puisse adapter ses actions.

La réutilisation des connaissances est étudiée dans la littérature, mais beaucoup moins que plusieurs autres étapes intervenant avant la réutilisation en elle-même [100] que nous décrivons par la suite. Du point de vue des processus à forte intensité de connaissances, les fragments de processus permettent de s'appuyer sur l'expérience des individus dans leurs activités pour extraire un sous-ensemble de tâches et activités afin de les réutiliser lors de la conception et de l'exécution.

Le défi de la réutilisation dans le cas de la gestion des connaissances connaissances

Les travaux présentés dans [68] étudient la réutilisation de connaissances sous sa forme de processus avec des rôles et des dépôts de connaissances, pour en déduire quatre types de situations de réutilisation possibles. Le processus général de gestion des connaissances se divise en quatre étapes :

1. Capturer ou documenter les connaissances (« *capturing or documenting knowledge* ») : documentation de(s) l'activité(s) par des moyens automatiques ou induits par l'activité(s) elle(s)-même(s) (archivage automatique de messages échangés, par exemple), ou par des structures prévues à la conception (champs obligatoires à remplir, par exemple).

2. Empaqueter les connaissances à réutiliser (« *packaging knowledge for reuse* ») : les documents générés sont nettoyés puis mis en forme pour pouvoir être indexés dans un dépôt. Les informations concernant le contexte sont ajoutées, le contenu des documents est filtré et nettoyé, et des modèles de classification sont développés.

3. Distribuer ou disséminer les connaissances (« *distributing or disseminating knowledge* ») : partage des connaissances de façon passive (dépôt alimenté régulièrement par les nouvelles expériences, ou envoi régulier d'une note d'informations, par exemple) ou active (organisation de réunions détaillant un retour d'expérience particulier, par exemple), et activités de facilitation (aide à l'utilisation

ou à l'identification de besoins en réutilisation de connaissances, ou mise en place de bonnes pratiques, par exemple).

4. Réutiliser les connaissances (« *reusing knowledge* ») : Cette dernière étape se décompose elle-même en quatre activités. Tout d'abord, *définir la question* (« *defining the search question* ») afin de mobiliser les connaissances et experts les plus aptes à y répondre. Puis, *chercher et localiser les experts ou les expertises* (« *the search for, and location of, experts or expertise* »), pour pouvoir ensuite *sélectionner l'expert ou le conseil approprié d'un expert* (« *selection of an appropriate expert or of expert advice* »). Enfin, *appliquer les connaissances* (« *applying the knowledge* ») à la situation actuelle, c'est-à-dire en recontextualisant les connaissances récupérées du dépôt. Cette dernière étape s'appuie sur deux notions spécialisées sur les connaissances : le *rappel* (« *recall* ») qui permet de savoir si une information a été indexées et la retrouver, et la *reconnaissance* (« *recognition* ») qui permet de savoir si une information répond aux besoins de l'utilisateur et s'applique aux connaissances visées. Mais aussi sur deux notions spécialisées sur l'expertise humaine : l'*identification* (« *identification* ») d'experts sur un domaine en particulier, et la *sélection* (« *selection* ») de l'expert le plus approprié pour la question.

Trois rôles sont également mobilisés lors de l'exécution de ce processus de réutilisation [68]. Ces rôles peuvent être tenus par des individus distincts, ou par le même :

— Producteur de connaissances (« *knowledge producer* ») : la personne qui documente les connaissances en enregistrant les connaissances explicites, ou en transformant les connaissances tacites en connaissances explicites (extériorisation).
— Intermédiaire de connaissances (« *knowledge intermediary* ») : la personne qui adapte les connaissances pour les réutiliser grâce à l'élucidation, le nettoyage, l'indexation, et l'empaquetage. Cette personne est également chargée de distribuer les connaissances au sein de l'organisation.
— Consommateur de connaissances (« *knowledge consumer* ») : la personne qui retrouve et adapte le contenu des connaissances passées pour les appliquer à son cas.

De nombreux types de dépôts existent dont les caractéristiques les distinguent les uns des autres [68]. La plus simple des distinctions s'appuie sur la nature des objets entreposés : les *documents* (texte, audio, vidéo) n'étant pas toujours stockés de la même manière que les *données* (structures de données, systèmes de gestion de base de données relationnelles / SGBDR). Les travaux dans [17] distinguent les *connaissances externes*, concernant l'environnement extérieur à l'organisation, les *connaissances internes structurées*, tels que les documents et données, et enfin les *informations informelles*, comme les notes de réunions ou les mails. Une autre distinction présentée dans [130] concerne la nature des connaissances : les *connaissances générales*, tel que l'état des connaissances scientifiques actuelles, et les *connaissances spécifiques*, tels que les informations et le vocabulaire internes à une organisation.

Dans le contexte actuel où les données et informations sont accessibles quasiment n'importe où, ces dépôts peuvent être assimilés à d'autres formes que les *entrepôts de données* (« *date warehouses* ») ou SGBD traditionnels, comme par exemple les nombreux *wikis* [100].

Ce processus, les rôles, et les dépôts représentent un modèle général de gestion des connaissances dans lequel l'étape de réutilisation intervient relativement tard. La réutilisation de connaissances implique évidemment de créer et stocker des connaissances, avant de les partager ou de les transférer et éventuellement de les adapter. Quatre types de situations de réutilisation des connaissances ont été identifiés et sont présentés dans une typologie [68] :

— *Producteurs collaboratifs* (« *Shared Work Producers* ») : les individus travaillant dans des équipes proches les unes des autres produisent, stockent, partagent et retrouvent leurs propres connaissances pour les réutiliser. Bien qu'ils n'aient que peu de difficultés à réutiliser leurs connaissances, lorsque les équipes en question stockent beaucoup trop de connaissances, il est parfois difficile et long de retrouver précisément celles recherchées. Le problème se pose également lorsque ces connaissances ne sont pas numérisées (des brouillons, par exemple) ou qu'elles sont mal renseignées lors de leur enregistrement (contexte mal détaillé).

— *Praticiens collaboratifs* (« *Shared Work Practitioners* ») : les individus affectés à une même fonction mais éloignés géographiquement, dans l'organigramme, ou bien même dans des organisations distinctes, partagent néanmoins des connaissances qu'ils réutilisent. Les principales difficultés concernent la sélection des connaissances les plus adaptées au problème (les consommateurs n'arrivent pas à évaluer quelles connaissances répondent au mieux à leurs besoins), la réputation des contributeurs (certains contributeurs seront préférés à d'autres), enfin le contexte des connaissances (les consommateurs de connaissances peuvent ne pas connaître ni comprendre le contexte d'origine).

— *Novices cherchant une expertise* (« *Expertise-Seeking Novices* ») : les individus étrangers à un domaine ont besoin d'experts pour les aider à réutiliser des connaissances. De nombreuses difficultés sont rencontrées par les novices, dont la première est le jargon (la question posée ne contient pas le ou les termes spécifiques qui permettraient de comprendre le problème et retrouver immédiatement les connaissances adaptées), suivie du contexte des informations (un novice peut mélanger le contexte et les informations), et enfin les connaissances doivent être présentées de façon accessible aux non-experts (renvoyer des débutants vers un manuel technique complet ne leur permet pas d'identifier précisément la connaissance recherchée).

— *Explorateurs tiers de connaissances* (« *Secondary Knowledge Miners* ») : les individus effectuant de l'exploration de données peuvent également réutiliser des connaissances et en générer de nouvelles. Ces utilisateurs ne connaissent pas toujours le contexte exact dans lequel les données ont été extraites ni l'objectif

initial. Il est donc conseillé de fournir une formation solide aux utilisateurs effectuant de l'exploration de données afin qu'ils maîtrisent les limites des ensembles qu'ils manipulent.

La plupart des réponses pour dépasser ces limites concernent la documentation : documenter pour soi-même afin de pouvoir retrouver, comprendre, et réutiliser ses propres connaissances, documenter pour les pairs du domaine afin de les orienter vers les connaissances les plus adaptées, et enfin documenter pour les novices afin de leur permettre d'adapter les connaissances à leur cas (en retirant l'excès de contexte avant de les leur fournir) et d'éviter l'excès de données pouvant entraîner de mauvaises interprétations. Concrètement, dans les projets informatiques actuels, on peut retrouver cette documentation dans les commentaires d'un code, dans les notes des outils de gestion de versions, voire dans les wikis.

D'autres travaux [88] décrivent trois méthodes de réutilisation observées dans le cas des projets informatiques d'une société de conseils :

— Le *verbatim* (« *verbatim* ») : les connaissances sont directement réutilisées sans aucune adaptation (en répétant par exemple rigoureusement les mêmes étapes). Cette méthode ne fonctionne pas toujours étant donné que les situations varient constamment et qu'il est difficile de retrouver un contexte suffisamment proche pour que les connaissances soient parfaitement réutilisables telles quelles. Cependant, l'usage du verbatim à des fins heuristiques a fonctionné (par exemple estimer la durée d'exécution d'une tâche à partir d'exécutions passées).

— La *synthèse* (« *synthesis* ») : plusieurs sources de connaissances sont combinées afin de construire une solution adaptée au cas présent (plusieurs expériences similaires passées sont analysées ainsi qu'une liste de bonnes pratiques, afin de résoudre un nouveau problème). Cette méthode s'appuie à la fois sur des connaissances explicites et les connaissances tacites de l'individu construisant sa solution. L'intérêt majeur repose sur le fait que des réponses proches existent déjà, et que l'on ne construit pas une réponse à partir de rien.

— La *création* (« *creation* ») : une nouvelle solution est construite à partir de rien en effectuant des sessions de réflexion ou des réunions, c'est-à-dire en s'appuyant sur la sagesse collective. Cette méthode est surtout utilisée dans les situations difficiles, et lorsque les autres méthodes ne fonctionnent pas. Les réunions permettent d'exposer les problèmes et exigences à respecter pour en déduire de nouvelles solutions et leurs conséquences grâce à l'expérience de chaque individu (la simple combinaison de solutions passées n'est pas suffisante, et il est nécessaire de reprendre le problème depuis le début).

L'étude menée dans [3] a montré que les solutions déployées servaient principalement à accumuler les connaissances et n'étaient jusque là que peu utilisées en pratique. De plus, selon [100], les chercheurs et les praticiens s'intéressent plutôt à la collecte, au stockage, et au transfert de connaissances plutôt qu'à la réutilisation. Plus précisément,

l'analyse faite par [100] explique que les études réalisées avec une méthodologie orientée *behavior* visent à expliquer et comprendre les phénomènes, mais elles ne donnent aucune recommandation pour les praticiens. À l'inverse, les solutions construites avec une méthodologie *science du design* sont trop orientées vers les fonctionnalités, et elles ne vérifient donc pas les conséquences sur l'organisation ou les comportements des individus.

Du point de vue de la gestion des connaissances, la réutilisation des connaissances est donc un problème encore ouvert. Peu de travaux de recherche s'y sont spécifiquement intéressés, et les solutions techniques proposent des fonctionnalités plutôt liées aux activités avant la réutilisation (c'est-à-dire l'acquisition, le stockage, et le transfert).

Le défi de la réutilisation de fragments de processus dans le cas des processus à forte intensité de connaissances

Dans le domaine des processus à forte intensité de connaissances, la réutilisation de fragments est également un des défis actuellement étudiés [8]. Les travaux de [29][28] donnent une définition des *fragments de processus* en indiquant qu'il s'agit des activités, et des connaissances qui y sont attachées, auxquelles chaque participant au processus est assigné selon son propre contexte local. La notion *locale* est importante dans le sens où chaque fragment est lié à la fois à une partie d'un processus (un sous-ensemble d'activités), à un individu exécutant les activités impliquées, et aux connaissances mobilisées à l'exécution de ces activités. Lors de l'inscription d'un étudiant à l'université par exemple, l'étudiant dispose de son propre fragment sur l'expérience utilisateur, les personnels administratifs des inscriptions vérifient et répondent, et enfin, les personnels administratifs des unités des formation et de recherche attribuent un groupe à chaque étudiant selon les filières. La figure 2.3 expose les trois fragments de ce processus d'inscription. Chaque fragment mobilise donc des connaissances spécifiques : générer les pièces de son propre dossier, vérifier la validité des pièces et cas particuliers puis envoyer les documents réponses, construire les groupes d'étudiants selon les exigences de la formation et des salles disponibles. L'objectif des travaux présentés dans [28] visent à proposer à un expert de réutiliser des fragments de processus structurés lors de la conception des processus, mais également lors de leur exécution.

Une amélioration de l'usage de ces fragments de processus a été proposée dans les travaux présentés dans [14] en s'inspirant des techniques de CBR [104][62] afin de gérer des cas, et non plus des processus structurés. Le CBR s'appuie sur un schéma en plusieurs étapes rappelant énormément celui appliqué en gestion des connaissances :

1. une personne décrit tout d'abord au système le problème auquel elle est confrontée,
2. le système recherche les cas similaires dans sa base de cas,
3. le système sélectionne le cas le plus proche afin de le réutiliser,
4. le système adapte le cas retenu au contexte actuel,

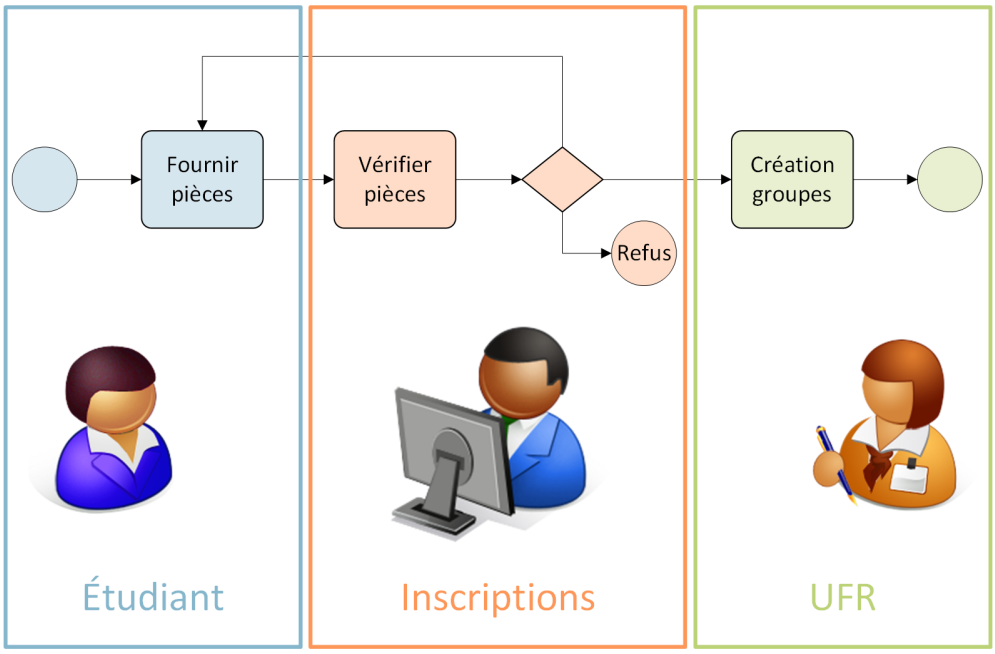


FIG. 2.3 – Exemple d'un processus d'inscription d'étudiant et de ses trois fragments

5. si la personne valide le cas adapté, celui-ci est retenu et est conservé à son tour pour un possible ré-usage futur.

Les systèmes de gestion des cas s'étant développés dans un environnement statique et contrôlé, ils ne sont pas adaptés aux situations exigeant de la flexibilité [78]. Ainsi, l'*Adaptive Case Management* (ACM) [107] s'est développé pour pouvoir s'affranchir des contraintes et permettre aux participants d'adapter le système aux cas rencontrés et aux évènements émergents. La figure 2.4 illustre la différence entre le point de vue classique impératif où un processus est modélisé de façon détaillée, et les traces d'exécutions contiennent la progression des instances, tandis que le point de vue cas s'intéresse aux intervenants qui les traitent en consultant, ajoutant, modifiant, ou supprimant des données au fur et à mesure.

L'ACM est un candidat naturel au traitement des processus à forte intensité de connaissances, de par sa nature à être dirigé par les données et les connaissances, mais également car il s'intéresse aux instances de ces processus, c'est-à-dire les cas. La différence majeure entre l'ACM et la gestion des connaissances provient du point de vue choisi : il s'agit non pas d'étudier comment les connaissances ou données sont traitées pour en proposer d'autres et les réutiliser, mais bien de traiter des cas et les résoudre grâce à d'anciens cas avec leurs connaissances et données. L'élément central est le cas, qui sera résolu en s'appuyant sur les connaissances des experts et les données qu'ils produisent et manipulent. La réutilisation des cas passés, ainsi que des connaissances et données qu'ils embarquent, permet donc de simplifier la gestion des processus à forte intensité de connaissances.

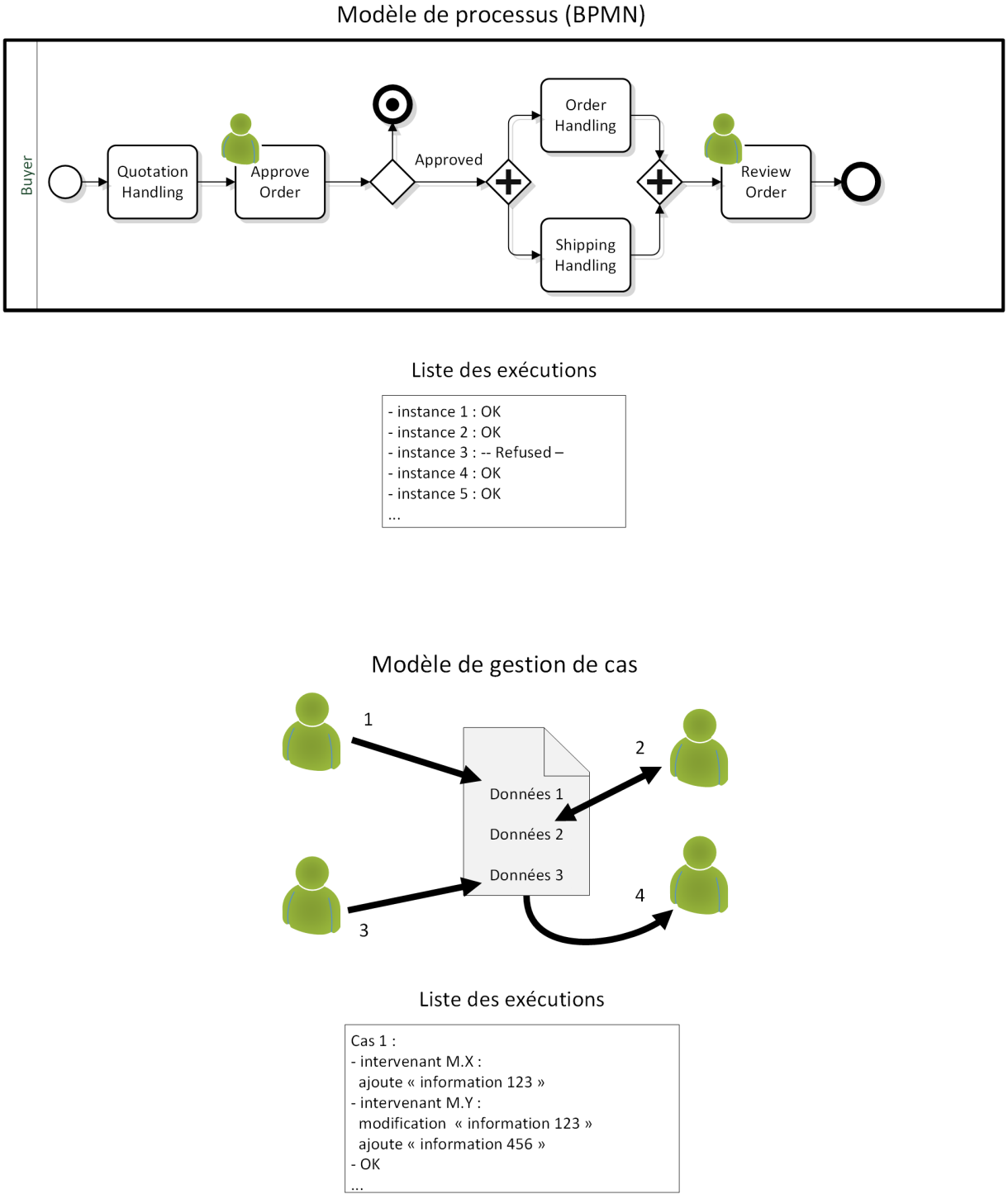


FIG. 2.4 – Comparaison des point de vues impératif (BPMN) et gestion de cas

2.2 Techniques d'analyse de données

Lors de la construction d'un cours, un enseignant utilise des connaissances implicites, grâce à son expérience personnelle, pour sélectionner les notions enseignées et les enchaîner de façon logique et la plus intuitive possible pour les étudiants. L'enchaînement de notions étant lui-même un ensemble de connaissances implicites, on constate que la production de connaissances à première vue explicites repose en réalité sur de nombreuses connaissances implicites (à la fois utilisées lors du processus de création, et dans la production finale). La réutilisation implique donc des traitements spécifiques pour analyser les connaissances embarquées dans les processus à forte intensité de connaissances et les documents qui y sont attachés. Dans cette section, nous présentons plusieurs techniques permettant de traiter ces connaissances et les données sur lesquelles elles s'appuient.

Les documents étant composés de textes en langage naturel, il est tout d'abord nécessaire de les analyser avec des techniques adaptées issues du domaine du *traitement automatique du langage*. Une fois les textes analysés, et leur contenu extraits, on peut y rechercher des connaissances implicites grâce à l'*analyse de concepts formels* et deux *métriques dédiées*. Une première métrique (l'*impact mutuel*) permet d'estimer la pertinence des documents insérés en entrées, tandis qu'une autre (la *similarité conceptuelle*) est utilisée pour construire des *clusters* de termes. Ces clusters représentent les séances sous forme d'ensemble de notions les plus proches selon les connaissances implicites exploitées par l'analyse de concepts formels.

2.2.1 Traitement Automatique du Langage

Comme nous l'avons vu, l'une des spécificités des KIP est de manipuler (ou faire intervenir de façon plus subtile) des connaissances, tout comme ACM se concentre sur les données des cas. Ces connaissances et données se transmettent par des signes et des langages étudiés par plusieurs domaines comme la sémiologie ou sémiotique [41]. L'approche présentée dans cette thèse visant en particulier l'enseignement supérieur, donc des activités proprement humaines où les textes sont prédominants, nous faisons appel au domaine du *traitement automatique du langage* (TAL), ou *natural language processing* (NLP) en anglais. Le "*triangle sémiotique (le terme, le concept, l'objet)*" [131] propose un point de vue où le texte (composé de termes) exprime par une vue de l'esprit (les concepts) les choses et phénomènes (les objets) observés, ce qui correspond à ce qu'un travailleur du savoir exprimerait dans les documents par rapport à ce qu'il observe ou fait. Les travaux de cette thèse utilisent le mot *terme* pour désigner la représentation textuelle d'un concept (« *unit of thought* » [56], une unité de la pensée) comme décrit dans la norme ISO 25964-1 [56][109]. Un *terme* correspond donc à un ou plusieurs mots, par exemple "*base de données*" est un *terme* composé de trois *mots*. Dans le domaine particulier de l'enseignement, nous emploierons *notions* ou *sujets* pour parler des unités enseignées aux étudiants.

Afin de manipuler les mêmes *termes* dans l'ensemble des documents traités, il est nécessaire de comprendre les concepts et standardiser leurs représentations textuelles.

Pour cela, nous employons une succession d'outils pour effectuer l'étiquetage morpho-syntaxique des mots, puis désambiguïser le sens et les lier aux *entités* d'une base de connaissances. Les *termes* sont donc les signifiants des concepts (signifiés) contenus dans les bases de connaissances.

Certaines classes grammaticales de mots ne permettant pas de déterminer des notions que nous souhaitons réutiliser, il est nécessaire de les filtrer. L'étiquetage morpho-syntaxique, ou *part-of-speech tagging* (POS tagging) en anglais, permet de déterminer la classe grammaticale des mots [122][2]. La plupart des étiqueteurs découpent tout d'abord les textes en tokens en détectant la ponctuation ou d'autres limites de fin de mot, puis recherchent les ambiguïtés en comparant les mots reconnus et leurs formes dans un lexique (les articles sont immédiatement détectables comparés aux noms et verbes dont les formes changent et peuvent induire des ambiguïtés : *restes* peut désigner un nom ou un verbe), enfin les mots ambigües sont analysés pour déterminer la classe la plus probable [122]. Pour étiqueter les mots et supprimer ceux ne portant pas de concepts utiles, par exemple les articles ou prépositions, nous utilisons TreeTagger [101][102]. Celui-ci repose sur des arbres de décision pour déterminer la classe grammaticale la plus probable en fonction du contexte.

Afin de retrouver les termes ainsi que les notions qui leurs sont rattachées, il est nécessaire d'utiliser un outil désambiguïsant les mots, ou *word sense disambiguation* (WSD) en anglais, puis faisant les liens avec les entités nommées, ou *entity linking* (EL) en anglais. L'annotation sémantique est une famille de techniques de TAL permettant d'identifier des entités nommées dans des textes, et de les lier aux entités d'une base de connaissances [95]. Ces techniques font partie du domaine de l'extraction d'informations qui permet le stockage et la réutilisation de ces connaissances à partir de textes en langue naturelle [95]. Afin de reconnaître ces entités, et éventuellement leurs synonymes, une étape de désambiguïsation des mots est requise. Nous avons opté pour BabelFy [77][76], un outil de désambiguïsation et d'annotation sémantique, lié à BabelNet [81], un réseau sémantique multilingue manipulant des concepts et des entités nommées lui-même lié à WordNet [72] et Wikipedia [120][70].

L'usage de Wikipedia comme base de connaissances peut paraitre déraisonnable scientifiquement de par la qualité de certains articles, mais celle-ci étant la plus grande base de connaissances collaborative et multilingue [81], il devient possible d'exploiter des entités nommées de la vie courante (par exemple des personnalités publiques, des marques, ou encore des entreprises). Dans [11] il est également présenté comment l'usage de Wikipedia améliore les résultats de la désambiguïsation des entités nommées par rapport à des encyclopédies moins complètes. La quantité d'articles, de langues, l'aspect collaboratif permettant de la maintenir à jour avec une cohérence vis-à-vis des évènements quotidiens, et sa simplicité la rendent particulièrement adaptée pour cette tâche.

Les travaux présentés dans [81] visent à construire un réseau sémantique multilingue nommé BabelNet. Un réseau sémantique sert à représenter des concepts exprimés par des mots du langage naturel et des phrases sous forme de nœuds reliés par des arcs représentants les relations sémantiques entre eux [103]. L'approche utilisée

pour BabelNet consiste à lier la plus grande encyclopédie multilingue (Wikipedia) au lexique informatisé le plus populaire (WordNet) [81]. Les entités de Wikipedia sont automatiquement liées aux mots et expressions stockés dans WordNet dans plusieurs langues. Les entités non liées, ou dont les traductions sont manquantes, sont ensuite traitées avec des outils de traduction automatique, ou *Machine Translation* en anglais, sur des ensembles de données issus de Wikipedia et SemCor [73]. Le réseau sémantique ainsi créé est un des plus performants [81].

Afin de comprendre les unités manipulées par BabelNet, il est nécessaire de décrire succinctement WordNet. WordNet [72] est une base de données lexicale initialement dédiée à l'anglais, mais étendue à quelques autres langues, qui manipule les concepts à l'aide de *synsets* (*synonym set*, un ensemble de synonymes) [81]. En l'interrogeant sur un mot, celle-ci renvoie chaque synset disposant du mot en précisant les sens et les classes grammaticales associées. Lorsque BabelNet est interrogé sur un concept au travers du (ou des) mot(s) le représentant, celui-ci renvoie également un synset (appelé dans ce cas un *Babel synset*) en précisant l'article Wikipedia associé à l'entité nommée et éventuellement le synset WordNet lorsque celui-ci existe. Afin d'obtenir l'équivalent du (ou des) mot(s) dans d'autres langues, les traductions proposées par Wikipedia sont collectées et liées aux Babel synsets, puis les traductions manquantes sont recherchées avec les techniques de traduction automatique dans les corpus de SemCor. Chaque Babel synset dispose également d'un identifiant unique au format $bn{:}xxxxxxxxY$, où les huit x sont des chiffres et Y une lettre correspondant à la classe grammaticale. Ainsi, BabelNet regroupe un ensemble de concepts et d'entités nommées sous forme de Babel synsets composés de mots liés aux entités nommées stockées dans Wikipedia.

L'outil permettant d'extraire les entités nommées d'un texte s'appelle *BabelFy* [77][76]. Celui-ci permet de retrouver les Babel synsets (et leurs identifiants) associés aux concepts et entités manipulés dans le texte. BabelFy vise à combiner la désambiguïsation des mots (WSD) et la reconnaissance d'entités nommées (EL), afin de retrouver l'ensemble des entités nommées possibles pour chaque mot ou ensemble de mots. Un morceau de texte peut permettre d'extraire plusieurs entités (les entités nommées et mentions nominales) comme le montre l'exemple dans [77] : pour « *Major League Soccer* », on retrouve l'entité *Major League Soccer*, mais également les mentions nominales *major league*, *league*, et *soccer*.

Afin de fonctionner, BabelFy construit tout d'abord un ensemble de signatures sémantiques à partir de l'ensemble des concepts et entités nommées du réseau sémantique BabelNet. Cette opération ne se déroule qu'une seule fois (ou au plus après chaque mise à jour de BabelNet). Pour chaque texte inséré, son contenu est étiqueté grammaticalement, puis, toutes les suites possibles de un à cinq mots consécutifs contenant au moins un nom (afin de pouvoir se lier à une entité dans BabelNet) sont créées. Un ensemble d'entités candidates pour chacune de ces suites est constitué. Les entités candidates de l'ensemble du texte sont réunies dans un graphe où les arcs représentent les signatures candidates entre chaque suite de mots. Un algorithme retire petit à petit les entités candidates dont les degrés sont les plus faibles, et obtient ainsi un ensemble de candidats fortement liés. Les entités candidates les moins liées les unes avec les autres, car d'un domaine différent, sont donc supprimées des réponses. Pour

chacune des entités candidates retenues, trois scores sont établis. Les travaux dans [90] les décrivent ainsi :

— Le *score de désambiguïsation* (*disambiguation score* en anglais) correspond à la confiance établie entre le (ou les) mot(s) d'origine et le Babel synset choisi. Il s'agit du score qu'établirait une étape de désambiguïsation des mots (WSD) seule, c'est-à-dire si la création de lien vers des entités nommées (EL) n'était pas effectuée en même temps.

— Le *score de cohérence* (*coherence score* en anglais) correspond au degré de connexion du Babel synset retenu par rapport aux autres Babel synset dans le graphe final représentant le texte inséré.

— Le *score de pertinence* (*relevant score* en anglais) correspond à la pertinence du concept dans le texte inséré. Celui-ci est calculé uniquement selon la position du nœud dans le graphe, sans que cela ne soit lié au Babel synset.

Pour le cas de l'enseignement supérieur, ces techniques permettent donc de lire les supports de cours et d'en extraire les notions abordées, c'est-à-dire les entités nommées qui les composent. Ces entités nommées étant standardisées au moyen des bases de connaissances utilisées par BabelFy et BabelNet, il est possible de chercher des points communs entre les documents insérés.

2.2.2 Analyse de Concepts Formels

L'*Analyse de Concepts Formels* (ACF) [125][126], ou *Formal Concept Analysis* (FCA) en anglais, est une méthode d'analyse de données. Elle permet d'analyser des données décrivant les relations entre des objets et leurs attributs [6] en « *visualisant les structures, implications, et dépendances inhérentes* » [127]. Ces objets et leurs attributs sont rassemblés dans un *treillis* composé de *concepts formels*, c'est-à-dire un graphe avec des propriétés particulières. La notion de *concepts formels* n'est pas totalement étrangère à la définition présentée dans la sous-section précédente, principalement car l'ACF peut aussi être définie comme une « *mathématisation de la compréhension philosophique des concepts* » (« *is a mathematization of the philosophical understanding of concept* » [127]). En effet, d'après [6], l'objectif de l'ACF est de pouvoir décrire des *concepts* humains sous forme de *concepts formels* tels que « voiture à quatre roues motrices » ou « nombres divisibles par 3 et 4 », afin d'en déduire des implications telles que « tous les nombres divisibles par 3 et 4 sont divisibles par 6 ». Des liens avec d'autres domaines de recherche sont également présentés par l'un des fondateurs de l'ACF dans [126]. Du point de vue du *treillis* (pouvant être considéré comme un graphe), un *concept formel* est un nœud contenant simultanément un ensemble d'objets, et un ensemble d'attributs.

Dans le cadre de cette thèse, l'ACF va donc permettre de manipuler les notions issues des documents (représentées par les termes récupérés avec l'aide des techniques de TAL) pour les analyser et en extraire des implications. Précisément, nous analysons les termes en tant qu'objets, et les supports de cours où ils apparaissent en tant qu'attributs.

Stratégies de binarisation

L'ACF repose sur un treillis de Galois sur lequel diverses interprétations sont effectuées pour découvrir des connaissances [58]. Pour construire ce treillis, il est nécessaire de fournir un *contexte formel* rassemblant les liens entre les objets et leurs attributs. Le *contexte formel* étant une matrice binaire, il est nécessaire d'appliquer une transformation lorsque les données analysées forment une matrice multivaluée. Dans le cas de l'enseignement supérieur, les matrices sont constituées d'occurrences de termes dans des supports de cours, donc de valeurs pouvant être supérieures à 1. Une stratégie de binarisation définit un algorithme qui transforme les valeurs d'une matrice de l'intervalle $[0, +\infty[$ vers la paire $\{0, 1\}$.

Il existe actuellement plusieurs stratégies pour permettre à l'ACF de présenter différentes informations contenues dans une matrice multivaluée [59][58]. Nous pouvons citer les stratégies les plus évidentes que sont les stratégies dites *directe* et *inverse* :

1. Stratégie Directe : il s'agit de la stratégie la plus simple, les valeurs non-nulles sont remplacées par des 1, et les valeurs nulles sont remplacées par des 0. L'objectif est simplement de montrer les liens entre les objets et attributs sans aucune pondération. Dans notre cas, on peut donc voir quels termes sont présents dans quels documents.

2. Stratégie Inverse : il s'agit de la stratégie inverse à la stratégie directe, les valeurs non-nulles sont remplacées par des 0, et les valeurs nulles sont remplacées par des 1. L'objectif est de montrer la matrice inverse (d'où son nom) et les relations engendrées entre les objets et attributs. Dans notre cas, on peut donc voir quels termes sont absents dans quels documents.

Les stratégies directe et inverse sont illustrées en figure 2.5, on notera que les stratégies peuvent être simplifiées en supprimant les lignes et colonnes ne contenant que des 0. Bien que ces stratégies semblent triviales, la stratégie directe est néanmoins utile pour exposer des informations concernant les termes les plus importants ainsi que la pertinence des documents insérés comme nous le verrons plus tard en sous-section 3.3.2.

Trois autres stratégies prenant en compte la pondération sont aussi présentées dans [59][58]. Celles-ci se basent sur la prise en compte de *l'intensité de la relation entre chaque objet et chaque attribut, via le calcul d'une valeur de fréquence* [58]. Deux seuils sont fixés par un $\beta \in [0, 1]$. Les seuils sont équidistants de la valeur $0,5$ (indiquée par la ligne en pointillés rouges) et s'en éloignent au fur et à mesure que β augmente, comme illustré sur la figure 2.6. Ces seuils permettent de découper l'espace entre 0 et 1 en trois parties : les valeurs de fréquences hautes (au dessus du *seuil haut*) associées à la *stratégie à haute dépendance*, les valeurs de fréquences basses (en dessous du *seuil bas*) associées à la *stratégie à faible dépendance*, et les valeurs de fréquences moyennes (entre les *seuil haut* et *seuil bas*) associées à la *stratégie à dépendance moyenne*.

La figure 2.7 présente plusieurs exemples suivant les valeurs de β. On notera que lorsque β vaut 0, il n'y a qu'un seuil à $0,5$ découpant l'espace en seulement deux

Matrice multivaluée

	Cours 1	Cours 2	Cours 3
web	25	20	0
php	25	40	0
sql	25	40	100
mysql	25	0	0

Stratégie Inverse

	Cours 1	Cours 2	Cours 3
web	0	0	1
php	0	0	1
sql	0	0	0
mysql	0	1	1

	Cours 1	Cours 2	Cours 3
web	1	1	0
php	1	1	0
sql	1	1	1
mysql	1	0	0

Stratégie Directe

	Cours 2	Cours 3
web	0	1
php	0	1
mysql	1	1

Stratégie Inverse (simplifiée)

FIG. 2.5 – Exemple d'application des stratégies directe et inverse

parties (les valeurs à fréquences hautes et basses), et inversement, lorsque β vaut 1, les deux seuils sont à 0 et 1 découpant l'espace en seulement une partie (les valeurs à fréquences moyennes).

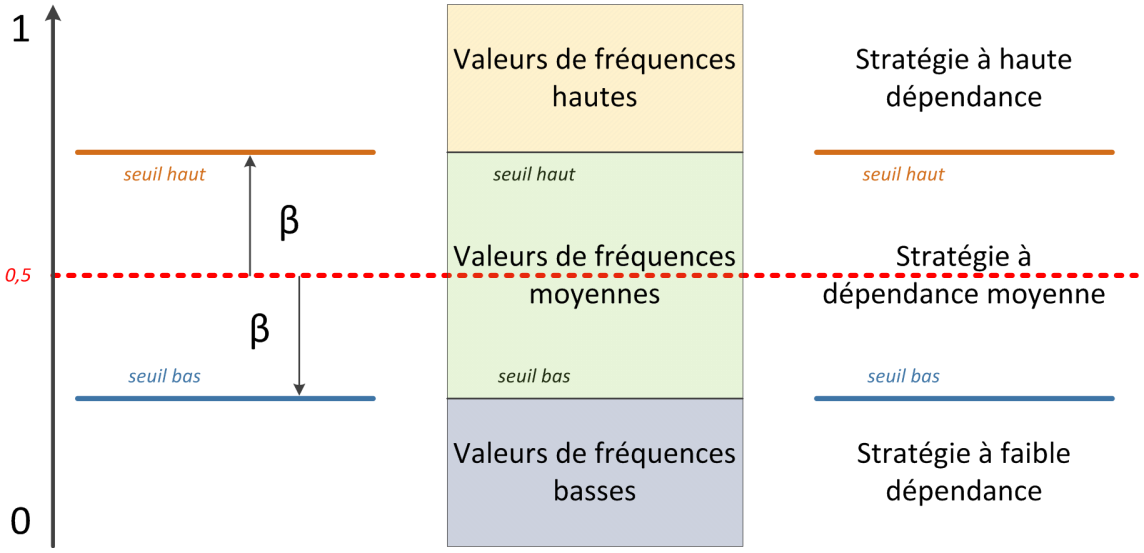


FIG. 2.6 – Les deux seuils générés par β découpent l'espace en trois parties

Selon le β choisi, les seuils vont se fixer pour permettre de délimiter les bornes des valeurs de fréquences requises pour chaque stratégie. Ensuite, la formule (2.1) permet d'obtenir la valeur de fréquence pour chaque proportion d'occurrences d'un terme dans un document de la matrice fournie. La valeur de fréquence ainsi obtenue est transformée en 0 ou 1 selon son positionnement entre les seuils (et donc par rapport

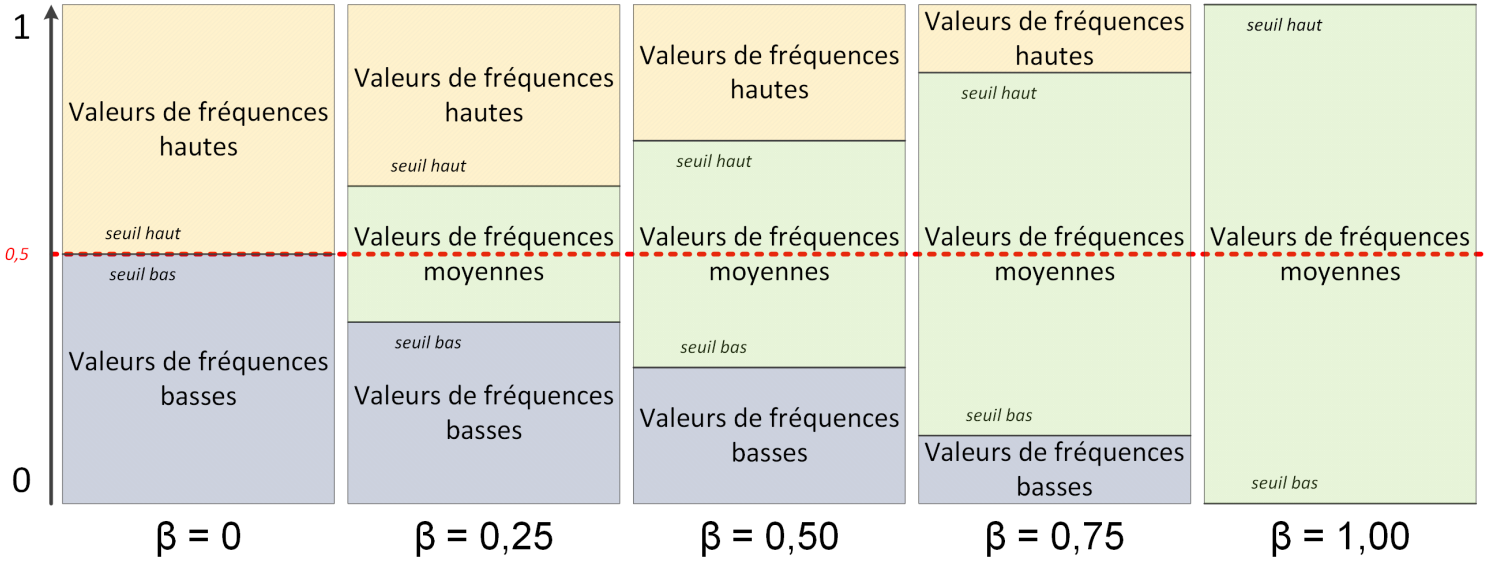


FIG. 2.7 – Différentes valeurs de β séparent l'espace en plusieurs parties

à la stratégie sélectionnée). Les seuils hauts et bas sont positionnés en suivant les formules (2.2) et (2.3). La matrice multivaluée est ainsi transformée en matrice binaire qui peut être utilisée comme un *contexte formel* par la suite. Afin de générer une matrice binaire, il faut donc fixer un β et sélectionner une stratégie parmi les suivantes :

- La *stratégie à haute dépendance* contient les termes dont la valeur de fréquence est supérieure au seuil haut. Il s'agit des termes qui apparaissent les plus fréquemment dans l'ensemble des documents.
- La *stratégie à faible dépendance* contient les termes dont la valeur de fréquence est inférieure au seuil bas. Il s'agit des termes qui apparaissent les moins fréquemment dans l'ensemble des documents.
- La *stratégie à dépendance moyenne* contient les termes dont la valeur de fréquence est inférieure au seuil haut tout en étant supérieure au seuil bas. Il s'agit des termes qui apparaissent ni trop fréquemment ni trop rarement dans l'ensemble des documents.

$$\textit{fréquence}(T, D) = \frac{\textit{Nb d'occurrences du terme } T \textit{ dans le document } D}{\textit{Nb d'occurrences du terme } T \textit{ dans tous les documents}} \tag{2.1}$$

$$\textit{seuil haut} = \textit{Moyenne des fréquences} + \beta \times \textit{Écart type des fréquences} \tag{2.2}$$

$$\textit{seuil bas} = \textit{Moyenne des fréquences} - \beta \times \textit{Écart type des fréquences} \tag{2.3}$$

La figure 2.8 illustre ces trois stratégies avec un β fixé à 0,50. On remarque que les valeurs non nulles de chaque ligne sont distribuées dans chacune des stratégies :

web et *php* voient leurs deux valeurs non nulles distribuées dans les stratégies basses et hautes, *sql* voit ses trois valeurs distribuées dans toutes les stratégies, tandis que *mysql* a son unique valeur non nulle rattachée à la stratégie moyenne.

Matrice multivaluée

	Cours 1	Cours 2	Cours 3
web	25	20	0
php	25	40	0
sql	25	40	100
mysql	25	0	0

Stratégie Basse (β = 0,50)

	Cours 1	Cours 2	Cours 3
web	0	1	0
php	1	0	0
sql	1	0	0
mysql	0	0	0

	Cours 1	Cours 2	Cours 3
web	0	0	0
php	0	0	0
sql	0	1	0
mysql	1	0	0

Stratégie Moyenne (β = 0,50)

	Cours 1	Cours 2	Cours 3
web	56%	44%	0%
php	38%	62%	0%
sql	15%	24%	61%
mysql	100%	0%	0%

Matrice de fréquences

	Cours 1	Cours 2	Cours 3
web	1	0	0
php	0	1	0
sql	0	0	1
mysql	0	0	0

Stratégie Haute (β = 0,50)

Fig. 2.8 – Exemple d'application des stratégies complexes avec un $\beta = 0,50$

Construction du treillis de Galois (PII.1.d) :

Le *contexte formel* généré avec les stratégies de binarisation contient maintenant des termes liés à des documents par des 0 et des 1. Il peut maintenant être transformé en un *treillis de Galois* pour en extraire des *concepts formels* qui serviront de *fragments* (de cas passés) réutilisables.

La construction d'un treillis de Galois depuis un contexte formel peut se réaliser avec plusieurs algorithmes [71]. Chaque nœud du treillis correspond à un *concept formel*. Un *concept formel* contient le maximum d'objets partageant un maximum d'attributs communs, c'est-à-dire que le maximum de termes sont rassemblés selon le maximum de documents auxquels ils sont liés. La figure 2.9 illustre le treillis de Galois issu d'un contexte formel (une version simplifiée du contexte formel est présente, afin de mieux observer les résultats sur le treillis). Les concepts formels aux deux extrémités haute et basse du treillis contiennent respectivement l'ensemble des objets (avec éventuellement le(s) attribut(s) commun(s) à tous les objets) et l'ensemble des attributs (avec éventuellement le(s) objet(s) commun(s) à tous les attributs).

Dans l'exemple, nous retrouvons donc l'ensemble des termes du contexte formel (qui ne partagent aucun document commun) dans le concept formel du haut, et l'ensemble des documents (qui ne partagent aucun terme commun) dans le concept formel du bas. En partant du concept formel contenant tous les objets, chaque attribut non nul du contexte formel est progressivement ajouté pour former le plus grand sous-ensemble

Contexte Formel

	C1	C2	C3	C4	C5
web	0	1	0	0	1
php	1	1	0	1	0
sql	0	0	1	1	0
mysql	0	0	0	1	1
apache	1	0	0	1	0
http	1	0	0	0	0
html	1	0	1	0	0

	C1	C2	C3	C4	C5
web		X			X
php	X	X		X	
sql			X	X	
mysql				X	X
apache	X			X	
http	X				
html	X		X		

Contexte Formel (simplifié)

Treillis de Galois

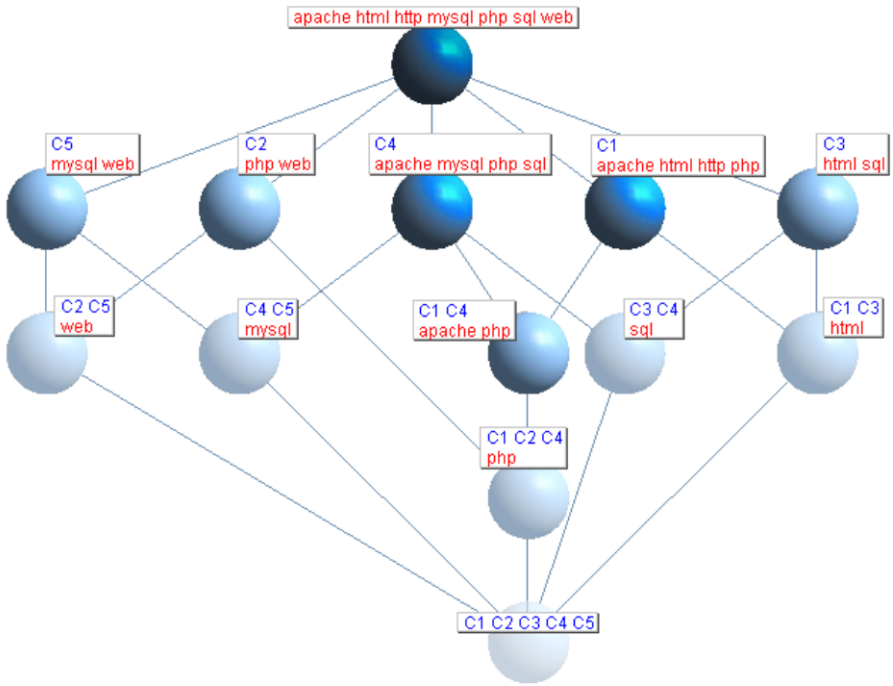
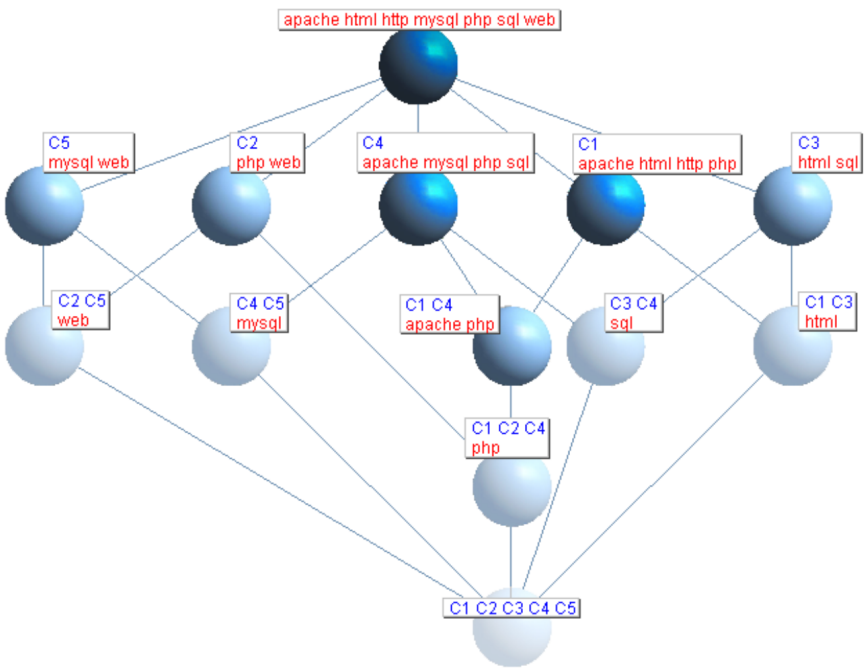


Fig. 2.9 – Contexte formel et son treillis de Galois

d'objets. Chaque niveau du treillis introduit ainsi un attribut supplémentaire (certains niveaux peuvent être vides). Dans notre exemple, nous voyons au niveau 0 l'ensemble des termes "*apache*", "*html*", "*http*", "*mysql*", "*php*", "*sql*", et "*web*", puis au niveau 1, un document est ajouté à chaque concept formel pour former le sous-ensemble le plus grand de termes partageant cet attribut commun : "*apache*", "*mysql*", "*php*", et "*sql*" ont le document "*C4*" en commun, mais "*apache*", "*html*", "*http*", et "*php*" ont le document "*C1*" en commun. En niveau 2, un sous-ensemble plus restreint de termes possède deux documents en commun : "*apache*" et "*php*" partagent les documents "*C1*" et "*C4*". Ensuite, au niveau 3, seul le terme "*php*" est partagé par trois documents "*C1*", "*C2*", et "*C4*". Finalement, au niveau 4, les cinq documents du contexte formel sont réunis en un concept formel sans aucun terme.

Dans l'exemple précédent, treize concepts formels ont été générés à partir du contexte formel pour construire le treillis de Galois (ceux-ci sont explicités sur la figure 2.10). Ces concepts formels regroupent donc des termes et les documents auxquels ils sont rattachés, c'est-à-dire, suite aux stratégies de binarisation, chaque concept formel rassemble des termes qui apparaissent ensemble dans un ou plusieurs documents selon leurs fréquences d'apparitions communes. Les concepts formels ainsi formés permettent de manipuler des informations plus abstraites grâce à la combinaison de termes et documents engendrée par le treillis de Galois.

Treillis de Galois



Concepts formels

Concept 0 (niveau 0) :	7 objets - 0 attribut
	apache html http mysql php sql web
Concept 1 (niveau 1) :	2 objets - 1 attribut
	mysql web - C5
Concept 2 (niveau 1) :	2 objets - 1 attribut
	php web - C2
Concept 3 (niveau 1) :	4 objets - 1 attribut
	apache mysql php sql - C4
Concept 4 (niveau 1) :	4 objets - 1 attribut
	apache html http php - C1
Concept 5 (niveau 1) :	2 objets - 1 attribut
	html sql - C3
Concept 6 (niveau 2) :	1 objet - 2 attributs
	web - C2 C5
Concept 7 (niveau 2) :	1 objet - 2 attributs
	mysql - C4 C5
Concept 8 (niveau 2) :	2 objets - 2 attributs
	apache php - C1 C4
Concept 9 (niveau 2) :	1 objet - 2 attributs
	sql - C3 C4
Concept 10 (niveau 2) :	1 objet - 1 attribut
	html - C1 C3
Concept 11 (niveau 3) :	1 objet - 3 attributs
	php - C1 C2 C4
Concept 12 (niveau 4) :	0 objet - 5 attributs
	C1 C2 C3 C4 C5

Fig. 2.10 – Treillis de Galois et ses concepts formels

Calcul des métriques du treillis (PII.1.e) :

L'ACF permet de calculer de nombreuses métriques sur le treillis généré [4][58] : stabilité des concepts, poids conceptuel des objets (respectivement des attributs), similarité conceptuelle des objets (respectivement des attributs), impact mutuel absolu entre un objet et un attribut, impact mutuel relatif entre un objet et un attribut, et beaucoup d'autres. Dans le cas de la méthode CREA, nous nous concentrons particulièrement sur l'impact mutuel relatif [58] qui permet de générer des graphiques utiles pour évaluer la qualité des données (présentés en sous-section 3.3.2), ainsi que la similarité conceptuelle des objets [58] nécessaire pour d'autres traitements ultérieurs (notamment lors du regroupement des termes présenté en sous-section 3.3.3).

Similarité Conceptuelle entre deux objets : La *similarité conceptuelle* permet de comparer deux objets (respectivement attributs) en tenant compte de leurs présences dans l'ensemble du treillis. C'est-à-dire, la métrique correspond à la fréquence d'apparition des objets dans les concepts formels du treillis. La formule calculant la similarité conceptuelle entre deux objets est la suivante :

$$Similarité\ conceptuelle(O_i, O_j) = \frac{Nombre\ de\ concepts\ contenant\ O_i\ \mathbf{et}\ O_j}{Nombre\ de\ concepts\ contenant\ O_i\ \mathbf{ou}\ O_j} \quad (2.4)$$

La similarité conceptuelle étant un rapport, elle prend des valeurs entre 0 et 1. La valeur maximale 1 indique que les deux objets (respectivement attributs) sont toujours présents ensemble dans les mêmes concepts formels du treillis, c'est-à-dire qu'il n'existe aucun concept formel ne contenant que l'un des deux objets (respectivement attributs), ou encore qu'ils partagent exactement les mêmes attributs (respectivement objets). L'exemple en figure 2.11 illustre le calcul de la similarité entre deux objets.

Il est possible de générer un tableau regroupant les similarités conceptuelles entre tous les objets (respectivement attributs) d'un treillis, ce qui a pour conséquence de former une *matrice de similarité conceptuelle*. La figure 2.12 illustre une matrice de similarité conceptuelle regroupant les similarités conceptuelles de tous les objets d'un treillis.

Dans le cadre de la méthode CREA, la matrice de similarité conceptuelle est une donnée d'entrée aux méthodes de *clustering* abordées plus tard en sous-section 3.3.3.

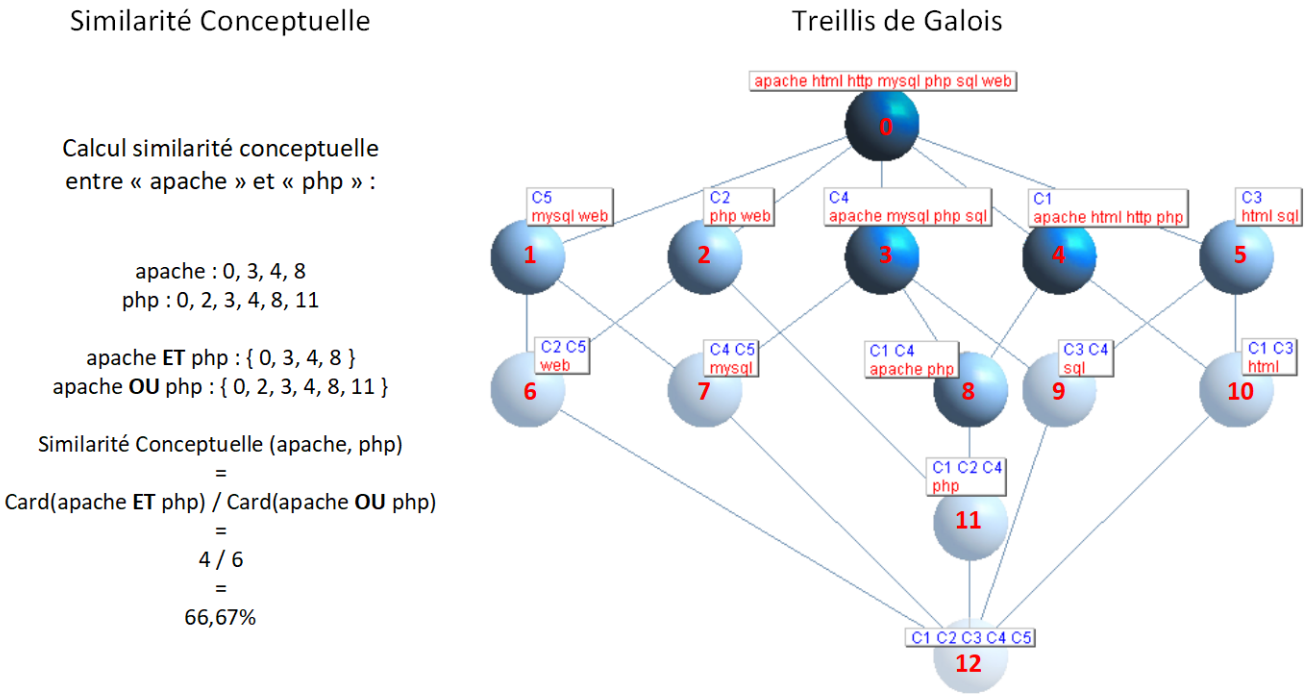


Fig. 2.11 – Calcul de la similarité conceptuelle entre deux objets

Matrice de Similarité Conceptuelle

	apache	html	http	mysql	php	sql	web
apache	100%	33%	50%	33%	67%	33%	14%
html	33%	100%	50%	14%	25%	33%	14%
http	50%	50%	100%	20%	33%	20%	20%
mysql	33%	14%	20%	100%	25%	33%	33%
php	67%	25%	33%	25%	100%	25%	25%
sql	33%	33%	20%	33%	25%	100%	14%
web	14%	14%	20%	33%	25%	14%	100%

Treillis de Galois

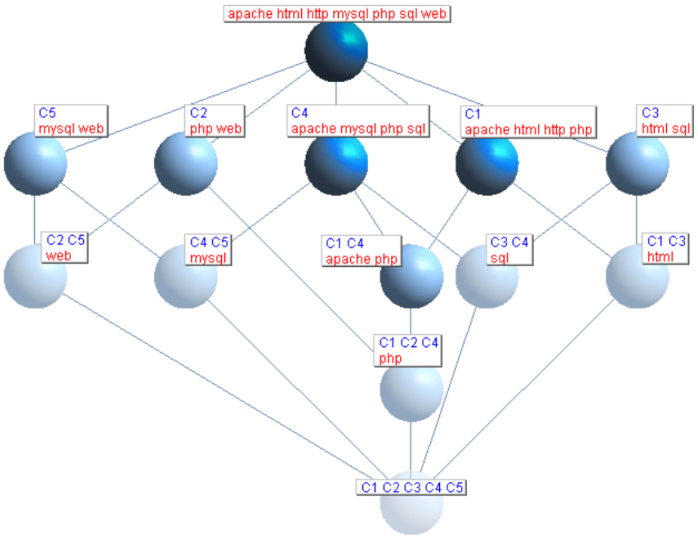


FIG. 2.12 – Matrice de similarité conceptuelle

Impact mutuel entre un objet et un attribut : L'*impact mutuel* analyse la "*force de la relation entre un objet O_i et un attribut A_j en fonction des concepts formels qui les associent*" [58]. Il est défini à partir de l'équation suivante :

$$\textit{Impact mutuel}(O_i, A_j) = \frac{\textit{Nombre de concepts contenant } O_i \textbf{ et } A_j}{\textit{Nombre de concepts contenant } O_i \textbf{ ou } A_j} \tag{2.5}$$

L'impact mutuel étant un rapport, il prend des valeurs entre 0 et 1. Plus la valeur de l'impact mutuel est élevée, plus l'attribut caractérise l'objet parmi tous les autres, et réciproquement : plus l'objet se définit particulièrement grâce à cet attribut. En d'autres termes, un impact mutuel proche de 1 implique que l'attribut apparait quasiment uniquement avec cet objet, et réciproquement, l'objet n'a quasiment pas d'autre attribut. À l'inverse, un impact mutuel proche de 0, mais non nul, implique que l'attribut est partagé par énormément d'objets, et réciproquement, que l'objet utilise quasiment tous les attributs existants. L'exemple en figure 2.13 illustre le calcul de l'impact mutuel entre un objet et un attribut.

Le tableau regroupant les impacts mutuels entre tous les objets et attributs d'un treillis s'appelle une *matrice d'impact mutuel*. La figure 2.14 illustre une matrice d'impact mutuel regroupant les impacts mutuels entre chaque objet et chaque attribut d'un treillis.

La matrice d'impact mutuel permet de générer un *graphe d'impact mutuel*. Le graphe d'impact mutuel a la particularité d'être bi-parti, en proposant des nœuds de la classe des objets, et des nœuds de la classe des attributs. Dans le cadre de la méthode CREA, ce graphe permet de déterminer visuellement deux informations importantes :

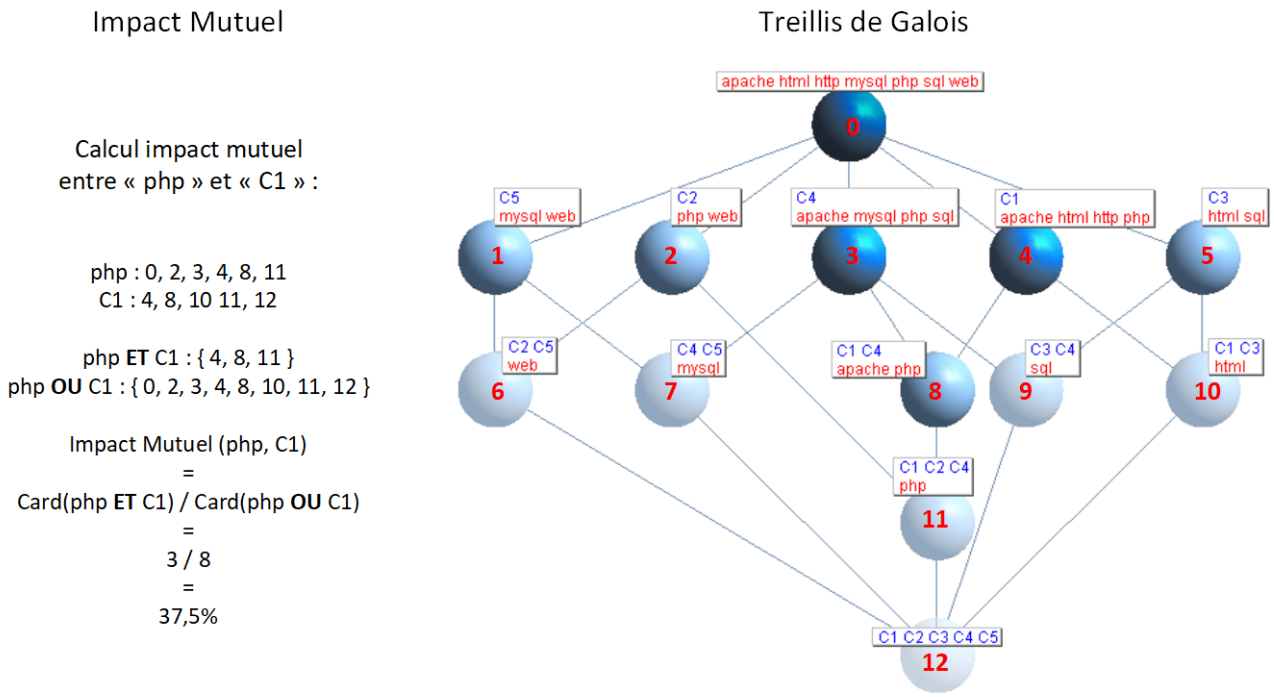


Fig. 2.13 – Calcul de l'impact mutuel entre un objet et un attribut

Matrice d'Impact Mutuel

	apache	html	http	mysql	php	sql	web
C1	29%	29%	17%	0%	38%	0%	0%
C2	0%	0%	0%	0%	25%	0%	33%
C3	0%	33%	0%	0%	0%	33%	0%
C4	25%	0%	0%	25%	33%	25%	0%
C5	0%	0%	0%	33%	0%	0%	33%

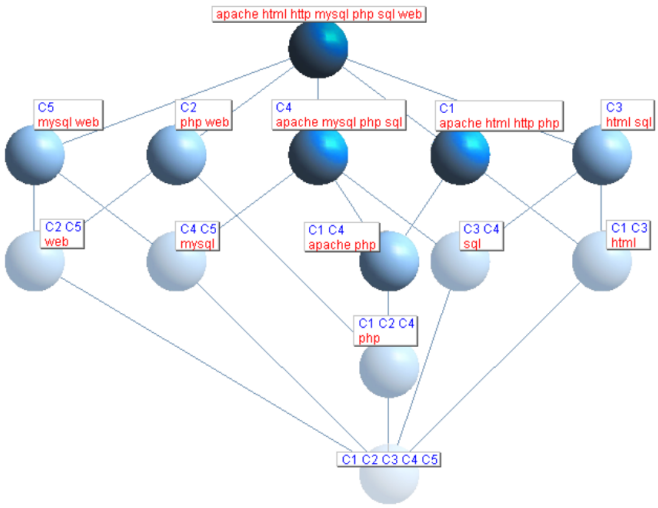


Fig. 2.14 – Matrice d'impact mutuel

les termes les plus représentatifs du corpus de documents rassemblés ; la pertinence des documents dans le contexte ainsi formé. La figure 2.15 illustre le graphe d'impact mutuel et fait un agrandissement de l'ensemble central où le document *C1* apparait parmi des termes.

Typiquement, un ou des documents excentrés et avec peu de liens peuvent être considérés comme peu pertinents, voire comme du *bruit* d'un point de vue statistique.

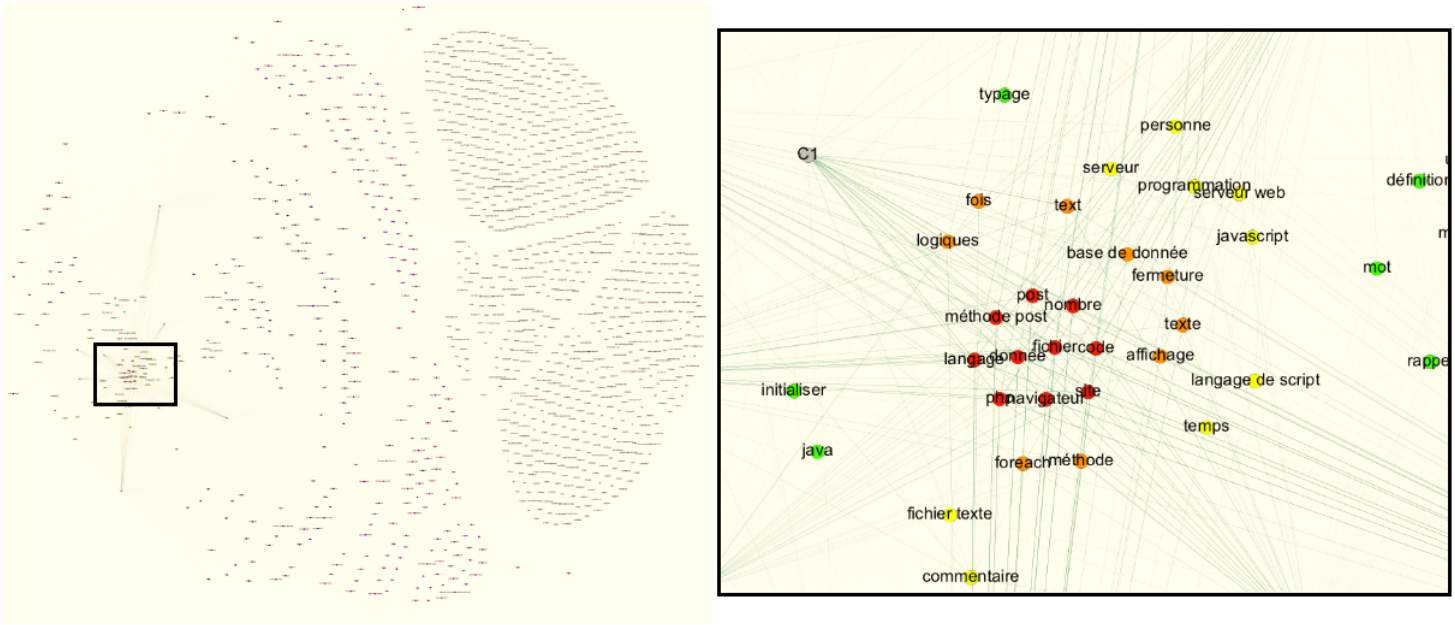


Fig. 2.15 – Graphe d'impact mutuel (généré avec Gephi en utilisant la spatialisation *Force Atlas* et la coloration par *partition* selon le *degré*) et agrandissement de la communauté centrale

Ces documents peu pertinents peuvent être supprimés de la matrice d'occurrences (voir les sous-section 3.3.1 et 3.3.2) pour améliorer la qualité des résultats, ou peuvent être conservés si l'utilisateur l'exige. La même réflexion peut être appliquée sur un ou des termes trop isolés (avec peu de liens). Une fois supprimés de la matrice d'occurrences, il suffit de relancer les traitements depuis l'étape *normalisation* (voir la sous-section 3.3.1). Cependant, si l'utilisateur estime que ces documents ou termes sont nécessaires, ceux-ci peuvent être conservés.

L'impact mutuel, et sa représentation graphique, mettent en évidence certaines qualités des documents et termes : un utilisateur comprend visuellement quels documents et termes sont importants du point de vue de l'ensemble des cas passés, mais surtout lesquels sont considérés comme peu importants ou peu pertinents. Il faut rappeler que les résultats de cette métrique dépendent directement de deux paramètres (les cas passés dont les documents ont été sélectionnés et donnés en entrée de la méthode, et le choix de la stratégie de binarisation) et indirectement de deux paramètres (la technique de TAL employée et sa base de connaissances). La faible représentativité d'un terme ou d'un document peut s'expliquer de plusieurs façons, mais il est important que l'utilisateur choisisse de conserver ou supprimer les termes et documents selon le cas qu'il cherche à construire.

Pour conclure cette sous-section, il est important de noter que d'autres travaux de thèse [110] ont développé et éprouvé une approche similaire à celle présentée dans cette thèse. *KESAM (Knowledge Extraction and Semantic Annotation Management)* est un outil basé sur l'usage d'annotation sémantique depuis une base de connaissances et d'ACF pour aider des experts à analyser des documents et leurs contenus textuels.

Cependant, ces travaux se limitent à présenter et manipuler le treillis construit, son contexte formel, et les liens avec les documents d'origine. Dans les travaux présentés dans cette thèse, nous utilisons en complément deux des métriques présentées dans [58] afin d'interpréter les données issues du treillis, pour non seulement détecter la pertinence des supports de cours en entrée, mais surtout pour générer des clusters de notions réutilisables.

2.2.3 Clustering

Dans le domaine de la fouille de données, plusieurs activités visent à regrouper et classifier les données dans des catégories (ou classes) selon des propriétés communes ou des critères de similarité [128]. Deux familles de méthodes cohabitent : les méthodes d'apprentissage supervisé (plutôt pour classifier), et les méthodes d'apprentissage non-supervisé (plutôt pour du clustering) [97]. L'apprentissage supervisé vise à entrainer un modèle à reconnaitre des catégories existantes (connues grâce à des données étiquetées) afin d'y classer les données d'entrée [128], ou éventuellement de découvrir de nouvelles catégories [60]. L'apprentissage non-supervisé vise à séparer les données dans un certain nombre de groupes selon des critères peu évidents, voire dans des structures cachées [128]. « *Le but du clustering est descriptif, là où la classification est prédictive* » [97][118] (« *The goal of clustering is descriptive, that of classification is predictive* »).

Appliquées à nos travaux pour extraire des fragments réutilisables, en particulier pour la recherche de patterns, les méthodes de clustering sont fortement adaptées [60]. Le clustering contient néanmoins des méthodes impliquant un entraînement préalable, mais nous ne les aborderons pas. Nous nous intéressons en particulier aux méthodes ne nécessitant aucun entraînement, afin de les laisser rechercher les similarités sans à priori. Il est difficile de définir la notion de *cluster* [31][32], nous nous contenterons donc de considérer qu'il s'agit d'un regroupement d'objets selon des caractéristiques communes, donc de cohésion interne/homogénéité et d'isolation externe comme décrit par [15] et [42].

D'après [36], on peut diviser les méthodes de clustering en deux grandes familles dont les *méthodes de partitionnement* et les *méthodes hiérarchiques* (d'autres façons d'organiser sont présentées dans [97]). Ces deux familles sont décrites ainsi dans [97] :

— Les méthodes de partitionnement placent tout d'abord les données d'entrée dans un ou plusieurs clusters, puis déplacent les données d'un cluster à l'autre afin d'obtenir un placement optimal. Généralement, l'utilisateur doit indiquer le nombre de clusters désirés. On peut citer parmi ces algorithmes : *K-Means*, *K-medoids*, *DBSCAN*, *OPTICS*, *Minimal Spanning Tree*, ...

— Les méthodes hiérarchiques construisent progressivement les clusters, soit en divisant l'ensemble des données, soit en les agrégeant au fur et à mesure (les modifications deviennent définitives : un objet manipulé ne peut plus se déplacer). Ces méthodes reposent elles-mêmes sur deux principes : un fonctionnement

ascendant (chaque objet démarre dans son propre cluster, chaque itération regroupe deux clusters selon une métrique de distance jusqu'à n'en obtenir qu'un seul), un fonctionnement descendant (tous les objets démarrent dans le même cluster, chaque itération sépare un cluster en deux). Ces méthodes proposent une représentation graphique appelée *dendrogramme*. La *classification ascendante hiérarchique* (CAH), ou *hierarchical cluster analysis* (HCA) en anglais, est la principale méthode employée.

Certains algorithmes de clustering peuvent générer des classes recouvrantes (ou chevauchantes), c'est-à-dire qu'un même objet se retrouve simultanément dans plusieurs classes. Parmi ces techniques se trouve par exemple la *Classification Ascendante Pyramidale* et les pyramides [25] se rapprochant de la CAH et des dendrogrammes, ou encore OKM [13] (*Overlapping K-Means*) une extension de K-Means.

Dans le cadre de nos travaux, afin de créer des fragments réutilisables issus de cas passés, nous visons la construction de clusters (l'équivalent des fragments) à partir de termes extraits des documents (les données à proposer à l'utilisateur). Les traitements de l'ACF permettent d'obtenir une matrice de similarité (voir sous-section 2.2.2) à partir de laquelle nous pouvons déduire une similarité entre les objets contenus. Parmi les deux familles de méthodes de custering, certains algorithmes nécessitent un placement des points dans un espace (parfois en deux dimensions, parfois à N dimensions), d'autres uniquement les distances ou similarités entre les objets. Afin d'exploiter au mieux les distances entre objets, tout en évitant une distorsion trop élevée par l'usage du *multidimensional scaling* [64][129] réduisant à deux dimensions nos données, nous avons préféré l'utilisation de la *classification ascendante hiérarchique* (CAH).

Classification Ascendante Hiérarchique

La *classification ascendante hiérarchique* (CAH) est une méthode hiérarchique agglomérant un à un chaque objet aux autres. Par définition, cette méthode est donc non-recouvrante (un objet ne peut être que dans un seul cluster à la fois). À partir d'une matrice contenant les distances entre chaque objet, le fonctionnement général est le suivant [128] :

1. Mettre chaque objet dans son propre cluster
2. Calculer les distances entre les clusters (dépendant de la métrique choisie)
3. Fusionner les clusters les plus proches
4. S'il reste plus d'un cluster, répéter depuis l'opération 2
5. Sinon, découper les clusters au partir du niveau souhaité

La figure 2.16 illustre le dendrogramme construit à partir de la matrice de distance, puis la création des clusters lorsque ce dendrogramme est coupé à la hauteur 3. L'axe des abscisses représente les distances entre les éléments (et clusters après fusion), et l'axe des ordonnées représente les éléments à regrouper en clusters. Dans l'ensemble de la matrice de distance, les objets C et D sont les plus proches par rapport aux autres,

ils sont donc fusionnés en premiers. Ensuite, le recalcul de la matrice de distance fait que A se trouve être le plus proche du cluster contenant C et D, A est donc ajouté à ce cluster. Finalement, le dernier élément B est ajouté au cluster. Lorsque l'on découpe le dendrogramme à la hauteur 3 (la ligne rouge en pointillés), on obtient trois clusters.

Plusieurs métriques de distance existent afin de déterminer comment agréger ou découper les clusters. Le *saut minimum* (ou *single linkage* en anglais) consiste à considérer la distance entre deux clusters comme étant la distance des deux objets les plus proches de chacun des clusters [99]. À l'inverse, le *complete linkage* en anglais, consiste à considérer la distance entre deux clusters comme étant la distance des deux objets les plus éloignés de chacun des clusters [99]. D'autres métriques basées sur la moyenne des distances des points des clusters, ou encore sur les centroïdes des clusters permettent de déterminer les clusters les plus proches à fusionner.

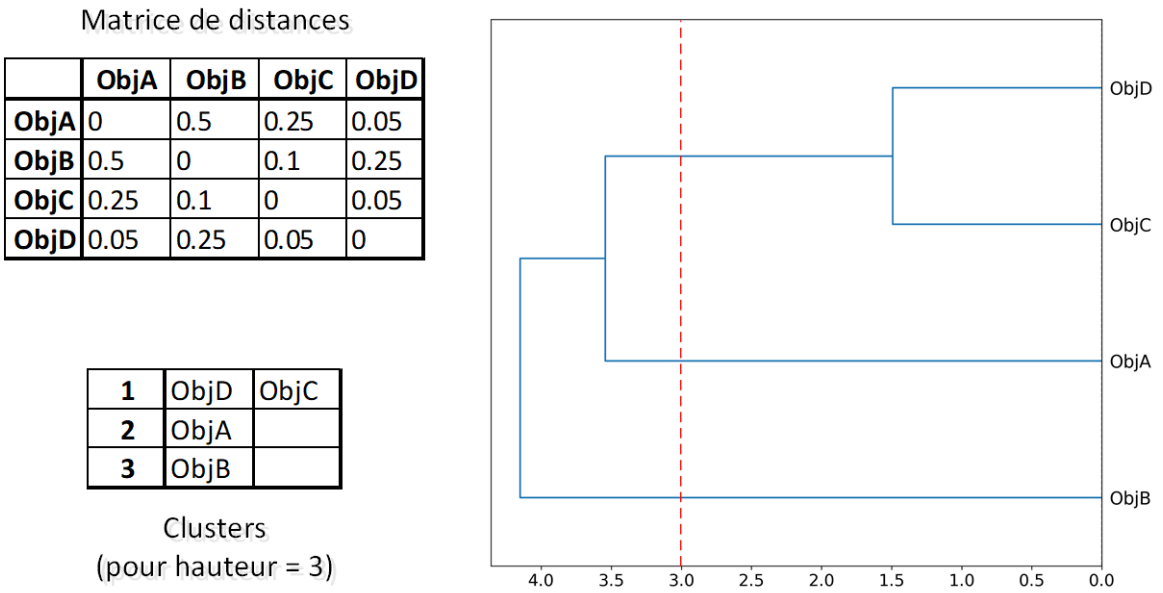


Fig. 2.16 – Exemple de construction du dendrogramme et des clusters à la hauteur 3 à partir d'une matrice de distance

2.3 Travaux connexes et similaires

Afin de surmonter les défis rencontrés par les processus à forte intensité de connaissances, et contribuer à la réutilisation de fragments de processus et de leurs connaissances, plusieurs solutions ont été proposées dans la littérature. Nous présentons maintenant ces solutions et leurs limites, puis nous positionnons notre propre solution.

2.3.1 Travaux connexes sur les processus à forte intensité de connaissances

La réutilisation et les fragments appliqués aux processus à forte intensité de connaissances se retrouvent dans plusieurs travaux. L'usage de *patterns*, c'est-à-dire des motifs qui se répètent, dans les outils BPM appliqués à des modèles de workflows (BPMN, EPC, ...) est connu et permet de proposer des ensembles de tâches et activités liées et prêtes à l'emploi.

Les travaux présentés dans [7] visent à expérimenter des patterns dans d'autres types de modèles, en particulier des patterns de buts appliqués au *STate-oriented Business Process Modeling* (SToBPM), où les états sont représentés sous forme d'espace multidimensionnel. Chaque dimension représente un paramètre, donc un point représente un résultat possible d'une instance de processus. Si la dimension temporelle est ajoutée, une courbe permet de déterminer l'avancement du processus dans le temps. Afin de représenter le processus et ses résultats de façon plus adaptée, des formulaires sont préférés en lieu et place de courbes dans des espaces multidimensionnels. Les buts (et sous-buts) visés par le concepteur, permettant de choisir les patterns dans une bibliothèque, sont construits sur des *décisions* (la finalité à atteindre), des *problèmes* (auxquels appliquer les décisions), des *motivations* (raisons pour lesquelles une décision a été prise, et/ou avancement du problème), des *preneurs de décisions* (organisation, groupe, ou individu), et éventuellement une *date limite* (gestion temporelle). Un concepteur de processus choisit donc des patterns, s'il le souhaite, en indiquant les buts recherchés dans la bibliothèque. Il dispose d'une grande liberté pour réutiliser ou non ces patterns, car l'utilisateur n'est pas guidé mais seulement assisté.

Ces travaux illustrent une autre façon de construire des patterns : au lieu de proposer des ensembles d'activités, le concepteur va rechercher des buts à atteindre et les activités nécessaires pour les valider. La technique de modélisation est cependant beaucoup trop rigide pour être utilisée telle quelle pour des enseignements : bien qu'il soit nécessaire de s'inspirer de sources existantes, le nombre de dimensions finales est inconnu (tous comme le nombre d'états possibles), et les formulaires (ou le découpage en décisions, problèmes, motivations, preneurs de décisions, date limite) ne sont pas du tout adaptés à l'extraction de connaissances.

Dans [132], on trouve cette fois l'usage de *compliance patterns* [30] générés à partir de textes légaux pour permettre à des utilisateurs de construire et exécuter des processus respectant les réglementations en place. La législation sur l'industrie alimentaire a été manuellement analysée afin de sélectionner plusieurs articles, les organiser en catégories et sous-catégories, et y retrouver des patterns fréquents. Les exigences réglementaires ont ensuite été extraites et transformées en séquences de code, afin de pouvoir les préciser à l'aide des patterns. Enfin, les patterns ont été formalisés sous

forme de formules de *logique linéaire temporelle* (ou *linear temporal logic*/LTL en anglais).

Bien que la méthode puisse être appliquée à d'autres domaines, celle-ci n'est pas encore totalement supportée par les outils informatiques et nécessite plusieurs opérations manuelles. De plus, il s'agit de patterns construits à partir de textes de loi en langage naturel, donc d'exigences à respecter. Les supports de cours ayant comme objectif de transmettre des connaissances explicites, ceux-ci n'ont pas vocation à contraindre la conception ou l'exécution d'un processus. On pourrait tout de même imaginer construire des patterns indiquant dans quel ordre aborder quelles notions lors d'un cours, mais cela implique de s'appuyer sur des textes explicitant ces informations sous forme d'assertions claires et non pas d'un enchaînement de notions comme dans les supports de cours classiques.

Parmi les approches CBR, les travaux présentés dans [14] proposent un langage de modélisation de fragments de processus qui vise à générer et perfectionner manuellement des cas. Ce langage, nommé *Business Process Feature Model* (BPFM), permet de représenter les cas et leur contenu de façon plus complète que d'autres langages. Plus spécifiquement, le BPFM est un langage déclaratif capable de représenter des processus partiellement structurés et ordonnés grâce à des contraintes, ainsi que des types de données complexes. Le point le plus intéressant est sa capacité à représenter des variations de processus. Les processus sont représentés sous forme d'arbres dont les nœuds correspondent aux sous-processus, et les feuilles aux activités atomiques. Les contraintes permettent de déterminer si les activités et sous-processus peuvent ou doivent être utilisés dans les variants du processus, et s'ils peuvent ou doivent être exécutés dans ces variants. BPFM permet également d'indiquer si des variations apparaissent lors de l'exécution du processus, et de retenir les liens de parentés entre les variations et les modèles initiaux. Afin de conserver les connaissances liées au processus, l'*Ontology-Based Case-Based Reasoning* [69] (OBCBR) est utilisée pour conserver les éléments décrivant l'objectif de chaque activité, les rôles de chaque intervenant, et l'auteur de la description du cas. La réutilisation d'instances passées s'effectue de deux manières : lors de la phase de conception en indiquant les éventuelles variations, lors de l'exécution en modifiant selon les besoins chaque activité ou sous-processus.

Bien que ce langage offre une plus grande liberté d'exécution grâce aux contraintes et aux variations lors de l'exécution, les cas représentés restent fortement orientés du côté des processus métiers structurés : on représente un processus par un arbre de sous-processus et d'activités qui s'enchaînent. Dans le cadre de la construction d'un cours, le syllabus pourrait être représenté sous forme d'arbre où les chapitres et sections formeraient différents niveaux, cependant, les connaissances que l'on cherche à transmettre ne sont pas de simples activités qu'il suffit de mentionner pour pouvoir les valider. La réutilisation de cours passés implique d'assister à chaque cours ou de lire intégralement le contenu pour le modéliser, puis l'intégrer au modèle BPFM. La difficulté réside dans l'association de certaines notions les unes avec les autres : il n'est pas possible en l'état de découvrir facilement les contraintes dans l'ordre des notions entre chaque variation de cours, car il s'agit de connaissances implicites à chaque enseignant.

Afin de mieux renseigner les gabarits réutilisables, les travaux dans [112] explorent l'usage de l'*Acte de Langage* (ou *Speech Act Theory* en anglais). L'excès de gabarits, et le manque d'informations ou leur trop courte description pour les discerner clairement les uns des autres, implique que l'utilisateur doit rechercher lui-même l'objectif de chaque gabarit pour sélectionner le plus adapté à son cas. Afin de retrouver plus rapidement le gabarit le plus adapté, les interactions et micro-processus sont analysés. Un micro-processus est composé d'actes de coordination (équivalent de méta-données décrivant ce que le micro-processus vise à faire, qui l'exécute, envers quelle(s) autre(s) personne(s), le support, l'horodatage, le contenu, et d'autres annotations) et d'actes de production (création d'un artefact). Les micro-processus s'appuient sur les actes de langage pour décrire toutes les informations des actes de coordination. L'intérêt des micro-processus repose sur la possibilité laissée aux utilisateurs du métier de créer leurs propres gabarits selon les interactions qu'ils utilisent le plus souvent sans être perdus parmi d'immenses bibliothèques de gabarits. Les connaissances tacites sont ainsi plus facilement mises à contribution dans les nombreuses méta-données des actes de coordination : les utilisateurs du métier emploient leur vocabulaire et modélisent leurs activités depuis leur propre point de vue, ils sont ainsi plus facilement capables de retrouver et réutiliser leurs propres gabarits.

Bien que ces travaux n'embarquent pour le moment qu'un petit sous-ensemble d'objets par rapport à ce que la gestion de cas préconise (documents, contacts, communications, ...), laisser les utilisateurs créer eux-mêmes des gabarits réutilisables à partir de leurs propres connaissances simplifie l'usage de la solution présentée. Une limitation soulevée en conclusion dans [113] indique que cette solution ne peut fonctionner que si les utilisateurs acceptent de renseigner chacun des micro-processus qu'ils utilisent. Dans le cas de la construction de cours, on retrouve une limitation très proche : pour pouvoir réutiliser les supports de cours d'autres enseignants, il faut que tous les enseignants utilisent le même outil pour y insérer leurs connaissances. Bien que PowerPoint et ses équivalents libres soient très fréquents, ceux-ci sont très rarement utilisés avec l'ensemble des fonctionnalités qui se rapprochent de la contribution décrite dans [112] (par exemple, les schémas sont souvent de simples images importées, et les méta-données des objets PowerPoint ne sont quasiment jamais renseignées). Cependant, le vocabulaire du domaine et certaines activités (tels que les travaux dirigés ou pratiques) sont présents dans les diapositives qui peuvent être réutilisées à postériori par d'autres enseignants.

Du point de vue de la gestion des connaissances appliquée à la gestion de projets, une solution sur la réutilisation des connaissances est proposée dans [100]. Celle-ci se compose de trois parties :

— Un double-cycle dont la boucle choisie dépend du contexte de l'organisation et de l'équipe. Récapitulatif : à la fin de chaque projet, l'équipe se réunit pour transférer son expérience aux autres équipes. Préparation : au démarrage du projet ou d'une étape importante (c'est-à-dire *pendant* le projet), l'équipe se réunit pour réfléchir à la conception de la solution à partir de l'expérience de chaque membre.
— De nouveaux rôles sont définis pour retenir et transférer les expériences des équipes. L'expert en leçons (*lessons learned expert* en anglais) qui dispose de

compétences pour traiter les connaissances et expériences issues de la réalisation du projet. L'expert du sujet (*topic expert* en anglais) qui dispose de connaissances et d'expérience sur le sujet traité par le projet.

— Un processus centré sur les connaissances appliqué à la gestion de projet (*knowledge-centric project management process* en anglais) s'appuyant sur le PMBOK© du *Project Management Institute*.

Dans le contexte récapitulatif, l'équipe doit se souvenir des évènements clés qui ont eu un impact sur le résultat du projet (succès ou échec). Chaque évènement est ensuite analysé pour en déterminer les causes et les retenir. L'ensemble de ces connaissances est enregistré dans un dépôt partagé, puis un membre de l'équipe est chargé de surveiller l'apparition de ces évènements lors des projets suivants. À l'inverse, dans le contexte de préparation, les risques sont évalués en amont grâce à l'expérience de chaque membre (ou des sources de connaissances les plus adaptées au projet) et peuvent être pris en charge par les techniques de gestion des risques. Le processus général proposé s'appuie en pratique sur plusieurs itérations de la double boucle. Lors du démarrage du projet, la boucle de *préparation* est effectuée une première fois. À chaque étape clé, un cycle *récapitulatif* et un cycle de *préparation* sont effectués pour tirer les leçons nécessaires de l'étape terminée et continuer à progresser vers l'étape suivante. Enfin, lorsque le projet se termine, un cycle *récapitulatif* est effectué pour conserver toutes les connaissances. Les deux rôles d'experts interviennent soit pour gérer le processus dans son ensemble (expert en leçons), soit pour traiter les connaissances liées à l'étape en cours (expert du sujet).

Cette solution et son processus permettant d'apprendre des expériences passées sont intéressants, mais il faut néanmoins l'adapter au domaine visé. Dans le cadre de l'enseignement et des cours, on retrouve ce processus sous une forme plus simplifiée : l'enseignant construit tout d'abord son cours en amont (éventuellement en s'appuyant sur l'expérience des précédents semestres où il l'a enseigné), à la fin de chaque séance il note les difficultés rencontrées avec les groupes, puis adapte ses séances suivantes. Un autre point de vue de plus haut niveau est aussi accessible : une phase exclusivement préparatoire s'appuyant sur les expériences passées (les siennes ou celles de ses collègues) sera exécutée, puis, à chaque semestre où il enseignera de nouveau son cours, il pourra effectuer un récapitulatif du semestre précédent et préparer le suivant.

2.3.2 Positionnement de la méthode CREA

Comme nous venons de le voir, plusieurs solutions sur la réutilisation de fragments de processus (contenant des connaissances tacites) ont été proposées dans la littérature mais ne peuvent pas directement être utilisées dans le cas de la construction d'un cours dans l'enseignement supérieur. Le processus proposé dans [100] se divise en deux étapes : préparation puis récapitulatif. La préparation implique de s'inspirer des expériences passées pour prévoir les difficultés rencontrées. On retrouve partiellement cette vision dans le raisonnement à base de cas où un (ou des) cas passé(s) ser(ven)t de base pour le cas courant. Il s'agit donc de s'appuyer sur les expériences passées, et leurs productions, pour répondre au problème actuel. La méthode CREA, *Case REuse and Adaptation*, que nous proposons s'intéresse particulièrement à l'extraction et l'analyse

de connaissances à partir de documents issus d'instances passées pour proposer des scénarios de réutilisation possibles.

Les supports de cours produits par les enseignants sont assimilables à une explicitation des connaissances où subsiste malgré tout une certaine expertise tacite : le choix précis des notions et de leur enchaînement est directement lié à l'enseignant et sa propre expérience qu'il compte mettre à profit des étudiants. Les documents ainsi produits par les enseignants peuvent donc être traités comme des artefacts de cas passés. Ces documents étant rédigés en langage naturel, pour pouvoir en extraire des informations et connaissances réutilisables, il est nécessaire d'employer tout d'abord des techniques de TAL. BabelFy et BabelNet permettant de s'appuyer sur des bases de connaissances très populaires et communes à l'humanité, leur usage répond parfaitement aux besoins de la gestion des connaissances concernant l'usage d'un vocabulaire commun. Le point de vue des cas employé par ACM et CBR implique d'analyser et réutiliser les connaissances des cas, pour cela, nous employons l'ACF qui permet de découvrir des connaissances tacites parmi les documents et les textes les composant (en particulier les liens entre eux). Ce couple base de connaissances partagée et ACF pour analyser des documents ayant déjà été approuvés par des travaux de thèse [110], leur usage à des fins de réutilisation du contenu de ces documents est donc tout à fait approprié.

Bien qu'il soit impossible de manipuler directement les connaissances tacites, il est néanmoins possible de les manipuler indirectement grâce aux métriques présentées dans une autre thèse [58], en particulier avec l'impact mutuel et la similarité conceptuelle. Une fois l'extraction et l'analyse réalisées, il est nécessaire de ré-assembler les informations et connaissances pour qu'un enseignant (l'expert) puisse les adapter à son cours (son cas), pour cela, nous employons des techniques de clustering sur la similarité conceptuelle des termes dans les documents. L'enseignant se voit proposer des groupes de notions pertinents issus de précédents cours. Afin que la pertinence soit maximale, il est nécessaire que l'enseignant insère lui-même uniquement des supports de cours traitant du même sujet. Cependant, si quelques documents traitent de sujets trop éloignés, l'impact mutuel sera capable d'identifier ces documents et indiquer à l'enseignant lesquels devraient être retirés.

La méthode CREA s'insère donc parfaitement dans la partie réutilisation de la gestion des connaissances et de la gestion de cas (ACM et CBR en particulier dans cette thèse) : des connaissances sont retrouvées, préparées, sélectionnées, et réutilisées. L'enseignant est ensuite libre d'adapter les clusters proposés à son propre cas. Il peut néanmoins supprimer des notions qui ne l'intéressent pas, et relancer une partie de la méthode pour obtenir de nouveaux résultats plus proches de ses attentes.

3. MÉTHODE CREA : Case REuse and Adaptation

Dans ce chapitre, nous détaillons la méthode CREA en trois sections. Tout d'abord, une présentation générale de la méthode est faite pour expliquer les objectifs visés, le cadre général de travail qui a été utilisé lors de la conception, et le fonctionnement global de la méthode. Les deux sections suivantes traitent plus en détails des phases de pré-traitement sémantique, visant à extraire les termes représentant les connaissances réutilisables des documents d'entrée, et d'analyse structurelle qui génère les regroupements de termes représentants la structure, ainsi que des métriques concernant les termes et documents d'entrée.

Sommaire

3.1 Présentation générale de la méthode et cadre de travail

3.1.1 Objectifs de la méthode

La méthode CREA, ou *Case REuse and Adaptation*, vise à répondre à plusieurs défis exposés en section 2.1.2, en particulier la réutilisation de fragments issus de cas passés pour former des nouveaux cours, et la prise en compte du contexte au travers de la vérification de la pertinence des documents. Dans les travaux de cette thèse, les cas passés correspondent aux cours (et les supports associés) générés par d'autres enseignants, voire tous les documents actuellement disponibles traitant du sujet (articles de recherche, présentation technique, etc). Les fragments réutilisables correspondent aux notions abordées dans les cours. Précisément, il s'agit des termes associés à ces notions regroupés sous forme de clusters. Un utilisateur souhaitant traiter un nouveau cas se verra proposer des fragments issus de plusieurs documents générés lors d'exécutions de précédents cas grâce à la méthode CREA. Ces fragments lui permettent d'adapter ses décisions en se basant sur des données qui ont été générées lors de cas passés qu'il a sélectionnés. Ces cas passés ont été gérés par d'autres utilisateurs qui leur ont associé des données en s'appuyant sur leur propre expertise (ou connaissances tacites).

L'objectif de la méthode CREA est donc de profiter de ces expériences passées pour proposer les meilleurs regroupements (ou *clusters*) de données aptes à répondre au nouveau cas, le tout de façon la plus automatisée possible. La construction d'un nouveau cas, et l'évaluation de sa pertinence par rapport au sujet traité, s'effectuent en analysant les notions présentes dans les documents au travers des termes les constituant. Des métriques permettent d'illustrer l'absence ou la présence de termes dans les documents et les écarts entre documents. Ainsi, il est possible d'aider l'utilisateur à améliorer la qualité des regroupements en exploitant une des métriques pour conserver (ou supprimer) des documents ou des termes. Un second usage de cette métrique permet également d'évaluer la pertinence de son propre support de cours en le comparant à de nombreux autres.

Dans le domaine de l'enseignement supérieur et de la recherche, la méthode CREA permet de générer automatiquement à partir de supports de cours existants un syllabus de cours découpé en autant de séances que l'enseignant souhaite. Des documents de nature plus variée, tels que des articles de recherche ou des présentations techniques, peuvent y être adjoints. Un enseignant souhaitant construire un nouveau cours doit uniquement fournir en entrée de la méthode CREA le nombre de séances visées et des supports de cours existant, afin de pouvoir obtenir une proposition de structure de syllabus en sortie. Précisément, les notions essentielles des supports de cours sont extraites, analysées, puis ré-assemblées sous formes de clusters pour pouvoir les proposer à l'enseignant. Le nombre de clusters fourni est égal au nombre de séances demandées par l'enseignant. L'évaluation de la pertinence des supports de cours fournis par l'enseignant est possible grâce à une métrique explicitant quelles notions centrales sont partagées par tous les documents, ou éventuellement quels supports sont éloignés, voire sont hors sujet. L'ordonnancement des clusters formant la structure du syllabus est encore en cours d'évaluation, elle est néanmoins discutée en section 5.2.3 afin de montrer tout le potentiel de la méthode CREA.

3.1.2 Cadre de travail

La réutilisation de fragments issus de cas passés implique de se positionner selon le point de vue de la gestion de cas, et non pas seulement des processus à forte intensité de connaissances : chaque cas est unique, bien que des similitudes puissent exister. L'autre intérêt de ce point de vue est qu'il est dirigé par les données, ce qui permet de se concentrer sur celles-ci pour pouvoir exposer les meilleures informations aux utilisateurs impliqués et les aider à prendre les meilleures décisions. Le traitement d'un cas, jusqu'à sa résolution, produit donc des données réutilisables par la suite.

Dans le cadre de l'enseignement supérieur, les enseignants sont considérés comme les utilisateurs. Chaque enseignant a sa propre expertise qui diffère d'un individu à l'autre, qu'ils aient suivi le même parcours ou non. Les cas étudiés concernent les cours universitaires, et particulièrement l'aide à la construction de supports de cours. Chaque cas traité (instance de processus) produit donc un support de cours (les données du cas) dont les fragments réutilisables correspondent aux notions sélectionnées par l'enseignant. Ces supports de cours générés consolident donc l'ensemble des notions à aborder en cours magistral, en travaux dirigés, voire en travaux pratiques. Bien que ces supports de cours puissent être constitués d'images, sons, vidéos ou de nombreux autres médias variés, nous nous concentrons exclusivement sur les cours dont le contenu est majoritairement constitué de texte. Ces textes peuvent facilement être traités automatiquement et découpés en termes dont le ré-agencement permet de construire des clusters rassemblant plusieurs notions à enseigner. Ces clusters de termes permettent directement de proposer une structure de syllabus qu'un enseignant peut réutiliser partiellement ou complètement lorsqu'il construit un nouveau cours. Les niveaux de granularité retenus des données traitées correspondent aux supports de cours texte contenant des notions à aborder, dont les termes sont manipulés (sélectionnés, filtrés, standardisés, ...) pour former des clusters assimilables à la structure d'un syllabus de cours.

3.1.3 Fonctionnement général

La méthode CREA se divise actuellement en deux phases principales de pré-traitement sémantique (PI) puis d'analyse structurelle (PII), comme illustré sur la figure 3.1. Une troisième phase d'analyse temporelle est encore en cours d'évaluation et est présentée en section 5.2.3. L'atout majeur de CREA réside dans l'automatisation complète de chacune des phases la composant, bien qu'il soit possible de paramétrer plus finement certaines étapes pour améliorer la qualité des résultats. La phase de pré-traitement sémantique (PI) vise à extraire les termes des documents donnés en entrée, tandis que la phase d'analyse structurelle (PII) expose tout d'abord des métriques sur la pertinence des documents entre eux (l'impact mutuel présenté en sous-section 2.2.2), puis construit les clusters de termes explicitant la structure des documents fournis afin de pouvoir les réutiliser.

La phase de pré-traitement sémantique (PI) est divisée en cinq étapes qui s'appuient sur plusieurs techniques de traitement automatique du langage présentées en section 2.2.1. Ces étapes permettent d'extraire les termes des documents sélectionnés,

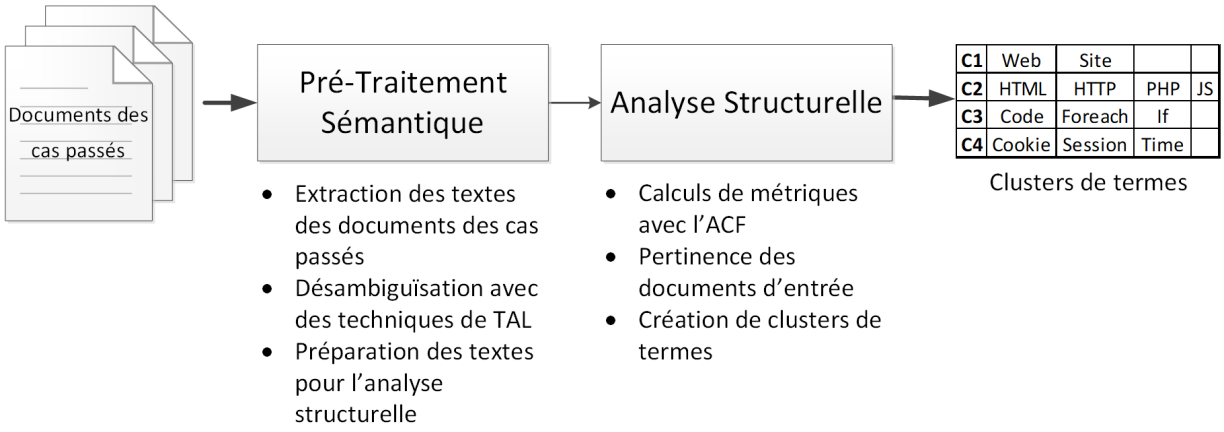


Fig. 3.1 – Les deux principales phases de la méthode CREA

puis de lier ces termes à des entités reconnues dans des bases de connaissances grâce à une étape de désambiguïsation. Ces bases de connaissances permettent de standardiser les entités manipulées dans l'ensemble du corpus de textes en faisant abstraction des synonymes mais également des langues utilisées dans les documents. À l'issue de ces cinq étapes, une liste de termes standardisés est générée pour chacun des documents. Les traitements suivants s'appuient soit sur ce format en *liste*, soit sur un format en *matrice d'occurrences* de termes dans des documents comme illustré sur la figure 3.2. Appliqué au cas de l'enseignement il s'agit donc de générer, pour chacun des supports de cours sélectionnés, une liste de termes standardisés et reconnus dans des bases de connaissances représentant les notions abordées dans le cours.

Listes d'occurrences

Cours 1	Cours 2	Cours 3
web	web	sql
php	php	
web	sql	
php	php	
sql		
mysql		

Matrice d'occurrences

	web	php	sql	mysql
Cours 1	2	2	1	1
Cours 2	1	2	1	0
Cours 3	0	0	1	0

Fig. 3.2 – Exemple de listes et de matrice d'occurrences de termes générées à l'issue de la phase de pré-traitement sémantique (PI)

La phase d'analyse structurelle (PII) est divisée en trois grandes étapes qui s'appuient sur l'analyse de concepts formels, présentée en section 2.2.2, et les méthodes de clustering, présentées en section 2.2.3. Ces étapes permettent d'analyser les documents et les termes les composant, afin d'évaluer leur pertinence et d'en extraire les regroupements de termes les plus utiles à l'utilisateur. Dans le domaine de l'enseignement, il s'agit donc de générer des métriques évaluant la qualité de la composition des supports de cours entre eux, afin de permettre à l'enseignant de supprimer des supports qui ne conviennent pas à ses exigences ou éventuellement des notions qu'il ne souhaite pas aborder, et pouvoir en extraire des clusters de termes représentant

une structure de cours réutilisable. Par exemple, si un enseignant souhaite construire un cours de développement web à l'aide de PHP en s'appuyant sur une dizaine de supports de cours, mais dont l'un traite en fait de Java, un graphique lui permettra de constater qu'un des supports de cours insérés est hors sujet et doit être vérifié pour éventuellement le retirer. La figure 3.3 montre que les supports C6 et CJA sont beaucoup plus éloignés des autres, et devraient être vérifiés manuellement. Ensuite, lorsque le corpus de documents est considéré par l'enseignant comme convenable, des clusters peuvent être générés selon le nombre de séances souhaité. La figure 3.4 montre huit clusters générés à partir de plusieurs supports de cours de PHP (les clusters ne sont pas ordonnés), l'enseignant peut ainsi s'appuyer dessus pour préparer ses huit séances de cours.

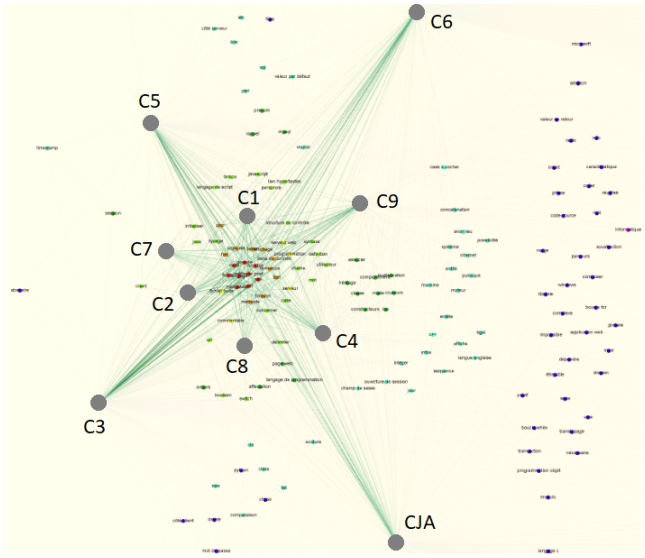


FIG. 3.3 – Exemple de graphe d'impact mutuel explicitant que les supports C6 et CJA sont éloignés des autres

1	php	code	fois	post	jour	foreach	cle	classe	class	mysqli	
2	page web	navigateur	serveur web	texte	concerner	délimiter	utilisateur	associer	personne	machine	mysql
3	url	langage	case	fermeture	session	chaîne	entête	avoir accès			
4	fichier	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client				
5	typage	mot	moteur	affiche	transaction	visiteur					
6	base de donnée	insert	varchar	null							
7	xml	configuration	composer	doctype							
8	donnée	text	méthode post	programmation	site	langage de script	list	méthode	timestamp	files	

FIG. 3.4 – Exemple de huit clusters générés avec la méthode CREA pour huit séances à partir de supports de cours sur PHP

La figure 3.5 explicite les étapes des deux phases tout en indiquant les données générées au fur et à mesure. Comme indiqué précédemment, on observe que la première phase se concentre explicitement sur la préparation des documents fournis en entrée, afin de produire une matrice représentant l'ensemble des documents sélectionnés et les termes les composants. La deuxième phase analyse en profondeur cette matrice pour en

extraire des métriques qui servent à aider l'utilisateur à comprendre les données qu'il manipule et améliorer la qualité de sa base documentaires, le tout pour générer une structure réutilisable issue d'une liste de clusters de termes résumant ces documents.

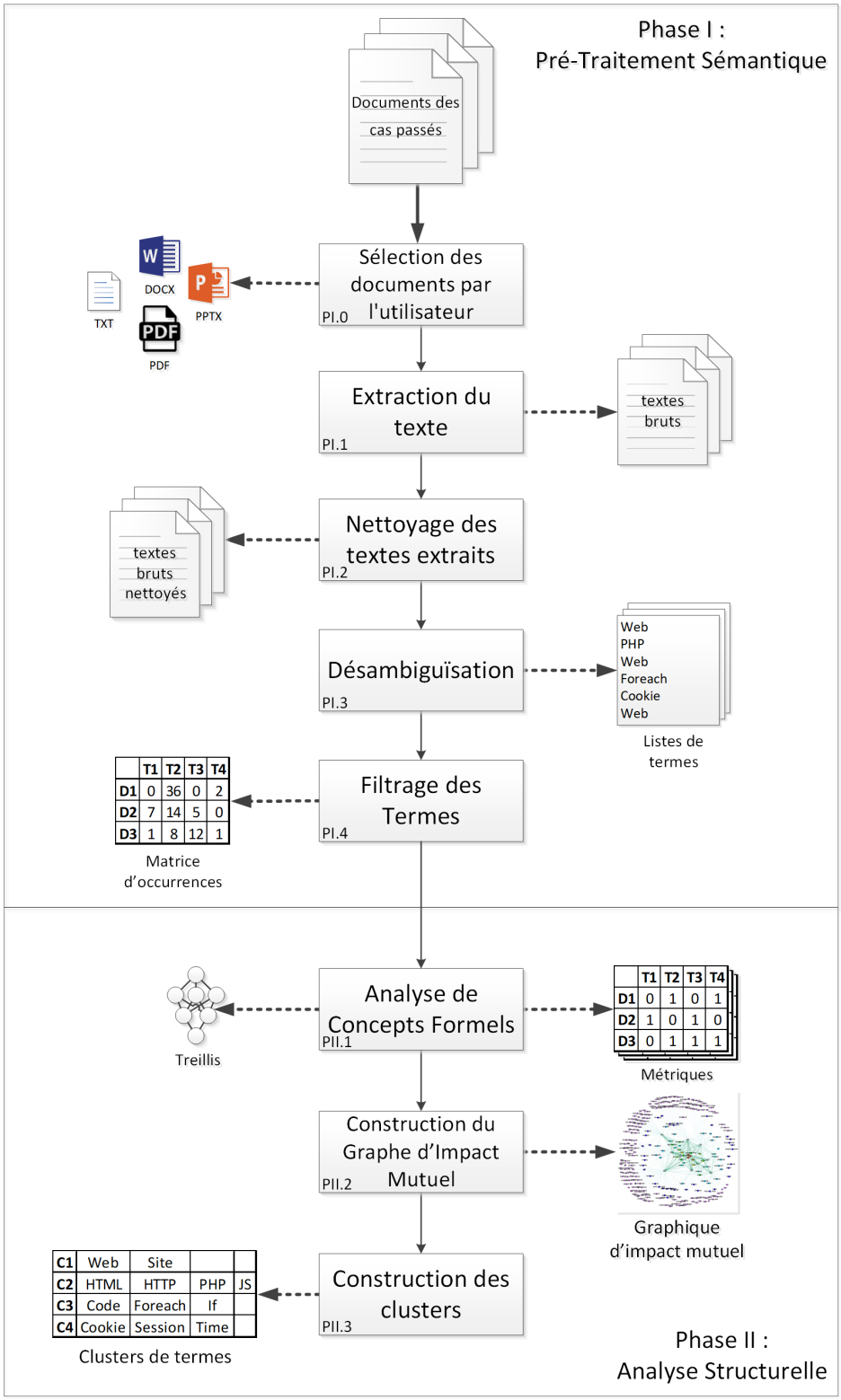


Fig. 3.5 – Les deux phases détaillées de la méthode CREA

3.2 Pré-traitement sémantique : extraction des termes

La première phase (PI) de la méthode CREA, le pré-traitement sémantique, a pour objectif d'extraire une liste de termes à partir de documents fournis en entrée, et de les rassembler sous forme d'une matrice d'occurrences afin de les organiser plus tard sous forme de clusters. Dans le contexte de l'enseignement supérieur, ces termes représentent les notions abordées dans les supports de cours fournis en entrée. Cette première phase se déroule en cinq étapes successives illustrées par la figure 3.6.

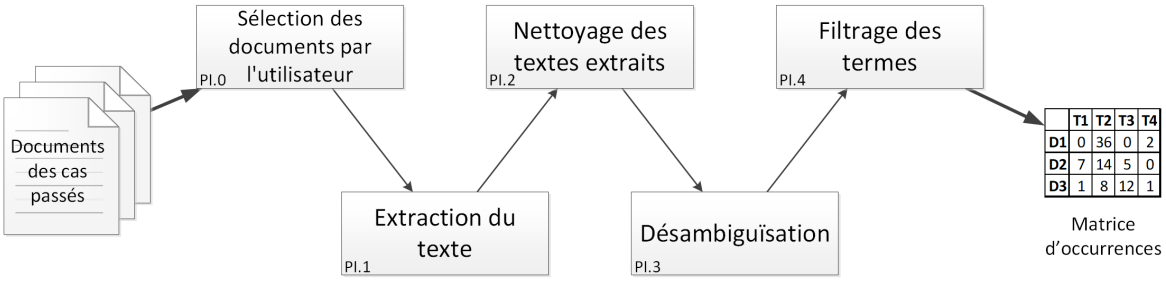


Fig. 3.6 – Les étapes de la phase de pré-traitement sémantique

— Sélection des documents par l'utilisateur (PI.0) : Tout d'abord, l'enseignant sélectionne des supports de cours dont les titres lui semblent pertinents et dont le contenu est majoritairement composé de texte.

— Extraction du texte (PI.1) : Une étape d'extraction du texte s'appuie sur la reconnaissance optique de caractères (ou *OCR*) pour pouvoir transformer les supports de cours en texte brut dont les termes seront reconnaissables par la suite.

— Nettoyage des textes extraits (PI.2) : Les textes extraits sont ensuite nettoyés afin de supprimer les caractères mal reconnus ainsi que les classes grammaticales de mots dont la présence augmente le bruit dans la suite des traitements.

— Désambiguïsation (PI.3) : Les textes nettoyés sont ensuite désambiguïsés et suivent un traitement d'annotation sémantique afin d'en extraire les termes directement liés à des entités reconnues dans des bases de connaissances.

— Filtrage des Termes (PI.4) : Enfin, les termes dont le score de désambiguïsation est trop faible sont retirés, afin de ne garder que les termes dont le sens est reconnu avec suffisamment de confiance.

3.2.1 Sélection des documents par l'utilisateur (PI.0)

La *sélection des documents par l'utilisateur* est l'étape préliminaire où l'utilisateur sélectionne des documents selon ses exigences. Le but de cette étape est de choisir les documents aptes à être analysés avec les techniques de TAL. Afin de construire le corpus documentaire le plus adapté au sujet visé et le plus propice à la réutilisation, trois exigences sont à respecter concernant les documents : le contenu doit être en relation avec le sujet de cours visé (le titre et quelques extraits du contenu peuvent donner un indice de pertinence), le contenu doit être suffisamment conséquent (un

document trop court n'apportera que peu de notions, ni ne contiendra suffisamment de relations entre ces notions pour produire un résultat utile), et enfin, le contenu doit être principalement au format textuel. En effet, le traitement de formats spécifiques de représentation du contenu (par exemple des images, des tableaux, des listings de code, etc) est hors du cadre de ces travaux, et nous en discutons dans la conclusion.

Dans le contexte de l'enseignement supérieur, les supports de cours sous forme de manuscrits ou de diapositives contenant du texte, sont adaptés, tout comme les articles de recherche.

3.2.2 Extraction du texte (PI.1)

L'*extraction du texte* vise à transformer les formats variés d'encodage des textes, en texte brut. Le but de cette étape est d'extraire les textes des documents insérés. Les documents dont le texte est déjà numérisé sous forme de texte brut peuvent être utilisés tels quels. En revanche, les textes imprimés ou les scans sous forme d'images ont besoin de passer par des outils de reconnaissance optique de caractères. Nous ne détaillerons pas ces opérations étant donné que les nouveaux supports sont principalement produits et stockés sous format numérique, et que de plus en plus de projets de numérisation des bibliothèques sont lancés depuis les années 1990 [10]. Il existe encore d'autres formats numériques qui encapsulent les textes (PDF, DOCX, ...). Certains outils permettent d'extraire et reconstituer ces textes encapsulés, comme PDFtoText [1].

3.2.3 Nettoyage des textes extraits (PI.2)

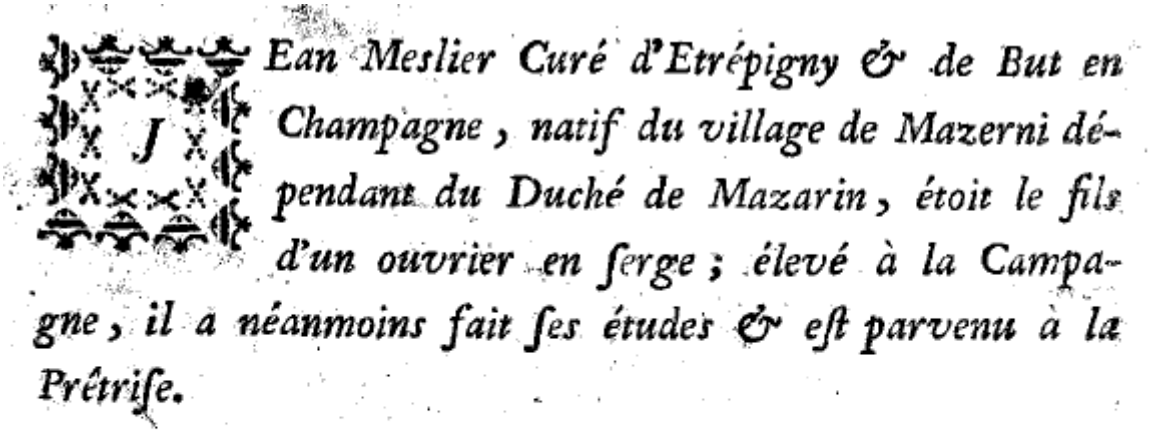
Le *Nettoyage des textes extraits* a pour objectif d'améliorer la qualité des données en les nettoyant (l'équivalent du *data cleansing* en anglais), c'est-à-dire de supprimer les caractères non-affichables ainsi que les mots inutiles pour les traitements suivants. L'extraction du texte brut des documents générant parfois des coquilles dans les textes, le but de cette étape est d'améliorer la qualité des données pour les traitements automatiques suivants. La figure 3.7 illustre en pratique certaines difficultés des logiciels de reconnaissance optique de caractères dans un cas extrême, où un nettoyage du texte brut est requis. Des informations peu pertinentes pour le cas présent peuvent se retrouver en grande quantité dans le texte (par exemple les pronoms et articles fréquemment employés, ou les en-têtes sur chaque page). Afin de limiter l'impact sur les traitements suivants, cette étape se fait relativement tôt dans la méthode.

Pour notre implémentation, nous avons choisi de supprimer certaines classes grammaticales de mots, comme présenté dans le tableau 3.1. Les résultats de l'extraction peuvent varier selon les documents insérés et peuvent potentiellement être améliorés avec un nettoyage manuel ultérieur.

Afin d'étiqueter puis filtrer les classes grammaticales, nous utilisons TreeTagger [101][102]. Le domaine étudié dans nos expériences étant l'informatique, nous avons choisi de conserver certaines classes qui pourraient paraitre superflues. Par exemple, la ponctuation est nécessaire pour pouvoir parler des différences de traitement par les guillemets

1. Projet PDFtoText

simples (') et doubles (") dans certains langages de programmation. En outre, l'absence de certaines prépositions et certains adjectifs entraîne des résultats assez négatifs dans les traitements suivants (« *base de données* » étant transformé en « *base* », ce qui fait perdre tout son sens au terme). La liste des classes TreeTagger que nous avons conservées ou supprimées pour la langue française est présentée dans le tableau 3.1.




```
EanᴧMeslkr Curé d'Etrépigny & de But en J vf Champagne,
natif du village de Mazerni dé-M x;><îvJ* pendant du
Duché de Mazarin, êtoit le fils d'un ouvrier en ferge t
élevé à la Campagne >il a néanmoins fait fes études & ejt
parvenu à la

e.
```

Fig. 3.7 – Exemple d'un cas difficile pour un logiciel de reconnaissance optique de caractères (extrait du Testament de Jean Meslier - 1762)

Conservées	**Supprimées**
abréviation (ABR)	adverbe (ADV)
adjectif (ADJ)	article (DET:ART)
interjection (INT)	pronom possessif (DET:POS)
nom propre et mots inconnus (NAM)	conjonction (KON)
nom commun (NOM)	pronom (PRO)
numéro (NUM)	pronom démonstratif (PRO:DEM)
préposition (PRP)	pronom indéfini (PRO:IND)
ponctuation (PUN)	pronom personnel (PRO:PER)
ponctuation de citation (PUN:cit)	pronom possessif (PRO:POS)
point final ($SENT$)	pronom relatif (PRO:REL)
symbole (SYM)	préposition plus article (PRP:det)
verbe - conditionnel (VER:$cond$)	
verbe - futur (VER:$futu$)	
verbe - impératif (VER:$impe$)	
verbe - imparfait (VER:$impf$)	
verbe - infinitif (VER:$infi$)	
verbe - participe passé (VER:$pper$)	
verbe - participe présent (VER:$ppre$)	
verbe - présent (VER:$pres$)	
verbe - passé simple (VER:$simp$)	
verbe - imparfait du subjonctif (VER:$subi$)	
verbe - présent subjonctif (VER:$subp$)	

Tab. 3.1 – Classes grammaticales conservées ou supprimées

3.2.4 Désambiguïsation (PI.3)

La *désambiguïsation* est l'étape clé du pré-traitement sémantique. En effet, la phase suivante d'analyse structurelle manipulant des documents et les termes communs entre ces documents, il est nécessaire de standardiser ces termes en retrouvant les concepts (au sens du triangle sémiotique [131]) qui leurs sont attachés. Cette étape permet donc à la fois de désambiguïser les termes selon le contexte, mais aussi de les lier à un concept particulier. Ainsi, les problèmes de polysémie [2] et de synonymie [3] sont résolus grâce au contexte du document.

À l'issue de cette étape, une liste de termes désambiguïsés avec l'identifiant unique de concept associé est fournie.

Nous avons implémenté cette étape avec BabelFy [77][76], un outil de désambiguïsation et d'annotation sémantique, lié à BabelNet [81], un réseau sémantique multilingue manipulant des concepts et des entités nommées lié à WordNet [72] et Wikipedia. BabelFy est capable de déduire quel entité nommée ou concept est manipulé pour chacun des termes, et ce dans une multitude de langues. Son autre atout réside dans son lien avec BabelNet, et donc l'encyclopédie Wikipedia : les concepts et entités de la vie courante (technologies, produits commerciaux, ...) sont reconnaissables dans les textes, y compris lorsque ceux-ci sont spécifiques à un domaine bien particulier. BabelFy cherche donc les liens les plus probables entre un terme et un concept ou une entité nommée, puis il indique l'identifiant unique retenu et différents scores liés à ses traitements.

La figure 3.8 illustre parfaitement l'utilité de BabelFy par rapport à d'autres outils plus traditionnels concernant les termes particuliers à certains domaines. On peut voir dans le formulaire d'entrée que le terme « *PHP* » est reconnu par le navigateur web, mais pas « *MySQLi* » (le correcteur orthographique intégré le faisant remarquer). Cependant, grâce à son accès à une base de connaissances suffisamment large, BabelFy arrive à reconnaître l'extension peu connue des non-initiés qu'est « *MySQLi* ». Dans le cadre de cours d'informatique, tous les termes clés sont donc automatiquement extraits des supports de cours insérés, en plus d'être identifiés de façon unique.

Parmi les choix de paramétrage de BabelFy, nous avons opté pour des correspondances exactes (*MatchingType.EXACT_MATCHING*) et les candidats obtenant les meilleurs scores (*ScoredCandidates.TOP*), afin d'obtenir les sens dont BabelFy est le plus sûr. BabelFy n'autorisant pas des textes de plus de 10.000 caractères, les textes transmis sont équitablement découpés pour distribuer autant que possible le plus de caractères dans le moins de groupes possibles, tout en respectant la limite et en conservant entiers les mots aux extrémités. L'intérêt de transmettre de grands groupes de mots est que le contexte est mieux compris par BabelFy, et donc la qualité de ses résultats en est améliorée. L'essentiel étant d'éviter l'utilisation d'une simple division euclidienne qui ne transmettrait que dans un seul cas des groupes parfaitement égaux, et inversement, dans beaucoup d'autres cas le dernier groupe serait relativement petit.

2. « *La plupart [des mots] sont polysémiques, c'est-à-dire pourvus de plusieurs sens* » [41]
3. « *Les synonymes sont des mots qui, appartenant à la même classe grammaticale, ont à peu près la même signification* » [41]

Fig. 3.8 – Exemple de désambiguïsation et d'annotation sémantique avec BabelFy

3.2.5 Filtrage des termes (PI.4)

La dernière étape, le *filtrage des termes*, consiste à améliorer la qualité des listes de termes désambiguïsés en supprimant les termes hors sujet. L'intérêt est de ne proposer à l'utilisateur que des termes en lien avec le domaine, et donc de produire des fragments contenant des termes pertinents pour le cas géré. BabelFy met à disposition plusieurs scores suite à ses traitements. Le score de cohérence mesure en particulier la connectivité d'un terme avec les autres termes du texte transmis [90] (d'où l'intérêt d'envoyer les plus grands groupes de mots possibles de façon équitable).

Nous avons empiriquement constaté que beaucoup de termes inutiles pouvaient être facilement éliminés en fixant un score minimum de cohérence à atteindre. Le score de cohérence correspond au niveau de connectivité du terme désambiguïsé par rapport aux autres termes du même texte fourni à BabelFy. Les termes dont le score de cohérence est strictement supérieur à $0,05$ sont conservés. Ce score élimine malgré tout quelques occurrences de termes intéressants, mais celles-ci restent relativement faibles par rapport à la quantité de termes hors sujet.

À l'issue de cette étape nous obtenons pour chaque document une liste de termes désambiguïsés et en lien avec le sujet. Ces listes sont fusionnées sous forme de matrice d'occurrences, afin d'obtenir un format utile pour la phase suivante.

Étant donné que plusieurs documents peuvent contenir les mêmes termes désambiguïsés, on crée une matrice rassemblant les occurrences des termes contenus dans chaque document. La figure 3.9 illustre une matrice d'occurrences où l'on peut voir dix documents d'entrée (Cours 1 à 10) et quatre termes (web, php, sql, mysql) avec leurs identifiants dans la base de connaissances BabelNet. On peut voir en pratique que le terme « *php* », dont l'identifiant unique est « *bn*:01753580*n* », apparait *15* fois dans le *Cours 1* et *53* fois dans le *Cours 2*. Cette matrice est un pré-requis pour la phase suivante appliquant l'ACF.

	web - bn:00080772n	php - bn:01753580n	sql - bn:02266432n	mysql - bn:01225760n
Cours 1	31	15	2	0
Cours 2	2	53	0	0
Cours 3	6	70	3	6
Cours 4	5	4	0	2
Cours 5	6	67	0	3
Cours 6	8	17	0	0
Cours 7	28	36	0	0
Cours 8	5	57	0	0
Cours 9	11	25	2	2
Cours 10	24	15	1	3

Fig. 3.9 – Exemple de matrice d'occurrences

3.3 Analyse structurelle : métriques de qualité et extraction des clusters

La deuxième phase (PII) de la méthode CREA, l'analyse structurelle, a pour objectif d'extraire des clusters de termes en identifiant les relations entre les termes et documents traités lors de la phase précédente. Lors de la création des documents, des connaissances ont été mobilisées afin d'organiser de manière intelligible (connaissances tacites) des notions à transmettre (connaissances explicites) à d'autres individus. Pour le domaine de l'enseignement, il s'agit d'identifier les liens entre les notions abordées dans les supports de cours afin de réorganiser ces notions sous forme d'un syllabus, ou plus concrètement, sous forme de clusters de termes. Cette phase se déroule en trois étapes successives illustrées par la figure 3.10.

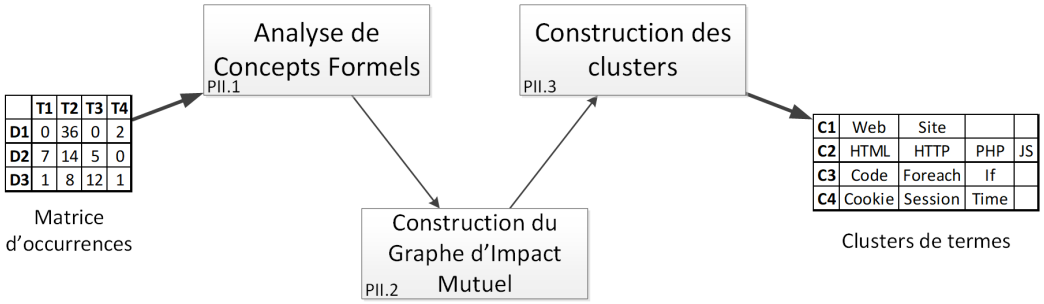


Fig. 3.10 – Les étapes de la phase d'analyse structurelle

— Analyse de Concepts Formels (PII.1) : Les techniques d'analyse de concepts formels servent à analyser les liens entre les termes et les documents les contenant afin de calculer deux métriques (l'impact mutuel et la similarité conceptuelle). Ces métriques nous permettent d'évaluer la pertinence des documents entre eux à partir des termes les plus fréquents, mais aussi d'établir la similarité des termes entre eux.

— Construction du Graphe d'Impact Mutuel (PII.2) : L'impact mutuel entre les termes et les documents est représenté graphiquement afin de mesurer la pertinence des documents entre eux tout en affichant un ensemble de termes explicitant les principaux sujets abordés dans ces documents. Cette métrique permet d'indiquer quels documents devraient être retirés pour obtenir les meilleurs regroupements de termes.

— Construction des clusters (PII.3) : À partir de la similarité conceptuelle des termes entre eux, des clusters de termes sont formés pour présenter à l'utilisateur les fragments réutilisables pour le cas traité.

3.3.1 Analyse de Concepts Formels (PII.1)

L'Analyse de Concepts Formels, présentée en sous-section 2.2.2, est un ensemble de techniques permettant de « *découvrir et [de] structurer des connaissances* » [58] à partir d'une matrice rassemblant des *objets* et leurs *attributs*. Appliquée au domaine de l'enseignement, avec une matrice de supports de cours et de termes les composant,

il s'agit de rechercher les relations entre ces termes et les supports de cours afin d'en extraire des métriques de similarité conceptuelle et d'impact mutuel. Ces métriques permettent ensuite d'obtenir une vision globale de l'ensemble du corpus documentaire, et d'en déduire les sujets centraux à partir des termes, mais également la pertinence des supports de cours par rapport à ces sujets centraux. Le principe général de l'ACF se déroule en plusieurs sous-étapes illustrées par la figure 3.11.

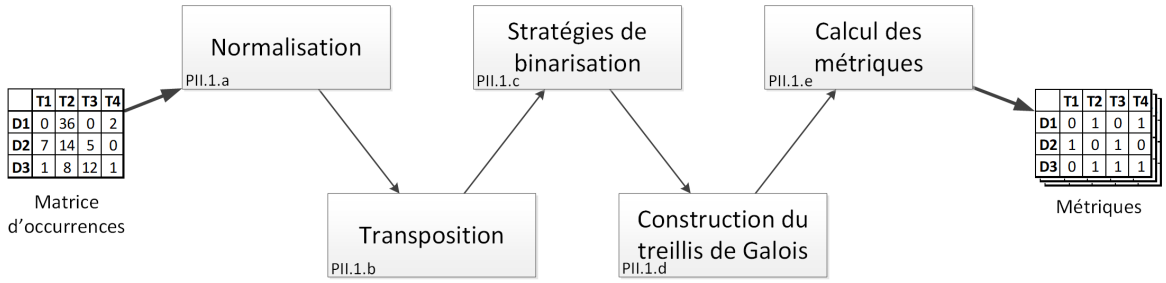


Fig. 3.11 – Les sous-étapes de l'analyse de concepts formels

Normalisation (PII.1.a) :

Afin de binariser la matrice d'occurrences pour en faire un *contexte formel* nous avons choisi de prendre en compte la taille des documents. Un document contenant des milliers de termes au milieu de documents contenant une centaine de termes chacun créera une disproportion. Afin de limiter l'effet de ce biais, une première étape vise à normaliser les quantités proportionnellement à la taille des documents : les occurrences des termes sont transformées en proportions d'occurrences dans chacun des documents. La figure 3.12 illustre l'étape de normalisation.

Matrice d'occurrences

	web	php	sql	mysql
Cours 1	10	10	10	10
Cours 2	1	2	2	0
Cours 3	0	0	1	0

Matrice normalisée

	web	php	sql	mysql
Cours 1	25	25	25	25
Cours 2	20	40	40	0
Cours 3	0	0	100	0

Fig. 3.12 – Exemple de normalisation d'une matrice d'occurrences

Transposition (PII.1.b) :

Le sens de lecture des lignes et colonnes de la matrice est important pour les stratégies de binarisation, l'ACF, et l'interprétation des fragments qui en seront extraits. L'ACF construisant un treillis avec des *objets* caractérisés par des *attributs* issus d'un *contexte formel* (une matrice binaire), et les stratégies se basant sur ces caractéristiques pour générer ce *contexte formel*, il est important que la matrice d'entrée soit correctement présentée : les *objets* forment les lignes, et les *attributs* forment les colonnes. En l'état, les documents (objets) sont caractérisés par les termes (attributs) les composant. Afin de changer de point de vue, et mettre en avant les connaissances

réutilisables, c'est-à-dire les liens entre termes et documents, il est nécessaire de transposer la matrice pour permettre de caractériser les termes (objets) selon les documents (attributs) dans lesquels ils apparaissent. La figure 3.13 illustre l'étape de transposition.

Matrice normalisée

	web	php	sql	mysql
Cours 1	25	25	25	25
Cours 2	20	40	40	0
Cours 3	0	0	100	0

Matrice normalisée transposée

	Cours 1	Cours 2	Cours 3
web	25	20	0
php	25	40	0
sql	25	40	100
mysql	25	0	0

Fig. 3.13 – Exemple de transposition pour caractériser les termes selon les documents où ils apparaissent

Stratégies de binarisation (PII.1.c) :

La matrice d'occurrences étant normalisée et transposée pour mettre en avant les termes caractérisés par les documents dans lesquels ils apparaissent, il est maintenant possible d'appliquer une stratégie permettant de transformer les occurrences en valeurs binaires pour obtenir un *contexte formel* nécessaire à la construction du treillis dans les étapes suivantes. Une stratégie de binarisation définit un algorithme qui transforme les valeurs d'une matrice de l'intervalle $[0, +\infty[$ vers la paire $\{0, 1\}$.

Il existe actuellement plusieurs stratégies [59][58] pour permettre à l'ACF de présenter différentes informations contenues dans une matrice multivaluée. Ces stratégies sont présentées en sous-section 2.2.2. Les métriques visées par la méthode CREA impliquent de générer les contextes formels de plusieurs de ces stratégies. La stratégie *directe* permet d'avoir une vision d'ensemble du corpus utile pour évaluer l'ensemble des sujets centraux et la pertinence des documents entre eux (les valeurs non-nulles sont remplacées par des 1, et les valeurs nulles sont remplacées par des 0). La stratégie *haute* utilisant un $\beta = 1.00$ permet de ne retenir que les termes dont les valeurs de fréquences d'apparition dans les documents sont les plus hautes. Cette stratégie permet de retenir les termes les plus fréquents pour l'ensemble du corpus, mais également pour certains ensemble de documents.

La figure 3.14 illustre ces deux stratégies (Directe et Haute avec un β fixé à 1,00). On remarque que les valeurs non nulles de chaque ligne sont distribuées différemment entre les deux stratégies : *sql* apparaissant dans tous les cours de façon non négligeable à chaque fois, c'est-à-dire que tous les cours parlent de *sql* et l'un lui est dédié, il est le seul terme à être conservé sur la stratégie haute. Cet exemple, plutôt extrême, sert à bien distinguer les objectifs différents des deux stratégies : la stratégie directe permet de voir que le cours *C3* ne dispose que d'une seule notion, et devrait donc être retiré du corpus, à l'inverse, la stratégie haute récupère en effet le sujet central de **l'ensemble** du corpus documentaire, qui est *sql*. Un enseignant doit donc s'assurer

tout d'abord avec la stratégie directe que les documents insérés sont cohérents, puis, une fois les documents correctement sélectionnés, il peut appliquer la stratégie haute pour en extraire les notions essentielles.

Matrice initiale transposée

	Cours 1	Cours 2	Cours 3
web	10	1	0
php	10	2	0
sql	10	2	1
mysql	10	0	0

Stratégie Directe

	Cours 1	Cours 2	Cours 3
web	1	1	0
php	1	1	0
sql	1	1	1
mysql	1	0	0

	Cours 1	Cours 2	Cours 3
web	25	20	0
php	25	40	0
sql	25	40	100
mysql	25	0	0

Matrice normalisée transposée

	Cours 1	Cours 2	Cours 3
web	0	0	0
php	0	0	0
sql	0	0	1
mysql	0	0	0

Stratégie Haute (β = 1,00)

Fig. 3.14 – Exemple d'application des stratégies directe et haute avec un $\beta = 1,00$

Construction du treillis de Galois (PII.1.d) :

Le *contexte formel* généré avec les stratégies de binarisation contient maintenant des termes liés à des documents par des 0 et des 1. Il peut donc être transformé en un *treillis de Galois* composé de *concepts formels* qui serviront à produire l'impact mutuel et la similarité conceptuelle. Le *treillis* est un graphe dont chaque nœud correspond à un *concept formel*. Chaque *concept formel* rassemble des *objets* et leurs *attributs* (et inversement, des *attributs* et les *objets* auxquels ils sont rattachés). Dans notre cas, il s'agit donc de créer des concepts formels contenant des termes et les documents où ils apparaissent. La construction d'un treillis est présentée plus en détails en sous-section 2.2.2.

Nous avons utilisé la bibliothèque *Concepts*[4] en Python pour automatiser la construction des treillis à partir des matrices précédentes.

Calcul des métriques du treillis (PII.1.e) :

L'ACF permet de calculer l'impact mutuel et la similarité conceptuelle sur le treillis généré [58]. Ces métriques sont présentées en sous-section 2.2.2. Dans le cas de la méthode CREA, nous nous concentrons particulièrement sur l'impact mutuel, qui permet de générer des graphiques utiles pour l'amélioration de la qualité des données, et la similarité conceptuelle permettant la construction des clusters de termes.

4. Page du projet *Concepts* pour Python

Similarité Conceptuelle entre objets : La *similarité conceptuelle* permet de comparer deux objets (respectivement attributs) en tenant compte de leur présence dans l'ensemble du treillis. C'est-à-dire, est-ce que les deux objets apparaissent souvent ensemble dans les concepts formels ? Pour un enseignant, cette métrique permet de retrouver quelles notions sont les plus similaires entre elles pour les regrouper. Techniquement, dans le cadre de la méthode CREA, la matrice de similarité conceptuelle permet d'établir la similarité entre l'ensemble des termes afin de pouvoir créer des clusters. Suite à nos expérimentations (présentées dans la sous-section 4.2.2), nous avons fixé les paramètres de cette métrique en utilisant le treillis représentant la matrice de stratégie haute avec un $\beta = 1.00$. En effet, nous avons empiriquement constaté que la stratégie haute permet d'obtenir les termes les plus représentatifs de l'ensemble des documents, donc d'établir une vue globale des notions abordées, et $\beta = 1.00$ filtre suffisamment les termes pour obtenir les meilleurs résultats.

Impact mutuel entre un objet et un attribut : L'*impact mutuel* analyse la "*force de la relation entre un objet et un attribut en fonction des concepts formels qui les associent*" [58]. Cette métrique permet de visualiser quels documents partagent le plus de notions, et réciproquement, quelles notions sont les plus présentes dans les documents. Pour un enseignant, il s'agit d'identifier les notions clés de l'ensemble du corpus documentaire, mais aussi quels documents sont les moins représentatifs (afin de les retirer). En assemblant les mesures d'impact mutuel entre tous les objets et attributs, une *matrice d'impact mutuel* est formée. Cette matrice d'impact mutuel permet de générer un *graphe d'impact mutuel* dans l'étape suivante. Afin d'avoir une vision d'ensemble, elle est générée à partir du treillis représentant la matrice de stratégie directe.

3.3.2 Construction du Graphe d'Impact Mutuel (PII.2)

La *Construction du Graphe d'Impact Mutuel* permet de visualiser la matrice d'impact mutuel et en déduire plusieurs informations importantes pour l'utilisateur. Le graphe d'impact mutuel a la particularité d'être bi-parti, en proposant des nœuds de la classe des objets, et des nœuds de la classe des attributs. Dans la méthode CREA, il s'agit donc de nœuds représentant des termes et des documents.

Une communauté de termes qui apparaissent visuellement au centre du graphe permet d'identifier clairement les termes les plus employés dans l'ensemble du corpus. Cet ensemble central est déduit du degré de connexion des nœuds de la classe des termes : plus le nœud d'un terme est connecté à des nœuds de cours, plus il est situé au centre (et réciproquement, moins il est connecté, plus il est éloigné). À partir de cette visualisation, il est possible de comprendre quels sujets sont abordés par l'ensemble des documents, ou au contraire par certains documents précis. La visualisation permet également de voir si quelques documents sont complètement ou partiellement hors sujet : les documents hors sujet étant peu reliés aux termes apparaissant dans l'ensemble central, leurs nœuds sont particulièrement excentrés.

Pour notre usage, l'utilisateur peut donc déduire si le corpus documentaire correspond à ses attentes selon les notions abordées (via les termes au centre), et éven-

tuellement si certains documents devraient être retirés dans le cas où ceux-ci sont partiellement ou complètement hors sujet.

Lors de nos expérimentations, nous avons utilisé le logiciel *Gephi* [5] pour visualiser les matrices d'impact mutuel sous forme de graphe. Précisément, nous avons utilisé la spatialisation *Force Atlas* qui est un algorithme basé sur les forces [117] (« *force-based* » ou encore « *force-directed* » en anglais).

La *visualisation* cherche à projeter sur un plan des graphiques de points et de lignes [117] (« *points-and-lines charts* » en anglais). La *spatialisation* est une forme de visualisation qui s'intéresse particulièrement à l'espace créé par la projection des données : l'espace de sortie n'est plus considéré comme une contrainte (réduisant le nombre de dimensions), mais bien comme une donnée de sortie liée aux objets projetés. Les algorithmes basés sur les forces permettent de placer les nœuds d'un graphe en respectant un principe : les nœeuds se repoussent mutuellement avec une *force de répulsion*, et les arcs attirent les nœuds avec une *force d'attraction* [117]. Ces algorithmes tendent à espacer les nœuds faiblement liés, et au contraire, à rapprocher les nœuds fortement liés [117]. L'algorithme *Force Atlas* est une spatialisation propre à Gephi dont l'objectif est de « *permettre une interprétation rigoureuse des graphes [...] le plus directement et lisiblement possible malgré un temps d'exécution assez long* » [39]. Il peut s'appliquer sur des graphes comptant de 1 à 10000 nœuds avec une complexité en $O(N^2)$ [38][39]. *Force Atlas* permet à l'utilisateur de sélectionner plusieurs paramètres tels que les forces d'attraction, de répulsion, et quelques autres afin de placer les nœuds [57].

La figure 3.15 illustre le graphe d'impact mutuel entre des termes et des documents. Sur la gauche, le graphe dans son ensemble est affiché. Sur la droite, un zoom est effectué sur les termes apparaissant au cœur du graphe. Les nœuds en rouge sont les termes connectés à tous les cours, les nœuds en orange sont connectés à tous les cours sauf 1, les nœuds en jaune sont connectés à tous les cours sauf 2, et ainsi de suite. Les nœuds gris représentent les supports de cours. En lisant ce graphe, on peut voir que les termes qui apparaissent au cœur du graphe sont *post*, *méthode post*, *nombre*, *langage*, *donnée*, *fichier*, *code*, *php*, *navigateur*, *site*. La combinaison de certains termes permet de comprendre assez vite qu'il s'agit de cours sur du développement web, en particulier de PHP.

La figure 3.16 illustre cette fois la distance entre les documents. Typiquement, le cours *C6* semble très éloigné de l'ensemble, tout comme le cours *C3*, et légèrement *C5*. Il apparait donc judicieux de se pencher plus en détails sur ces derniers pour vérifier leur pertinence pour le corpus documentaire rassemblé.

Au delà du manque de pertinence d'un document par rapport aux termes, un utilisateur peut également décider d'écarter un document si certains termes contenus ne devraient pas apparaître dans les clusters finaux. À l'issue de cette étape, l'utilisateur peut donc décider de supprimer certains documents et ré-exécuter l'analyse de concepts formels à partir de la matrice épurée des documents en question, ou de continuer avec les données déjà générées.

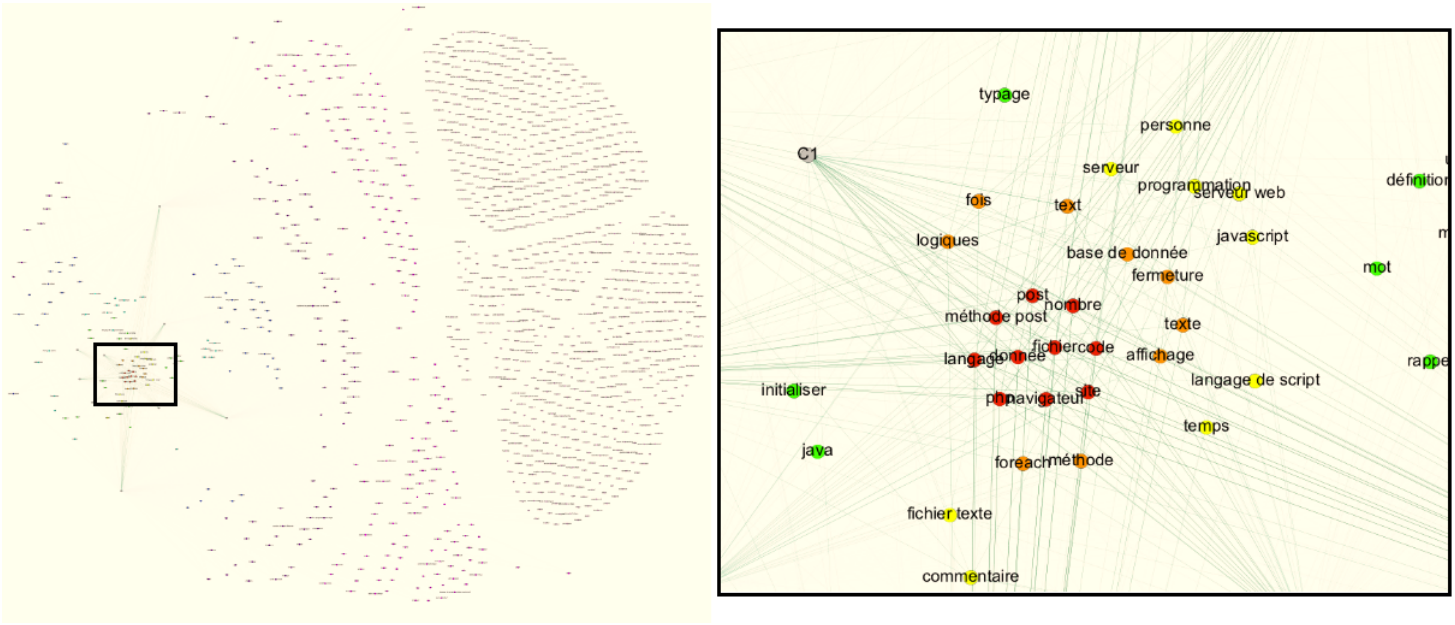


FIG. 3.15 – Graphe d'impact mutuel (généré avec Gephi en utilisant la spatialisation *Force Atlas* et la coloration par *partition* selon le *degré*)

FIG. 3.16 – Graphe d'impact mutuel (généré avec Gephi en utilisant la spatialisation *Force Atlas* et la coloration par *partition* selon le *degré*)

3.3.3 Construction des clusters (PII.3)

La *construction des clusters* est l'ultime étape construisant des clusters à partir de la matrice de similarité conceptuelle des termes. Ces clusters correspondent aux fragments réutilisables, c'est-à-dire dans le contexte de l'enseignement, les notions à aborder ensemble et pouvant former un syllabus. Plusieurs techniques de clustering sont présentées en sous-section 2.2.3. Selon l'implémentation employée, il est parfois nécessaire de transformer la matrice de similarité en une *matrice de dissimilarité*.

Dans nos travaux, nous avons opté pour la *classification ascendante hiérarchique* (ou CAH) afin d'obtenir des clusters non-recouvrants permettant de ne placer chaque terme que dans un seul cluster, tout en indiquant le nombre de clusters que l'utilisateur souhaite. En pratique, l'implémentation de CAH inclue dans la bibliothèque SciPy [119] en Python exige une matrice de distance ou à minima une matrice avec la dissimilarité entre chacun des objets. Ainsi, pour transformer notre matrice de similarité en matrice de dissimilarité (composée de valeurs entre 0 et 1), nous avons utilisé la méthode la plus simple présentée dans [93] consistant à soustraire chaque valeur de similarité à la valeur de similarité maximale, c'est-à-dire soustraire chacune des valeurs à 1 (les 1 devenant des 0, et ainsi de suite jusqu'aux 0 devenant des 1). La formule (3.1) illustre ce calcul.

$$\textit{dissimilarité}(\textit{objet } A, \textit{objet } B) = 1 - \textit{similarité}(\textit{objet } A, \textit{objet } B) \tag{3.1}$$

Afin d'appliquer la CAH, nous avons utilisé les bibliothèques scikit-learn [86] et SciPy [119]. Précisément, nous avons appliqué trois traitements successifs afin de générer les clusters finaux.

1. Nous standardisons tout d'abord les valeurs grâce à *sklearn.preprocessing.scale()* en laissant les paramètres par défaut.
2. Nous appliquons ensuite le traitement effectuant l'agglomération des clusters individuels avec *scipy.cluster.hierarchy.linkage()*. Nous demandons l'usage de la méthode de Ward (`method='ward'`) avec une métrique euclidienne pour ce traitement. (`metric='euclidean'`)
3. Nous utilisons enfin *scipy.cluster.hierarchy.fcluster()* pour récupérer la liste des clusters. Nous demandons de rechercher un maximum de clusters possibles (`criterion='maxclust'`) en visant 8 clusters (`t=8`).

La figure 3.17 illustre un résultat en huit clusters, c'est-à-dire pour huit séances. Les clusters n'étant pas ordonnés, c'est à l'enseignant que revient ce choix. Une technique d'ordonnancement temporelle, encore en cours d'évaluation, est proposée en sous-section 5.2.3.

Dans tous les cas, ces clusters constituent une base réutilisable de termes dont l'utilisateur peut se servir pour traiter son nouveau cas.

1	php	code	fois	post	jour	foreach	cle	classe	class	mysqli	
2	page web	navigateur	serveur web	texte	concerner	délimiter	utilisateur	associer	personne	machine	mysql
3	url	langage	case	fermeture	session	chaîne	entête	avoir accès			
4	fichier	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client				
5	typage	mot	moteur	affiche	transaction	visiteur					
6	base de donnée	insert	varchar	null							
7	xml	configuration	composer	doctype							
8	donnée	text	méthode post	programmation	site	langage de script	list	méthode	timestamp	files	

Fig. 3.17 – Liste de huit clusters générés

4. ÉVALUATION ET VALIDATION DE LA MÉTHODE CREA

Dans ce chapitre, nous évaluons la méthode CREA décrite précédemment et discutons ces résultats. Les expérimentations sont effectuées selon la démarche méthodologique de *design science* [49][48] décrite par la suite. Nous présentons un protocole d'évaluation basé sur plusieurs validations (structurelles, fonctionnelles, et par retour d'expérience), puis nous l'exécutons. Enfin, nous discutons des conclusions de ces expérimentations, des limites de la méthode CREA, et de la méthodologie d'évaluation en elle-même.

Sommaire

4.1 Méthodologie d'évaluation

La *science du design* [49][48][87][85], ou *design science* en anglais, décrit une méthodologie de recherche permettant de développer et évaluer un artefact visant à répondre à une question de recherche. Cette méthodologie de recherche appliquée au système d'informations s'appuie sur trois boucles d'activités (pertinence, rigueur, et design) décrites dans [48] et [85], et illustrées par la figure 4.1. Chaque boucle décrit un ensemble d'activités à réaliser successivement pour produire un artefact de meilleure qualité à chaque itération :

— La *boucle de pertinence* permet de déterminer les problèmes ou opportunités d'un domaine d'étude en particulier, et les critères nécessaires à évaluer (dans le cadre de recherche plus général [49], il s'agit des critères permettant de valider les tests fonctionnels). Cette boucle est itérée autant de fois que nécessaire tant que les critères fonctionnels ne sont pas validés.

— La *boucle de rigueur* permet de choisir les théories et méthodes les plus adaptées [85] pour répondre à la question de recherche, et donc déterminer les solutions déjà proposées dans le domaine de recherche. Cette boucle permet de construire et alimenter une base de connaissances, elle se rapproche d'une revue de la littérature ou d'un état de l'art.

— La *boucle de design* s'appuie sur deux activités en particulier : la conception et l'évaluation [85]. Ces activités dépendent des résultats des deux autres boucles afin de déterminer quoi réaliser (par rapport aux exigences), quels résultats sont visés (par rapport aux critères de validation), et enfin qu'est-ce que ces travaux apportent à la base de connaissances.

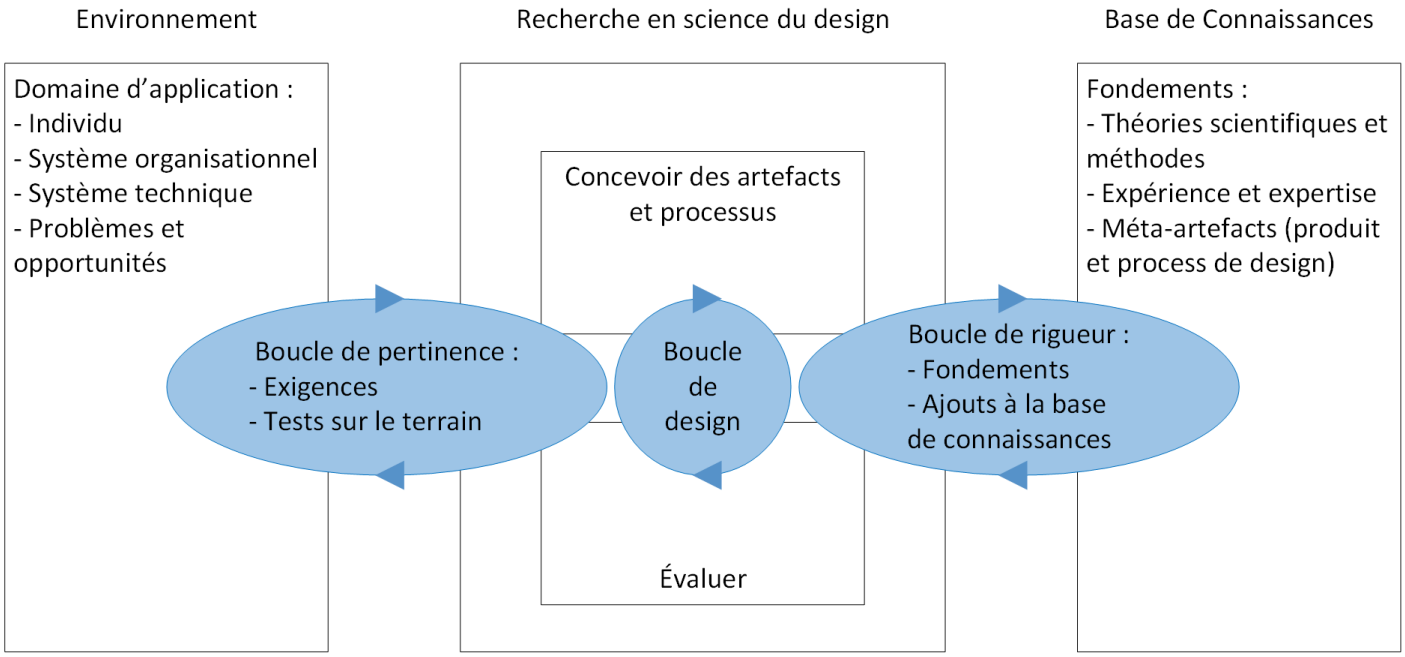


Fig. 4.1 – Les trois boucles de la science du design présentées dans [48] et [85]

Pour répondre à notre question de recherche, nous avons effectué une méthodologie de recherche s'inspirant particulièrement de la science du design. Tout d'abord, nous avons réalisé une étude du domaine des processus à forte intensité des connaissances et l'avons publié dans un article [8]. Celle-ci a permis de mettre en évidence six défis récurrents dans le domaine des processus à forte intensité de connaissances en étudiant les différents fondements théoriques et réponses actuellement apportées par la recherche. Cette étude a donc contribué aux boucles de pertinence et de rigueur en construisant un premier cadre. Afin d'évaluer les résultats d'une contribution pouvant relever certains des défis présentés, nous avons choisi un domaine particulièrement adapté à la manipulation de connaissances et rencontrant actuellement des difficultés suite à la crise du COVID-19 : l'enseignement supérieur et la recherche. Notre premier objectif a été de proposer une méthode permettant d'aider à construire le cours le plus pertinent possible pour un sujet donné en s'appuyant sur la réutilisation de fragments de processus. Notre stratégie de validation pour la méthode CREA consiste à la réalisation et l'analyse de cinq scénarios de création de cours ainsi qu'un retour d'expérience basé sur des entretiens.

Un premier scénario servant de cas de référence a été réalisé afin d'établir une première itération de la boucle de design et vérifier les critères attendus pour les domaines de recherche et d'application. Ce premier scénario vise à rassembler plusieurs supports de cours traitant d'un même sujet, afin de pouvoir en extraire des fragments réutilisables (des notions à présenter à chaque séance de cours). La boucle de design pouvant fonctionner seule, ce premier scénario nous a permis de fixer certains paramètres au sein de la méthode construite. Ces paramètres sont présentés au fur et à mesure dans le chapitre 3. De plus, plusieurs directions ont été testées et abandonnées étant donné la faible qualité des résultats apportés.

Nous avons ensuite étendu nos objectifs en ajoutant une détection de la pertinence des supports de cours par rapport à un sujet donné. Un deuxième scénario a donc été construit en insérant parmi les données d'entrée un document traitant d'un sujet un peu plus éloigné, mais pas complètement hors sujet. En comparaison du premier scénario, celui-ci doit mettre en évidence que le document supplémentaire est moins pertinent que les autres, et donc qu'il n'est pas conforme.

Un troisième scénario a permis d'étudier l'impact de l'augmentation du nombre de documents sur les métriques de sorties et les clusters générés. Deux fois plus de documents ont été insérés en entrée pour observer les effets. Ces trois premiers cas ont surtout permis d'effectuer une validation structurelle de la méthode CREA que nous avons construite, tout en s'assurant de l'absence de problèmes fonctionnels sur le fond.

Un quatrième scénario a permis d'étudier la correction du bruit sur un support et les conséquences sur le graphe d'impact mutuel au fur et à mesure. Certains documents précédemment utilisés ont été utilisés et l'un d'entre eux a été *corrigé* en lui retirant des sections hors sujet.

Un cinquième scénario a permis d'étudier l'impact de la langue sur les clusters finaux et de la nature des documents sur leur pertinence, tout en s'assurant que la méthode n'a pas été spécialisée sur le cas de référence et ses dérivés. Plusieurs documents de nature diversifiée en anglais (supports de cours, d'articles de recherche, et de pages web) ont été insérés en entrée.

Enfin, deux questionnaires ont été réalisés sur le cas de référence en interrogeant des experts du sujet traité pour s'assurer que les résultats sont exploitables par des utilisateurs. Nous présentons dans cette thèse le résultat des multiples itérations de ces trois boucles et les cinq scénarios qui ont permis cela.

Cette méthodologie de recherche manipule elle-même de nombreuses connaissances et réutilise des contributions passées, ou s'en inspire grandement pour avancer plus loin. Comme tout processus de recherche, il s'agit bien d'un processus à forte intensité de connaissances qui contribue à répondre à un problème de processus à forte intensité de connaissances (condition nécessaire). La contribution proposée dépendant elle aussi des connaissances implicites de l'utilisateur, ainsi que des connaissances intégrées aux documents et aux bases de connaissances utilisées comme nous l'avons présenté dans les chapitres précédents, il n'y a donc pas de contradiction théorique.

4.2 Protocole d'évaluation

4.2.1 Scénarios d'évaluation

Cinq scénarios ont été préparés pour pouvoir valider la méthode CREA dans plusieurs situations. Ces scénarios visent à construire un cours en 8 séances, donc générer 8 clusters.

— Scénario n°1 (cas référence) : Ce scénario de référence doit pouvoir fournir des clusters de termes pertinents pour un cours de développement web en PHP. Les clusters de termes générés doivent au moins aider un enseignant à parler des bases du développement du web avec HTML, PHP, et un peu de base de données avec MySQL. 9 supports de cours de PHP sont utilisés, 6 sont au format diapositives et 3 au format texte long. Tous les documents fournis en entrée sont pertinents, bien qu'ils aient leurs spécificités. Ceux-ci sont détaillés dans les paragraphes suivants afin de mieux comprendre leurs spécificités et ce qui est attendu dans le graphe d'impact mutuel. Nous avons également réalisé une étude comparative depuis notre propre point de vue expert sur ce cas de référence. Enfin, nous avons demandé à 5 informaticiens de construire 8 clusters à partir des termes retenus par la stratégie haute et $\beta = 1.00$ pour pouvoir les comparer à ceux produits par la méthode CREA. Nous leurs avons ensuite demandé de donner leur avis sur les clusters générés par la méthode CREA. Ainsi, ce scénario contribue à répondre aux hypothèses H1, H2, H3, H4a-b, H5, H6a-b-c, et H8a-b-c en faisant l'extraction des termes d'un premier corpus documentaire traitant d'un même sujet, puis en produisant un premier graphe d'impact mutuel et des clusters. Ces données seront comparées à celles produites dans les autres scénarios afin de valider ou invalider les hypothèses.

— Scénario n°2 : Ce scénario vise à vérifier la résistance au bruit en insérant un cours sur un sujet éloigné dans le corpus initial. Les 9 supports de cours de PHP du scénario n°1 sont utilisés, et sont complétés d'un cours de Java au format texte long. Le document Java traitant de développement, il sera suffisamment proche des cours de PHP pour avoir quelques liens sémantiques, mais le cœur du sujet n'étant pas le développement web, ces liens sémantiques seront limités et il devrait apparaitre comme éloigné des autres documents. Ce scénario se limite à la génération du graphe d'impact mutuel pour permettre à un enseignant de constater qu'un des documents insérés est trop peu pertinent par rapport aux autres, et qu'il faudrait donc le retirer. Ce deuxième scénario contribue à répondre aux hypothèses H1, H2, H3, H7b, et H8a-b-c en faisant l'extraction des termes d'un deuxième corpus documentaire dont un document est incorrect, puis en produisant un second graphe d'impact mutuel devant faire apparaître une incohérence.

— Scénario n°3 : Ce scénario vise à vérifier la robustesse de la méthode en doublant le nombre de documents par rapport au corpus initial. 18 supports de cours de PHP sont donc utilisés. 9 proviennent du scénario n°1 (6 au format diapositives et 3 au format texte long), et 9 autres sont ajoutés (5 au format diapositives et

4 au format texte long). Ce troisième scénario contribue à répondre aux hypothèses H1, H2, H3, H4a-b, H5, H6a-b-c, et H8a-b-c en faisant l'extraction des termes d'un troisième corpus documentaire traitant d'un même sujet mais avec beaucoup plus de documents, puis en produisant un troisième graphe d'impact mutuel et des clusters.

— Scénario n°4 : Ce scénario vise à vérifier que la méthode est fonctionnellement correcte en s'assurant que l'amélioration des documents en entrée produit un graphe d'impact mutuel reflétant ces améliorations. 7 supports de cours de PHP sont utilisés parmi les 18 du scénario n°3, en particulier, nous ne sélectionnons que ceux au format texte long. L'un de ces supports étant considéré comme assez peu cohérent par rapport aux autres, nous étudions succinctement ses différences pour pouvoir le *corriger* deux fois de suite en supprimant des sections hors sujet. Afin de s'assurer de la validité des corrections, et de leurs conséquences, chaque test est également reproduit une deuxième fois en insérant le support de cours Java parmi les données d'entrée. Ce quatrième scénario contribue à répondre aux hypothèses H1, H3, H7b, et H8a-b-c en étudiant l'évolution du graphe d'impact mutuel suite aux modifications dans le contenu des documents.

— Scénario n°5 : Ce scénario a trois objectifs distincts. Il vise tout d'abord à s'assurer que la méthode n'a pas été construite pour favoriser la détection de cours PHP, étant donné que le cas de référence a été utilisé pour fixer les paramètres tout au long des travaux de cette thèse. Le second objectif est de valider que la partie traitement automatique du langage supporte d'autres langues, en particulier l'anglais. Le troisième objectif est de confirmer que les documents pouvant être analysés ne se limitent pas à des supports de cours, mais peuvent être étendus à d'autres formats (page web, articles de recherche, livres, ...). Nous avons donc testé la méthode CREA sur 13 documents en anglais traitant des Statecharts [44]. Afin de gérer l'anglais, les configurations de TreeTagger et BabelFy ont été adaptées. Ce cinquième scénario contribue à répondre aux hypothèses H1, H2, H3, H4a-b, H5, H6a-b-c, H7a-c, et H8a-b-c en faisant l'extraction des termes d'un quatrième corpus documentaire beaucoup plus hétérogène traitant d'un sujet différent des précédents, puis en produisant un cinquième graphe d'impact mutuel et des clusters.

	H1	H2	H3	H4			H5	H6			H7			H8			H9
				a	b	c		a	b	c	a	b	c	a	b	c	
S1	•	•	•	•	•		•	•	•	•				•	•	•	
S2	•	•	•									•		•	•	•	
S3	•	•	•	•	•		•	•	•	•				•	•	•	
S4	•		•									•		•	•	•	
S5	•	•	•	•	•		•	•	•	•	•		•	•	•	•	

Tab. 4.1 – Scénarios et hypothèses visées

4.2.2 Validations structurelles, fonctionnelles, et par retour d'expérience

Afin d'évaluer la méthode CREA, nous avons effectué plusieurs validations structurelles et fonctionnelles. Nous nous assurons ainsi que chaque composant fonctionne, et que l'ensemble de la méthode CREA génère des résultats utiles pour un enseignant.

Validations structurelles

La validation structurelle du point de vue design science est décrite ainsi dans [49] : « *Perform coverage testing of some metric (e.g., execution paths) in the artifact implementation* », c'est-à-dire qu'elle vérifie certaines métriques dans l'artefact implémenté. Elle est plus spécifiquement comparée au fonctionnement « *White Box* » dans lequel les composants d'un artefact sont connus avec précision, et sont testés un à un.

Pour effectuer la validation structurelle, nous comparons plusieurs valeurs par rapport aux documents d'entrée :

— Nous comptons le nombre de *mots* (suite de caractères séparés par des espaces, de la ponctuation, ou des retours à la ligne) et de *termes* (ensemble de mots formant un concept ou une entité nommée dans le réseau sémantique BabelNet lors de l'étape de désambiguïsation (PI.3)) dans chaque document, et nous les comparons au nombre de mots et termes conservés suite aux étape de nettoyage des textes (PI.2) et de filtrage des termes (PI.4). Les pourcentages de mots et de termes conservés sont également indiqués. Cette métrique permet de comparer les documents entre eux en cas de disproportions dans les quantités de mots ou de termes, et d'observer les conséquences sur les résultats intermédiaires et finaux.

— Nous comptons le nombre de *termes uniques* (entités nommées et concepts retrouvés dans le réseau sémantique BabelNet lors de l'étape de désambiguïsation (PI.3), sans prendre en compte les occurrences) retenus et exclus dans chaque document, et nous le comparons au nombre de termes uniques retenus et exclus suite à l'étape de filtrage des termes (PI.4). Les pourcentages de termes uniques retenus et exclus, et le nombre de termes communs aux deux listes sont également indiqués. Cette métrique permet de comparer les documents entre eux en cas de disproportions dans les quantités de termes uniques, et d'observer les conséquences sur les résultats intermédiaires et finaux. Les termes communs servent également à mesurer le nombre de termes dont le contexte détecté par BabelFy ne correspond pas à celui qui serait attendu ou dont la désambiguïsation n'est pas sûre.

— Nous comptons la quantité de termes uniques retenus pour les quatre stratégies de binarisation (directe, moyenne, haute, basse) appliquées avec cinq valeurs de β allant de 0.00 à 1.00 par pas de 0.25. Cette métrique permet de vérifier la quantité de termes uniques retenus par les stratégies, et estimer la valeur de β la plus adaptée pour sélectionner les termes les plus importants/retirer les termes pouvant provoquer du bruit dans les traitements suivants. Elle vise également

à confirmer que les stratégies directe et haute conservent respectivement le plus de termes et une quantité limitée mais suffisante de termes pour les traitements suivants.

— Nous comptons la quantité et les proportions de « 0 » et de « 1 » dans les matrices générées par les quatre stratégies de binarisation et les cinq valeurs de β. Cette métrique permet de s'assurer que les treillis générés ne seront ni trop vides (beaucoup trop de « 0 » forment quelques concepts formels aux extrémités du treillis uniquement), ni trop pleins (beaucoup trop de « 1 » forment toutes les combinaisons possibles de documents et de termes), mais feront suffisamment de liens entre les termes et les documents pour que des concepts formels intéressants soient formés (c'est-à-dire qu'il existe des concepts formels mélangeant plusieurs documents *et* plusieurs termes, sans produire toutes les combinaisons possibles).

— Nous analysons les clusters générés avec la stratégie haute afin de confirmer que le choix du β à 1.00 est adapté. Les clusters générés sont comparés pour connaître le β générant les meilleurs clusters (c'est-à-dire ceux contenant le moins de bruit, et dont l'organisation est la plus correcte pour un expert). Cette analyse est également réalisée avec la validation fonctionnelle étant donné qu'elle concerne la sortie finale.

Validations fonctionnelles

La validation fonctionnelle du point de vue design science est décrite ainsi dans [49] : « *Execute artifact interfaces to discover failures and identify defects* », c'est-à-dire qu'elle vérifie si l'artefact présente des défaillances à l'usage. Elle est plus spécifiquement comparée au fonctionnement « *Black Box* » dans lequel les composants d'un artefact sont inconnus, et seules les entrées et sorties peuvent être contrôlées.

Pour effectuer la validation fonctionnelle, nous analysons cette fois les données produites en sortie de la méthode pour chaque cas :

— Nous analysons le graphe d'impact mutuel formé grâce aux métriques du treillis avec la stratégie directe lors de l'étape calcul des métriques du treillis (PII.1.e) afin d'évaluer si la méthode détecte correctement quels documents portent sur le même sujet, et lesquels en sont éloignés. La validation étant visuelle, nous observons donc quels documents sont suffisamment rapprochés pour former un groupe au centre du graphique, et quels documents sont au contraire éloignés du centre.

— Nous analysons les clusters générés avec la stratégie haute afin de confirmer que la méthode fonctionne et produit des clusters pertinents pour la réutilisation. Cette validation étant purement qualitative, nous nous sommes permis d'évaluer avec notre propre expertise les résultats. Pour confirmer la pertinence des clusters, nous avons également demandé à des informaticiens leurs avis sur un cas servant de référence.

Validation par retour d'expérience

Nous avons également réalisé des entretiens en demandant un avis extérieur sur la qualité des clusters générés par la méthode CREA. Pour cela, 5 informaticiens ont été interrogés avec deux questionnaires successifs (le deuxième n'étant envoyé qu'une fois le premier récupéré) à propos des résultats du scénario n°1 (les 9 supports de cours de développement web en PHP). 2 informaticiens sont novices en PHP (faible expérience en développement web avec PHP ou en administration d'un serveur web utilisant PHP). 3 informaticiens sont experts en PHP (nombreuses expériences en développement web, déploiement d'applications, et administration d'applications utilisant PHP). Parmi les 3 experts en PHP, 2 ont une expérience dans l'enseignement de l'informatique et en particulier des langages de programmation. L'objectif de diviser en deux questionnaires séparés est de récupérer tout d'abord l'avis individuel de chacun (en utilisant leurs connaissances tacites) sur l'organisation du cours idéal avec les termes uniques obtenus, avant de demander un avis critique sur les résultats de la méthode CREA.

Afin de connaître plusieurs dispositions possibles du point de vue humain, le premier questionnaire demande d'organiser les 60 termes issus du scénario n°1 en 8 clusters. Les consignes suivantes ont été envoyées avec une feuille excel contenant l'ensemble des termes :

```
ORGANISER LES 60 TERMES EN 8 SEANCES DE COURS
(voir onglet/feuille "Clusters" tout en bas)
CONTRAINTES :
 - au moins 1 terme par séance (une séance est représentée par un cluster)
 - chaque terme doit être au plus ET au moins dans une séance
```

Le deuxième questionnaire présente tout d'abord les clusters du cas de référence avant de poser 4 questions. Les 2 premières nous intéressent particulièrement pour cette expérience, les 2 autres servant pour des travaux ultérieurs sur l'axe temporel (voir sous-section 5.2.3 en conclusion). Les questions sont posées comme suit :

1. Notez cette proposition de regroupements :

   (0) Les regroupements n'ont aucun sens pour un cours de PHP [aucun terme ne va avec un autre dans un même cluster pour faire un cours de PHP]

   (1) Seuls quelques regroupements ont un peu de logique [beaucoup de clusters contiennent des termes qui n'ont aucun lien entre eux pour faire un cours de PHP]

   (2) Les termes sont vaguement liés, ce serait très difficile d'en faire un cours de PHP [les clusters nécessitent tous au moins plusieurs modifications pour être utilisables pour faire un cours de PHP]

   (3) Les termes sont plutôt liés, mais il y a beaucoup trop de bruit pour pouvoir facilement construire un cours PHP [quelques clusters ont besoin d'être profondément modifiés pour faire un cours de PHP]

(4) Les termes sont quasiment tous liés, il y a un peu de bruit, mais il est possible de construire un cours de PHP [les clusters sont majoritairement bons, quelques-uns ont besoin de quelques modifications pour faire un cours de PHP]

(5) Les termes sont correctement liés, un cours peut directement en être produit [aucune modification n'est requise : un cours de PHP peut directement être produit]

2. Expliquez votre choix :
3. À quel moment placeriez-vous chacun des clusters pour faire un cours de PHP ?

   Début :

   Milieu :

   Fin :
4. Dans quel ordre très précis placeriez-vous chacun des clusters pour faire un cours ? (si vous avez un doute sur certain(s), indiquez si vous avez un doute entre plusieurs endroits OU si vous n'arrivez absolument pas à le placer)

   Ordre :

   Doute 1 : (hésitation entre plusieurs moments pour le placer)

   Doute 2 : (aucune idée du moment où le placer)

Les réponses au premier questionnaire sont comparées avec l'indice de Rand [94] et l'indice de Rand ajusté [53] afin de déterminer les différences entre chacune des réponses. Ces deux indices permettent de comparer deux partitions, c'est-à-dire comparer dans deux cas la façon dont des éléments ont été regroupés dans des clusters.

L'indice de Rand [94] produit une mesure de similarité en comparant des paires de points entre chacune des deux partitions. Cette mesure produit un résultat allant de 1 (les deux partitions ont strictement les mêmes clusters contenant exactement les mêmes points) à 0 (les deux partitions n'ont strictement aucune paire de points regroupée dans les mêmes clusters, par exemple comme expliqué dans [94] : une première partition a tous ses points regroupés dans un unique cluster, et une deuxième partition a autant de clusters qu'il y a de points). Pour calculer l'indice de Rand, on s'intéresse particulièrement à quatre types de paires de points dans chacune des deux partitions testées [53][105] :

— a : les paires de points appartenant dans les deux cas aux mêmes clusters (exemple figure 4.2 : *Terme 2* et *Terme 4* sont ensemble dans les deux cas)

— b : les paires de points appartenant à des classes différentes dans le premier cas, et à la même classe dans le deuxième cas (exemple figure 4.2 : *Terme 1* et *Terme 3* sont dans les clusters I et II dans le cas n°1, puis dans le cluster A dans le cas n°2)

— c : les paires de points appartenant à la même classe dans le premier cas, et à des classes différentes dans le deuxième cas (exemple figure 4.2 : *Terme 1* et *Terme 2* sont dans le cluster I dans le cas n°1, puis dans les clusters A et B dans le cas n°2)

— d : les paires de points appartenant à des classes différentes dans les deux cas (exemple figure 4.2 : *Terme 1* et *Terme 6* sont dans des clusters différents dans les deux cas)

Un autre point de vue permet de considérer les types de paires a et d comme étant en *accord* (les paires de points sont ensemble dans les deux cas, ou séparées dans les deux cas), et les types de paires b et c comme étant en *désaccord* (les paires de points sont ensemble dans un cas, et séparées dans l'autre cas). L'indice de Rand se calcule précisément par le rapport décrit dans l'équation (4.1), où a, b, c, d correspondent aux nombres de paires de chaque type. On notera que l'indice de Rand est symétrique.

$$RI(Cas\ n°1, Cas\ n°2) = \frac{a+d}{a+b+c+d} \tag{4.1}$$

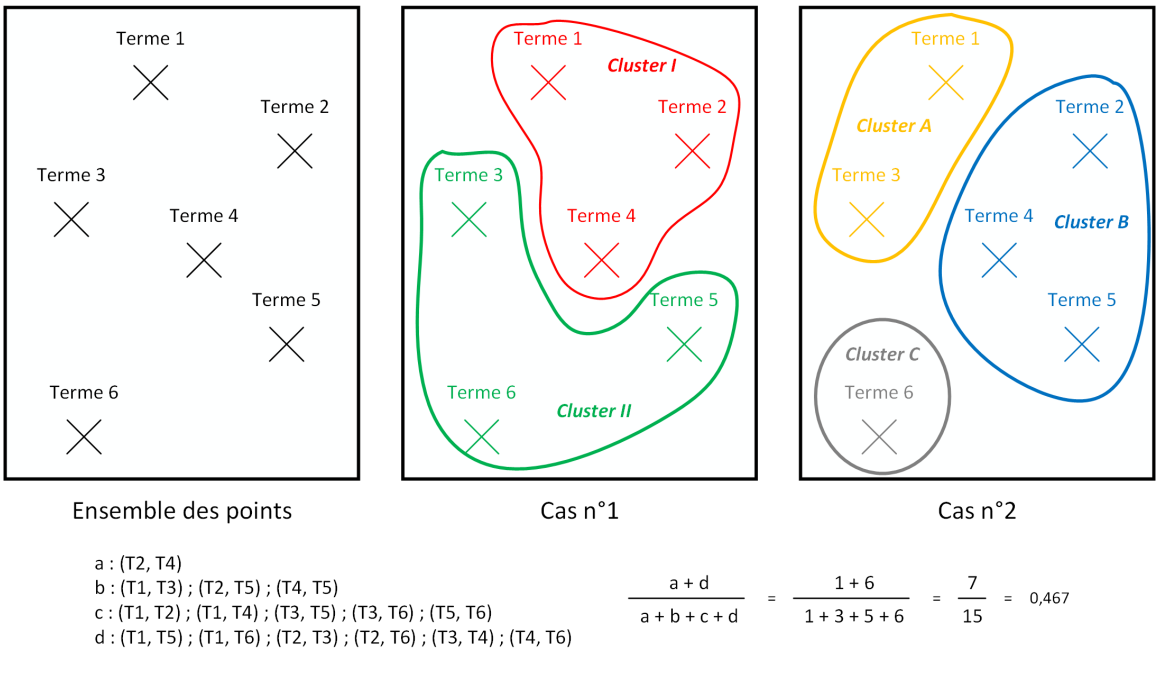


Fig. 4.2 – Exemple de points, de deux partitions différentes, et du calcul de l'index de Rand

L'indice de Rand ajusté [53] vise à prendre en compte le fait que certaines paires de points considérées comme en accord ne le sont que par hasard. Le point de vue utilisé par cet indice est légèrement différent : il teste la possibilité que les deux partitions aient été construites au hasard, tout en s'appuyant sur la considération qu'une des partitions est considérée comme correcte (« *ground truth* ») et que l'autre est testée pour connaître sa similarité. L'indice de Rand ajusté corrige également certains défauts de l'indice de Rand original, notamment le fait que la valeur générée s'approche rapidement de 1 lorsque le nombre de clusters augmente [92]. Concernant l'interprétation, un indice de Rand ajusté proche de 1 implique que les deux partitions sont identiques, et un indice proche de 0 indique que les partitions ont été générées aléatoirement [92] (les valeurs peuvent également être négatives). Nous ne détaillerons pas la construction de l'indice de Rand ajusté dans cette thèse, mais il est intéressant de noter que celui-ci est également symétrique.

4.3 Déroulement des expérimentations

Dans cette section, nous présentons tout d'abord les 9 supports de cours du scénario 1 plus en détails, étant donné que ce scénario sert de référence pour plusieurs expériences. Nous disposons ensuite succinctement l'ensemble des supports utilisés lors des différentes validations. Les résultats bruts de chaque étape de la méthode CREA, ainsi que les résultats finaux (graphe d'impact mutuel, et clusters de termes), sont finalement exposés.

4.3.1 Présentation détaillée des documents du cas de référence

Document :	**C1**
Titre :	*Cours HTML/PHP*
Auteur :	E.Coquery
Type de document :	diapositives
Année :	2008
Nb pages/diapositives :	46
En-têtes et/ou en-pieds :	Oui
Mots clés :	HTML, liens hypertexte, formulaires, PHP, opérateurs, tableaux, structure de contrôle, fonctions, GET, POST, MySQL, sessions
Niveau du public :	Débutant
Commentaires :	C1 est divisé en 3 sections : les pages webs statiques/dynamiques, HTML, puis PHP. Quelques schémas illustrent le cours, mais restent en quantité limité. Les diapositives contiennent beaucoup d'exemples de code, mais également les mots clés pour un tel cours.

Document :	**C2**
Titre :	*PHP*
Auteur :	T.Lecroq
Type de document :	diapositives
Année :	2009
Nb pages/diapositives :	64
En-têtes et/ou en-pieds :	Oui
Mots clés :	XHTML, formulaires, PHP, tableaux, instructions de contrôle, chaînes de caractères, fonctions, transfert de fichiers, lecture, écriture, cookies, sessions, XHTML 2.0, X/HTML 5
Niveau du public :	Débutant
Commentaires :	C2 est divisé en 7 sections se concentrant majoritairement sur PHP.

Document :	**C3**
Titre :	*Cours PHP - Versions 4.x et 5.x*
Auteur :	S.Rohaut
Type de document :	texte
Année :	2006
Nb pages/diapositives :	93
En-têtes et/ou en-pieds :	Oui
Mots clés :	PHP, ASP, CGI, HTML, syntaxe, variables, opérateurs, expression, structure de contrôle, fonctions, formulaires, GET, POST, dates, MySQL, PhpMyAdmin, MysqlCC, SQL, types colonnes, types tables, index de tables, fichiers, bases de données, sessions, cookies, images, POO, PHP 4, PHP 5
Niveau du public :	Moyen
Commentaires :	C3 est divisé en 17 chapitres. Le cours introduit surtout les nouveautés de PHP 5 par rapport à PHP 4 (particulièrement la programmation orientée objet).

Document :	**C4**
Titre :	*Programmation Spécifique : Programmation Web*
Auteur :	V.Pagé
Type de document :	diapositives
Année :	2007
Nb pages/diapositives :	99
En-têtes et/ou en-pieds :	Oui
Mots clés :	HTML, client web, serveur web, CSS, PHP, fonctionnement, instructions, variables, chaînes de carcatères, tableaux, fonctions, inclusion de fichiers, variables prédéfinies, portée des variables, formulaires, cookies, sessions, fichiers, base de données, Postregress, images, GD, sécurité
Niveau du public :	Débutant
Commentaires :	C4 est divisé en 2 parties : l'une sur HTML, et l'autre sur PHP et MySQL. Le cours se termine sur une mention de la gestion d'images avec GD (mais le contenu est absent), et une mention sur la sécurité.

Document :	**C5**
Titre :	*Cours PHP Accéléré*
Auteur :	G.Rozsavolgyi
Type de document :	texte
Année :	2018
Nb pages/diapositives :	115
En-têtes et/ou en-pieds :	Oui
Mots clés :	caractéristiques, historique, installation, configuration, HTML, PHP, bases du langage, tableaux, inclusion de fichiers, POO, base de données, SQL, PDO, transactions, connexions persistantes, sessions, cookies, HTTP, XML, MVC, templates, frameworks, Composer, Symfony, Flex, tests, TDD, web services, REST, API, Git
Niveau du public :	Intermédiaire
Commentaires :	C5 aborde quasiment toutes les notions possibles en une quarantaine de sections. Le document contient quelques mentions sur des feuilles de TD.

Document :	**C6**
Titre :	*PHP, une initiation*
Auteur :	D.Gonzalez
Type de document :	texte
Année :	2009
Nb pages/diapositives :	160
En-têtes et/ou en-pieds :	Non
Mots clés :	langages informatiques, langages webs, PHP, Javascript, bases du langage, fonctions, formulaires, HTML, chaînes de caractères, tableaux, connecteur, PDO, instructions, SQL, inclusion de fichiers, gestion de l'identification des utilisateurs, sécurité, sessions
Niveau du public :	Débutant
Commentaires :	C6 est divisé en 4 chapitres : le cours, le cours hors programme, les corrigés des exercices, et des études de cas. Le document démarre sur une présentation du document lui-même, puis aborde les notions du cours. Le cours contient un peu de bruit au début avec diverses mentions de copyright, licences, et du projet GNU, puis les façons de retrouver le document. *Seul le premier chapitre contient le cours.* Un mini projet est présenté dans les pages 59 et 60. De la page 65 à 93, des corrigés d'exercices sont présents. Puis, de la page 93 à 149, des projets sont présentés.

Document :	**C7**
Titre :	*Le Langage PHP, Ou comment concevoir des sites dynamiques*
Auteur :	C.Escazut
Type de document :	diapositives
Année :	2017
Nb pages/diapositives :	37
En-têtes et/ou en-pieds :	Oui
Mots clés :	bases du langage, PHP, intégration, HTML, variabes, structures conditionnelles, structures itératives, tableaux, fonctions, formulaires, variables super globales, sessions, cookies
Niveau du public :	Débutant
Commentaires :	C7 est le document le plus court, sa structure se divise en 7 sections. Quelques images montrent des exemples de code.

Document :	**C8**
Titre :	*Création d'un site Web dynamique PHP*
Auteur :	M.Kirsch Pinheiro
Type de document :	diapositives
Année :	2017
Nb pages/diapositives :	81
En-têtes et/ou en-pieds :	Oui
Mots clés :	bases du langage, PHP, intégration, HTML, interprétation, variable, type, opérateurs, tableaux, POO, classes, objets, constructeurs, formulaires, GET, POST, instructions de contrôle, fonctions, conditions, boucles, importation de fichiers, bases de données, SQL, MySQL, PDO, MySQLi, sessions, cookies
Niveau du public :	Débutant
Commentaires :	C8 est divisé en 6 sections. Le document ayant été construit avec PowerPoint et des cadres se superposant les uns aux autres, les textes extraits automatiquement contiennent parfois des erreurs dans les mots (plusieurs mots sont fusionnés en un). Un exercice termine le cours.

Document :	**C9**
Titre :	*Base de donnée 2 : PHP and Mysql*
Auteur :	L.Dounas
Type de document :	diapositives
Année :	2018
Nb pages/diapositives :	131
En-têtes et/ou en-pieds :	
Mots clés :	SQL, web, PHP, installation, commentaires, types de données, opérateurs, date, constantes, tableaux, conditions, boucles, fonctions, intégration, HTML, chaînes de caractères, formulaires, GET, POST, portée des variables, super-globales, inclusion de fichiers, redirections automatiques, sessions, cookies, POO, calsse, objet, héritage, polymorphisme, constructeur, visibilité des méthodes, static, opérateur : :, abstraction, constantes de classe, final, SGBD, PhpMyAdmin, MySQL, MySQLi, PDO, erreurs, try catch, transactions, ACID, moteurs de stockage
Niveau du public :	Débutant
Commentaires :	C9 est divisé en 11 sections. Plusieurs exercices sont présents au sein du document, ainsi que des références à la fin.

4.3.2 Présentation succincte des documents

Comme indiqué dans le protocole d'évaluation en section 4.2, les documents C1 à C9 sont utilisés dans les trois premiers scénarios. Le document CJA correspond au document Java utilisé dans le scénario n°2. Le scénario n°3 s'appuie sur les documents C1 à C9 et C11 à C19. Tous ces documents sont en français.

	ID	**Format**	**Nb Pages**	**Auteur**	**Titre**
S1	C1	Diapositives	46	E.Coquery	Cours HTML/PHP
	C2	Diapositives	64	T.Lecroq	PHP
	C3	Texte	93	S.Rohaut	Cours PHP - Versions 4.x et 5.x
	C4	Diapositives	99	V.Pagé	Programmation Spécifique : Programmation Web
	C5	Texte	115	G.Rozsavolgyi	Cours PHP Accéléré
	C6	Texte	160	D.Gonzalez	PHP, une initiation
	C7	Diapositives	37	C.Escazut	Le Langage PHP, Ou comment concevoir des sites dynamiques
	C8	Diapositives	81	M.Kirsch Pinheiro	Création d'un site Web dynamique PHP
	C9	Diapositives	131	L.Dounas	Base de donnée 2 : PHP and Mysql
S2	CJA	Texte	88	A.Morelle	LE LANGAGE JAVA - Petit mémento de syntaxe & éléments de programmation
S3	C11	Texte	129	E.Vandeput	Développer une application avec PHP et MySQL
	C12	Diapositives	134	J.Gaulmin	Programmer en PHP
	C13	Diapositives	74	L.Pouilloux	Projets Web - L3STEP
	C14	Diapositives	99	T.Fressin	Développement web
	C15	Texte	113	D.Hadjadj	Initiation à la programmation de page web en PHP
	C16	Diapositives	137	V.Sans	Programmation Web : PHP
	C17	Texte	30	R.Mokadem	Cours introductif au PHP
	C18	Diapositives	82	V.Ricard	PHP (ET MYSQL)
	C19	Texte	159	B.Liaudet	PHP

Tab. 4.2 – Liste des 19 cours sélectionnés pour les scénarios n°1-2-3 portant sur PHP, ainsi que le cours de Java

Le scénario n°4 vise à vérifier si une amélioration de la qualité des documents en entrée propage également une amélioration des résultats jusqu'au graphe d'impact mutuel. Dans ce scénario, les 7 supports de cours au format texte travaillant sur PHP issus des scénarios n°1 et 3 sont réunis (C3, C5, C6, C11, C15, C17, C19), puis, le document C6 est *corrigé* en lui retirant certains chapitres ne concernant pas le cours ou pas directement (présentation de projets et « hors programme »). Afin d'évaluer l'impact de cette correction de façon plus générale, le cours de Java (CJA) est également introduit lors d'un test pour évaluer le nouvel écart. Ce scénario est donc particulièrement orienté validation fonctionnelle.

	ID	**Format**	**Nb Pages**	**Auteur**	**Titre**
S4	C3	Texte	93	S.Rohaut	Cours PHP - Versions 4.x et 5.x
	C5	Texte	115	G.Rozsavolgyi	Cours PHP Accéléré
	C6	Texte	160	D.Gonzalez	PHP, une initiation
	C11	Texte	129	E.Vandeput	Développer une application avec PHP et MySQL
	C15	Texte	113	D.Hadjadj	Initiation à la programmation de page web en PHP
	C17	Texte	30	R.Mokadem	Cours introductif au PHP
	C19	Texte	159	B.Liaudet	PHP
	CJA	*Texte*	*88*	*A.Morelle*	*LE LANGAGE JAVA - Petit mémento de syntaxe & éléments de programmation*

Tab. 4.3 – Liste des 7 documents au format texte long pour le scénario n°4 traitant de PHP

Le scénario n°5 vise cette fois à utiliser des documents en anglais traitant des Statecharts. Dans ce scénario précis, les documents A1 à A5 concernent des publications académiques, tandis que les documents C1 à C8 concernent des supports de cours et des publications de l'industrie ou grand public. Tous ces documents sont en anglais.

	ID	**Format**	**Nb Pages**	**Auteur**	**Titre**
S5	A1	Article	44	D.Harel	Statecharts : A visual formalism for complex systems
	A2	Article	17	D.Harel	On visual formalisms
	A3	Article	11	E.Kushnareva et al.	Modeling crisis management process from goals to scenarios
	A4	Chapitre	75	R.Keller	12. Finite-State Machines
	A5	Article	15	S.Van Mierlo et al.	Introduction to statecharts modeling, simulation, testing, and deployment
	C1	Page Web	1	E.Mogensen	Welcome to the world of Statecharts
	C2	Diapositives	63	M.A.Martínez Ibáñez	Statecharts
	C3	Diapositives	75	S.Van Mierlo et al.	An Introduction to Statecharts Modelling and Simulation
	C4	Diapositives	26	B.Franke	Embedded Systems Lecture 4 : Statecharts
	C5	Diapositives	17	H.Vangheluwe	Statecharts
	C6	Diapositives	18	M.Felici	Statechart Diagrams
	C7	Wikipedia	13	Wikipedia (17 mai 2020)	Finite-state machine
	C8	Diapositives	44	M.Di Natale	Statecharts (hierarchical FSMs)

Tab. 4.4 – Liste des 9 documents en anglais sélectionnés pour le scénario n°5 traitant des Statecharts

4.3.3 Validation structurelle

Nous présentons maintenant les résultats bruts pour la validation structurelle. Comme indiqué dans le protocole présenté en section 4.2, nous comparons tout d'abord le nombre de mots suite à l'étape de nettoyage des textes (PI.2), et le nombre de termes uniques suite à l'étape de filtrage des termes (PI.4). Puis nous comptons le nombre de termes uniques pour les quatre stratégies de binarisation avec plusieurs valeurs de β, et les proportions de « 0 » et de « 1 » avec ces stratégies et β. Enfin, nous présentons les clusters générés avec la stratégie haute et plusieurs valeurs de β pour les scénarios n°1 et n°5 uniquement.

Résultats du nettoyage des textes (PI.2) et du filtrage des termes (PI.4)

Les premières étapes de la méthode CREA sont appliquées aux cours. Les résultats de l'étape du nettoyage des textes (PI.2) sont résumés dans le tableau 4.5 pour les scénarios n°1-2-3, et dans le tableau 4.6 pour le scénario n°5. Les textes extraits des supports de cours sont analysés au fur et à mesure de la première phase. Ces tableaux indiquent le nombre de mots et de termes conservés au fil des étapes de nettoyage des textes (PI.2) et de filtrage des termes (PI.4) La différence entre les *mots* des deux premières étapes et les *termes* des deux dernières étapes provient du fait que c'est l'étape de désambiguïsation (PI.3) qui agrège certains mots en un seul terme (par exemple, « *base de données* » est comptabilisée comme 3 mots par l'étape de nettoyage des textes (PI.2), particulièrement car l'outil TreeTagger assigne des classes grammaticales aux mots, avant que BabelFy retrouve l'entité associée à cet ensemble de mots).

Pour les scénarios n°1-2-3, les résultats montrent des proportions de conservation des mots entre 76% et 91% à l'issue de la phase de nettoyage des textes par TreeTagger. Les classes sélectionnées dans TreeTagger permettent de conserver une majorité de mots dans les documents (effet attendu), et nous constatons qu'aucun document ne subit d'exclusion disproportionnée. Le scénario n°5 travaillant en anglais et sur des documents dont la nature varie beaucoup plus, ses proportions restent malgré tout assez proches : conservation des mots entre 69% et 95%. Le cas extrême de C5 s'explique par le fait qu'il s'agit d'une présentation avec très peu de termes et beaucoup de schémas : le format diapositives implique souvent de réduire le vocabulaire à quelques noms et verbes essentiels, ce qui correspond exactement à ce que nous cherchons à extraire. À l'inverse, C1 est une page web assez courte contenant beaucoup de mots des classes grammaticales exclues. Les proportions restent donc stables dans tous les scénarios, il n'y a pas d'écart majeur entre les documents.

Concernant la phase de filtrage des termes dans les scénarios n°1-2-3, on observe des taux entre 17% et 35% pour les proportions de conservation des termes. La quantité de termes conservés semble très faible en comparaison des quantités de mots conservés, mais la grande majorité de termes exclus ont des scores de désambiguïsation à 0, indiquant que BabelFy n'a pas jugé ces termes comme cohérents dans leur contexte.

En analysant de plus près les termes exclus, on s'aperçoit que les supports de cours au format texte ont éliminé beaucoup de termes inutiles (des noms propres, quelques en-têtes, ...), mais à l'inverse, dans les supports au format diapositives, des termes utiles ont été éliminés (tels que *web*, *css*, *tableau*). Bien que des termes utiles soient éliminés dans certains cas, une grande majorité d'entre eux (*php*, *client*, *serveur*, ...) sont conservés. Ce découpage est prévisible du fait du traitement automatique et des scores de cohérence à 0.

Dans le cas du scénario n°5 en anglais, certains documents ont des taux de rétention des termes assez élevés (jusqu'à 47%). Deux explications nous semblent plausibles : la taille réduite de certains documents (moins de 1.000 termes reconnus, dans des documents initialement plutôt courts) implique des écarts beaucoup plus grands à la moindre variation, mais surtout, BabelNet étant un réseau sémantique basé sur des sources accessibles en ligne et la langue anglaise étant majoritaire sur internet, il est beaucoup plus aisé de retrouver des concepts et entités nommées avec.

	ID	**Mots (PI.1)**	**Mots (PI.2)**	**%**	**Termes (PI.3)**	**Termes (PI.4)**	%
S1	C1	2.806	2.314	82%	1.288	270	21%
	C2	3.060	2.704	88%	990	275	28%
	C3	26.631	20.874	78%	11.244	2.620	23%
	C4	5.516	4.688	85%	2.139	532	25%
	C5	18.921	16.286	86%	5.936	1.530	26%
	C6	35.752	28.222	79%	13.278	2.667	20%
	C7	2.613	2.309	88%	1.199	283	24%
	C8	6.833	6.219	91%	1.744	355	20%
	C9	8.075	6.693	83%	3.132	781	25%
S2	CJA	23.025	17.781	77%	9.283	2.535	27%
S3	C11	30.517	23.384	77%	12.788	2.237	17%
	C12	14.978	12.524	84%	5.994	2.086	35%
	C13	2.055	1.798	87%	839	219	26%
	C14	3.951	3.486	88%	1.705	524	31%
	C15	34.317	27.021	79%	14.209	3.322	23%
	C16	7.997	7.279	91%	3.144	1.101	35%
	C17	10.836	8.186	76%	5.034	1.181	23%
	C18	3.467	3.089	89%	1.393	362	26%
	C19	27.038	21.396	79%	10.366	2.608	25%

Tab. 4.5 – Statistiques de filtrage de la phase I pour les scénarios n°1-2-3

	ID	Mots (PI.1)	Mots (PI.2)	%	Termes (PI.3)	Termes (PI.4)	%
S5	A1	14.855	10.565	71%	6.420	1.922	30%
	A2	10.900	7.822	72%	4.559	1.507	33%
	A3	3.506	2.583	74%	1.960	724	37%
	A4	17.698	13.377	76%	7.086	2.423	34%
	A5	6.542	4.759	78%	3.238	1.066	33%
	C1	5.204	3.574	69%	2.298	664	29%
	C2	2.486	1.835	74%	1.253	293	23%
	C3	959	848	88%	499	209	42%
	C4	1.560	1.229	79%	675	201	30%
	C5	786	743	95%	350	166	47%
	C6	1.306	970	74%	643	165	26%
	C7	4.535	3.601	79%	2.840	1.293	46%
	C8	1.829	1.419	78%	939	248	26%

Tab. 4.6 – Statistiques de filtrage de la phase I pour le scénario n°5

Quantités de termes uniques issus du filtrage des termes (PI.4)

Afin de mesurer la diversité des termes retenus par ces premiers traitements, nous effectuons une analyse des termes conservés à l'issue de l'étape de filtrage des termes (PI.4). Le tableau 4.7 illustre cette analyse pour les scénarios n°1-2-3, et permet de constater que certains termes sont à la fois supprimés et conservés (les *termes communs* appartiennent simultanément aux catégories *retenus* et *exclus*). Par exemple, dans le cours C1, il s'agit de 7 termes « *lien* » (bn:00045513n), « *langage* » (bn:00049910n), « *texte* » (bn:00076732n), « *web* » (bn:00080778n), « *choisir* » (bn:00084931v), « *spécifier* » (bn:00086464v), « *dernières* » (bn:00105773a). Ce phénomène provient du fait que nous utilisons le score de cohérence que BabelFy alloue à chaque terme selon le contexte détecté. Des termes trop techniques peuvent devenir incohérents dans des phrases génériques, et inversement, des termes trop génériques peuvent devenir incohérents dans des phrases techniques. On notera également que la proportion de termes communs aux listes de termes retenus et de termes exclus est un peu plus élevée (entre $4,6\%$ et $6,3\%$) dans le cas de documents au format texte (C3, C5, C6), par rapport au format diapositives (entre $1,2\%$ et $2,4\%$). Comme nous l'avions déjà fait remarquer, ce résultat est prévisible du fait que les diapositives utilisent généralement un vocabulaire beaucoup plus restreint et concentré sur les notions abordées. Concernant le scénario n°2 et le cours de Java (*CJA*) en particulier, plusieurs termes techniques intéressants ont tout de même été exclus (tels que « *api* », « *mise à jour* », « *classe java* », « *vector* »), mais cela correspond statistiquement aux résultats des 9 précédents supports de cours. Le scénario n°3 confirme également ces statistiques avec les 9 documents supplémentaires.

	ID	Termes uniques (PI.3)	Termes uniques retenus (PI.4)	%	Termes uniques exclus (PI.4)	%	Termes communs (PI.4)	%
S1	C1	360	84	23%	283	79%	7	1,9%
	C2	375	97	26%	283	75%	5	1,3%
	C3	1.597	501	31%	1.181	74%	85	5,3%
	C4	614	151	25%	475	77%	12	2,0%
	C5	1.269	373	29%	955	75%	59	4,6%
	C6	2.032	609	30%	1.551	76%	128	6,3%
	C7	259	70	27%	192	74%	3	1,2%
	C8	339	83	24%	264	78%	8	2,4%
	C9	772	239	31%	550	71%	17	2,2%
S2	CJA	1.590	513	32%	1.169	74%	92	5,7%
S3	C11	1.734	500	29%	1.334	77%	100	5,8%
	C12	1.068	328	31%	779	73%	39	3,7%
	C13	363	111	31%	258	71%	6	1,7%
	C14	620	168	27%	461	74%	9	1,5%
	C15	1.803	558	31%	1.358	75%	113	6,3%
	C16	587	177	30%	420	72%	10	1,7%
	C17	956	318	33%	669	70%	31	3,2%
	C18	472	133	28%	346	73%	7	1,5%
	C19	1.268	401	32%	948	75%	81	6,4%

Tab. 4.7 – Statistiques des termes uniques filtrés pour les scénarios n°1-2-3

Résultats des stratégies de binarisation (PII.1.c) des scénarios n°1 et n°3

À l'issue de la première phase de pré-traitement sémantique, la normalisation est effectuée en calculant des proportions de termes dans chaque cours, puis la transposition afin de passer de cours décrits par des termes à des termes présents dans des cours. Le calcul des stratégies de binarisation (voir en sous-section 3.3.1) permet ensuite d'obtenir plusieurs visions des cours insérés et des termes contenus. Les quantités de termes uniques générées par chacune des stratégies pour les scénarios n°1 et n°3 sont présentées dans les tableaux 4.8 et 4.9 en suivant un pas de 0.25 pour les évolutions de β. Ces tableaux permettent d'observer dans quelle mesure les termes sont conservés et éliminés par chacune des stratégies et β, afin de sélectionner la stratégie et le β les plus appropriés pour les traitements suivants.

On remarque tout d'abord que les quantités de termes varient très peu entre les stratégies *Directe* et *Moyenne*, quelle que soit la valeur de β. Les quantités de termes des stratégies *Haute* et *Basse* évoluent de façon similaire, mais la stratégie *Basse* perd de plus en plus de termes avec l'accroissement de β. On s'aperçoit que les termes retenus en $\beta = 0.00$ et $\beta = 0.25$ pour les stratégies *Basse* et *Haute* sont strictement les mêmes (seules les valeurs de la matrice binaire changent comme expliqué par la suite).

β	Stratégies			
	Directe	**Moyenne**	**Haute**	**Basse**
0.00	1.279	1.206	74	74
0.25		1.217	74	74
0.50		1.240	74	54
0.75		1.265	74	25
1.00		1.279	60	7

Tab. 4.8 – Quantités de termes uniques des stratégies et Bêtas pour le scénario n°1 (référence)

β	Stratégies			
	Directe	**Moyenne**	**Haute**	**Basse**
0.00	1.951	1.835	116	116
0.25		1.848	116	116
0.50		1.908	115	60
0.75		1.940	115	23
1.00		1.951	106	6

Tab. 4.9 – Quantités de termes uniques des stratégies et Bêtas pour le scénario n°3

Les tableaux 4.10 et 4.11 présentent cette fois les quantités de « 0 » et de « 1 » dans chacune des matrices générées par les stratégies pour des β évoluant par paliers de 0.25. On y voit que les matrices des stratégies *Directe*, *Moyenne*, et *Haute* sont relativement stables dans les proportions de « 0 » et de « 1 » (environ 80 $\sim$ 85% de « 0 » et 20 $\sim$ 15% de « 1 » pour le scénario n°1, et environ 85 $\sim$ 90% de « 0 » et 15 $\sim$ 10% de « 1 » pour le scénario n°3). La stratégie *Basse* produit des matrices avec beaucoup plus de « 1 » pour des valeurs de β faible ($\sim$ 40% pour $\beta = 0.00$ dans les deux scénarios) et s'approche progressivement de la tendance ($\sim$ 25% pour $\beta = 1.00$ dans les deux scénarios). Cette instabilité semble liée à la quantité de termes uniques en forte baisse selon la valeur de β. Une comparaison des termes uniques entre la stratégie *Haute* $\beta = 0.00$ et $\beta = 1.00$ dans le scénario n°3 montre que les termes retirés sont tous du bruit, excepté l'un d'entre eux (« *symfony* ») qui est plutôt connu et utile, mais est effectivement absent de plusieurs supports de cours anciens (le framework symfony étant publié pour la première fois courant 2005, sa démocratisation a nécessité un certain temps).

Cette analyse vient compléter la précédente en analysant non pas seulement la quantité de termes retenus, mais la quantité et la qualité des concepts formels pouvant être produits. Comme indiqué dans le protocole, un excès de « 0 » ne formera quasiment aucun concept formel (donc aucune connaissance ne pourra être exploitée), et un excès de « 1 » formera quasiment toutes les combinaisons possibles de termes et de documents sans mettre en avant les combinaisons les plus pertinentes pour la réutilisation.

Qualitativement parlant, dans le cadre des stratégies de binarisation, utiliser la stratégie *Directe* pour avoir une vision d'ensemble sera très similaire à la stratégie *Moyenne* quelle que soit la valeur de β. On pourra préférer la stratégie *Haute* à la

stratégie *Basse* de par sa stabilité, mais aussi par la théorie impliquant que seules les hautes fréquences sont conservées, donc les termes les plus fréquents dans certains cours. La stratégie *Directe* semble donc adaptée pour construire le graphe d'impact mutuel, et la stratégie *Haute* pour les clusters de termes.

β	Stratégies							
	Directe		**Moyenne**		**Haute**		**Basse**	
	0	1	0	1	0	1	0	1
0.00	9.304	2.207	9.036	1.818	549	117	394	272
0.25			9.112	1.841	556	110	410	256
0.50			9.194	1.966	569	97	342	144
0.75			9.314	2.062	574	92	172	53
1.00			9.388	2.123	471	69	48	15

β	Stratégies							
	Directe		**Moyenne**		**Haute**		**Basse**	
	0	1	0	1	0	1	0	1
0.00	81%	19%	83%	17%	82%	18%	59%	41%
0.25			83%	17%	83%	17%	62%	38%
0.50			82%	18%	85%	15%	70%	30%
0.75			82%	18%	86%	14%	76%	24%
1.00			82%	18%	84%	16%	76%	24%

Tab. 4.10 – Quantités et proportions de « 0 » et de « 1 » des stratégies et Bêtas pour le scénario n°1 (référence)

β	Stratégies							
	Directe		**Moyenne**		**Haute**		**Basse**	
	0	1	0	1	0	1	0	1
0.00	30.217	4.901	29.213	3.817	1.834	254	1.258	830
0.25			29.397	3.867	1.857	231	1.285	803
0.50			29.961	4.383	1.854	216	778	302
0.75			30.291	4.629	1.881	189	331	83
1.00			30.401	4.717	1.743	165	53	19

β	Stratégies							
	Directe		**Moyenne**		**Haute**		**Basse**	
	0	1	0	1	0	1	0	1
0.00	86%	14%	88%	12%	88%	12%	60%	40%
0.25			88%	12%	89%	11%	62%	38%
0.50			87%	13%	90%	10%	72%	28%
0.75			87%	13%	91%	9%	80%	20%
1.00			87%	13%	91%	9%	74%	26%

Tab. 4.11 – Quantités et proportions de « 0 » et de « 1 » des stratégies et Bêtas pour le scénario n°3

Résultats des clusters (PII.3) des scénarios n°1 et n°5

L'étape finale de clustering s'appuie sur la matrice de similarité conceptuelle (vue en sous-section 3.3.1). L'implémentation choisie utilisant *scikit-learn* [86] et le module *scipy.cluster.hierarchy* de *sci-py* [119], il est nécessaire de transformer la matrice de similarité en matrice de distance, ou à minima en matrice de dissimilarité (voir en sous-section 3.3.3). Étant donné que nous voulons créer une structure haut niveau pour un syllabus, nous nous orientons vers les résultats produits par la stratégie *Haute* afin de n'utiliser que les termes les plus fréquents et reconnus comme étant des notions générales. Nous utilisons donc la matrice de similarité issue de la stratégie *Haute*, que nous transformons en matrice de dissimilarité. Une fois ce traitement préparatoire réalisé, nous demandons à l'algorithme de *classification ascendante hiérarchique* de générer un ensemble de 8 clusters afin de pouvoir préparer 8 séances de cours.

Les figures 4.3 et 4.4 illustrent les 8 clusters générés pour les scénarios n°1 et n°5 à partir des matrices de similarité conceptuelle obtenue avec la stratégie *Haute* pour chaque valeur de β de 0.00 à 1.00 par pas de 0.25.
Les tableaux 4.12 et 4.13 listent les tailles des plus grands clusters pour chaque valeur de β afin d'analyser les variations.

La lecture des clusters du scénario n°1 se fait en associant les termes contenus dans chacun d'entre eux. Par exemple en $\beta = 1.00$, le cluster n°6 contient : *base de données, insert, varchar, null*, ce qui correspond à des termes typiquement du monde de la base de données (*insert* servant d'ordre dans les requêtes SQL, *varchar* étant un type pour les colonnes, *null* étant la gestion des colonnes ne contenant pas de valeur). Cet exemple avec *null* illustre également que le sens d'un terme peut varier selon son contexte. Ici, il prend tout son sens pour de la base de données, mais dans le contexte de PHP, il existe également plusieurs méthodes pour gérer l'absence de valeur dans une variable, l'absence de déclaration d'une variable, etc... dont *null* est une de ces méthodes. On comprend qu'un cluster devient un contexte contribuant à orienter les choix de l'enseignant, ou à lui proposer une grille de lecture possible des termes.

En répétant cette lecture des clusters du scénario n°1, nous avons estimé que les clusters en $\beta = 1.00$ sont ceux de meilleure qualité. En effet, les termes provoquant du bruit ont été automatiquement effacés par rapport aux autres valeurs de β, mais aussi, les clusters prennent beaucoup plus de sens à la lecture.

1. *php, code, fois, post, jour, foreach, cle, classe, class, mysqli* : on peut parler de l'accès aux valeurs des variables transmises en *post* avec des *foreach*/*fois* sur les *cle*, et commencer à parler du connecteur *mysqli*.
2. *page web, navigateur, serveur web, texte, concerner, délimiter, utilisateur, associer, personne, machine, mysql* : un ensemble très large de notions sont abordées (*serveur web*, *machine*, *mysql*) tout en parlant de notions côté *utilisateur* (*page web*, *navigateur*, *texte*). On peut penser à une séance d'introduction donnant une vision large de comment fonctionne le web, et comment les utilisateurs y naviguent.

3. *url, langage, case, fermeture, session, chaîne, entête, avoir accès* : la gestion des *sessions* (et leur *fermeture*) permettant d'*avoir accès* à des données ou non, puis, les *url* (et le passage des valeurs au format *texte* dans les *entêtes* http).
4. *fichier, commentaire, case à cocher, interpréter, côté serveur, serveur, côté client* : le web fonctionne avec un *côté serveur* et un *côté client*. Le *serveur interprète* les *fichiers* PHP, et ignore les *commentaires*.
5. *typage, mot, moteur, affiche, transaction, visiteur* : pour gérer les *transactions* sur un site, y compris quand les *visiteurs* créent un panier, il faut choisir le bon *moteur* de stockage sur la base de données.
6. *base de données, insert, varchar, null* : une séance dédiée à la *base de données* n'est pas du tout à exclure dans le cadre d'un cours de développement web en PHP. On peut y aborder tous les outils pour *insérer* des données, dont certaines colonnes seront des *varchar* dont il faut absolument préciser la taille, et d'autres *null* seront optionnelles.
7. *xml, configuration, composer, doctype* : le html est lié au format *xml* qui impose de déclarer le *doctype* du document.
8. *donnée, text, méthode post, programmation, site, langage de script, list, méthode, timestamp, file* : la séance parle du fait que PHP est un *langage de script* dont l'usage habituel est la *programmation* de *site*. On peut y aborder les *list*, expliquer ce qu'est un *timestamp* sur des *file*, et démarrer le fonctionnement de la *méthode post* pour transférer des *données*.

Ces propositions d'interprétation varient d'un individu à l'autre, étant donné qu'elles s'appuient sur l'expérience de chacun, c'est-à-dire les connaissances tacites. C'est d'ailleurs pour cela qu'il s'agit d'un KIP dont la créativité individuelle est importante, et qui exige une certaines expérience de la part de l'enseignant sur le sujet étudié. Les 8 séances proposées sous formes de clusters permettent donc à un enseignant d'avoir un aperçu des structures existantes, tout en lui proposant une organisation possible de notions auxquelles il ajoutera ses propres liens logiques. Ces clusters n'étant pas ordonnancés temporellement, l'enseignant doit les réordonner (nous proposons néanmoins en conclusion en section 5.2.3 une méthode d'ordonnancement encore expérimentale).

Une vérification succincte des clusters générés par la stratégie *Basse* avec $\beta = 0.00$ a été réalisée, mais les résultats confirment que les clusters n'ont que très peu d'intérêt pour la construction d'un cours en comparaison de ceux de la stratégie *Haute*. Bien que les termes retenus soient similaires à ceux de la stratégie haute, leur organisation dans les clusters diffère et est beaucoup moins pertinente (il nous était beaucoup plus difficile de produire une interprétation pour la plupart des clusters). De plus, comme indiqué précédemment en 4.3.3), plus le β est élevé, moins il y a de termes, rendant impossible la construction de 8 clusters dans certains cas (seulement 7 termes étant disponibles dans le cas $\beta = 1.00$).

Dans le scénario n°5, les 13 documents fournis en entrée génèrent 111 termes pour des β de 0.00 à 0.75 inclus, puis 83 pour $\beta = 1.00$. Tout d'abord, la quantité de documents fournis en entrée semble avoir un impact sur la quantité de termes présents

dans les clusters (ce qui était partiellement indiqué par le nombre de termes uniques vu en sous-section 4.3.3). Ensuite, on remarque que certains clusters varient peu à chaque nouveau pas de β. Par exemple, le cluster n°2 reste le même de $\beta = 0.00$ à 0.75, il subit plusieurs modification et devient le cluster n°1 à $\beta = 1.00$. De même pour le cluster n°1 qui varie légèrement à chaque pas, et devient le cluster n°2 en $\beta = 1.00$ en perdant cette fois beaucoup de termes. Concernant la qualité, on constate que les clusters contiennent beaucoup plus de termes issus des exemples classiques utilisés dans le domaine des statecharts (*traffic light*, *light*, *city*, *new york*, *police*, *automobiles*, *sec*, *second*, *timers*, ...) par rapport au scénario n°1. Ce point n'est pas totalement négatif, car il s'agit effectivement d'exemples incontournables pour présenter le sujet. Plus le β augmente, plus les clusters deviennent courts et donc plus simples à lire. Le nombre de termes représentant des exemples se réduit (un même exemple devient illustré non plus par plusieurs termes mais par un seul) et permet de revenir à une lecture plus abstraite des concepts et de quelques exemples typiques. L'usage des clusters issus du $\beta = 1.00$ est donc plus approprié pour une lecture de haut niveau afin de former un syllabus.

$\beta = 0.00$	20 termes	
$\beta = 0.25$	18 termes	
$\beta = 0.50$	12 termes	(2 clusters)
$\beta = 0.75$	13 termes	
$\beta = 1.00$	11 termes	

Tab. 4.12 – Quantités de termes dans le(s) plus grand(s) cluster(s) pour le scénario n°1 (référence)

$\beta = 0.00$	24 termes	(2 clusters)
$\beta = 0.25$	33 termes	
$\beta = 0.50$	28 termes	
$\beta = 0.75$	29 termes	
$\beta = 1.00$	24 termes	

Tab. 4.13 – Quantités de termes dans le(s) plus grand(s) cluster(s) pour le scénario n°5

Stratégie Haute (β = 0,00)

1	url	case	fermeture	france	session	entête	valeur	généralité	autre	avoir accès
2	navigateur	fichier	client	base de donnée	donnée	php	délimiter	post	nombre	langage
	zones	personne	méthode post	mysql	site	langage de script	foreach	list	timestamp	chaîne
3	page web	serveur web	texte	concerner	utilisateur	associer	machine	salaire		
4	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client	pointeur	casse		
5	fois	jour	cle	opérateur	classe	mysqli	bienvenue	paris		
6	text	typage	mot	moteur	affiche	transaction	visiteur			
7	code	programmation	méthode	echo	files	class				
8	xml	configuration	insert	varchar	composer	doctype	null			

Stratégie Haute (β = 0,25)

1	url	langage	case	fermeture	france	session	entête	valeur	généralité	autre	avoir accès
2	page web	navigateur	serveur web	texte	concerner	utilisateur	associer	machine	salaire		
3	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client	pointeur	casse			
4	fois	jour	cle	opérateur	classe	mysqli	bienvenue	paris			
5	fichier	client	base de donnée	donnée	php	délimiter	post	nombre	text	zones	personne
	méthode post	mysql	site	langage de script	typage	foreach	list				
6	code	programmation	méthode	timestamp	chaîne	echo	files	class			
7	mot	moteur	affiche	transaction	visiteur						
8	xml	configuration	insert	varchar	composer	doctype	null				

Stratégie Haute (β = 0,50)

1	code	fois	post	jour	foreach	cle	opérateur	classe	class	mysqli	bienvenue	paris
2	url	langage	case	fermeture	france	session	entête	valeur	généralité	autre	avoir accès	
3	page web	navigateur	serveur web	texte	concerner	utilisateur	associer	machine	mysql	salaire		
4	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client	pointeur	casse				
5	programmation	méthode	timestamp	chaîne	echo	files						
6	base de donnée	méthode post	list	xml	configuration	insert	varchar	composer	doctype	null		
7	mot	moteur	affiche	transaction	visiteur							
8	fichier	client	donnée	php	délimiter	nombre	text	zones	personne	site	langage de script	typage

Stratégie Haute (β = 0,75)

1	php	code	fois	post	jour	foreach	cle	opérateur	classe	class	mysqli	bienvenue	paris
2	url	langage	case	fermeture	france	session	entête	valeur	généralité	autre	avoir accès		
3	page web	navigateur	serveur web	texte	concerner	utilisateur	associer	machine	mysql	salaire			
4	fichier	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client	pointeur	casse				
5	typage	mot	moteur	affiche	transaction	visiteur							
6	programmation	méthode	timestamp	chaîne	echo	files							
7	client	donnée	délimiter	nombre	text	zones	personne	méthode post	site	langage de script	list		
8	base de donnée	xml	configuration	insert	varchar	composer	doctype	null					

Stratégie Haute (β = 1,00)

1	php	code	fois	post	jour	foreach	cle	classe	class	mysqli	
2	page web	navigateur	serveur web	texte	concerner	délimiter	utilisateur	associer	personne	machine	mysql
3	url	langage	case	fermeture	session	chaîne	entête	avoir accès			
4	fichier	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client				
5	typage	mot	moteur	affiche	transaction	visiteur					
6	base de donnée	insert	varchar	null							
7	xml	configuration	composer	doctype							
8	donnée	text	méthode post	programmation	site	langage de script	list	méthode	timestamp	files	

Fig. 4.3 – Clusters issus de la stratégie *Haute* pour β de 0.00 à 1.00 par pas de 0.25 pour le scénario n°1 référence

Stratégie Haute (β = 0,00)

1	science	computer	statecharts	systems	state	diagrams	behavior	example	kind	way	time
	number	state transition	triggering	possible	semantics	automata	history	effect	simultaneous	light	definition
	police	traffic light									
2	david harel	units	determinism	syntax	formal syntax	operational semantics	higraphs	large number	interaction diagram	deterministic finite state automata	fsa
3	automobiles	hierarchical	defined	types	multiple	stateflow	inner	deep	fsms	state actions	
4	language	simulation	nature	workflows	plant	code	autonomous	red light			
5	complex	problem	information	concept	sec	require	second	merely	substates	entity	composite
	transition state	composite states	simple	atomic	activity diagrams	basic					
6	occurring	america	possibility	conditional	condition c	logical value	summary				
7	communication	reactive systems	tool	orthogonality	broadcast	c d	timers	current	include	superstate	
8	mathematics	state machine	complex system	networks	set	output	main	achieved	application	reg	subset
	management	finite	new york	automata theory	sequence	goal	scenario	city	strategy	transducer	bit
	gate	flip-flop									

Stratégie Haute (β = 0,25)

1	science	computer	statecharts	systems	mathematics	state	state machine	diagrams	communication	behavior	example
	kind	way	time	set	output	tool	number	state transition	triggering	orthogonality	possible
	semantics	automata	history	effect	simultaneous	light	second	timers	definition	police	traffic light
2	david harel	units	determinism	syntax	formal syntax	operational semantics	higraphs	large number	interaction diagram	deterministic finite state automata	fsa
3	automobiles	hierarchical	defined	types	multiple	stateflow	inner	deep	fsms	state actions	
4	language	simulation	nature	workflows	plant	code	autonomous	red light			
5	complex	problem	information	concept	sec	require	merely	substates	entity	composite	transition state
	composite states	simple	atomic	activity diagrams	basic						
6	occurring	america	possibility	conditional	condition c	logical value	summary				
7	reactive systems	broadcast	c d	current	include	superstate					
8	complex system	networks	main	achieved	application	reg	subset	management	finite	new york	automata theory
	sequence	goal	scenario	city	strategy	transducer	bit	gate	flip-flop		

Stratégie Haute (β = 0,50)

1	science	computer	statecharts	systems	mathematics	state	state machine	diagrams	behavior	example	kind
	way	set	number	state transition	triggering	orthogonality	possible	semantics	automata	history	effect
	simultaneous	light	second	timers	definition	police					
2	david harel	units	determinism	syntax	formal syntax	operational semantics	higraphs	large number	interaction diagram	deterministic finite state automata	fsa
3	automobiles	hierarchical	defined	types	multiple	stateflow	inner	deep	fsms	state actions	
4	language	simulation	nature	workflows	plant	code	autonomous	red light			
5	communication	reactive systems	time	broadcast	c d	current	include	superstate			
6	complex	problem	information	concept	sec	require	merely	substates	entity	composite	transition state
	composite states	simple	atomic	activity diagrams	basic						
7	occurring	america	possibility	conditional	condition c	logical value	summary				
8	complex system	networks	output	tool	main	achieved	application	reg	subset	management	finite
	new york	automata theory	sequence	goal	scenario	city	strategy	transducer	traffic light	bit	gate
	flip-flop										

Stratégie Haute (β = 0,75)

1	science	computer	statecharts	systems	mathematics	state	state machine	behavior	problem	example	kind
	set	number	state transition	triggering	orthogonality	possible	semantics	concept	automata	history	effect
	second	timers	definition	transition state	police						
2	david harel	units	determinism	syntax	formal syntax	operational semantics	higraphs	large number	interaction diagram	deterministic finite state automata	fsa
3	automobiles	hierarchical	defined	types	multiple	stateflow	inner	deep	fsms	state actions	
4	diagrams	complex	information	simultaneous	substates	entity	composite	composite states	activity diagrams	basic	
5	language	simulation	nature	light	workflows	plant	code	autonomous	red light		
6	communication	reactive systems	time	broadcast	c d	current	include	superstate			
7	occurring	america	possibility	conditional	condition c	logical value	summary				
8	complex system	networks	way	output	tool	main	achieved	sec	require	application	merely
	reg	subset	management	finite	new york	automata theory	sequence	goal	scenario	city	strategy
	transducer	traffic light	bit	gate	flip-flop	simple	atomic				

Stratégie Haute (β = 1,00)

1	david harel	state	units	determinism	syntax	formal syntax	higraphs	large number	interaction diagram	deterministic finite state automata	
2	science	computer	systems	example	number	state transition	possible	semantics	automata	effect	definition
3	communication	reactive systems	time	broadcast	c d	history	current				
4	statecharts	state machine	problem	orthogonality	concept	sec	require	second	merely	simple	
5	diagrams	complex	information	simultaneous	substates	entity	composite	composite states			
6	language	simulation	nature	light	workflows	code	autonomous				
7	automobiles	hierarchical	defined	types	multiple	stateflow					
8	mathematics	complex system	networks	way	set	output	tool	triggering	main	occurring	america
	achieved	application	possibility	subset	management	finite	new york	automata theory	sequence	strategy	traffic light
	bit	summary									

FIG. 4.4 – Clusters issus de la stratégie *Haute* pour β de 0.00 à 1.00 par pas de 0.25 pour le scénario n°5

4.3.4 Validation fonctionnelle

Nous présentons maintenant les résultats bruts pour la validation fonctionnelle. Comme indiqué dans le protocole présenté en section 4.2, nous présentons tout d'abord les graphes d'impact mutuel formés avec les différents scénarios, puis les clusters de termes pour chacun des scénarios. Le scénario n°2 visant à vérifier la résistance au bruit, et le scénario n°4 visant à comparer l'écart en cas de corrections, nous ne nous intéressons qu'aux graphes d'impact mutuel dans ces cas. Les réponses de 5 informaticiens sont également présentées : les clusters qu'ils ont tout d'abord construits, puis leurs avis concernant les clusters produits par le cas de référence.

Graphes d'impact mutuel (PII.2)

Le graphe d'impact mutuel permet à un enseignant d'obtenir une vue d'ensemble des supports de cours insérés et leurs termes afin de disposer d'une carte des notions les plus récurrentes et des documents partageant le plus ces notions. La stratégie *Directe* est employée pour générer le treillis de Galois (voir sous-section 3.3.1) étant donné qu'elle conserve l'ensemble des termes présents et permet donc d'avoir la vue la plus complète possible. Pour rappel, le treillis généré permet de construire la matrice d'impact mutuel (voir sous-section 3.3.1) qui est illustrée par le graphe d'impact mutuel. Le graphe d'impact mutuel est construit sur Gephi [5] avec l'algorithme de spatialisation *Force Atlas* en insérant une *force de répulsion* entre *1.000* et *15.000* (selon le nombre de supports insérés), ainsi qu'une *force d'attraction* entre *5* et *10* (pour la même raison). Les figures 4.5 et 4.6 illustrent le graphe généré et la répartition concentrique des termes selon leur degré pour le scénario n°1. Les cercles les plus extérieurs contiennent les termes connectés à peu de cours, à l'inverse, l'ensemble central en rouge contient les termes connectés aux 9 cours.

Dans le scénario n°1, en observant de plus près l'ensemble central on peut y lire les termes *post*, *méthode post*, *nombre*, *langage*, *donnée*, *fichier*, *code*, *php*, *navigateur*, *site*. Dans le cadre de supports de cours abordant le développement web en PHP, ces termes sont très pertinents. Seuls « *langage* » et éventuellement « *nombre* » semblent moins techniques, mais peuvent tout de même s'insérer dans le cadre de la gestion *multi-lingue* d'un site ou du *langage de programmation* PHP, ainsi que du type de données *entier* ou *integer*. Les termes en orange (donc connectés à 8 des 9 cours) sont eux aussi adaptés : « *base de données* », « *foreach* », « *méthode* ». En jaune (les termes connectés à 7 cours), on retrouve également des termes pertinents : « *serveur web* », « *client* », « *langage de script* ». De même sur les termes en vert (connectés à 6 cours) : « *mysql* », « *session* », « *switch* ». Plus on s'éloigne de l'ensemble central, plus on retrouve de termes peu adaptés à la construction d'un syllabus, voire, n'ont aucun sens par rapport au sujet. Typiquement, les nœuds les plus éloignés contiennent le bruit non effacé par les précédentes étapes de nettoyage.

En comparant ensemble central avec les termes uniques retenus dans la matrice générée par la stratégie *Haute*, on remarque que ces termes sont pour la grande majorité connectés à au moins 5 cours. Plusieurs termes sont néanmoins absents, mais ceux-ci peuvent soit être sous-entendus par d'autres termes présents (« *sql* » pouvant être

retrouvé via « *mysql* » par hyponymie, ou encore « *constructeur* » pouvant être retrouvé via « *classe* » par hyperonymie), soit sont des notions très spécifiques (« *switch* » et d'autres termes liés au langage de développement lui-même, donc des notions peu adaptées à une vision d'ensemble d'un cours). La stratégie *Haute* reflète cet ensemble central et les principales notions abordées dans les différents supports de cours.

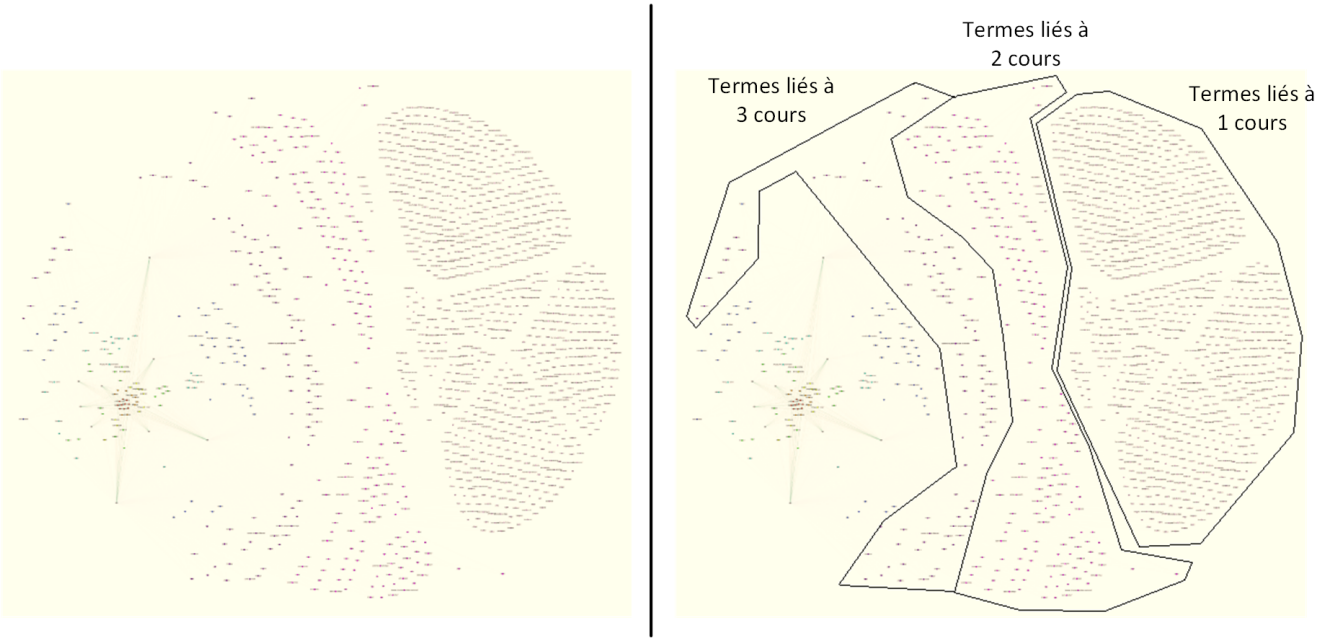


Fig. 4.5 – Graphe d'impact mutuel des 9 cours [scénario n°1] : Vision d'ensemble éloignée et annotations

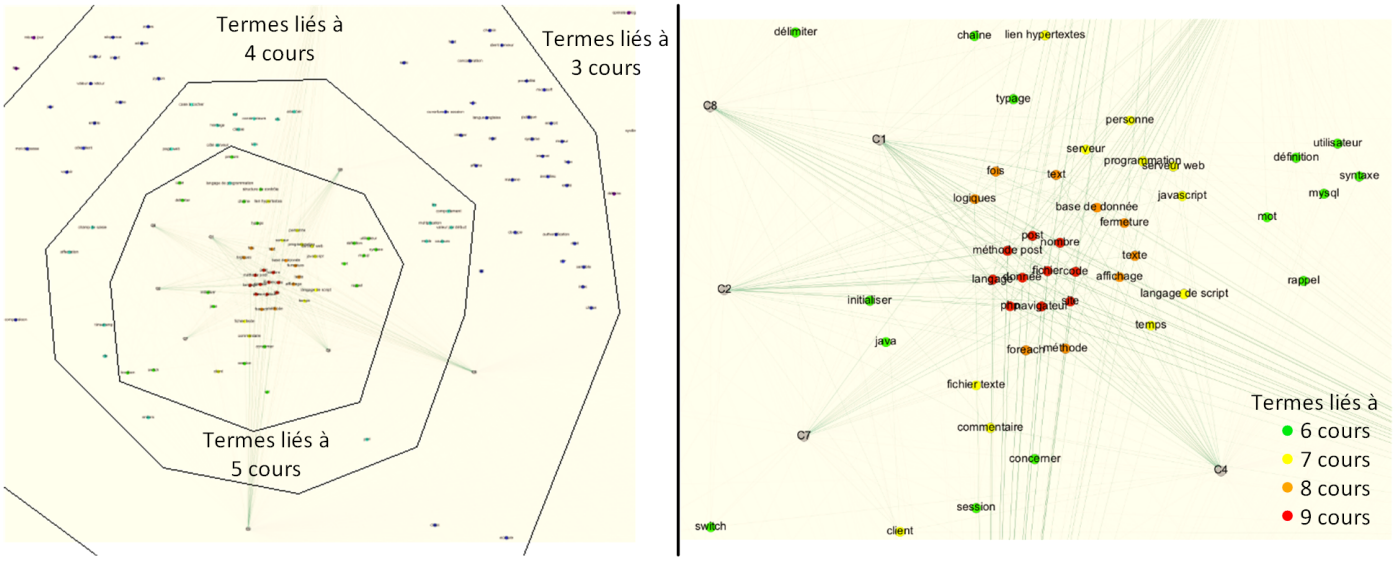


Fig. 4.6 – Graphe d'impact mutuel des 9 cours [scénario n°1] : Vision d'ensemble rapprochée et annotations

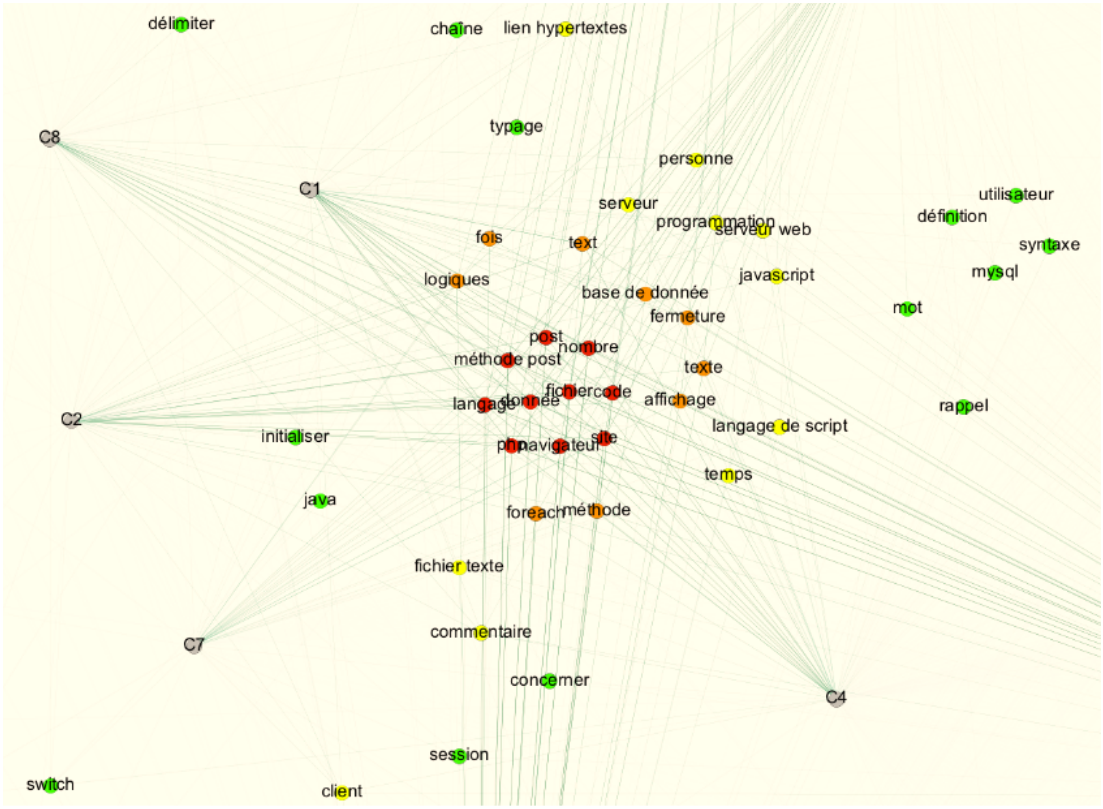


Fig. 4.7 – Graphe d'impact mutuel des 9 cours [scénario n°1] : Zoom sur l'ensemble central

L'ensemble central correspondant à ce qu'un enseignant s'attend à extraire de supports de cours de PHP, celui-ci peut évaluer la qualité des supports (quelles notions souhaite-t-il retrouver ou éviter, quelle distance maximale entre les supports accepte-t-il, ...). Une analyse sur les supports de cours est illustrée par la figure 4.8. On observe surtout que les supports au format texte sont les plus éloignés de l'ensemble central. Les supports au format diapositives sont au contraire plus rapprochés de l'ensemble central. Comme précédemment indiqué, cette différence peut s'expliquer par la plus forte quantité de vocabulaire employée dans les supports au format texte, par rapport au format diapositives. Les supports au format texte disposant de beaucoup plus de termes uniques, ces derniers forment de grands ensembles dédiés à chacun des supports. Ces ensembles étaient déjà sous-entendus dans le tableau 4.7 grâce aux stratégies filtrant de nombreux termes selon leurs fréquences, et sont maintenant visibles sur le graphe d'impact mutuel.

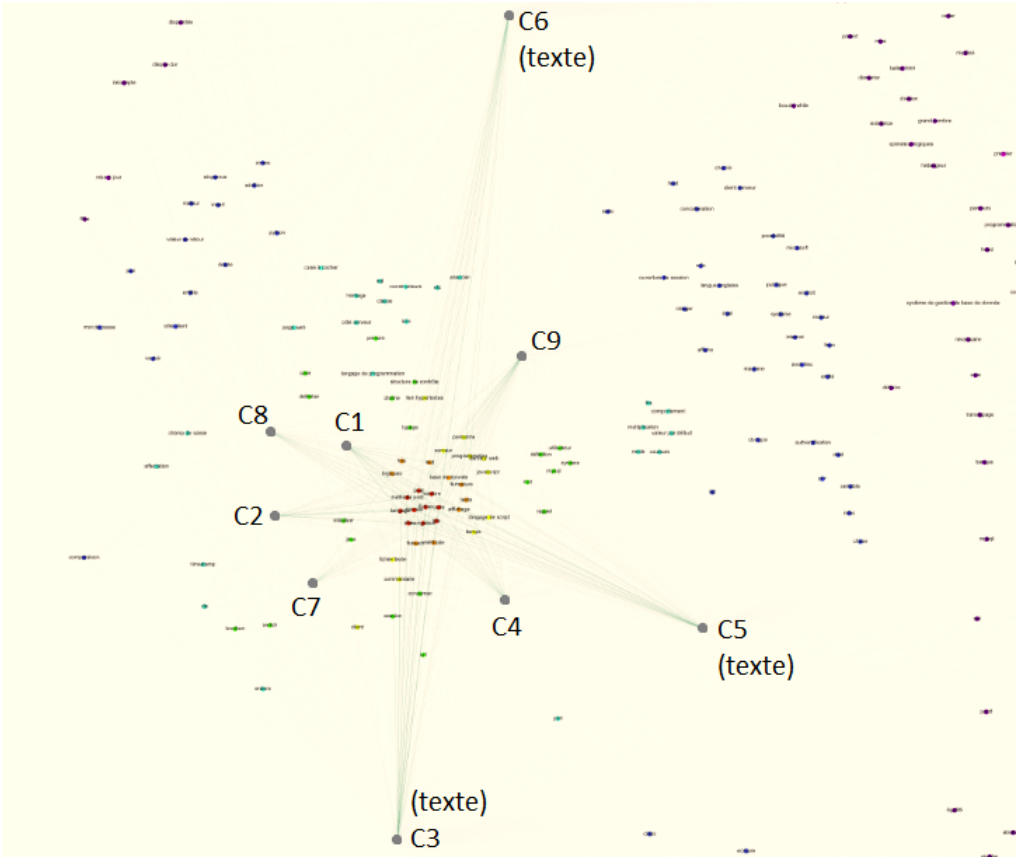


FIG. 4.8 – Graphe d'impact mutuel des 9 cours [scénario n°1] : Zoom sur le positionnement des cours

Le scénario n°2 vise à insérer un support de cours traitant d'un autre sujet. Nous nous intéressons donc aux variations entre les graphiques du scénario n°1 et celui-ci. La figure 4.9 illustre le graphe d'impact mutuel dans son ensemble. La figure 4.10 illustre l'ensemble central. Enfin, la figure 4.11 illustre le positionnement des documents.

Dans l'ensemble central, les termes les plus fréquents (en rouge) sont maintenant *donnée*, *code*, *nombre*, *fichier*, *langage*, *méthode post*, *site*, *navigateur*, *post*. En comparant avec la liste précédente (figure 4.7), on peut voir l'unique différence suivante : *donnée*, *code*, *nombre*, *fichier*, *langage*, *méthode post*, *site*, *navigateur*, *post*, ~~*php*~~. Celle-ci est confirmée par le contenu du cours de Java abordant les notions sus-citées, excepté PHP. Le terme *php* est maintenant coloré en orange (donc relié à tous les cours, sauf 1) mais est le plus éloigné de la communauté centrale par rapport à tous les autres noeds oranges : il se retrouve à la même distance que les nœuds jaunes (reliés à tous les cours, sauf 2) voire à la même distance que certains verts clairs (reliés à tous les cours, sauf 3). On s'aperçoit que PHP n'est pas ce qui unit l'ensemble des supports de cours, mais bien les notions générales de programmation, et particulièrement ceux du développement web.

Fig. 4.9 – Graphe d'impact mutuel des 10 cours [scénario n°2] : Vision d'ensemble éloignée

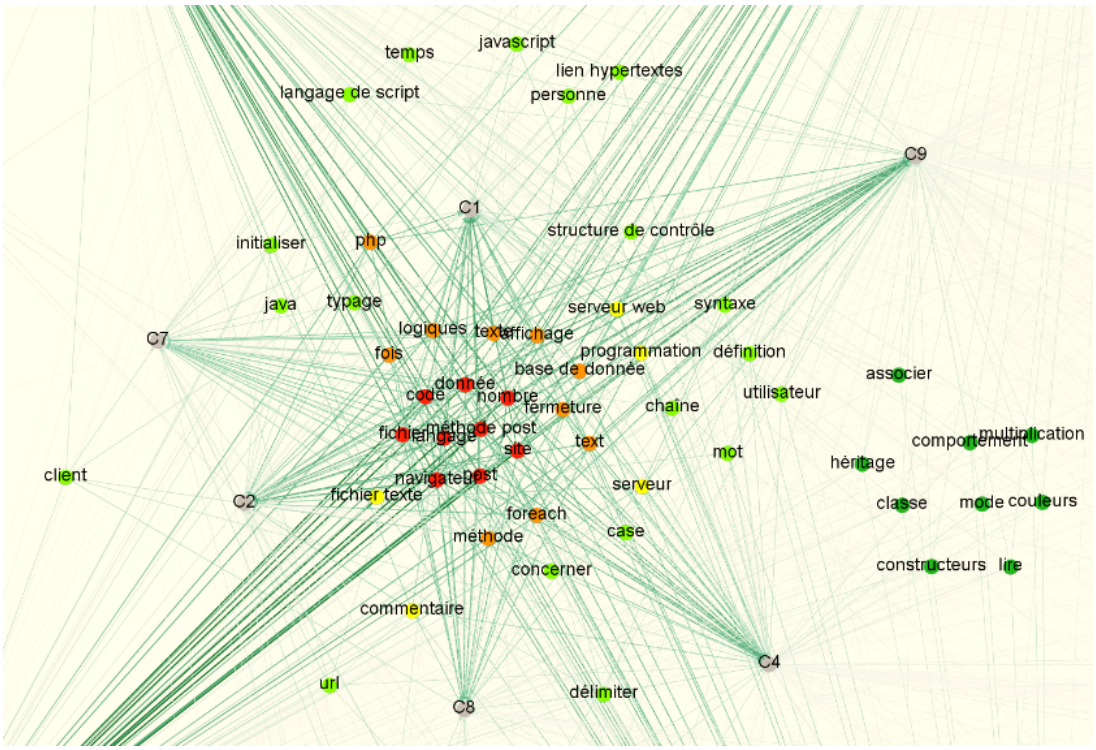


FIG. 4.10 – Graphe d'impact mutuel des 9 cours [scénario n°2] : Zoom sur l'ensemble central

En observant le placement des cours sur la figure 4.11, on s'aperçoit que le cours de Java (CJA) est éloigné de l'ensemble central, au moins autant que C6. Là où C3 et C5 étaient éloignés de l'ensemble central dans le cas n°1 (voir figure 4.8), ils apparaissent maintenant plus proches. L'introduction d'un support de cours éloigné a contribué à partiellement rapprocher certains cours qui jusque là pouvaient paraitre anormalement éloignés. Un enseignant s'apercevra donc inévitablement que les supports C6 et CJA devraient être vérifiés de plus près pour s'assurer de leur rapport avec le sujet qu'il souhaite enseigner. Selon l'éloignement sémantique des supports, on peut supposer que la distance dans le graphe sera comparable : insérer un cours de philosophie risque d'être extrêmement éloigné et donc de rapprocher tous les autres cours (on peut néanmoins opposer l'idée qu'un enseignant n'insèrera pas de documents totalement hors sujet). Chaque support éloigné doit donc être évalué manuellement pour vérifier sa pertinence, et éventuellement recalculer le graphe d'impact mutuel sans celui-ci.

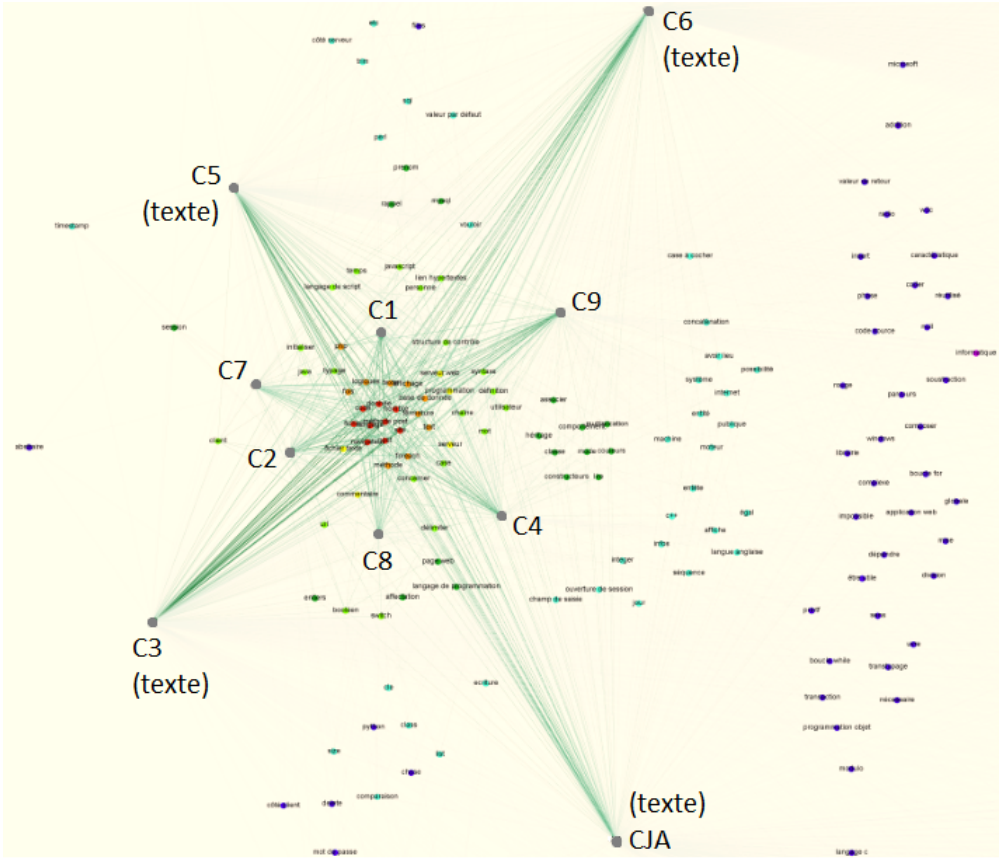


FIG. 4.11 – Graphe d'impact mutuel des 10 cours [scénario n°2] : Zoom sur le positionnement des cours

Le scénario n°3 vise à s'assurer du maintien des résultats précédents malgré le doublement du nombre de documents insérés par rapport au scénario n°1 (18 documents insérés contre seulement 9). La figure 4.12 illustre le graphe d'impact mutuel dans son ensemble. La figure 4.13 illustre l'ensemble central. Enfin, la figure 4.14 illustre le positionnement des documents.

Étant donné que le nombre de cours est doublé, le nombre de groupes de termes l'est lui aussi. Comme nous l'avons observé pour le scénario n°1, chaque document dispose d'un ensemble de termes qui lui est propre : en doublant le nombre de documents, nous avons ajouté autant d'ensembles de termes uniques que de documents en plus. On retrouve en haut à gauche l'ensemble central, puis plus bas vers la droite, chaque groupe de termes dont le nombre de connexions se réduit par itération jusqu'à 1.

En analysant l'ensemble central, illustré par la figure 4.13, on trouve 6 termes reliés à l'ensemble des supports de cours : *navigateur*, *site*, *fichier*, *langage*, *php*, *code*. Certains termes ont été légèrement déplacés plus en extérieur en perdant 1 relation par rapport au cas n°1 : *nombre*, *méthode post*, *texte*, *donnée*. Visuellement, l'ensemble central est resté stable (peu de variation des termes), et le sujet principal est toujours correctement remonté par le graphe. Enfin, les termes perdant 2 relations par rapport au cas n°1 (*langage de script*, *serveur web*, *javascript*, *serveur*, *base de donnée*, *text*) constituent eux aussi des notions importantes. Les termes avec des degrés plus faibles restent néanmoins utiles pour un cours de développement web en PHP (ceux en vert et bleu). On notera que quelques termes provoquant du bruit comme dans le cas n°1 (*fois*, *mot*, et *définition* (qui peut éventuellement être interprété comme un *define* de constante)) sont toujours présents. Le doublement du nombre de documents n'a donc pas modifié le sujet central détecté par le graphe d'impact mutuel.

L'ensemble central étant stable malgré le dédoublement du nombre de supports de cours insérés, nous nous intéressons maintenant à la pertinence des supports avec la figure 4.14. En observant le positionnement des documents, on se rend compte que la plupart des supports forment eux aussi un ensemble homogène. C6, qui était déjà éloigné dans les scénarios n°1 et n°2, est confirmé dans son éloignement, mais est rejoint par C11. Les supports textes sont les plus éloignés. C17, qui était le support avec le moins de mots parmi les supports au format texte, d'après le tableau 4.5, se trouve éloigné comme les autres.

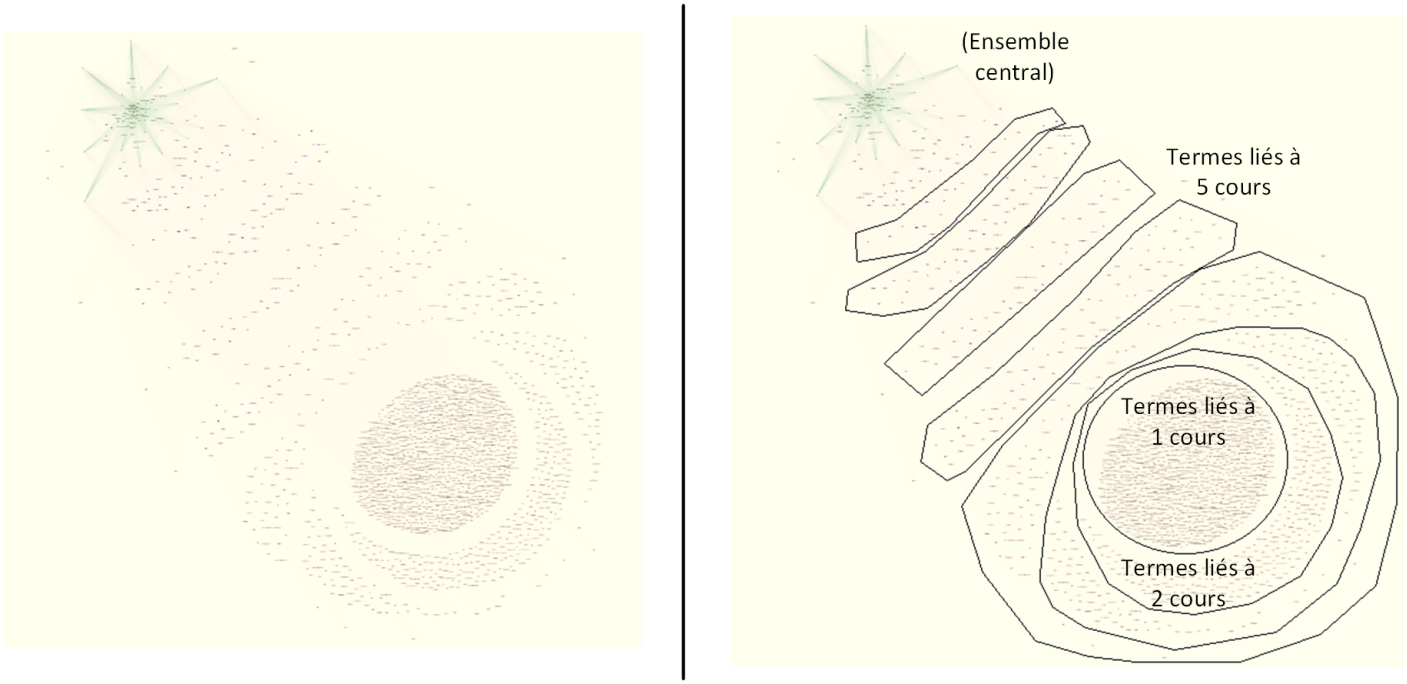


FIG. 4.12 – Graphe d'impact mutuel des 18 cours [scénario n°3] : Vision d'ensemble éloignée et annotations

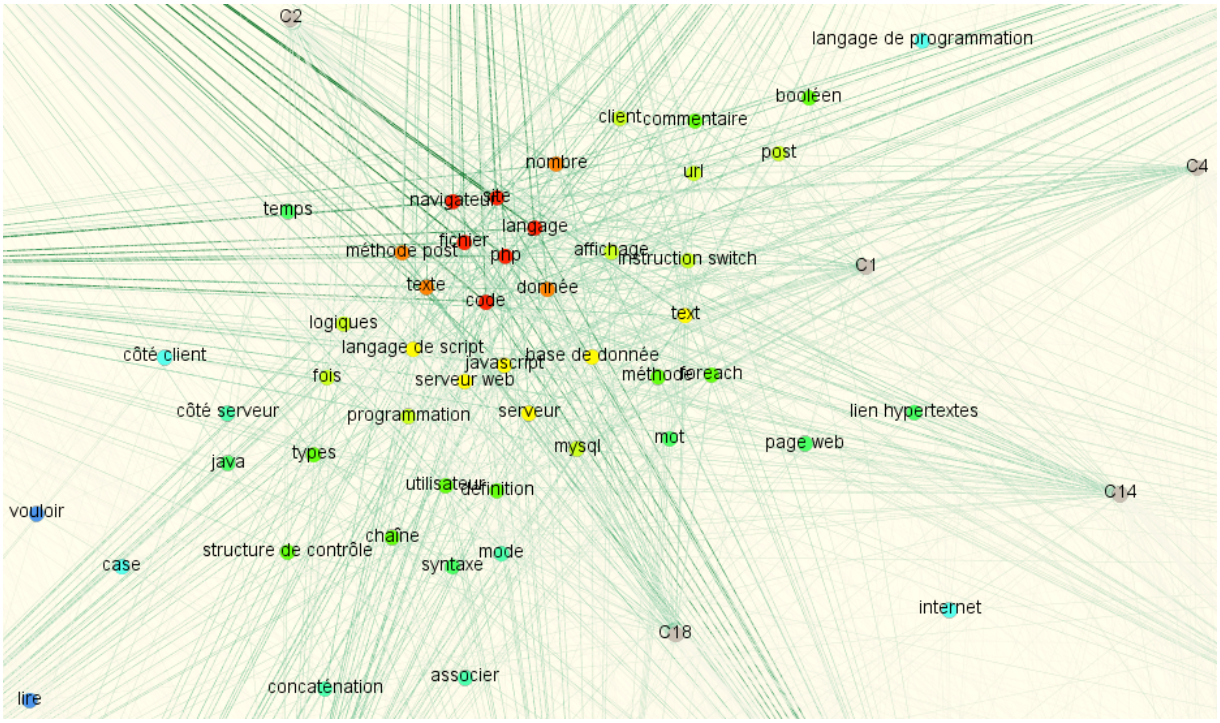


FIG. 4.13 – Graphe d'impact mutuel des 18 cours [scénario n°3] : Zoom sur l'ensemble central

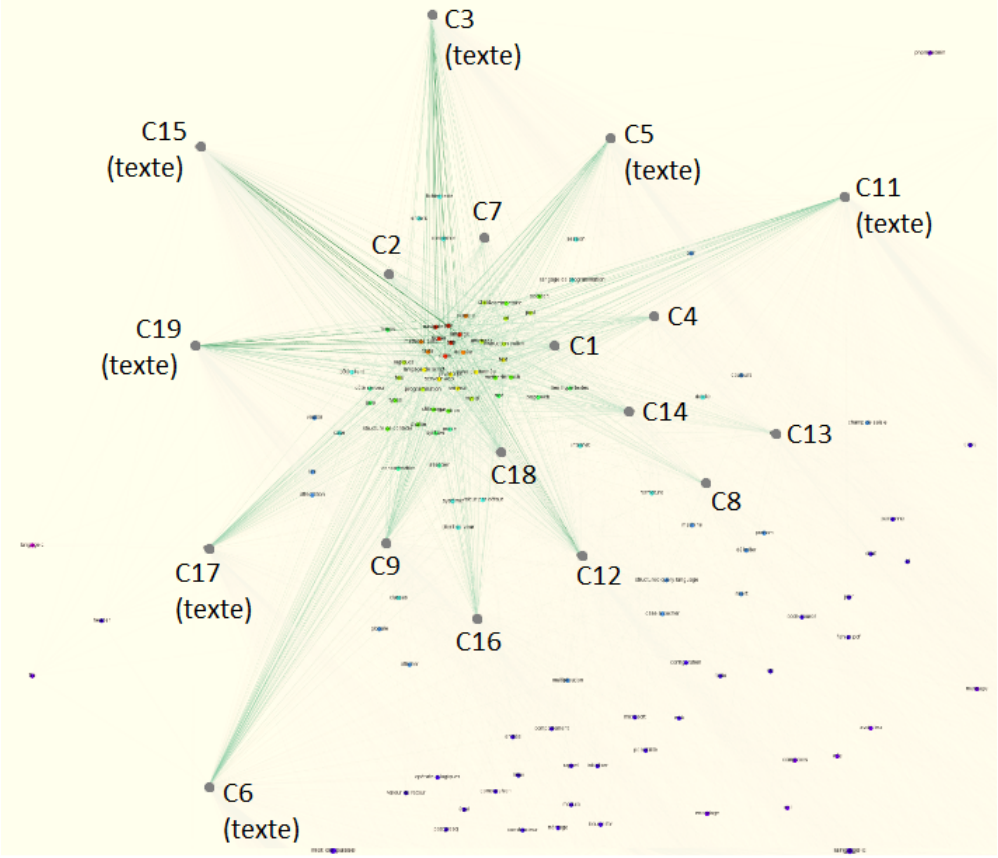


FIG. 4.14 – Graphe d'impact mutuel des 18 cours [scénario n°3] : Zoom sur le positionnement des cours

Le scénario n°4 vise à vérifier que la méthode est fonctionnellement correcte en s'assurant que l'amélioration des documents en entrée produit un graphe d'impact mutuel reflétant ces améliorations. Pour cela, le document C6 (qui était éloigné sur les précédents graphes d'impact mutuel) est successivement corrigé. Ce scénario étudie six versions successives :

1. Corpus documentaire des 7 supports de cours au format texte traitant de PHP
2. Suppression du chapitre dédié aux projets (pages 93 - 150, soit 103 pages restantes) dans C6
3. Suppression du chapitre déclaré comme « hors programme » (pages 41 - 65, soit 79 pages restantes) dans C6
4. Ajout du support Java à la version initiale contenant les 7 supports au format texte
5. Ajout du support Java à la version sans projets
6. Ajout du support Java à la version sans projets ni « hors programme »

Le premier graphe d'impact mutuel présentant les 7 supports de cours originaux au format texte est représenté par la figure 4.15. On constate que C6 est effectivement le document le plus éloigné de l'ensemble central. On notera que C5, C11, et C15 sont également assez éloignés, mais moins que C6.

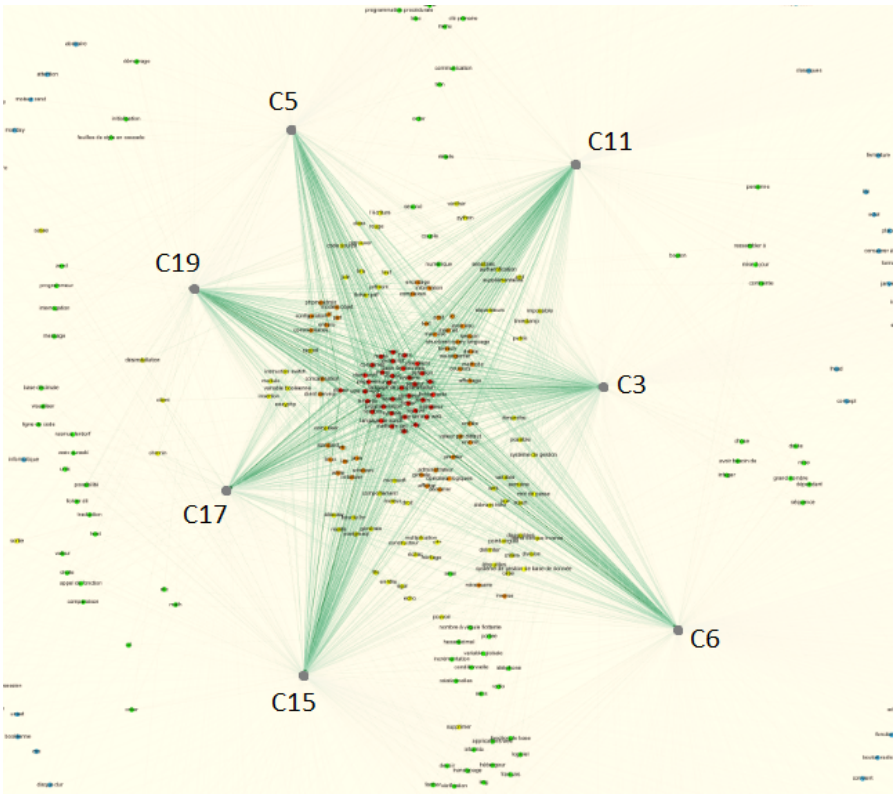


FIG. 4.15 – Graphe d'impact mutuel des 7 supports de cours originaux au format texte traitant de PHP [scénario n°4] : Zoom sur le positionnement des documents

Le deuxième graphe d'impact mutuel présentant les 7 supports de cours au format texte, mais dans lequel C6 a été corrigé en lui retirant le chapitre sur les projets étudiants, est représenté par la figure 4.16. On constate que C6 s'est très nettement rapproché de l'ensemble central, et que les documents les plus éloignés sont devenus C5, C11, et C15. La correction a donc eu un impact positif pour le document C6 : celui-ci est maintenant beaucoup plus proche du sujet présenté dans le corpus. Du point de vue global, le graphe est visuellement plus équilibré : C6 étant devenu assez proche de l'ensemble central, les autres documents sont un peu plus difficiles à distinguer au niveau de leurs distances avec l'ensemble central.

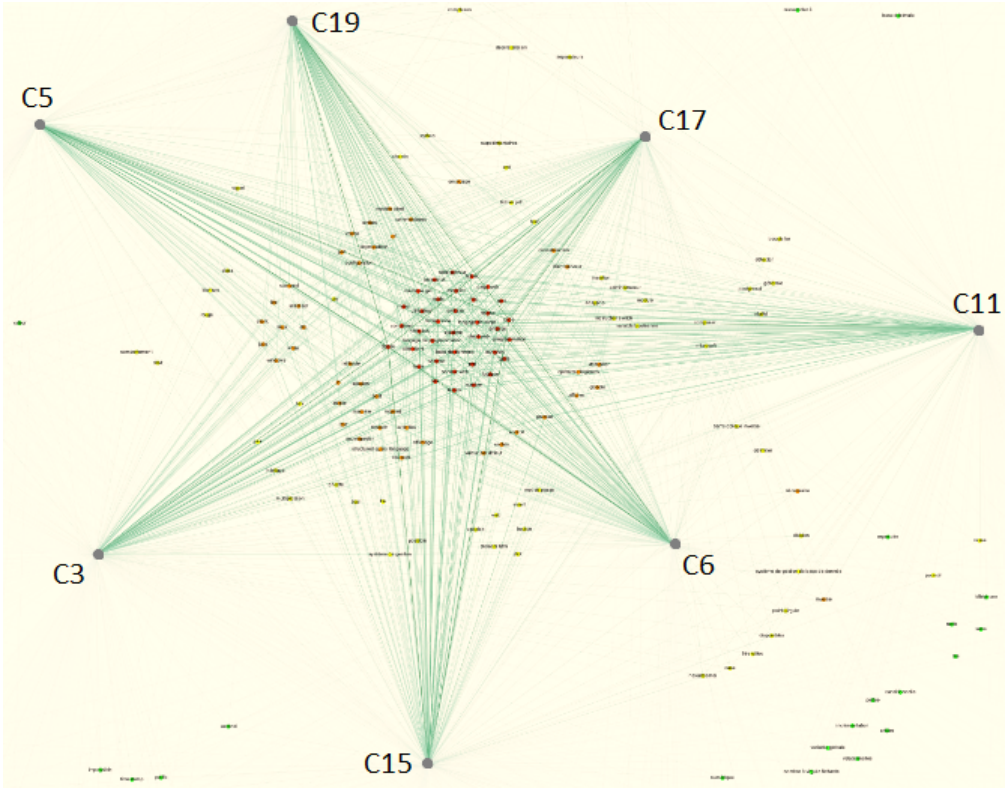


Fig. 4.16 – Graphe d'impact mutuel des 7 supports de cours au format texte traitant de PHP, dont C6 a été corrigé en retirant le chapitre traitant des projets étudiants [scénario n°4] : Zoom sur le positionnement des documents

Le troisième graphe d'impact mutuel présentant les 7 supports de cours au format texte, mais dans lequel C6 a été corrigé en lui retirant le chapitre sur les projets étudiants ainsi que le chapitre déclaré comme « hors programme », est représenté par la figure 4.17. On constate que C6 s'est encore un peu plus rapproché de l'ensemble central, et au contraire, C11 est devenu le document le plus éloigné de l'ensemble central. La correction a donc encore eu un impact positif pour le document C6, mais le graphe s'est visuellement déséquilibré en repoussant particulièrement C11.

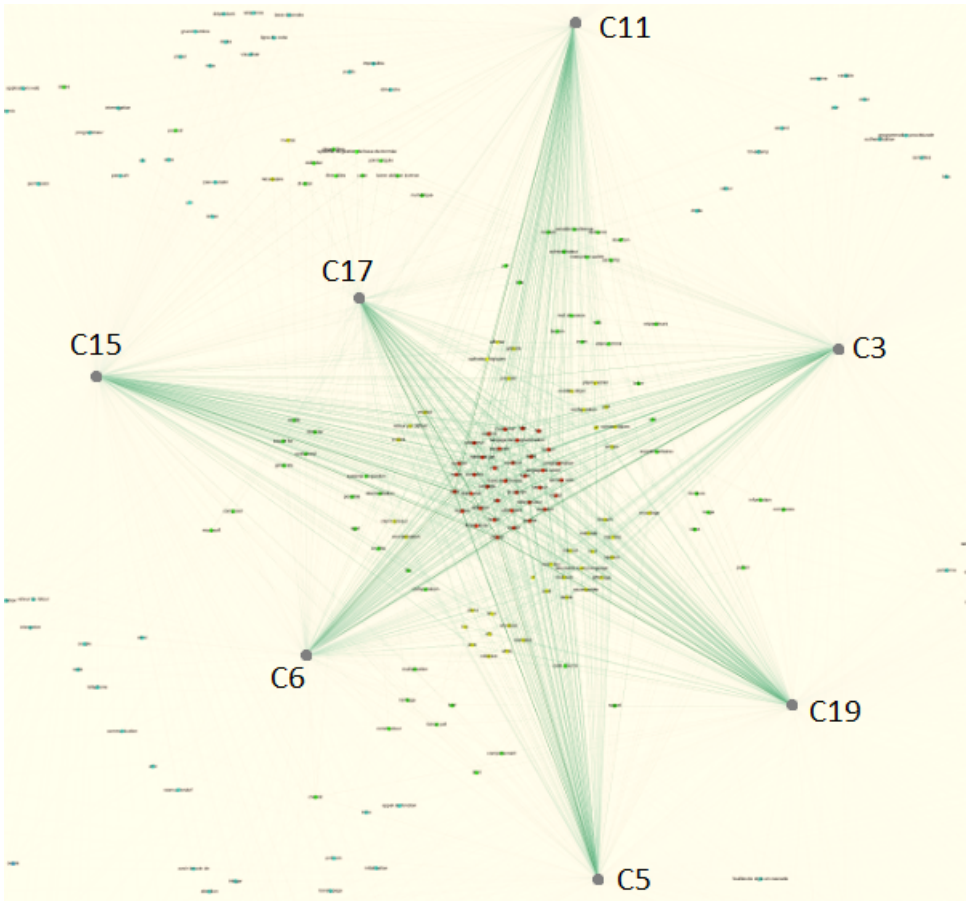


Fig. 4.17 – Graphe d'impact mutuel des 7 supports de cours au format texte traitant de PHP, dont C6 a été corrigé en retirant le chapitre traitant des projets étudiants ainsi que le chapitre hors programme [scénario n°4] : Zoom sur le positionnement des documents

Le quatrième graphe d'impact mutuel présentant les 7 supports de cours au format texte ainsi que le support Java (CJA), est représenté par la figure 4.18. On remarque que C6 et CJA sont les plus éloignés, particulièrement C6. Bien que C6 traite de PHP, l'excès de chapitres hors sujet (et particulièrement la quantité de termes) en fait un document beaucoup plus éloigné des autres.

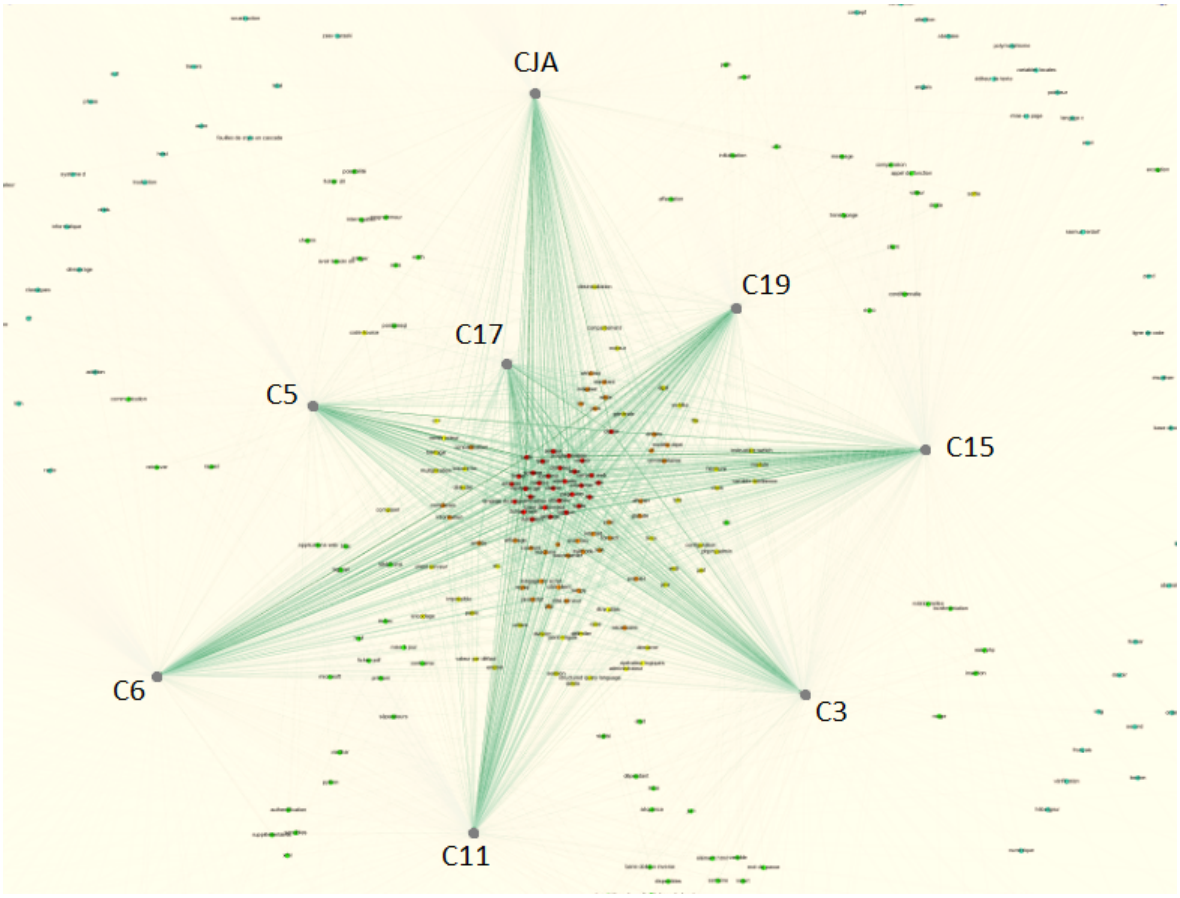


Fig. 4.18 – Graphe d'impact mutuel des 7 supports de cours au format texte traitant de PHP complétés du support Java [scénario n°4] : Zoom sur le positionnement des documents

Le cinquième graphe d'impact mutuel présentant les 7 supports de cours au format texte ainsi que le support Java (CJA), mais dans lequel C6 a été corrigé en lui retirant le chapitre sur les projets étudiants, est représenté par la figure 4.19. On constate que C6 s'est très rapproché de l'ensemble central, mais seuls CJA et C11 sont devenus les plus éloignés. Visuellement, la forme du graphe s'approche d'un parallélogramme, voire d'un losange : hormis CJA et C11, les autres documents sont presque équidistants de l'ensemble central (le cas parfait aurait évidemment été un cercle). La correction de C6 a donc créé un corpus qui visuellement est beaucoup plus homogène que dans les autres cas.

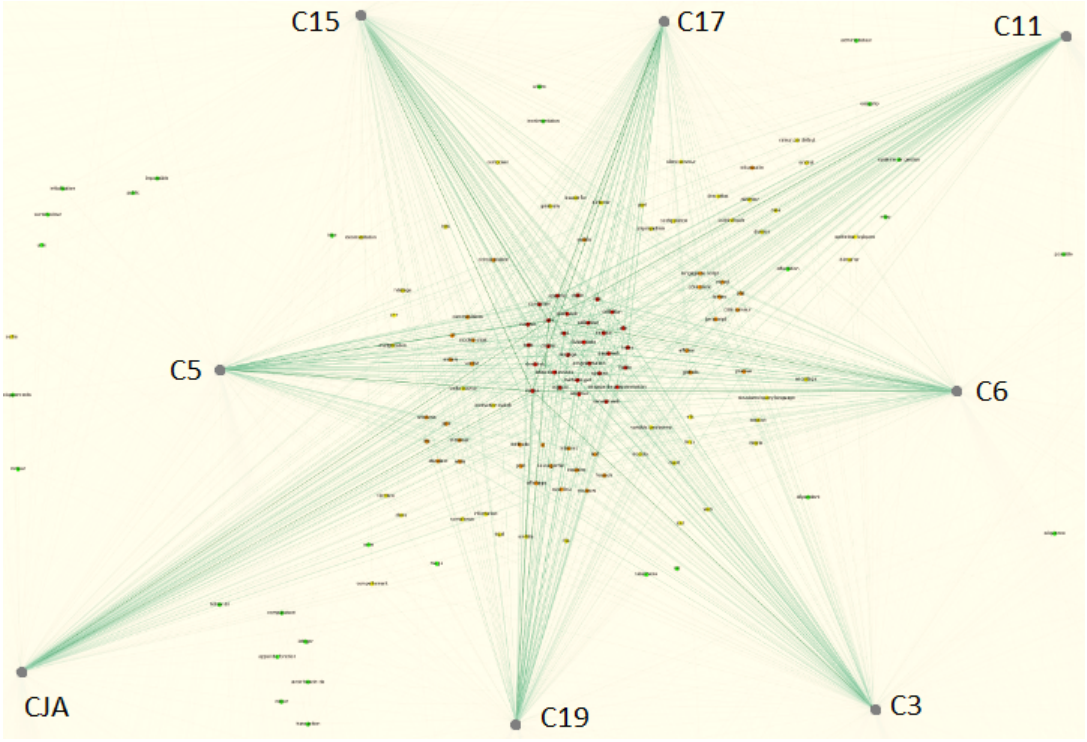


Fig. 4.19 – Graphe d'impact mutuel des 7 supports de cours au format texte traitant de PHP complétés du support Java, dont C6 a été corrigé en retirant le chapitre traitant des projets étudiants [scénario n°4] : Zoom sur le positionnement des documents

Le sixième graphe d'impact mutuel présentant les 7 supports de cours au format texte ainsi que le support Java (CJA), mais dans lequel C6 a été corrigé en lui retirant le chapitre sur les projets étudiants ainsi que le chapitre déclaré comme « hors programme », est représenté par la figure 4.20. On constate que C6 s'est nettement rapproché de l'ensemble central, et est devenu le document le plus proche. C11 et CJA sont les plus éloignés, ainsi que C15. La double correction a donc bien eu un effet sur le graphe, mais graphiquement, la forme aperçue dispose d'un sommet particulièrement décalé vers l'intérieur.

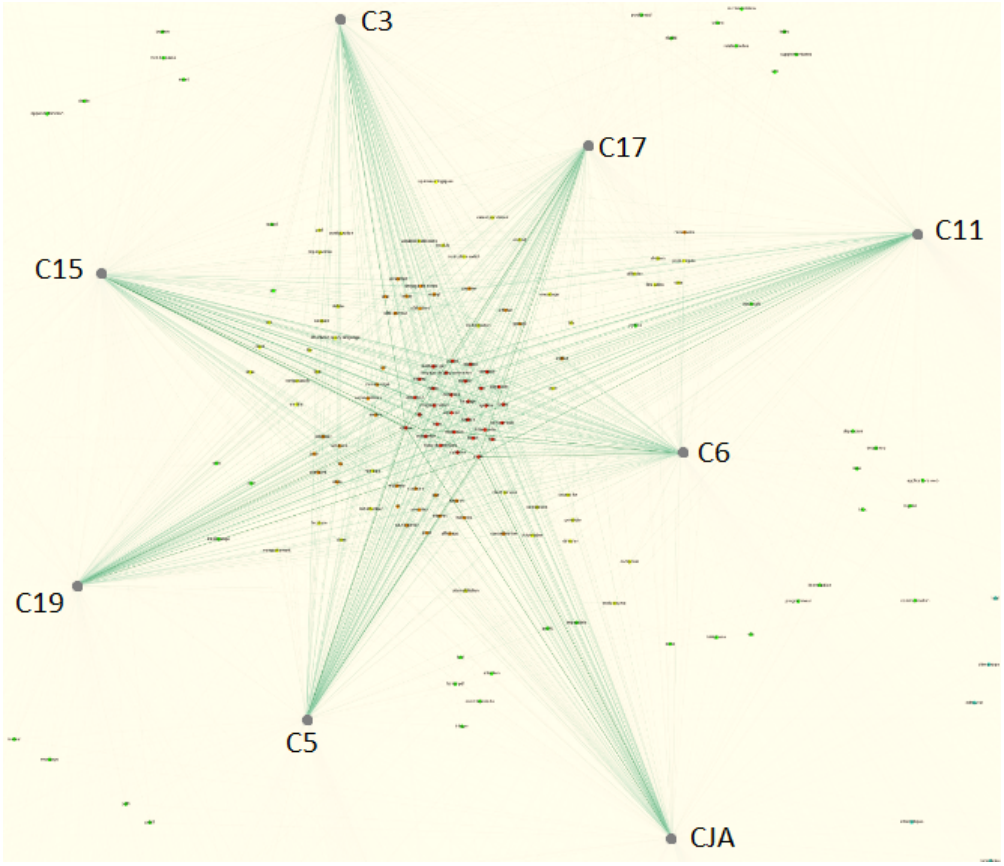


Fig. 4.20 – Graphe d'impact mutuel des 7 supports de cours au format texte traitant de PHP complétés du support Java, dont C6 a été corrigé en retirant le chapitre traitant des projets étudiants ainsi que le chapitre hors programme [scénario n°4] : Zoom sur le positionnement des documents

Le scénario n°5 vise à s'assurer que la méthode CREA fonctionne sur d'autres sujets d'intérêts, et dans une autre langue. La figure 4.21 illustre l'ensemble central. La figure 4.22 illustre le positionnement des documents.

L'ensemble central du graphe d'impact mutuel, visible sur la figure 4.21, présente seulement 2 termes communs à l'ensemble des 13 documents : « *statecharts* » et « *state machine* ». Le sujet étudié est donc correctement détecté dans l'ensemble des documents. On remarque que « *time* » est le troisième terme le plus connecté. Celui-ci est lié aux exemples employant régulièrement ce terme (on peut citer l'exemple détaillant le fonctionnement d'une montre digitale), mais également au domaine lors de la gestion des transitions et évènements dans le temps. Plusieurs autres termes importants du domaine sont visibles en périphérie du noyau en rouge : « *triggering* », « *state transition* », « *substates* », « *process* », « *diagrams* », « *automata* », « *semantics* », « *systems* », « *complex* ». Bien que l'ensemble central soit plus éparse que nos précédents exemples, il n'en reste pas moins explicite et correct concernant le contenu des documents sélectionnés.

Concernant les 13 documents, la figure 4.22 montre que les articles sont plus éloignés que les autres types de documents. Les documents les plus éloignés sont aussi les plus longs en quantité de mots et de termes : A4 étant le chapitre de livre et A1 étant un article particulièrement long. A2, A5, et C7 sont également parmi les plus gros en quantité de termes, par rapport aux cours composés de quelques centaines de termes. L'article A3 est plus difficile à expliquer : sur la quantité de termes, il est comparable à C1, mais il se retrouve graphiquement presque aussi éloigné que A1 et A4. Il s'agit d'un article **utilisant** les statecharts, mais détaillant également une étude de cas et quelques autres formalismes différents des statecharts. Cette non spécialisation sur les statecharts, contrairement à l'ensemble des autres documents, pourrait être à l'origine de cet écart.

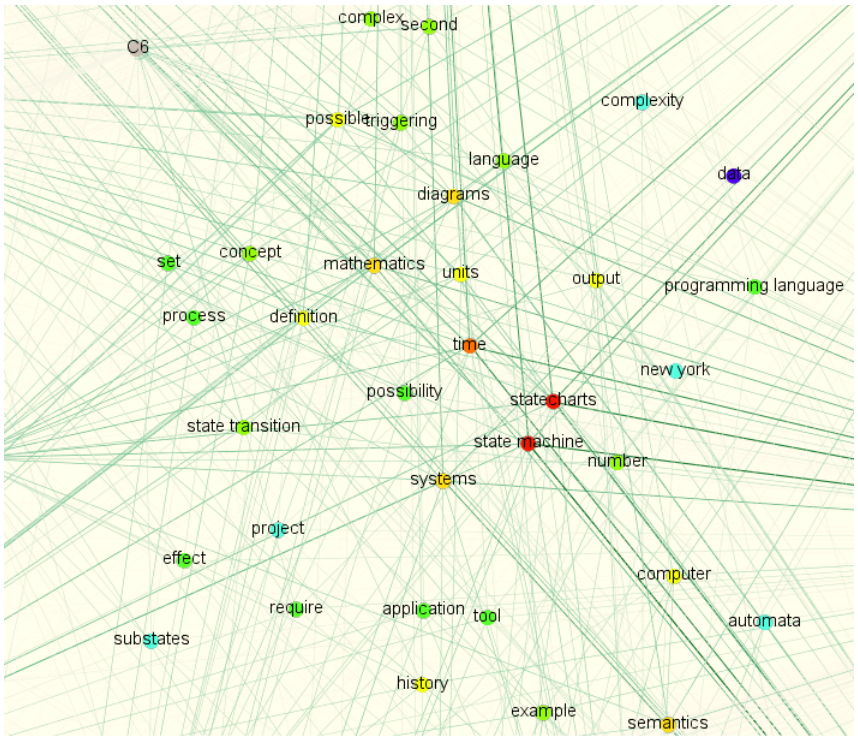


FIG. 4.21 – Graphe d'impact mutuel des 8 cours et 5 articles sur les statecharts [scénario n°5] : Zoom sur l'ensemble central

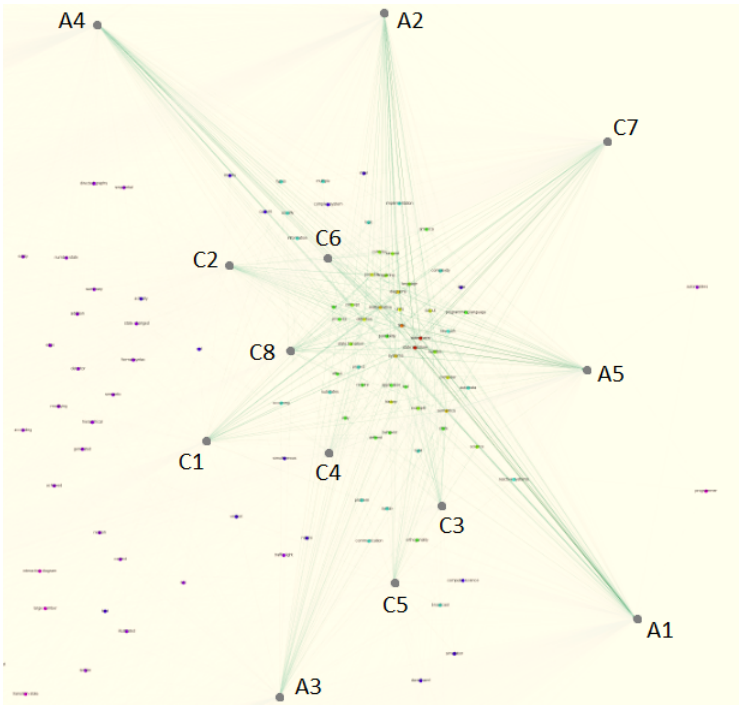


FIG. 4.22 – Graphe d'impact mutuel des 8 cours et 5 articles sur les statecharts [scénario n°5] : Zoom sur le positionnement des documents

Clusters de termes (PII.3)

Les supports étant en lien avec le sujet, et l'ensemble central illustrant ce sujet par l'intermédiaire de plusieurs termes, nous nous intéressons maintenant aux clusters générés avec la matrice de similarité conceptuelle. Comme expliqué lors de la validation structurelle en 4.3.3, il apparait que $\beta = 1.00$ est la valeur la plus adaptée pour générer des clusters pertinents pour un enseignant. Nous générons donc les 8 clusters des scénarios n°1 et 3 comme illustré sur la figure 4.23, et les 8 clusters du scénario n°5 comme illustré sur la figure 4.24.

La figure 4.23 compare les clusters générés avec les 18 supports de cours par rapport au scénario n°1 composé de 9 supports de cours. On remarque le nombre plus élevé de termes dans le scénario n°3, précisément 106 termes, contre 60 termes dans le scénario n°1. Certains termes qui étaient éliminés en augmentant β à 1.00 dans le scénario n°1 réapparaissent (« *salaire* », « *france* », « *bienvenue* », « *paris* », ...). Les clusters apparaissent disproportionnés : le cluster 3 contenant 34 termes, et le cluster 8 contenant 31 termes, contre une moyenne de 7 termes pour les autres cluster du scénario n°3. Ces anomalies pourraient s'expliquer par la valeur de β devenue trop faible, et donc ne filtrant plus les termes en amont. Des expériences supplémentaires testant des valeurs de β supérieures à 1.00 sont nécessaires pour pouvoir confirmer cela.

En analysant la qualité des clusters proposés, des liens logiques entre des termes peuvent être faits (par exemple « *fermeture* », « *session* », « *avoir accès* » : pour parler des sessions, les accès, et la fermeture de session), mais les clusters 3 et 8 sont trop disproportionnés pour pouvoir être utilisés en l'état. Il est difficile d'utiliser ces clusters tels quels, bien que toutes les étapes précédentes aient montré des résultats encourageants. Une quantité maximale de mots par clusters serait utile, une meilleure répartition des termes parmi les clusters, une limite de termes uniques par les stratégies, augmenter β proportionnellement au nombre de supports de cours insérés, ou encore un nombre précis de supports en entrée pourraient être envisagés.

Dans le scénario n°5, quelques termes forment du bruit (par exemple « *c d* », ou encore « *america* »), ou sont directement issus d'exemples (« *traffic light* », « *automobiles* »). Le cluster 5 semble relativement intéressant en rassemblant « *diagrams* », « *simultaneous* », « *substates* », et « *composite states* » : on retrouve bien l'idée des sous-états et états composites qui se visualisent parfaitement bien sur des diagrammes. Le cluster 1 rassemble des termes utiles pour une introduction : « *david harel* », « *state* », « *determinism* », « *formal syntax* », « *deterministic finite state automata* ». Le cluster 3 rend compte de l'aspect temporel et réactif utile : « *reactive systems* », « *time* », « *current* ». Ces résultats rappellent ceux du scénario n°3 où des bribes de clusters pouvaient être réutilisées, mais la grande quantité de termes dans quelques clusters empêche d'avoir une vue optimale des clusters. Cependant, ce cas montre que la langue et le format des documents ne sont pas si bloquant pour la méthode CREA. De plus, le domaine a été plutôt bien reconnu par BabelFy, bien qu'il soit beaucoup plus spécialisé que le développement web en PHP.

Stratégie Haute (β = 1,00) – cas n°3

1	case à cocher	passwd	programmation	web	méthode	programmation web	attribut	head
2	côté serveur	côté client	casse	interpréter	pointeur	exe		
3	javascript	gras	lien hypertextes	nombre	sélectionner	affectation	structure de contrôle	définition
	types	entités	valeur par défaut	composer	list	mode	configuration	xml
	bits	conditionnelle	pouvoir	mot	information	varchar	easyphp	insert
	impossible	fonctionnalité	visiteur	null	rue	doctype	file	moteur
	transaction	exception						
4	internet	système	menu	sauvegarder	couleurs	id		
5	langage	fermeture	session	avoir accès	france	valeur		
6	page web	serveur web	texte	concerner	associer	machine	salaire	
7	post	jour	class	foreach	bienvenue	paris	opérateurs	mysqli
8	navigateur	fichier	client	base de donnée	donnée	php	code	commentaire
	délimiter	fois	url	text	zones	utilisateur	syntaxe	case
	personne	méthode post	mysql	site	langage de script	afficher	nécessaire	serveur
	chaîne	timestamp	cle	classes	header	echo	wamp	

1	php	code	fois	post	jour	foreach	cle	classe	class	mysqli	
2	page web	navigateur	serveur web	texte	concerner	délimiter	utilisateur	associer	personne	machine	mysql
3	url	langage	case	fermeture	session	chaîne	entête	avoir accès			
4	fichier	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client				
5	typage	mot	moteur	affiche	transaction	visiteur					
6	base de donnée	insert	varchar	null							
7	xml	configuration	composer	doctype							
8	donnée	text	méthode post	programmation	site	langage de script	list	méthode	timestamp	files	

Stratégie Haute (β = 1,00) – cas n°1

Fig. 4.23 – Clusters issus de la stratégie *Haute* pour $\beta = 1.00$ pour les scénarios n°1 (référence) et n°3

Stratégie Haute (β = 1,00)

1	david harel	state	units	determinism	syntax	formal syntax	higraphs	large number	interaction diagram	deterministic finite state automata	
2	science	computer	systems	example	number	state transition	possible	semantics	automata	effect	definition
3	communication	reactive systems	time	broadcast	c d	history	current				
4	statecharts	state machine	problem	orthogonality	concept	sec	require	second	merely	simple	
5	diagrams	complex	information	simultaneous	substates	entity	composite	composite states			
6	language	simulation	nature	light	workflows	code	autonomous				
7	automobiles	hierarchical	defined	types	multiple	stateflow					
8	mathematics	complex system	networks	way	set	output	tool	triggering	main	occurring	america
	achieved	application	possibility	subset	management	finite	new york	automata theory	sequence	strategy	traffic light
	bit	summary									

Fig. 4.24 – Cluster issu de la stratégie *Haute* pour $\beta = 1.00$ pour le scénario n°5

4.3.5 *Validation par retour d'expérience*

Afin d'obtenir un avis extérieur sur la qualité des clusters générés nous avons interrogé 5 informaticiens à propos des résultats du scénario n°1 servant de référence. Les réponses aux deux questionnaires sont maintenant présentés.

Réponses aux deux questionnaires

Les résultats du premier questionnaire sont présentés sur la figure 4.25 avec le cluster de référence du scénario n°1 tout en haut. Afin d'évaluer la similarité des réponses des informaticiens aux clusters générés par la méthode CREA, nous avons calculé l'indice de Rand et l'indice de Rand ajusté entre chacune des partitions. Étant donné que certaines réponses ne contiennent pas tous les termes, nous avons calculé une version où les termes oubliés sont tous regroupés dans un unique cluster (le cluster des termes oubliés), mais également une version où chaque terme oublié est dans son propre cluster.

Précisément, il faut noter les oublis suivants :

— Non-expert n°1 a oublié 4 termes : « *texte* », « *typage* », « *varchar* », « *timestamp* »

— Expert n°3 a oublié 1 terme : « *class* »

— Expert n°4 a oublié 1 terme : « *list* »

Les tableaux 4.14 et 4.15 présentent les résultats de la version où les termes oubliés sont regroupés dans un unique cluster. Les tableaux 4.16 et 4.17 présentent les résultats de la version où les termes oubliés sont chacun dans leur propre cluster.

On notera tout d'abord que ces résultats varient très peu entre les deux versions. Les résultats montrent des indices de Rand élevés ($> 0,750$) dans l'ensemble des comparaisons, ce qui indique que le placement des paires de termes est plutôt en accord entre l'ensemble des participants et la méthode CREA. Cependant, l'indice de Rand ajusté est très faible, voire négatif dans certains cas, ce qui indique que les partitions sont considérées comme construites aléatoirement.

Étant donné que les clusters ont bien été construits avec un objectif pédagogique en tête, ou à minima une certaine logique, ces résultats présentent au contraire la complexité de retrouver des liens entre les termes et leurs placements dans des clusters : toute la difficulté concernant la visualisation des connaissances implicites est exposée ici. De plus, nous ne pouvons pas affirmer avoir une partition plus *correcte* qu'une autre (la « *ground truth* ») tout est question de contexte et d'expérience lors de l'interprétation des clusters de termes. Une métrique de similarité seule n'est donc pas suffisante, il est nécessaire d'associer un contexte pour évaluer la pertinence des regroupements et leurs similarités.

	Méthode CREA	Non-expert n°1	Non-expert n°2	Expert n°3	Expert n°4	Expert n°5
Méthode CREA	1,0					
Non-expert n°1	0,809	1,0				
Non-expert n°2	0,768	0,789	1,0			
Expert n°3	0,782	0,798	0,786	1,0		
Expert n°4	0,788	0,821	0,807	0,805	1,0	
Expert n°5	0,767	0,784	0,764	0,794	0,799	1,0

Tab. 4.14 – Indice de Rand appliqué au scénario n°1 et aux réponses des 5 informaticiens *(les termes non reportés sont mis dans un unique cluster)*

	Méthode CREA	Non-expert n°1	Non-expert n°2	Expert n°3	Expert n°4	Expert n°5
Méthode CREA	1,0					
Non-expert n°1	0,080	1,0				
Non-expert n°2	0,009	0,053	1,0			
Expert n°3	-0,014	0,008	0,073	1,0		
Expert n°4	0,005	0,117	0,160	0,069	1,0	
Expert n°5	-0,022	0,006	0,030	0,084	0,102	1,0

Tab. 4.15 – Indice de Rand ajusté appliqué au scénario n°1 et aux réponses des 5 informaticiens *(les termes non reportés sont mis dans un unique cluster)*

	Méthode CREA	Non-expert n°1	Non-expert n°2	Expert n°3	Expert n°4	Expert n°5
Méthode CREA	1,0					
Non-expert n°1	0,812	1,0				
Non-expert n°2	0,768	0,785	1,0			
Expert n°3	0,782	0,799	0,786	1,0		
Expert n°4	0,788	0,821	0,807	0,805	1,0	
Expert n°5	0,767	0,787	0,764	0,794	0,799	1,0

Tab. 4.16 – Indice de Rand appliqué au scénario n°1 et aux réponses des 5 informaticiens *(les termes non reportés sont chacun dans leur propre cluster)*

	Méthode CREA	Non-expert n°1	Non-expert n°2	Expert n°3	Expert n°4	Expert n°5
Méthode CREA	1,0					
Non-expert n°1	0,085	1,0				
Non-expert n°2	0,009	0,028	1,0			
Expert n°3	-0,014	0,001	0,073	1,0		
Expert n°4	0,005	0,105	0,160	0,069	1,0	
Expert n°5	-0,022	0,010	0,030	0,084	0,102	1,0

Tab. 4.17 – Indice de Rand ajusté appliqué au scénario n°1 et aux réponses des 5 informaticiens *(les termes non reportés sont chacun dans leur propre cluster)*

Dans le deuxième questionnaire, la première question vise à noter la qualité des clusters générés par la méthode CREA. 1 seul non-expert en PHP a répondu avec la note de 3, tous les autres ont répondu avec la note de 4, faisant une moyenne de $3,8/5$.

Dans les remarques concernant cette note, les non-experts ont pointé l'ordre d'enseignement des notions (« *Selon le niveau de départ de chacun certains modules risquent de les perdre par manque de connaissance. Exemple XML avant de connaitre le php ou l'HTML* »), mais aussi le regroupement logique de certaines d'entre elles (« *Ne connaissant pas bien le PHP je ne suis pas le meilleur des critiques à ce sujet. J'aurais eu tendance à faire un groupe d'introduction des notions du PHP puis un groupe avec des éléments comme les chaînes de caractères, afficher la chaîne à l'écran et récupérer une information de l'URL* »).

Les experts en PHP ont quant à eux estimé que les clusters sont plutôt pertinents, malgré du bruit à plusieurs niveaux. L'un d'entre eux estime que l'ordre des termes dans les clusters provoque ce bruit (« *Tous les clusters ont une cohérence propre, parfois difficile à extraire (mais c'est lié à l'ordre des termes)* »), un autre est d'avis que quelques termes en trop le provoque (« *1 ou 2 notions par cluster n'ont pas de lien évident à mon sens* »), le dernier pointant plutôt que 3 clusters ne sont pas assez cohérents (« *cinq clusters sur huit contiennent des termes fortement liés de mon point de vue. Les 3 autres ont besoin d'un peu de remaniement* »).

La qualité des clusters est donc globalement correcte, il n'y a pas de notion complètement hors sujet, mais plutôt un bruit limitant l'usage tel quel des clusters. Ceci correspond à ce que nous pouvons attendre d'un outil guidant un enseignant, tout en le laissant décider des notions particulières à conserver ou éliminer pour ses besoins, et surtout, pour ceux de son public.

REFERENCE – Stratégie Haute (β = 1,00)

1	php	code	fois	post	jour	foreach	cle	classe	class	mysqli	
2	page web	navigateur	serveur web	texte	concerner	délimiter	utilisateur	associer	personne	machine	mysql
3	url	langage	case	fermeture	session	chaîne	entête	avoir accès			
4	fichier	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client				
5	typage	mot	moteur	affiche	transaction	visiteur					
6	base de donnée	insert	varchar	null							
7	xml	configuration	composer	doctype							
8	donnée	text	méthode post	programmation	site	langage de script	list	méthode	timestamp	files	

Informaticien Non-Expert n°1

1 intro	concerner	délimiter	langage	machine	navigateur	page web	programmation	site	
2 Install	configuration	côté serveur	fichier	files	moteur	serveur	utilisateur		
3 PHP	class	code	commentaire	entête	fois	foreach	méthode	php	url
4 SQL	base de donnée	chaîne	classe	cle	insert	mot	mysql	null	post
	session	text							
5 Lien PHP SQL	associer	côté client	donnée	fermeture	list	transaction			
6 Rendu USER	affiche	case	case à cocher	interpréter	personne	serveur web	visiteur		
7 Modifier conf	avoir accès	méthode post	mysqli						
8 Pousser	composer	doctype	jour	langage de script	xml				

Informaticien Non-Expert n°2

1	associer	base de donnée	cle	donnée	insert	machine	mysql	mysqli	serveur	transaction
2	configuration	langage	php	programmation						
3	affiche	chaîne	code	commentaire	fois	interpréter	jour	langage de script	null	text
	texte	timestamp	typage	varchar						
4	avoir accès	class	classe	foreach	list					
5	côté client	côté serveur	serveur web	session	url					
6	doctype	entête	fermeture	fichier	files	xml				
7	case	case à cocher	méthode	méthode post	mot	moteur	navigateur	page web	personne	post
	site	utilisateur	visiteur							
8	composer	concerner	délimiter							

Informaticien Expert n°3

1	code	concerner	associer	commentaire	interpréter	langage	langage de script	machine	php	programmation	utilisateur
2	affiche	chaîne	donnée	fois	mot	navigateur	serveur	texte	typage		
3	fermeture	fichier	files	foreach	list	null	text				
4	cle	côté client	doctype	entête	page web	serveur web	site	url	visiteur	xml	
5	case	case à cocher	délimiter								
6	classe	méthode	méthode post	post	session						
7	avoir accès	base de donnée	configuration	côté serveur	insert	jour					
8	composer	moteur	mysql	mysqli	personne	timestamp	transaction	varchar			

Informaticien Expert (a enseigné 3 ans) n°4

1	langage	page web	interpréter	affiche	navigateur	moteur	machine	site	url	
2	doctype	entête	xml	fichier	files	composer	concerner	côté client	côté serveur	
3	php	typage	langage de script	classe	programmation					
4	null	case	case à cocher	associer	commentaire	code	foreach	utilisateur	class	fois
5	méthode	méthode post	post	serveur	serveur web	session	visiteur	personne	avoir accès	
6	mysql	mysqli	transaction	donnée	base de donnée	configuration				
7	cle	varchar	text	texte	timestamp	insert	chaîne	délimiter		
8	jour	mot	fermeture							

Informaticien Expert (enseignant) n°5

1	affiche	code	commentaire	navigateur	page web	serveur	serveur web	site	texte
	url	visiteur	langage	composer					
2	programmation	php	machine	langage de script	interpréter	foreach	côté serveur	côté client	chaîne
3	case	case à cocher	configuration	donnée	list	null	post	méthode post	
4	fichier	files							
5	associer	avoir accès	class	classe	concerner	personne	typage	xml	méthode
6	base de donnée	cle	fermeture	moteur	mysql	mysqli	text	varchar	délimiter
7	fois	insert	jour	mot	timestamp	transaction	session	utilisateur	
8	doctype	entête							

Fig. 4.25 – Réponses au questionnaire n°1 par les 5 informaticiens

4.4 Discussions

Dans cette section nous analysons et discutons les conclusions des expériences précédemment réalisées. Nous identifions ensuite les limites de la méthode CREA et leurs impacts. Enfin, nous discutons la méthode d'évaluation et la validité des conclusions obtenues.

4.4.1 Analyse et discussions des résultats

Durant la validation structurelle, le nettoyage des mots de l'étape de *pré-traitement sémantique* a montré un certain écart entre les supports au format texte et ceux au format diapositives, confirmant partiellement l'hypothèse H7a. Les documents texte perdent plus de mots que les documents au format diapositives dans les scénarios n°1-2-3. L'origine de cette différence provient du fait que nous avons exclu des classes de mots qui sont généralement absentes des diapositives. En effet, le format diapositives incite généralement à limiter le vocabulaire aux classes grammaticales portant le plus de sémantique (noms et verbes) en évitant les classes grammaticales formant des liens (pronoms, adverbes, ...). Cette tendance se remarque également lors de l'analyse des termes uniques du *filtrage des termes* où les supports au format texte ont de plus grandes proportions termes communs (exclus et inclus) que les autres supports.

Moins significativement, les supports textes du scénario n°3 conservent moins de termes que leurs homologues au format diapositives, mais cette tendance ne semble pas s'appliquer aux supports texte du scénario n°1. Enfin, on retrouve cette distinction entre les deux types de supports dans le graphe d'impact mutuel. En effet, les proportions de termes uniques retenus pour les supports texte sont parmi les plus élevées (jamais inférieures à 29%), et on peut observer que les supports texte sont constamment aux extrémités dans chacun des graphes d'impact mutuel des scénarios n°1-2-3. Cette observation n'invalide pas pour autant les hypothèses H3 et H8b concernant la détection des textes peu pertinents selon la distance sur le graphe d'impact mutuel : le support Java est un des deux supports les plus éloignés de tous les autres.Le vocabulaire plus restreint des supports au format diapositives leur permet également de partager plus de termes que les supports au format texte, et donc d'apparaitre plus proches de l'ensemble central où les termes sont connectés au maximum de documents.

Plus les supports ont un vocabulaire diversifié et unique au document (c'est-à-dire un vocabulaire peu commun), plus ils disposeront de termes faiblement connectés, et donc plus ils apparaitront en retrait des autres. Cette analyse suppose qu'un document complètement hors sujet sera uniquement connecté à ses propres termes, et réciproquement, ses propres termes ne seront connectés à aucun autre document, excluant d'office les termes et le document de l'ensemble central.

Le pré-traitement sémantique effectué a également permis de constater qu'une majorité de termes pertinents et quelques termes peu pertinents ont été retenus, malgré quelques erreurs, dans ces premiers scénarios. Ainsi, l'hypothèse H4a est validée.

Bien que le scénario n°4 vise particulièrement l'hypothèse H7b, il contribue également à appuyer les hypothèses H3 et H8b tout en testant spécifiquement les documents au format texte. En effet, une fois le document C6 corrigé, c'est-à-dire qu'il est considéré comme *correct*, celui-ci est représenté sur le graphe d'impact mutuel comme étant

devenu pertinent par rapport au contexte. Ce degré de pertinence devient d'autant plus fort lorsque le support de cours Java est introduit. Ainsi, le graphe d'impact mutuel résultant est effectivement impacté par la présence de parties hors sujet (H7b), mais permet tout de même de visualiser l'écart entre les documents (H8b).

Inversement, l'hypothèse H8c impliquant que le graphe d'impact mutuel n'est pas une représentation suffisante pour visualiser avec le maximum de précision les écarts entre documents est également validée par ce scénario : les documents C6 d'origine et CJA sont les plus éloignés de l'ensemble central sans indication plus spécifique. Le graphe d'impact mutuel ne retranscrit pas que C6 dispose d'un chapitre parfaitement intégré au sujet général en complément d'au moins un chapitre hors sujet, là où CJA est intégralement hors sujet. L'utilisateur de la méthode CREA doit chercher et comprendre par lui-même les problèmes dans les documents.

Lors de l'application des stratégies de binarisation, nous espérions pouvoir exploiter les stratégies hautes et basses afin de pouvoir respectivement exposer un syllabus des notions générales abordées dans l'ensemble des documents (les documents s'intéressent surtout à ces notions), et obtenir une carte des notions spécifiques à chaque document (cette notion est spécifiquement abordée dans ce document). Cependant, bien que les termes conservés soient les mêmes dans les deux cas avec des valeurs basses de β, nous avons constaté une très forte diminution du nombre de termes avec la stratégie basse en augmentant la valeur β. Une autre différence concerne les proportions de « 0 » et de « 1 » : la stratégie basse dispose de beaucoup plus de « 1 » que les autres stratégies à valeur égale de β. Cela signifie que les termes conservés par la stratégie basse, et étiquetés d'un « 1 », apparaissent peu dans chacun des documents concernés. Le faible nombre de termes concernés par cette stratégie, et nos tentatives d'exploiter les graphes générés avec, ne permettent pas en l'état de produire une carte des notions spécifiques à chaque document.

À l'inverse, la stratégie haute conserve une proportion de « 1 » adaptée à la production de concepts formels au centre du treillis. Cette proportion produit plusieurs combinaisons entre les termes et documents, mais pas l'ensemble des combinaisons possibles. Le treillis ne peut pas produire d'informations en l'absence de combinaisons, mais à l'inverse, un excès de combinaisons ne permet pas de déterminer la pertinence des informations (tous les cas possibles étant extraits sans aucune pondération). De plus, la stratégie haute est théoriquement censée exposer les hautes fréquences d'apparitions de termes dans les documents, et donc de sélectionner les notions les plus abordées. En calculant la métrique de similarité conceptuelle sur le treillis, et en l'utilisant avec la classification ascendante hiérarchique par la suite, les clusters générés confirment cette intuition. Nous pouvons générer des clusters rassemblant les notions les plus abordées dans le corpus documentaire. Le paramétrage du β montre qu'une valeur élevée réduit la quantité de termes peu pertinents pour un cours. L'augmentation du nombre de supports en entrée, sans changer le nombre de séances demandées, augmente le nombre de termes dans les clusters, les rendant plus difficiles à lire. Cependant, les termes conservés restent pertinents concernant le sujet abordé. On peut supposer qu'il est nécessaire d'augmenter la valeur de β au delà de 1.00 dans ce cas précis.

Le scénario n°5 traitant d'un autre sujet confirme néanmoins que la stratégie haute, un β élevé, et la classification ascendante hiérarchique produisent des clusters utiles pour la construction d'un nouveau cours. De manière générale, le pré-traitement sémantique et le graphe ont correctement retranscrit le contexte des documents et les clusters ont agrégé des termes pertinents (H1, H2, H6a-c, H8a). L'hypothèse H7a semble en partie confirmée du fait que la nature des documents n'a pas spécifiquement changé les résultats : les clusters générés contiennent des termes logiques mais leur taille les rendent difficiles à interpréter. L'hypothèse H6b est donc invalidée étant donné que les scénario n°3 et n°5 montrent que la quantité de documents a un impact sur la lisibilité des clusters. Une étude plus approfondie est requise pour déduire le rapport idéal entre la quantité de termes, la quantité de documents, la valeur de β, et le nombre de clusters.

L'interprétation des clusters des scénarios n°1 et n°5 confirme que quelques termes sont liés sémantiquement, et proposent une organisation logique (H5). La transposition directe de certains clusters du scénario n°1 en sections de cours contribue à valider partiellement l'hypothèse H4b.

Le deuxième questionnaire a également confirmé ces résultats encourageants sur la qualité des clusters produits par la méthode CREA et les paramétrages actuels (H4b). Les remarques concernant le bruit produit par quelques termes et l'ordre des termes pointent la difficulté à lire les clusters. Afficher les clusters de termes dans un tableau, implique d'ordonner les termes dans les cases du tableau (H9). Cette contrainte impose une lecture séquentielle qui devient plus difficile lorsque le nombre de termes par cluster est élevé, et ne permet pas d'avoir une vision globale du cluster. Un travail sur la visualisation des données pourrait permettre d'améliorer l'affichage, et simplifier la réutilisation des notions par l'enseignant.

Les hypothèses validées (✓), invalidées (X), ou à confirmer ultérieurement (?) sont résumées dans le tableau 4.18.

H1	H2	H3	H4			H5	H6			H7			H8			H9
			a	b	c		a	b	c	a	b	c	a	b	c	
✓	✓	✓	✓	✓	?	✓	✓	X	?	?	✓	✓	✓	✓	✓	✓

Tab. 4.18 – Hypothèses validées (✓), invalidées (X), ou à confirmer ultérieurement (?)

4.4.2 Limites de la méthode CREA

Plusieurs limites ont été mises en évidence par les diverses expérimentations. Nous les discutons ici afin de mieux comprendre le contexte qui les a provoquées et proposer un cadre d'usage bien défini pour la méthode CREA. Nous étudions d'abord les limites structurelles, issues des tests de chaque étape et sous-étape, puis, nous étudions les limites fonctionnelles par rapport aux clusters produits.

Limites structurelles

En analysant les premières étapes de la méthode CREA, l'extraction du texte (PI.1) s'appuie sur les outils de reconnaissance optique des caractères. Nous avons utilisé le fonctionnement permettant d'extraire tous les caractères sans distinction de positionnement, ce qui fragilise l'étape de désambiguïsation (PI.3) : les en-têtes et en-pieds, les pages de garde, et autres textes autour du corps des documents se retrouvent mélangés au contenu sans distinction, ce qui modifie inévitablement le score de cohérence des termes. Un document avec beaucoup de méta-données visibles et répétées risque d'ajouter du bruit dans les termes désambiguïsés, et de le propager jusqu'aux clusters finaux. Un document dont le texte est difficilement numérisable génère le même problème qu'un document très court : peu de termes sont désambiguïsés, voire, le résultat est faux à cause du contexte insuffisant pour lever les ambiguïtés. Actuellement, il est nécessaire de fournir des documents contenant suffisamment de texte, si possible avec le moins de méta-données visibles (en-têtes, pages de garde, ...).

L'étape suivante de nettoyage des textes (PI.2) a été réglée à partir de notre scénario de référence où certains termes nous obligeaient à conserver des classes de mots (« *base de données* », en l'absence du « *de* », est considéré comme deux termes « *base* » [dans le sens de *fondement*] et « *donnée* »). Afin de reconnaitre certains termes spécifiques, comme par exemple la notion de *CRUD* (Create Read Update Delete) formée de quatre verbes, nous avons également choisi de conserver les verbes. Bien que certains verbes intéressants ont été extraits avec succès (par exemple *sauvegarder*), ce choix a également ajouté des verbes inutiles, voire provoquant du bruit (par exemple *concerner*). Cette étape dépend du vocabulaire du domaine, et nécessite d'être paramétrée selon le champ lexical visé.

BabelFy étant conçu pour reconnaitre des textes écrits par l'humain, il n'est pas adapté à la reconnaissance de code ou d'autres formats s'appuyant sur du texte. Typiquement, dans le cadre de cours d'informatique, du code peut être présent au milieu des explications afin de servir d'exemple ou d'illustration. Bien que certains langages comme COBOL ont tenté d'imiter la structure des phrases, BabelFy n'est pas capable de reconnaitre les langages de programmation et extrait uniquement les instructions disposant d'un article Wikipedia (telles que *printf*).

Les graphes d'impact mutuel (PII.2) générés illustrent que la méthode CREA est sensible aux formats des documents insérés. Les documents au format diapositives se retrouvent plus facilement autour de l'ensemble central du graphe, tandis que les documents au format texte en sont éloignés. Cela pourrait provenir du vocabulaire plus contrôlé des diapositives : en construisant les diapositives, plusieurs règles implicites nous incitent à limiter la quantité de mots et surtout à mettre en avant les termes importants. À l'inverse, un document au format texte permet de rédiger des phrases beaucoup plus complètes et donc avec un champ lexical bien plus large. Bien que les étapes de pré-traitement sémantique visent à lier les synonymes et séparer les homonymes, on peut observer des variations dans les quantités et proportion de termes désambiguïsés et filtrés selon le format des documents insérés. L'étape de normalisation des occurrences dans la matrice n'a aucun effet dans le cas où les synonymes ne

sont pas correctement fusionnés, et contribue à éloigner un peu plus les documents utilisant chacun un synonyme différent. L'enchaînement de ces limitations implique des métriques moins fiables après les opérations de l'analyse de concepts formels. Il est nécessaire d'approfondir les tests pour s'assurer de la raison provoquant cette distanciation.

Enfin, les clusters (PII.3) sont construits en utilisant la classification ascendante hiérarchique générant des clusters non-recouvrants. Nous ne faisons que forcer le nombre de clusters, mais pas l'équilibre des clusters en nombre de termes contenus. L'absence de cette contrainte a entraîné la génération de clusters peu, voire pas, utilisables à cause de l'excès de termes dans certains d'entre eux (deux clusters contenant plus de 30 mots et six clusters contenant moins de 10 constituent une disproportion où l'enseignant peut avoir du mal à comprendre les notions abordées dans certains d'entre eux). La quantité de termes fournis à la technique de clustering est importante, tout comme la quantité de documents en entrée. Il est donc nécessaire de surveiller ces valeurs, et éventuellement modifier le nombre de documents en entrée ou de termes manipulés. Faire évoluer la valeur de β au delà de 1.00 semble également être une piste d'amélioration pour la lecture des clusters de sortie. Une étude sur les relations entre le nombre de documents, la quantité de termes contenus, la valeur de β, et le nombre de clusters permettrait de surmonter ces difficultés en guidant l'enseignant vers des proportions et valeurs adaptées.

Limites fonctionnelles

La première limite concerne les données que nous avons utilisées : des supports de cours sur le développement web en PHP, donc du domaine informatique. Les outils informatiques étant mieux intégrés aux bases de connaissances informatisées que d'autres domaines (surtout dans le cas précis de BabelFy/BabelNet fondés sur la base de connaissances Wikipédia qui s'appuie elle-même sur les technologies PHP et MySQL présentées dans les cours insérés), il est normal que des cours d'informatique soient correctement reconnus. Le scénario n°5 montre néanmoins qu'un domaine beaucoup plus spécialisé au croisement de la gestion des processus et de la gestion des connaissances est capable d'être reconnu. Tant que le domaine est suffisamment renseigné dans BabelNet et Wikipedia, il n'y a pas de contre indication particulière.

Plusieurs des informaticiens interrogés dans le questionnaire (voir les sous-section 4.2.2 et 4.3.5) ont eux aussi relevé des difficultés liées au bruit causé par des termes peu liés au sujet étudié. Deux parmi eux ont indiqué que quelques clusters étaient peu cohérents, l'un précisant que quelques termes n'avaient pas de lien évident avec les autres. Un troisième a également indiqué que l'ordre des termes lui semblait important. La lecture des clusters pour en déduire des liens entre les termes dépend évidemment de chaque individu et implique un socle minimal de connaissances. Il est donc assez peu probable qu'un débutant du domaine puisse exploiter les données générées par la méthode CREA pour construire un cours. De plus, l'affichage sous forme de tableau impose une lecture séquentielle des termes, ce qui limite les possibilités d'interprétation et donc la créativité.

Dans la phase d'analyse structurelle, le choix des stratégies et de la valeur de β lors de l'analyse de concepts formels (PII.1) permet de sélectionner les termes selon leur fréquence. Nous espérions pouvoir sélectionner avec la stratégie haute les termes les plus fréquents pour chaque cours afin d'en déduire une carte listant l'ensemble des notions abordées dans le corpus, et avec la stratégie basse les termes les moins fréquents, donc ceux caractérisant le mieux certains cours. Les termes produits se sont avérés être les mêmes dans les deux stratégies, bien que les relations entre termes et documents soient différentes. Mais plus le β augmentait, plus la stratégie basse perdait de termes jusqu'à atteindre moins de 10 termes peu utiles. Seule la stratégie haute répond actuellement à nos prévisions, et il n'est donc pas possible en l'état de classifier les documents par spécificité.

Enfin, une dernière limite concerne les performances en général. L'implémentation actuelle des outils fait qu'insérer 18 documents en entrée nécessite plusieurs heures sur un ordinateur personnel pour calculer les stratégies, les treillis, et les métriques associées. Bien qu'il s'agisse d'une implémentation purement expérimentale dont certains traitements pourraient être optimisés, il est nécessaire de limiter le nombre de documents en entrée et le nombre de termes contenus. Cette limitation n'a pas été étudiée en détails, mais elle est liée aux relations entre le nombre de documents, la quantité de termes contenus, la valeur de β, et le nombre de clusters. Ceci conforte l'intérêt d'étudier ces relations.

4.4.3 Discussions sur la méthodologie d'évaluation

En proposant la méthode CREA pour répondre à un problème dans le domaine du système d'information, il était obligatoire de s'orienter vers une démarche de *design science* spécialisée [49][87]. Afin de produire un artefact conforme, nous avons développé, testé, et corrigé ou modifié diverses parties en plusieurs cycles, comme indiqué par le « *Information Systems Research Framework* » dans [49].

L'objectif de démontrer la possibilité de réutiliser des fragments de cas passés, dans le contexte des processus à forte intensité de connaissances, est impossible à totalement généraliser de par la nature même de ces processus mais aussi du point de vue des cas. Cependant, tester l'artefact sur plusieurs cas permet tout de même d'obtenir une indication sur son efficacité. Nous avons donc décidé de tester la méthode CREA dans plusieurs situations afin de vérifier ses limites et attester de son efficacité sur plusieurs cas proches. Cette méthode se rapproche des tests structurels(*white box*) et fonctionnels (*black box*). Les tests structurels visent à vérifier que chaque composant de l'artefact fonctionnent comme espéré en vérifiant des métriques précises dans ces composants. Les tests fonctionnels visent à vérifier que l'artefact répond correctement aux besoins sans connaitre le fonctionnement interne.

Nous avons préparé plusieurs scénarios pour observer les limites de la méthode CREA, s'assurer que les cas d'erreurs sont bien remontés, et vérifier si la méthode peut fonctionner sur plusieurs données suffisamment différentes. Un scénario de référence a tout d'abord été établi (le scénario n°1) afin de servir aux réglages de plusieurs

paramètres techniques, et obtenir des résultats suffisamment satisfaisants. Ensuite, un second scénario a introduit une anomalie (le scénario n°2) pour vérifier que la méthode réagit correctement à plusieurs niveaux en modifiant ses données de sortie. Un troisième scénario a augmenté le nombre de données correctes en entrée (le scénario n°3) pour s'assurer de la robustesse de la méthode, et confirmer les réglages des paramètres techniques. Un quatrième scénario s'est concentré sur la correction d'un document parmi ceux au format texte long (le scénario n°4) pour s'assurer que le retrait de chapitres et sections hors sujet améliorent bien la qualité du document sur le graphe d'impact mutuel. Enfin, un cinquième scénario a changé les données d'entrée, toutes étant valides (le scénario n°5), pour s'assurer que la méthode ne répondait pas uniquement au scénario de référence et à ses dérivés, mais également que la méthode est indépendante de la langue. Bien que ces tests ne soient ni exhaustifs, ni parfaits, nous avons voulu vérifier les composants de la méthode, et les cas courants qui peuvent être rencontrés.

Afin d'obtenir un point de vue externe, mais expert, sur les résultats du cas de référence, nous avons également interrogé des experts du domaine, dont certains avec une expérience dans l'enseignement. Plusieurs individus, avec leurs connaissances implicites distinctes, ont exprimé leur avis sur la qualité des données générées et émettre leurs remarques. Ces points de vues variés permettent également de conforter ou non l'efficacité de l'artefact. L'indice de Rand et l'indice de Rand ajusté ont cependant exposé les limites de l'usage de techniques ne prenant pas assez en compte le contexte pour évaluer les résultats.

Du point de vue des processus à forte intensité de connaissances, la méthode CREA n'étant pas uniquement un processus automatique, mais faisant intervenir les connaissances implicites de l'utilisateur plusieurs fois, nous traitons un problème KIP avec un KIP. En effet, l'utilisateur doit d'abord sélectionner des documents qui lui semblent pertinents au début de la méthode, ce qui implique une certaine quantité de connaissances de sa part concernant le domaine étudié. Ensuite, la méthode ne pouvant que le guider en lui proposant des fragments, il doit de nouveau faire appel à ses propres connaissances implicites pour créer des liens logiques entre les termes des clusters afin de faire émerger la solution adaptée à son cas.

Néanmoins, une remarque faite dans [100] indique que les recherches dans le domaine de la gestion des connaissances suivant le modèle du *design science* se concentrent particulièrement sur « *les fonctionnalités plutôt que sur les facteurs tels que le comportement des individus, la structure organisationnelle, la confiance ou les communications préférées* » (« *functionalities rather than also considering factors like individuals' behavior, organizational structure, trust or preferred communications* » [100]). Cette remarque s'applique en partie à la méthode CREA étant donné que l'objectif principal est de disposer de fonctionnalités précises (générer un graphe d'impact mutuel et des clusters de termes). L'expérience personnelle d'enseignement a tout de même été utilisée lors de la conception et le développement de la méthode : en effet, le comportement habituel nécessaire à la construction d'un cours implique inévitablement de rechercher des sources qui seront lues, analysées, et réutilisées.

4.4.4 Discussions sur la méthode CREA et les domaines de la gestion des connaissances et des processus à forte intensité de connaissances

Afin de replacer la méthode CREA dans le contexte de la recherche, nous discutons de son placement au sein des deux domaines étudiés dans le chapitre 2, à savoir la gestion des connaissances et les processus à forte intensité de connaissances.

La méthode CREA et la gestion des connaissances

Par rapport à la gestion des connaissances, plusieurs rôles présentés dans [68] peuvent être transposés sur les différents intervenants.

En admettant que les documents réutilisés sont produits par d'autres personnes agissant comme des « *producteurs de connaissances* » [68], la méthode CREA peut aider un enseignant lorsqu'il adopte la posture « *d'intermédiaire de connaissances* » [68] : en effet, les étapes de la méthode CREA suivent le cheminement (nettoyage, indexation, empaquetage sous forme de clusters) général, et l'enseignant peut paramétrer plus finement chacune des étapes pour ses besoins spécifiques. Les clusters de sortie de la méthode CREA permettent également à l'enseignant de devenir « *consommateur de connaissances* » [68] en obtenant une synthèse des connaissances qu'il doit adapter à la classe à laquelle il souhaite enseigner.

Les différentes situations également présentées dans [68] peuvent aussi être observées lors de l'utilisation de la méthode CREA.

La méthode CREA permet dans un premier temps, grâce au graphe d'impact mutuel, de répondre à la situation où « *un novice cherche une expertise* » [68]. En effet, l'ensemble central permet à un enseignant de s'assurer qu'il s'agit bien du sujet général qu'il vise, puis, en s'éloignant de cet ensemble, il peut découvrir des termes plus spécifiques et pointus. Un enseignant découvrant un nouveau domaine qu'il doit enseigner peut commencer à retrouver les termes clés du domaine en question.

Plus communément, un enseignant réutilisant des documents produits par d'autres enseignants ou des formateurs sera vu comme un « *praticien collaboratif* » [68]. Les autres enseignants et formateurs sont fonctionnellement très proches, mais ceux-ci sont dans des organisations distinctes (universités et industrie, par exemple). Les difficultés rencontrées seront donc que l'enseignant doit décider quelles connaissances répondent au mieux à son propre contexte.

L'enseignant peut également être vu comme un « *explorateur tiers de connaissances* » [68]. Les documents sélectionnés traitent à priori du sujet visé (par exemple grâce au titre), mais le contexte précis dans lequel ils ont été générés est peu connu (formation industrielle, articles et études scientifiques, ...). Les contextes beaucoup trop différents entre les producteurs et consommateurs rendent la partie *collaborative* caduque : un support peut être très ancien et employer des termes et concepts différents, mais qui restent cependant pertinents. Cet écart temporel, ou plus largement contextuel lorsque les organisations sont beaucoup trop différentes, oblige l'enseignant à bien étudier les limites des documents en question (voire de les retirer du corpus sélectionné).

Parmi les trois méthodes de réutilisation présentées dans [88], on peut affirmer

que la méthode CREA n'est ni un *verbatim* (trop de variations dans les situations), ni de la *création* étant donné qu'il s'agit de réutiliser l'existant, mais plutôt qu'elle aide un enseignant à effectuer une *synthèse* des connaissances en combinant plusieurs documents qu'il doit adapter au cas courant.

Ainsi, la méthode CREA s'insère effectivement dans plusieurs aspects de la gestion des connaissances, en particulier celui de la réutilisation de connaissances.

La méthode CREA et les processus à forte intensité de connaissances

Du point de vue des processus à forte intensité de connaissances, la méthode CREA est un outil permettant tout d'abord de visualiser le contexte des connaissances manipulées.

La reconnaissance des termes et concepts manipulés par les techniques de TAL, ainsi que leur filtrage puis leur organisation sous forme de graphe d'impact mutuel grâce à l'ACF, permettent de s'assurer de la pertinence du contexte d'exécution et des documents utilisés. Un utilisateur ayant oublié certaines informations clés pourra aisément les redécouvrir grâce à l'ensemble central, et éventuellement avec les anneaux un peu plus secondaires autour. Inversement, un utilisateur connaissant exactement les informations requises, mais souhaitant s'assurer de la pertinence des documents qu'il utilise pourra s'en assurer grâce au graphe d'impact mutuel. Fixer initialement les buts étant un pré-requis à l'exécution des processus à forte intensité de connaissances (voir chapitre 2 et [21] expliquant que les processus à forte intensité de connaissances sont « *orientés buts* » et « *dirigés par les connaissances* »), s'assurer de la qualité du contexte d'exécution est donc essentiel.

Parmi les six défis exposés dans [8], la méthode CREA permet d'apporter une réponse au troisième (« *Comment intégrer les informations de contexte lors de la conception d'un KIP ?* ») en proposant de vérifier le contexte au fur et à mesure de l'ajout et des corrections des documents. La véritable contrainte est actuellement technique et se situe sur le nombre de documents : seuls des processus à forte intensité de connaissances manipulant quelques documents (moins d'une vingtaine) peuvent être pris en charge dans des délais raisonnables.

Le sixième défi, principalement visé dans cette thèse, présenté dans [8] (« *Comment explorer et réutiliser des fragments de processus dans les KIPs ?* ») est traité au travers de la réutilisation de connaissances et des liens entre les termes et concepts. L'ACF permettant de retranscrire les liens entre des concepts [127], son usage sur les documents permet d'analyser et lier les termes dont les apparitions sont communes dans l'ensemble du corpus. La fabrication de ces liens s'appuie en pratique sur les connaissances implicites des auteurs des documents : chaque auteur produit des textes en manipulant des concepts et des termes associés selon sa construction intellectuelle. Ces liens sémantiquement compris par les lecteurs, mais indirectement manipulés par l'ACF, sont les fragments de processus que nous cherchons à réutiliser.

La similarité conceptuelle permet par la suite de mesurer ces liens entre les termes. Cette quantification sous forme de matrice peut être exploitée par d'autres techniques d'analyse de données, telles que le clustering, pour faire apparaître des regroupements

de termes. Ainsi, les fragments réutilisables produisent concrètement, grâce aux dernières étapes de la méthode CREA, des clusters lisibles par l'utilisateur.

Ces clusters présentent à l'utilisateur des termes intrinsèquement liés dans l'ensemble du corpus documentaire d'origine. Selon l'expérience de l'utilisateur, ces clusters lui suggèreront des assemblages utiles et logiques qu'il pourra réutiliser tels quels, ou au contraire, qu'il pourra modifier pour s'adapter à d'autres critères non pris en compte par les simples textes.

La méthode CREA est donc un outil permettant à la fois de s'assurer de la pertinence du contexte d'exécution d'un processus à forte intensité de connaissances à partir des documents manipulés, mais surtout de retrouver les connaissances implicites dans ces documents pour en matérialiser les liens dans des regroupements de termes qu'un utilisateur peut réutiliser lors de l'exécution de son nouveau cas.

5. CONCLUSION

Dans ce chapitre, nous effectuons une synthèses des contributions proposées dans cette thèse, nous rappelons les cas d'usages possibles, et nous discutons des menaces de validité. Nous proposons ensuite plusieurs perspectives, dont une extension succinctement testée de la méthode CREA pour pouvoir prendre en compte l'axe temporel et ordonnancer les clusters.

Sommaire

5.1 Synthèse

Dans cette thèse nous avons étudié la question de la réutilisation et de la pertinence du contexte dans le cadre de la gestion des connaissances et des processus à forte intensité de connaissances. Nous avons conçu et implémenté la méthode CREA permettant de réutiliser des documents existants afin de proposer des séances constituées de plusieurs notions à un enseignant. Cette méthode peut également être utilisée à d'autres fins comme vérifier la pertinence des documents vis-à-vis d'un sujet ou visualiser graphiquement les mots clés d'un corpus documentaire.

5.1.1 Rappel des contributions

La méthode CREA, Case REuse and Adaptation, vise à permettre de réutiliser des supports de cours et autres documents existants afin de proposer des ensembles de termes sous forme de séances, tout en s'assurant de la pertinence des documents sélectionnés. Pour cela, la méthode CREA s'appuie tout d'abord sur un pré-traitement sémantique des documents afin d'en extraire les notions les plus pertinentes pour l'enseignant sous forme de concepts et d'entités nommées issus du réseau sémantique BabelNet. Le résultat de ce pré-traitement sémantique permet ensuite d'effectuer des analyses plus poussées sur deux métriques en s'appuyant sur les liens entre les documents et les termes illustrant une partie des connaissances tacites exploitées par les auteurs desdits documents.

Afin de sélectionner les documents les plus adaptés, une première contribution concerne la visualisation des notions retrouvées dans l'ensemble du corpus documentaire et la pertinence des documents par rapport aux notions clés retrouvées. Le graphe illustré par la figure 5.1 permet à un enseignant de visualiser les notions les plus importantes dans son corpus documentaire (les nœuds rouges et oranges dans cet exemple), mais également la pertinence des documents (les nœuds gris indiquant la référence du document) sous forme de distance des uns par rapport aux autres et à l'ensemble central de notions. Cette cartographie s'appuie sur l'analyse de concepts formels et l'impact mutuel afin de comprendre les relations entre les termes précédemment extraits et les documents les contenant.

L'enseignant se voit proposer plusieurs séances, et les notions à aborder dans chacune d'elle, à l'issue de plusieurs traitements. Ces séances sont représentées sous forme de clusters de termes comme illustré par la figure 5.2. Pour cela, nous avons utilisé la stratégie haute de binarisation, l'analyse de concepts formels, le calcul de la similarité conceptuelle, et la classification ascendante hiérarchique afin d'obtenir des regroupements issus des termes les plus fréquents du corpus documentaire. Les clusters générés ne sont pas encore organisés temporellement et restent une proposition que l'enseignant peut encore adapter à son cas.

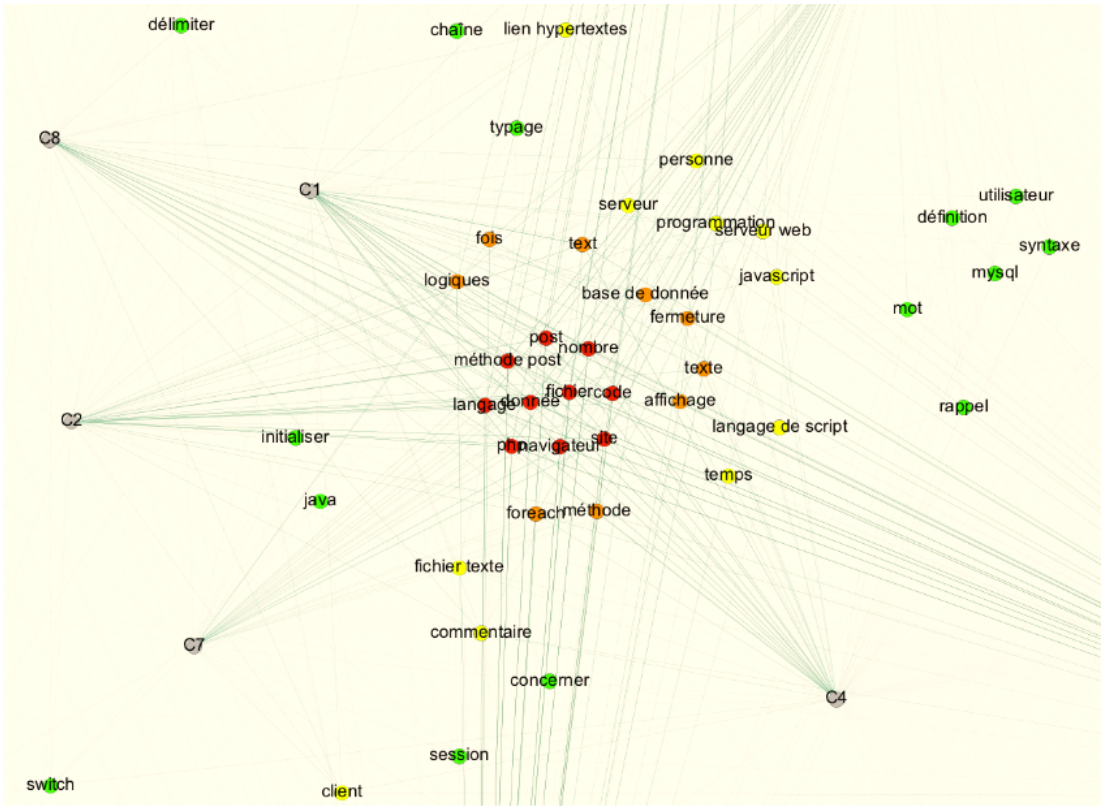


FIG. 5.1 – Graphe d'impact mutuel : notions centrales et documents

Stratégie Haute (β = 1,00)

1	php	code	fois	post	jour	foreach	cle	classe	class	mysqli	
2	page web	navigateur	serveur web	texte	concerner	délimiter	utilisateur	associer	personne	machine	mysql
3	url	langage	case	fermeture	session	chaîne	entête	avoir accès			
4	fichier	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client				
5	typage	mot	moteur	affiche	transaction	visiteur					
6	base de donnée	insert	varchar	null							
7	xml	configuration	composer	doctype							
8	donnée	text	méthode post	programmation	site	langage de script	list	méthode	timestamp	files	

FIG. 5.2 – Clusters de termes : proposition de découpage en séances

5.1.2 Usages possibles

La méthode CREA vise à guider un enseignant dans la conception d'un nouveau cours en proposant des clusters de notions issus de supports de cours existants, le tout en s'appuyant sur de nombreux outils automatiques. Cependant, d'autres cas d'usages s'appuyant sur la réutilisation en découlent également.

La méthode nécessite de réunir des supports de cours contenant du texte, et d'indiquer le nombre de séances voulues. Les supports insérés sont tout d'abord traités par plusieurs techniques de traitement automatique du langage pour en extraire des termes illustrant les notions abordées. Puis, une combinaison d'analyse de concepts formels et de clustering permet d'en extraire un graphe représentant les termes les plus représentatifs du corpus, la pertinence des documents de ce corpus, ainsi que des clusters regroupant les termes réutilisables par l'enseignant. L'enseignant obtient ainsi plusieurs ensembles de termes liés à partir desquels il peut construire son propre cours selon ses attentes et exigences.

Afin d'améliorer la qualité, un enseignant peut éliminer des documents isolés dans le graphe d'impact mutuel et relancer la méthode, ou raffiner manuellement les termes entre chacune des étapes de la méthode. Une contrainte imposant d'éliminer certains termes peut également être appliquée en supprimant ces termes durant les traitements, ou en retirant les documents contenant ces termes.

Un enseignant souhaitant obtenir une liste de termes clés d'un domaine peut donc s'appuyer sur le graphe d'impact mutuel généré par plusieurs documents représentant ce domaine. Les cas d'usages peuvent concrètement être vus ainsi :

1. Construction d'un nouveau cours à partir d'anciens cours
   - Entrées : *Documents* (supports de cours d'autres enseignants), *Nombre de séances*
   - Sorties : *Graphe d'impact mutuel* (cartographie des notions du corpus documentaire, et pertinence des documents), *Clusters de notions* (organisation des notions sous formes d'autant de séances que demandées)
2. Comparaison de son propre cours par rapport à d'autres cours et aux productions académiques
   - Entrées : *Documents* (son propre support de cours + supports de cours d'autres enseignants, articles de recherche, livres, ...)
   - Sorties : *Graphe d'impact mutuel* (cartographie des notions du corpus documentaire, et positionnement de son propre cours parmi les autres documents)
3. Découverte d'un domaine académique
   - Entrées : *Documents* (articles de recherche, livres, ...), *Paramètres des stratégies* (choix de la stratégie directe, haute, ou moyenne, et choix du β selon la granularité visée)
   - Sorties : *Graphe d'impact mutuel* (cartographie des notions du corpus documentaire)

4. Filtrage d'un corpus documentaire pour ne garder (ou retirer) que les documents traitant de certains thèmes précis
   - Entrées : *Documents*, *Notions et termes recherchés*
   - Sorties : *Graphe d'impact mutuel* (cartographie des liens entre les termes et les documents)
5. Aide à la composition de l'index d'un ouvrage
   - Entrées : *Document(s)*, *L'ouvrage rédigé, éventuellement les principales sources utilisées ou les plus influentes*
   - Sorties : *Liste de termes* (liste issue du pré-traitement sémantique, voire, filtrée suite à la stratégie haute de binarisation)

Au delà du domaine d'application de l'enseignement supérieur et de la recherche, la visualisation de la pertinence des documents peut être utilisée dans les situations où la sélection de documents basée sur les concepts et termes employés est requise, de même pour la visualisation des sujets abordés dans un ensemble documentaire. La prise de décision basée sur la pertinence de documents, l'absence/présence de termes dans ces derniers, ou les sujets d'intérêts d'un corpus documentaire grâce à la première contribution de la méthode CREA est donc tout à fait réaliste. À l'inverse, l'organisation sous forme de clusters des termes avec le fonctionnement actuel ne semblent pas transposable à d'autres usages contemporains : rechercher les maladies causant certains symptômes nécessite des traitements supplémentaires pour obtenir des clusters regroupant les combinaisons utiles, par exemple. Un usage original des premières étapes de la méthode consiste à aider un auteur en recherchant les termes clés dans son ouvrage, et éventuellement des sources qu'il utilise, afin de lui suggérer des mots-clés pour son index. La simple recherche de noms ou de verbes génèrera une liste particulièrement longue, mais BabelFy couplé à la stratégie de binarisation peuvent déterminer quels termes et concepts sont les plus appropriés. La méthode CREA dans son ensemble est donc actuellement dédiée à un usage appliqué à l'enseignement supérieur et à la recherche.

5.1.3 Menaces de validité

Sept types de menaces de validité sont présentés dans [34] : validité de la conclusion, validité interne, validité de la construction, validité externe et transférabilité, crédibilité, fiabilité, et confirmation.

La validité de conclusion concerne la mesure de changements significatifs dans les résultats grâce à l'usage de notre artefact. Étant donné que la méthode CREA n'a pas été comparée à d'autres méthodes traitant du même sujet, et qu'elle ne produit pas un cours en entier, ce type de validité est difficile à vérifier. Comme nous l'avons précédemment vu, la comparaison des partitions n'a pas donné de résultat satisfaisant de par la nature des termes et clusters manipulés : la combinaison de termes implique une interprétation avec un contexte. La comparaison des résultats produits par un humain variera indéniablement selon la personne qui fera l'évaluation et son niveau d'expertise (qui évolue lui aussi au cours du temps). Il est donc possible qu'une même personne n'ait pas le même avis sur les mêmes résultats produits à plusieurs instants différents.

La validité interne concerne les facteurs ayant un impact sur les résultats produits par l'artefact. Dans notre cas, nous avons étudié ces limites grâce aux différentes validations structurelles, puis nous avons discuté des différents problèmes lors des discussions. Typiquement, les types de documents (format diapositive ou texte) ont un impact sur les graphes d'impact mutuel, mais également sur la quantité de termes dans les clusters de sortie. Il est nécessaire de chercher à limiter le nombre de termes, voire, de remonter plus en amont et comprendre les relations entre le nombre de documents, la quantité de termes contenus, la valeur de β, et le nombre de clusters.

La validité de la construction s'intéresse au phénomène que l'artefact est censé traiter, ainsi qu'aux effets impactant les résultats. Nous nous sommes intéressés à la réutilisation des connaissances dans le cadre de la construction de cours, et nous avons effectivement constaté que les connaissances tacites ont un effet sur l'interprétation des résultats. Les différentes étapes de la méthode CREA visent à reproduire certaines manipulations humaines (extraire le sujet central des documents, indiquer quel document est hors sujet, etc) et produisent des résultats intéressants qui peuvent être encore améliorés. Nous avons bel et bien été confrontés aux difficultés liées aux connaissances tacites, tout en faisant une extraction et des filtrages pertinents concernant les termes. Notre artefact est donc valide de ce point de vue.

La validité externe et la transférabilité concernent la généralisation dans d'autres cas et situations. La généralisation (du point de vue fonctionnel) à l'ensemble des domaines d'études, voire, à l'ensemble des processus à forte intensité de connaissances n'a pas été démontrée. On peut tout de même supposer que BabelFy s'appuyant sur des bases de connaissances larges, tous les sujets seront probablement pris en charge tant qu'ils sont suffisamment documentés, mais nous ne pouvons pas l'assurer. Des scénarios touchant à des domaines moins industriels et beaucoup plus académiques pourraient être testés et évalués avec des experts du domaine (au lieu de PHP ou de *statecharts*, des documents traitant de philosophie pourraient être testés).

En dehors du cadre de la construction de cours, nous n'avons ni testé l'utilisation d'autres bases de connaissances plus spécifiques à l'industrie, ni l'ensemble des cas d'usages précédemment présentés.

La généralisation (du point de vue plus technique) à n'importe quel scénario n'est pas non plus démontrée. Les scénarios testés dans cette thèse ont permis d'achever plusieurs essais ouvrant une voie intéressante, mais pas totalement validée. En effet, la quantité de documents en entrée semble impacter les résultats et ne permet pas de totalement confirmer les hypothèses traitant des types de documents. Des scénarios testant tout d'abord uniquement des supports de cours sur un sujet actuellement maîtrisé, puis uniquement sur les articles de recherche ayant mené au développement et à la maturité de ce sujet permettraient de mieux observer les effets sur les graphes d'impact mutuel et les clusters résultant.

Ces scénarios souffrent néanmoins de l'évolution du vocabulaire au cours du temps : un sujet ayant mis plusieurs années à arriver auprès du grand public peut avoir vu énormément d'évolution dans son vocabulaire, dont certains termes pourraient être réutilisés dans un autre contexte proche (et donc avoir une autre définition), ou avoir tout simplement disparu. Par exemple, les « *moniteurs transactionnels* » (permettant de traiter des transactions) n'ayant plus exactement d'existence dans les systèmes classiques actuels (car intégrés aux « *systèmes de gestion de bases de données* » et à leurs « *connecteurs* »), leur mention dans les articles, ouvrages, ou anciens supports de cours ne sera pas toujours comprise par le lecteur, voire, la création de liens sémantiques sera plus complexe si la définition n'est pas parfaitement déclarée dans la base de connaissances.

Des scénarios utilisant des supports de cours récents sur un sujet contemporain, en comparant leurs résultats par rapport à ceux d'articles de recherche, pourraient être corrects. Il faut néanmoins s'assurer que les supports de cours respectent les exigences concernant la quantité minimale de texte. On peut également poser la question sur la construction de ces supports : étant donné qu'ils sont déjà des résumés des articles, ils pourraient tout simplement servir à confirmer la structure produite par la méthode CREA à partir des articles.

Concernant la langue des documents, nous n'avons pas testé au sein d'un même scénario l'usage de plusieurs langues distinctes. BabelFy permet déjà de retrouver un concept quel que soit sa langue, mais l'ensemble du document doit être dans la même langue (ou alors il faut manuellement tagger les mots). Bien qu'il semble très probable qu'un tel scénario produise des résultats corrects, une confirmation pratique pourrait être utile.

La crédibilité concerne la confiance dans les conclusions trouvées. La démarche effectuée étant inspirée du design science, plusieurs expériences ont permis d'affiner les résultats et mieux comprendre l'origine de certaines anomalies (par exemple l'impact de la taille des documents sur les graphes d'impact mutuel). Les expériences successives ont permis d'observer des améliorations notables lors de corrections, et indiquer des facteurs à surveiller à l'avenir.

La fiabilité concerne la cohérence des résultats. Comme nous l'avons expliqué, une confiance absolue dans la méthode CREA est actuellement impossible, mais les

résultats actuels et les fondements théoriques des outils utilisés indiquent des cas où les résultats sont cohérents et peuvent être consolidés avec d'autres recherches plus approfondies. Les graphes d'impact mutuel affichent des résultats cohérents et pertinents, et les facteurs les impactant sont de mieux en mieux identifiés. Les clusters nécessitent quant à eux d'autres travaux.

Enfin, la confirmation vise à connaître à quel point les conclusions ont été façonnés par les résultats d'expériences plutôt que par le chercheur et ses biais ou intérêts. L'interprétation des résultats s'appuyant totalement sur nos connaissances tacites, un biais existe. Afin de limiter ce biais, d'autres informaticiens ont également évalué une partie du cas de référence afin de déterminer si ces résultats sont cohérents. Bien qu'il soit impossible d'éviter certains biais d'interprétation, l'objectif de simplifier la lecture d'un corpus documentaire volumineux lors de la construction d'un cours est une motivation très intéressante pour l'ensemble des enseignants-chercheurs.

5.2 Perspectives et améliorations possibles

Les scénarios précédemment testés présentent à la fois des limites, mais également des menaces de validité. Plusieurs pistes de contournement ou de scénarios à réaliser sont maintenant présentés. Nous présentons également une extension à la méthode CREA lui permettant de prendre en charge l'organisation temporelle des clusters pour guider un enseignant jusqu'à l'ordonnancement de ses séances de cours, et ainsi répondre à la dernière hypothèse non testée (H4c).

5.2.1 Pré-traitement sémantique

Le pré-traitement sémantique est une étape reposant sur les techniques de traitement automatique du langage, mais aussi de reconnaissance optique des caractères. Afin de limiter le bruit dans la phase suivante, effacer automatiquement les différentes méta-données affichées dans les en-têtes (et autres pages informatives) serait un bénéfice. Tout comme reconnaitre automatiquement les sections de codes, d'exercice, ou de projet, et les retirer le plus tôt possible, ou au contraire, développer une extension de la méthode CREA dédiée à l'analyse et l'exploitation de ces autres formats.

Afin de fournir à BabelFy des données de qualité, nous avons utilisé TreeTagger pour éliminer des termes peu utiles lors de l'étape de nettoyage des textes extraits (PI.2). Cependant, il pourrait être intéressant de laisser BabelFy reconnaitre les entités nommées, et ensuite seulement d'effacer les termes correspondant exclusivement à certaines classes de mots. Par exemple, au lieu d'annoter les trois mots « *base* », « *de* », « *données* », et ensuite rechercher l'éventuelle entité, il pourrait être pertinent de rechercher d'abord les entités nommées, et ensuite effacer celles qui ne sont pas rattachées à une expression (« *et* » pouvant apparaitre seul, voire pourrait être filtré par BabelFy).

5.2.2 Analyse structurelle

Le calcul actuel des stratégies de binarisation est précédé de quelques étapes visant à répartir correctement les termes par documents pour pouvoir correctement gérer les documents de taille variable. Toutes ces transformations visent à générer un contexte formel que l'ACF manipulera. Or, comme nous l'avons vu, un β à 1.00 semble ne pas suffire à supprimer certains termes peu utiles lorsque l'on insère beaucoup plus de documents. Tester des valeurs supérieures de β pour trouver un coefficient proportionnel au nombre de documents ou de termes permettrait de réduire ce bruit. De manière plus générale, une étude des relations entre le nombre de documents, la quantité de termes contenus, la valeur de β, et le nombre de clusters permettrait de mieux comprendre comment construire les clusters optimaux pour un enseignant en quantité de termes, et contribuer à améliorer les performances à l'exécution de la méthode.

Une autre piste est celle de l'Analyse de Concepts Formels Flous [12][89] permettant de traiter des matrices contenant des valeurs entre $[0, 1]$ et non plus uniquement dans la paire $\{0, 1\}$. Les contextes formels flous générés pourraient apporter des métriques

plus précises concernant les proportions d'origine. Bien que cela supprime théoriquement l'usage des stratégies de binarisation, il est possible de continuer à n'utiliser ces stratégies que pour filtrer les termes de la matrice tout en conservant les valeurs d'origine associés à chacun : le calcul de la stratégie haute permet de sélectionner les termes à conserver depuis la matrice d'origine, afin de construire une nouvelle matrice multivaluée qui sera utilisée avec l'Analyse de Concepts Formels Flous.

Nous avons utilisé la classification ascendante hiérarchique afin de produire des clusters non-recouvrants, or, il existe d'autres techniques de clustering. Des essais mélangeant plusieurs techniques de clustering et une priorisation selon les clusters les plus générés ont été réalisés, et restent à étudier. Ceci permettrait de faire émerger des clusters reconnus par plusieurs méthodes différentes. Notre principale limitation concerne le nombre de dimensions à indiquer à certaines techniques de clustering : nombre de documents ? nombre de termes ? autre(s) métrique(s) ? En projetant nos données avec le *multidimensional scaling* [64][129], pour réduire à deux dimensions nos données, le stress de Kruskal [63][123] généré dépassait les seuils, indiquant une distorsion trop élevée des données pour pouvoir les exploiter. Cependant, la quantité de techniques de clustering nous laisse encore beaucoup de possibilités.

Enfin, les remarques des informaticiens interrogés lors de la présentation des clusters nous ont permis de constater que l'affichage sous forme de tableau n'est pas adaptée pour présenter plusieurs termes. Bien que l'idée de construire avec une phrase à partir des termes d'un cluster semble intéressante, ceci n'est pas recommandé : un cluster contient des termes, mais aucune contrainte grammaticale dans l'organisation de ces termes n'a été prévue. Cependant, améliorer la présentation des résultats au travers d'un affichage sous forme de cercle ou par l'usage de nuages de mots permettrait une lecture plus intuitive de l'ensemble des termes, et ainsi aider l'enseignant à mieux comprendre les notions véhiculées par chaque cluster.

5.2.3 Analyse temporelle : organisation des scénarios

Une piste de solution particulièrement prometteuse concernant l'ordonnancement temporel a été développée et succinctement évaluée. Comme vu dans le projet *Mail of Mine* [24][23], il est possible d'ordonner temporellement le déroulement de certaines activités à partir de traces d'exécutions. Dans le cadre de *Mail of Mine*, l'analyse temporelle s'effectue par la découverte de contraintes dans l'ordre des activités depuis des traces de processus à forte intensité de connaissances.

Dans la méthode CREA, afin d'ordonner nos fragments réutilisables, nous avons développé une troisième phase permettant d'ajouter une information temporelle. Cette extension est partiellement parallélisable étant donné qu'elle exploite les listes de termes non filtrées générées par BabelFy (l'étape de désambiguïsation) pour ses premières étapes, puis les clusters générés en fin d'analyse structurelle. Les figures 5.3 et 5.15 illustrent son fonctionnement et cette parallélisation partielle. À cause du fonctionnement actuel de l'extension, seuls les documents organisés temporellement peuvent permettre d'extraire une dimension temporelle (par exemple des chapitres

successifs introduisent des notions au fur et à mesure, tandis que des articles de recherche respectent en général un format où les notions sont expliquées à un endroit précis, puis des expériences ou démonstrations valident la contribution). La phase d'analyse temporelle extrait tout d'abord les étiquettes temporelles des termes, selon le nombre de séances visées, puis applique ces étiquettes aux clusters. Une sélection des étiquettes les plus fréquentes est effectuée pour chaque cluster, puis le point de vue *cluster* est basculé vers celui des *séances* afin de proposer des clusters pour chacune des séances. Les quatre étapes détaillées sont illustrées sur les figures 5.4 et 5.15. Les sous-sections suivantes donnent plus d'explications sur chacune de ces étapes.

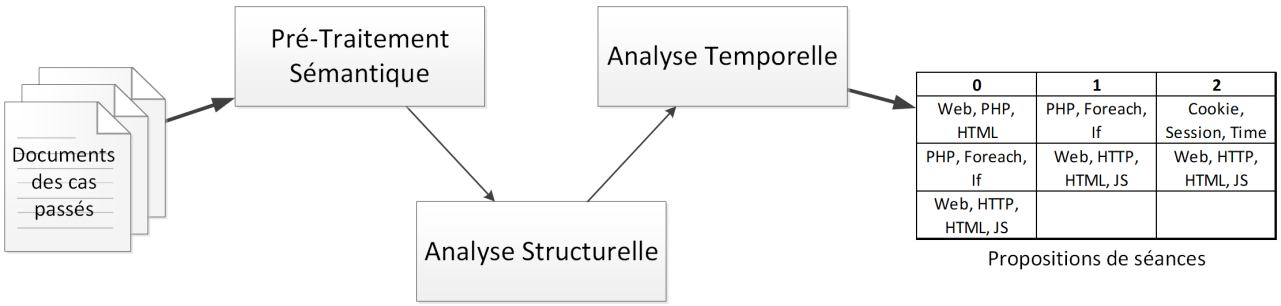


Fig. 5.3 – La méthode CREA étendue avec l'analyse temporelle

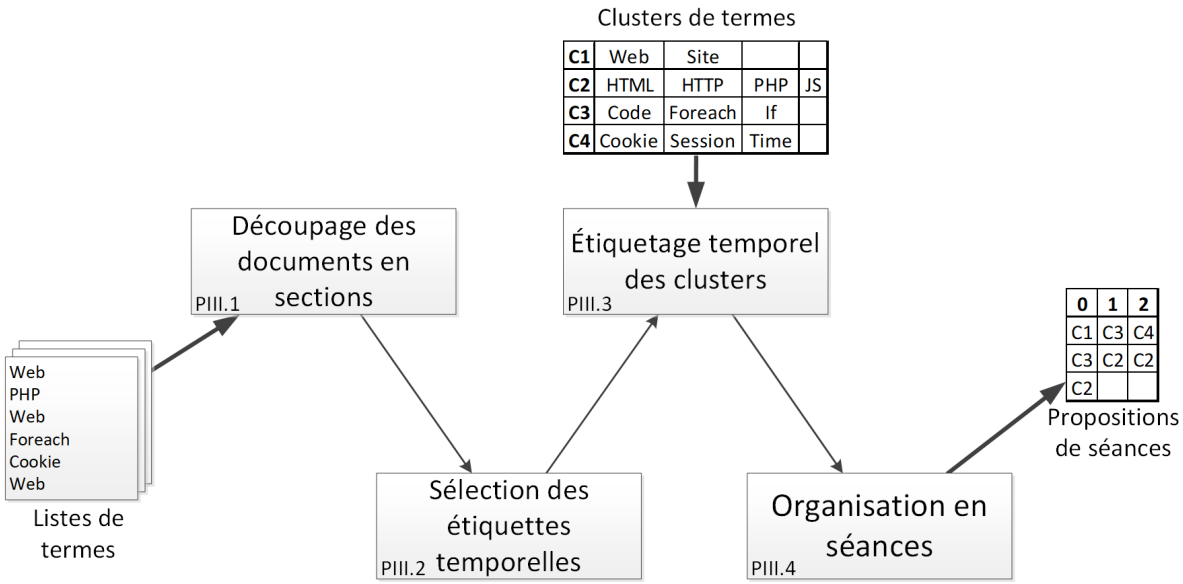


Fig. 5.4 – Les étapes de la phase d'analyse temporelle

Découpage des documents en sections (PIII.1)

Cette étape vise à identifier les proportions de termes par section afin de déduire par la suite quels termes sont spécifiques à certaines sections. Pour cela, tous les documents sont découpés en autant de sections que de séances exigées par l'utilisateur. Afin de conserver le meilleur ordonnancement des termes, nous avons décidé d'utiliser les termes issus de l'étape de désambiguïsation (PI.4) décrite en sous-section 3.2.4, et non pas du filtrage suivant. En effet, le filtrage supprimant les termes non pertinents dans l'intégralité de chaque document, la liste de termes est réduite, entraînant une déformation dans la structure des documents. Bien que les étapes précédents la désambiguïsation suppriment elles aussi des caractères et des mots, la désambiguïsation permet de manipuler des entités reconnues dans des bases de connaissances et donc standardisées dans l'ensemble des documents fournis par l'utilisateur, c'est-à-dire des listes de termes indépendants des problèmes de synonymes et homonymes.

Suite à l'étape de désambiguïsation, une liste de termes reconnus dans des bases de connaissances a été établie. L'utilisateur indique le nombre de séances qu'il souhaite, et la liste de termes est équitablement divisée en autant de sections. Le placement des termes dans la structure initiale est ainsi conservée, tout en permettant un découpage selon le nombre de séances visées. Ce traitement automatique ne nécessite aucune intervention humaine pour corriger ou indiquer où démarre ou finit chaque partie. Chaque occurrence de terme se voit donc attribuer le numéro de la section associée à sa position dans la liste. Étant donné que la nature des documents ne permet pas toujours un tel type de découpage, il est nécessaire de ne travailler que sur des documents respectant une organisation séquentielle. La figure 5.5 illustre le découpage en quatre sections (de 0 à 3) d'un document désambiguïsé sous forme d'une liste de huit termes.

Liste filtrée des termes d'un document

Terme	*Position*	*Identifiant*
php	(0,2)	bn:01753580n
php	(16,18)	bn:01753580n
document	(93,100)	bn:00028018n
document	(155,162)	bn:00028018n
compiler	(298,305)	bn:00021344n
web	(369,371)	bn:00080772n
faq	(458,460)	bn:00033653n
php	(478,480)	bn:01753580n

Liste filtrée et annotée des termes d'un document

Terme	*Position*	*Identifiant*	*N° Section*
php	(0,2)	bn:01753580n	0
php	(16,18)	bn:01753580n	0
document	(93,100)	bn:00028018n	1
document	(155,162)	bn:00028018n	1
compiler	(298,305)	bn:00021344n	2
web	(369,371)	bn:00080772n	2
faq	(458,460)	bn:00033653n	3
php	(478,480)	bn:01753580n	3

Fig. 5.5 – Découpage d'un document en quatre sections

Une fois les termes d'un document associés à une section, une proportion de présence de chaque terme dans chaque section est calculée. Cette proportion permet de connaître la distribution des termes dans chacune des sections d'un document, ce qui contribue à extraire le positionnement des termes les plus importants de l'ensemble des documents. Le calcul de la proportion des termes dans un document s'effectue comme indiqué dans l'équation (5.1).

$$proportion(T, P) = \frac{\sum \textit{ des occurrences du terme } T \textit{ dans la section } S}{\sum \textit{ des occurrences du terme } T \textit{ dans toutes les sections}} \times 100 \tag{5.1}$$

La figure 5.6 illustre les deux transformations successives de la liste filtrée et annotée de termes. Tout d'abord, les proportions d'occurrences des termes dans chaque section sont calculées : *php* apparait *2* fois dans la section *0* sur un total de *3* occurrences, ce qui donne une proportion d'occurrences de *66%* dans cette section, puis, *php* apparait *1* fois dans la section *3* sur un total de *3* occurrences, ce qui donne une proportion d'occurrences de *33%* dans cette section. Les proportions par termes et sections sont ensuite réunies en un seul tableau indiquant les proportions de termes dans chacune des sections. Le terme *compiler* apparait uniquement en section *2* (*100%* en *% S2*) avec *1* seule occurrence, tandis que *php* se trouve à *66%* en section *0* et *33%* en section *3* avec un total de *3* occurrences.

On obtient ainsi un tableau récapitulant les termes apparaissant dans chaque document, ainsi que plusieurs autres informations. Le nombre total d'occurrences de chaque terme et la proportion d'occurrences dans chaque section sont explicités. Ce tableau permet de voir immédiatement les termes les plus utilisés dans un document, ainsi que leur distribution parmi les sections visées.

Liste filtrée et annotée de termes

Terme	Identifiant	N° Section
php	bn:01753580n	0
php	bn:01753580n	0
document	bn:00028018n	1
document	bn:00028018n	1
compiler	bn:00021344n	2
web	bn:00080772n	2
faq	bn:00033653n	3
php	bn:01753580n	3

Proportions d'occurrences pour chaque terme

Terme	Identifiant	N° Section	Occ. Locales	Occ. Totales	% Section
compiler	bn:00021344n	2	1	1	**100%**
document	bn:00028018n	1	2	2	**100%**
faq	bn:00033653n	3	1	1	**100%**
web	bn:00080772n	2	1	1	**100%**
php	bn:01753580n	0	2	3	**66%**
php	bn:01753580n	3	1	3	**33%**

Terme	Identifiant	Occ. Totales	% S0	% S1	% S2	% S3
compiler	bn:00021344n	1	0%	0%	100%	0%
document	bn:00028018n	2	0%	100%	0%	0%
faq	bn:00033653n	1	0%	0%	0%	100%
web	bn:00080772n	1	0%	0%	100%	0%
php	bn:01753580n	3	66%	0%	0%	33%

Proportions d'occurrences de termes dans chaque section

Fig. 5.6 – Proportion de terme dans chacune des quatre sections du document

Sélection des étiquettes temporelles (PIII.2)

Les tableaux de proportions d'occurrences de termes dans chacun des documents sont fusionnés pour pouvoir obtenir une liste de termes dits *significatifs*. Les *termes significatifs* correspondent à des termes dont la fréquence d'apparition est supérieure à un seuil (détaillé par la suite) dans les mêmes sections de l'ensemble des documents. Ces termes sont donc suffisamment fréquents pour être importants et suffisamment concentrés dans quelques sections précises pour qu'il s'agisse de termes spécifiques à ces sections. Les *termes significatifs* permettent de déterminer les sections dans lesquelles certains termes pourraient être fixés, afin d'organiser temporellement les clusters contenant ces termes.

Les proportions sont analysées afin de mesurer la dispersion des termes dans l'ensemble des sections du document. Pour cela, le *coefficient de Gini* [40] est utilisé : « *L'indice (ou coefficient) de Gini est un indicateur synthétique permettant de rendre compte du niveau d'inégalité pour une variable et sur une population donnée. Il varie entre 0 (égalité parfaite) et 1 (inégalité extrême). Entre 0 et 1, l'inégalité est d'autant plus forte que l'indice de Gini est élevé* » [54]. En effet, pour comprendre si un terme est dispersé de façon égale dans toutes les sections, ou au contraire si celui-ci est concentré dans une ou quelques sections seulement, le coefficient de Gini est parfaitement adapté.

Explications du coefficient de Gini : Le coefficient de Gini d'une distribution correspond au double de l'aire issue de la différence entre la droite représentant une distribution parfaitement équitable (une diagonale) et la courbe de Lorenz de la distribution étudiée (formée par la somme successive des valeurs de la distribution) [80]. Le graphique 5.7 montre la distribution $5, 10, 40, 5, 40$ réordonnée en $5, 5, 10, 40, 40$ pour former une courbe de Lorenz (courbe bleue) passant sous la droite de la distribution équitable (droite rouge). Il existe plusieurs manières de calculer le coefficient de Gini [75]. Une de ces manières utilise l'équation de Kendal et Stuart (5.2) où n correspond au nombre d'éléments dans la distribution, et $y_1, y_2, ..., y_n$ correspond à chaque élément de la distribution [61][91]. La formule (5.3) issue de l'approche de Mookherjee et Shorrocks [74] étant plus simple à implémenter, celle-ci est retenue. Elle s'appuie sur la moyenne de la distribution (μ) ainsi que sur la somme des différences absolues des éléments de la distribution divisée par le carré du nombre d'éléments. Pour obtenir le coefficient de Gini, on divise la moyenne des différences absolues des éléments entre eux ($Moyenne(|y_1 - y_1|, |y_1 - y_2|, ..., |y_n - y_n|)$) par la moyenne des éléments ($Moyenne(y_1, y_2, ..., y_n)$), puis diviser le tout par 2.

$$G = \frac{(1/2n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j|}{(1/n) \sum_{i=1}^{n} y_i} \tag{5.2}$$

$$G = \frac{1}{2\mu n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j| \tag{5.3}$$

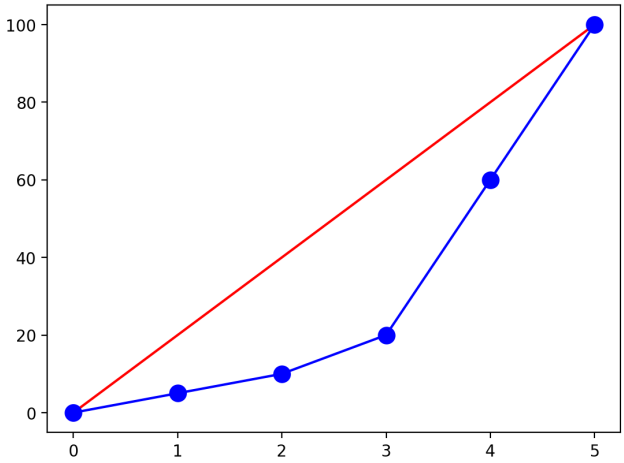


Fig. 5.7 – Exemple d'une courbe de Lorenz (en bleu) issue de la distribution $5, 10, 40, 5, 40$

Pour mesurer le degré de concentration d'un terme dans les différentes sections d'un document, on calcule donc le coefficient de Gini sur les proportions d'occurrences du terme dans chaque section. Les termes dont le coefficient de Gini est supérieur ou égal à $0,3$ sont considérés comme suffisamment concentrés pour être analysés plus en profondeur. Chaque proportion d'occurrences des termes conservés est ensuite comparée à la moyenne des proportions d'occurrences du terme considéré. Seules les proportions strictement supérieures à la moyenne sont considérées comme significatives. Ces choix empiriques proviennent du fait que le coefficient de Gini calcule un degré de concentration qui varie indépendamment de la moyenne (qui peut rester fixe). Un coefficient de Gini élevé sur une distribution où quelques valeurs dépassent la moyenne implique nécessairement que les autres valeurs sont beaucoup plus faibles, donc que les valeurs au dessus de la moyenne sont beaucoup plus grandes que les autres valeurs. À l'inverse, un coefficient de Gini faible implique dans tous les cas que les valeurs sont proches les unes des autres, y compris celles qui dépassent la moyenne. Nous avons fixé le seuil du coefficient de Gini à $0,3$ pour les besoins de la méthode CREA, car cette valeur sépare correctement les distributions trop peu concentrées des distributions suffisamment concentrées. La figure 5.8 montre trois exemples de proportions différentes pour cinq sections. Les histogrammes en jaune indiquent chaque valeur de la distribution, c'est-à-dire la proportion d'occurrences d'un terme dans chaque section (pour rappel, la courbe de Lorenz est formée par l'addition successive de ces valeurs). La somme des proportions étant toujours égale à 100, pour un document coupé en cinq sections la moyenne (droite verte en pointillés) sera toujours de 20.

— La première distribution $10, 15, 20, 25, 30$ donne un coefficient de Gini de $0,200$, ce qui indique que la distribution est plutôt équitable : le terme est présent dans toutes les sections avec une faible différence des occurrences.

— La deuxième distribution $0, 0, 0, 10, 90$ donne un coefficient de Gini de $0,760$, ce qui indique que la distribution est inéquitable : quelques sections concentrent toutes les occurrences du terme.

— La dernière distribution $5, 5, 10, 40, 40$ donne un coefficient de Gini de $0,424$, ce

qui indique une certaine iniquité entre les parties : le terme est présent dans toutes les sections, mais les occurrences sont plus élevées dans deux d'entre elles.

Dans cette dernière distribution, on voit effectivement que les proportions 40% et 40% sont particulièrement plus grandes que les trois autres proportions (respectivement 5%, 5%, et 10%). Empiriquement parlant, ceci permet d'affirmer qu'un terme se trouve majoritairement dans les deux sections associées à ces proportions. Le coefficient de Gini étant supérieur à $0,3$ dans deux des trois distributions (représentant un document chacune), on ne sélectionne que les sections dont les proportions (barres jaunes des histogrammes) sont supérieures à la moyenne (droite verte en pointillés). Ainsi, les termes sont considérés comme *significatifs* aux sections concernées dans les distributions des deux documents.

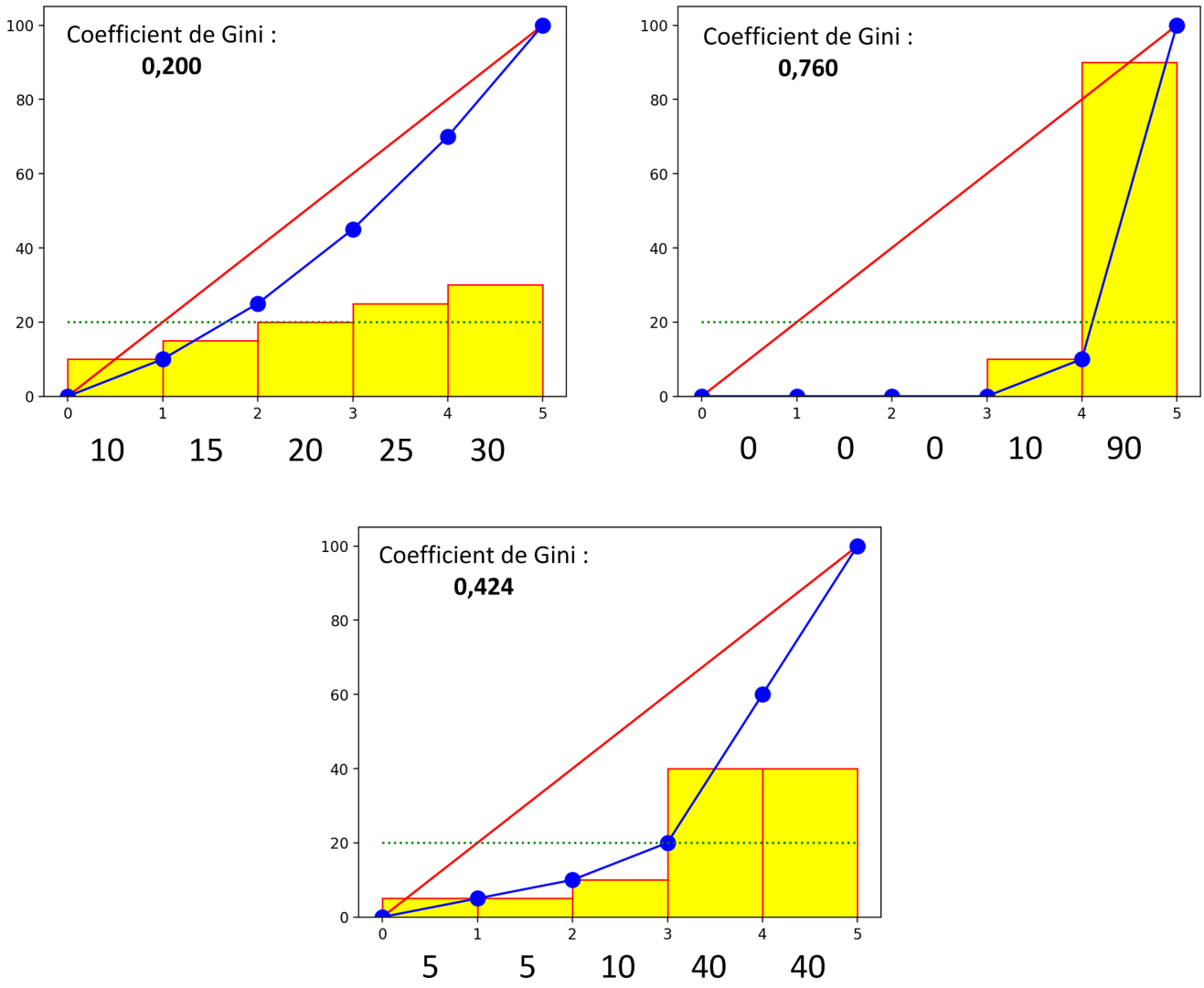


Fig. 5.8 – Trois exemples de proportions différentes sur cinq sections

Afin de considérer les termes comme *significatifs*, la forte concentration de chaque terme par partie par document ($G > 0,3$) n'est pas suffisante. Une moyenne des occurrences de l'ensemble des termes est calculée, et seuls les termes dont les occurrences

sont supérieures à la moyenne sont considérés pour la suite des traitements. Ainsi, les termes suffisamment présents dans le document et suffisamment concentrés dans certaines sections sont considérés comme des *termes significatifs*. Ce test est formalisé par la formule (5.4), où T correspond à un terme parmi l'ensemble de ceux du document (de 1 à n), et G correspond au coefficient de Gini.

$$EstSignificatif(T_x) = (Occurrences(T_x) > \frac{1}{n}\sum_{i=1}^{n} Occurrences(T_i)) \wedge (G(T_x) > 0,3) \tag{5.4}$$

La figure 5.9 illustre un exemple pas à pas du calcul des termes significatifs, et des sections auxquelles ils sont associés, pour un document coupé en quatre sections avec cinq termes.

1. Les proportions d'occurrences des termes dans chaque sections sont agrégées dans un unique tableau.
2. Le coefficient de Gini est calculé pour chaque terme. Le document étant coupé en quatre sections, la moyenne des proportions est donc de 25%. Seules les proportions supérieures à 25% associées à un terme dont le coefficient de Gini est supérieur à $0,3$ sont conservées (avec un 1).
3. La moyennes des occurrences de l'ensemble des termes est calculée pour déterminer quels termes sont fréquents ou non. Seuls les termes dont la valeur d'occurrences est supérieure à la moyenne seront conservés.
4. La liste finale des termes significatifs est produite : la colonne *significatif* indique avec un 1 que le terme est fortement concentré dans une ou quelques sections (indiquée(s) dans les colonnes *S0*, *S1*, *S2*, et/ou *S3*).

Dans cet exemple, on peut voir que les termes *compiler* et *document* sont concentrés dans certaines sections mais trop peu présents dans l'ensemble du document pour être retenus. Le terme *web* n'est pas assez concentré dans certaines sections pour être retenu. À l'inverse, les termes *faq* et *php* sont concentrés dans deux sections précises chacun (S0, S1, et S2) et suffisamment fréquents (au delà de la moyenne d'occurrences) pour être considérés comme significatifs dans ces sections pour ce document.

Ces traitements permettent d'obtenir pour chaque document une liste de termes significatifs dans chacune des sections. Afin d'ordonner par la suite les fragments réutilisables et former des scénarios plus concrets, une agrégation des termes significatifs de l'ensemble des documents est réalisée. Les termes sont reconnus par leur identifiant unique, puis sont agrégés selon les sections et documents dans lesquels ils sont considérés comme significatifs. Le tableau généré permet de connaître pour chaque terme dans quel(s) document(s) et section(s) il est considéré comme significatif. La figure 5.10 illustre comment les termes significatifs de trois documents sont rassemblés en un tableau. On peut voir que le terme « *size* » identifié par « *bn*:00016413*n* » est significatif pour la section *S2* dans le document *C3*, mais aussi pour la section *S3* dans les documents *C2* et *C3*. Cette organisation liant les termes, les documents, et les numéros de sections, se combine facilement aux clusters de termes précédemment créés : il devient possible d'ajouter une information temporelle dans les clusters, grâce aux numéros de sections associées aux termes et documents.

Proportions d'occurrences de termes dans chaque section

Terme	*Identifiant*	*Occ. Totales*	*% S0*	*% S1*	*% S2*	*% S3*
compiler	bn:00021344n	1	0%	0%	100%	0%
document	bn:00028018n	2	0%	100%	0%	0%
faq	bn:00033653n	7	66%	0%	0%	33%
web	bn:00080772n	8	25%	25%	25%	25%
php	bn:01753580n	10	50%	30%	10%	10%

Calcul du coefficient de Gini

Terme	*Identifiant*	*Coef. Gini*	*S0*	*S1*	*S2*	*S3*
compiler	bn:00021344n	0,75	0	0	1	0
document	bn:00028018n	0,75	0	1	0	0
faq	bn:00033653n	0,583	1	0	0	1
web	bn:00080772n	0	0	0	0	0
php	bn:01753580n	0,35	1	1	0	0

Calcul de la moyenne des occurrences

Terme	*Identifiant*	*Occ. Totales*	*Moyenne*	*S0*	*S1*	*S2*	*S3*
compiler	bn:00021344n	1	5,6	0	0	1	0
document	bn:00028018n	2	5,6	0	1	0	0
faq	bn:00033653n	7	5,6	1	0	0	1
web	bn:00080772n	8	5,6	0	0	0	0
php	bn:01753580n	10	5,6	1	1	0	0

Termes significatifs

Terme	*Identifiant*	*Significatif*	*S0*	*S1*	*S2*	*S3*
compiler	bn:00021344n	0	0	0	0	0
document	bn:00028018n	0	0	0	0	0
faq	bn:00033653n	1	1	0	0	1
web	bn:00080772n	0	0	0	0	0
php	bn:01753580n	1	1	1	0	0

Fig. 5.9 – Exemple de calcul des termes significatifs d'un document

Termes significatifs par document

Document C1

Terme	*Identifiant*	*Significatif*	*S0*	*S1*	*S2*	*S3*
compiler	bn:00021344n	0	0	0	0	0
document	bn:00028018n	0	0	0	0	0
faq	bn:00033653n	1	1	0	0	1
web	bn:00080772n	0	0	0	0	0
php	bn:01753580n	1	1	1	0	0

Document C2

Terme	*Identifiant*	*Significatif*	*S0*	*S1*	*S2*	*S3*
size	bn:00016413n	1	0	0	0	1
client	bn:00019764n	0	0	0	0	0
programme	bn:00021492n	1	1	0	0	0
web	bn:00080772n	1	1	1	0	0
php	bn:01753580n	1	1	0	0	1

Document C3

Terme	*Identifiant*	*Significatif*	*S0*	*S1*	*S2*	*S3*
size	bn:00016413n	1	0	0	1	1
client	bn:00019764n	1	0	0	1	0
document	bn:00028018n	1	0	0	0	1
web	bn:00080772n	1	1	1	0	0
php	bn:01753580n	1	1	1	0	0

Termes significatifs des documents

Terme	*Identifiant*	*Section*	*Fichiers*		
size	bn:00016413n	S2	C3		
size	bn:00016413n	S3	C2	C3	
client	bn:00019764n	S2	C3		
programme	bn:00021492n	S0	C2		
document	bn:00028018n	S3	C3		
faq	bn:00033653n	S0	C0		
faq	bn:00033653n	S3	C0		
web	bn:00080772n	S0	C2	C3	
web	bn:00080772n	S1	C2	C3	
php	bn:01753580n	S0	C1	C2	C3
php	bn:01753580n	S1	C1	C3	
php	bn:01753580n	S3	C2		

Fig. 5.10 – Exemple d'agrégation des termes significatifs de chaque document en un tableau

Étiquetage temporel des clusters (PIII.3)

Les clusters de termes précédemment formés, représentant des fragments réutilisables pour former une structure, peuvent maintenant être enrichis d'une information temporelle, ordonnancés. Cette étape vise à fusionner les clusters de termes avec l'information temporelle embarquée par les termes, pour pouvoir étiqueter temporellement les clusters et proposer un ordonnancement.

Nous avons tout d'abord testé le simple ajout des numéros de sections associées à chaque terme, de sorte à effectuer une *union* de toutes les sections proposées par les termes d'un même cluster. Cette *union* était cependant trop ambiguë dans nos essais, car quasiment tous les clusters étaient étiquetés d'à peu près toutes les sections. Puis, nous avons testé l'équivalent d'une *intersection* des sections proposées par les termes significatifs d'un même cluster (seuls les termes disposant d'une information temporelle sont pris en compte). Cette *intersection* était trop restrictive et n'étiquetait que les clusters contenant le moins de termes significatifs, engendrant des clusters peu voire pas du tout pertinents pour l'ordonnancement temporel. Finalement, nous avons préféré un *vote à la majorité* des sections des termes significatifs.

Les termes significatifs (i.e., ceux disposant d'une information temporelle) « votent » pour chacune de leurs sections. Chaque terme donne un point (« vote » pour) à chaque section dans laquelle il est considéré comme terme significatif. Les sections retenues sont celles disposant strictement de plus de la moitié des votes dans chacun des clusters. La figure 5.11 illustre ce vote à la majorité dans le cas d'un cours sur quatre séances (donc quatre clusters et quatre sections). Le terme significatif *liste* est étiqueté des quatre sections auxquelles il est rattaché *S0*, *S1*, *S2*, et *S3*. Le terme significatif *foreach* est étiqueté des deux sections auxquelles il est rattaché *S2* et *S3*. Les autres termes significatifs sont étiquetés eux aussi. Le cluster 1 contient *5* termes significatifs, c'est-à-dire qu'une section sera retenue si elle obtient strictement plus de $5 \div 2 = 2,5$ voix, donc en pratique au moins *3* voix. Le consensus se fait donc sur les sections *S0*, *S2*, et *S3*. La section *S1* ne disposant que de 2 voix, elle n'est pas retenue. Le cluster 4 contient *4* termes significatifs, c'est-à-dire qu'une section sera retenue si elle obtient strictement plus de $4 \div 2 = 2$ voix, donc en pratique au moins *3* voix. Chaque section n'obtenant que 2 voix, aucune n'est retenue, et le cluster n'est pas étiqueté temporellement. À l'issue de cette étape, les clusters qui le peuvent sont étiquetés temporellement.

Termes significatifs des documents

Terme	*Section*	*Fichiers*		
liste	S0	A0	A1	
liste	S1	A1	A2	
liste	S2	A2	A3	
liste	S3	A3		
foreach	S2	A2		
foreach	S3	A0	A1	A3
for	S0	A0	A1	
for	S2	A0	A2	
for	S3	A3		
boucle	S0	A3		
...	...	...	...	...
drupal	S3	A2		

Clusters de termes

ID Cluster	*Termes*					
1	liste	foreach	for	boucle	condition	if
2	texte	structure	entier	type		
3	web	hypertexte	lien	html		
4	php	installation	archive	drupal	wordpress	CMS

ID Cluster	*Sections*	*Nb Votes Minimum*	*Nb Termes Significatifs*	*Termes*					
1	S0,S2,S3	*3*	*5*	liste	foreach	for	boucle	condition	if
				S0,S1,S2,S3	S2,S3	S0,S2,S3	S0,S2	S0,S1,S2,S3	
2	S1	*2*	*3*	texte	structure	entier	type		
					S1	S0,S1	S1,S3		
3	S3	*2*	*2*	web	hypertexte	lien	html		
				S0,S2,S3			S1,S3		
4		*3*	*4*	php	installation	archive	drupal	wordpress	CMS
				S0,S1,S2,S3	S1	S0,S2	S3		

Clusters de termes étiquetés temporellement

FIG. 5.11 – Exemple d'étiquetage des clusters avec un vote strictement supérieur à la majorité absolue

Organisation en séances (PIII.4)

Les clusters étiquetés temporellement disposent d'une information sur leurs possibles placements dans le nouveau cas. Afin de proposer des scénarios possibles constitués de séances et des notions étudiées dans chacune, un changement de point de vue est nécessaire. Cette étape vise à passer du point de vue des clusters, vers celui des séances, pour permettre à l'utilisateur de visualiser les organisations possibles des séances et choisir celle qui lui convient.

L'utilisateur ayant précédemment indiqué le nombre de séances souhaitées, les clusters sont analysés pour ordonner ceux-ci. Pour chaque numéro de séance, les clusters disposant de l'étiquette temporelle associée au même numéro de section sont proposés en premier, puis, ceux ne disposant d'aucune information temporelle sont ensuite proposés. L'utilisateur dispose ainsi d'une vue globale de l'ensemble des séances à venir. Celui-ci est ainsi guidé et peut choisir quels clusters conserver ou retirer au fur et à mesure de l'exécution de son cas, ou précisément dans le contexte de l'enseignement, de son cours. La figure 5.12 illustre un exemple de réorganisation dans le cas de cours en quatre séances (donc quatre clusters et quatres sections réorganisés en quatre séances). Le cluster 4 n'étant pas étiqueté, il est proposé en dernier choix lors de toutes les séances. Le cluster 1 étant étiqueté aux sections *S0*, *S2*, et *S3*, il est proposé lors des séances 0, 2 et 3.

Une fois l'organisation en séances réalisée, l'enseignant dispose d'une vue globale d'un déroulé possible de cours par rapport à l'ensemble des supports sélectionnés en entrée et au nombre de séances demandées. Cette organisation des séances se calcule donc automatiquement d'un point de vue utilisateur, considérant nos différents réglages et choix empiriques sur l'ensemble de la chaîne de traitements.

Clusters de termes étiquetés temporellement

ID Cluster	*Sections*	*Nb Votes Minimum*	*Nb Termes Significatifs*	*Termes*					
1	S0,S2,S3	*3*	*5*	liste	foreach	for	boucle	condition	if
				S0,S1,S2,S3	S2,S3	S0,S2,S3	S0,S2	S0,S1,S2,S3	
2	S1	*2*	*3*	texte	structure	entier	type		
					S1	S0,S1	S1,S3		
3	S3	*2*	*2*	web	hypertexte	lien	html		
				S0,S2,S3			S1,S3		
4		*3*	*4*	php	installation	archive	drupal	wordpress	CMS
				S0,S1,S2,S3	S1	S0,S2	S3		

Réorganisation des clusters sous forme de séances

Séance 0	*Séance 1*	*Séance 2*	*Séance 3*
cluster 1	cluster 2	cluster 1	cluster 1
cluster 4	cluster 4	cluster 4	cluster 3
			cluster 4

Séance 0	*Séance 1*	*Séance 2*	*Séance 3*
liste, foreach, for, boucle, condition, if	texte, structure, entier, type	liste, foreach, for, boucle, condition, if	liste, foreach, for, boucle, condition, if
php, installation, archive, drupal, wordpress, CMS	php, installation, archive, drupal, wordpress, CMS	php, installation, archive, drupal, wordpress, CMS	web, hypertexte, lien, html
			php, installation, archive, drupal, wordpress, CMS

Clusters de termes organisés en séances

Fig. 5.12 – Exemple d'organisation en séances

Résultats préliminaires et discussion

Nous avons succinctement évalué cette extension et présentons maintenant quelques résultats partiels. Seul le scénario n°1 de référence vu en section 4.2.1 a été testé. Pour rappel, ce cas visait à construire un cours de développement web avec PHP sur 8 séances à partir de 9 supports existants, dont 3 au format texte long et 6 au format diapositives. Les résultats de l'étiquetage temporel des clusters (PIII.3) sont tout d'abord présentés sur la figure 5.13. Puis, les résultats de l'organisation sous forme de séances (PIII.4) sont présentés sur la figure 5.14.

ID	Sections	Nb Votes Minimum	Nb Termes Significatifs	Termes / Sections										
1	2,4,5,7	*6*	*10*	donnée 0,1,2,3,4,5,6,7	text 1,2,3,4,5,6	méthode post 0,1,2,3,4,5,6,7	programmation 0,2,5,7	site 0,4,5,7	langage de script 0,1,2,3,4,5,6,7	list 4	méthode 2,4,5,6,7	timestamp 2,3	files 2,4,7	
2		*3*	*4*	xml 4	configuration 0,1,6	composer 5,6,7	doctype 1,3,5							
3	6	*3*	*4*	base de donnée 0,2,3,6,7	insert 6	varchar 4,5	null 1,3,4,5,6							
4	4,7	*4*	*6*	typage 0,1,2,3,4,5,6,7	mot 1,2,3,4,5,6	moteur 7	affiche 0,1,2,3,4,5,6,7	transaction 7	visiteur 4					
5	0,1,2,4,5,6	*4*	*7*	fichier 0,1,2,3,4,5,6,7	commentaire 0,1,2,4	case à cocher 1,4,5,7	interpréter 0,2,3,7	côté serveur 0,4,5,6	serveur 0,1,2,3,4,5,6	côté client 0,6				
6	1,2,5,7	*4*	*7*	url 1,2,4,5,6,7	langage 0,1,2,3	case 1	fermeture	session 0,1,2,4,5,6,7	chaîne 0,1,2,3,4,7	entête 2,5,7	avoir accès 1,3,5,6,7			
7	0,1,2,3,5,6,7	*6*	*11*	page web 0,1,4,5,7	navigateur 0,1,2,3,4,5,6,7	serveur web 0	texte 0,1,2,3,4,5,6	concerner 1,2,3,6,7	délimiter 0,1,2,3,4,5,6,7	utilisateur 0,2,3,6,7	associer 2,5,6,7	personne 0,1,4,5,6,7	machine 6,7	mysql 0,1,2,3,6,7
8	0,1,2,3,4,5,6,7	*5*	*9*	php 0,1,2,3,4,5,6,7	code 0,2,4,5,6,7	fois 0,1,2,3,4,5,7	post 1,2,3,4,5,6,7	jour 0,2,3,4,7	foreach 1,2,3,4,5,7	cle 1,2	classe 0,1,2,4,6,7	class 4,5,6,7	mysqli	

Fig. 5.13 – Résultats de l'étiquetage temporel des clusters (PIII.3) du cas n°1 référence

Séance 0	Séance 1	Séance 2	Séance 3	Séance 4	Séance 5	Séance 6	Séance 7
4 clusters	5 clusters	6 clusters	3 clusters	5 clusters	6 clusters	5 clusters	6 clusters
fichier, commentaire, case à cocher, interpréter, côté serveur, serveur, côté client	fichier, commentaire, case à cocher, interpréter, côté serveur, serveur, côté client	donnée, text, méthode post, programmation, site, langage de script, list, méthode, timestamp, files	page web, navigateur, serveur web, texte, concerner, délimiter, utilisateur, associer, personne, machine, mysql	donnée, text, méthode post, programmation, site, langage de script, list, méthode, timestamp, files	donnée, text, méthode post, programmation, site, langage de script, list, méthode, timestamp, files	base de donnée, insert, varchar, null	donnée, text, méthode post, programmation, site, langage de script, list, méthode, timestamp, files
page web, navigateur, serveur web, texte, concerner, délimiter, utilisateur, associer, personne, machine, mysql	url, langage, case, fermeture, session, chaîne, entête, avoir accès	fichier, commentaire, case à cocher, interpréter, côté serveur, serveur, côté client	php, code, fois, post, jour, foreach, cle, classe, class, mysqli	typage, mot, moteur, affiche, transaction, visiteur	fichier, commentaire, case à cocher, interpréter, côté serveur, serveur, côté client	fichier, commentaire, case à cocher, interpréter, côté serveur, serveur, côté client	typage, mot, moteur, affiche, transaction, visiteur
php, code, fois, post, jour, foreach, cle, classe, class, mysqli	page web, navigateur, serveur web, texte, concerner, délimiter, utilisateur, associer, personne, machine, mysql	url, langage, case, fermeture, session, chaîne, entête, avoir accès	xml, configuration, composer, doctype	fichier, commentaire, case à cocher, interpréter, côté serveur, serveur, côté client	url, langage, case, fermeture, session, chaîne, entête, avoir accès	page web, navigateur, serveur web, texte, concerner, délimiter, utilisateur, associer, personne, machine, mysql	url, langage, case, fermeture, session, chaîne, entête, avoir accès
xml, configuration, composer, doctype	php, code, fois, post, jour, foreach, cle, classe, class, mysqli	page web, navigateur, serveur web, texte, concerner, délimiter, utilisateur, associer, personne, machine, mysql		php, code, fois, post, jour, foreach, cle, classe, class, mysqli	page web, navigateur, serveur web, texte, concerner, délimiter, utilisateur, associer, personne, machine, mysql	php, code, fois, post, jour, foreach, cle, classe, class, mysqli	page web, navigateur, serveur web, texte, concerner, délimiter, utilisateur, associer, personne, machine, mysql
	xml, configuration, composer, doctype	php, code, fois, post, jour, foreach, cle, classe, class, mysqli		xml, configuration, composer, doctype	php, code, fois, post, jour, foreach, cle, classe, class, mysqli	xml, configuration, composer, doctype	php, code, fois, post, jour, foreach, cle, classe, class, mysqli

Fig. 5.14 – Résultats de l'organisation en séances (PIII.4) du cas n°1 référence

Une première analyse de l'étiquetage temporel des clusters sur la figure 5.13 nous permet de constater tout d'abord qu'un seul des 8 clusters n'est pas étiqueté (13%), 2 clusters sont étiquetés avec une ou deux sections (25%), 2 clusters sont étiquetés avec quatre sections (25%), et 3 clusters sont étiquetés avec six sections ou plus (38%). Cela signifie que 4 clusters (50%) peuvent quasiment être placés sur n'importe quelle séance. Parmi les 4 autres clusters, 2 peuvent être placés dans la moitié des séances (25%). Seuls 2 clusters sont explicitement placés dans une ou deux séances (25%). Les résultats montrent qu'il est difficile de fixer automatiquement les clusters sur quelques séances précises avec la méthode actuelle. L'étiquetage tel quel donne des propositions d'organisation à l'utilisateur qui reste seul décideur.

Cependant, un résultat encourageant montre que le cluster n°3 contenant les termes « *base de données, insert, varchar, null* » semble explicitement placé en séance 6, c'est-à-dire à l'avant dernière séance. Ce cluster rassemble des termes qui sont tous apparentés au thème de la base de données. Les 3 cours au format texte l'abordent au moins au milieu, 2 d'entre eux en parlent en plus entre le début et le milieu, et 1 présente à la fin de nombreux projets utilisant parfois de la base de données. Concernant les 6 cours au format diapositives, 1 cours n'aborde pas ce thème, 4 cours l'abordent vers la fin, et 1 mentionne très succinctement quelques termes de ce thème à la fin. De manière générale, 6 cours en parlent au moins un peu à la fin (67%), 3 en parlent au milieu (33%), 2 en parlent vers le début (22%), et 1 ne l'aborde pas du tout (11%). D'un point de vue purement statistique, ce thème peut être proposé prioritairement vers la fin. Étant donné le faible nombre de tests, il peut aussi s'agir d'un cas extrême ou de bruit, qui normalement ne serait pas pris en compte.

Cette difficulté à placer les clusters limite l'exploitation telle quelle des résultats par un utilisateur. Nous proposons trois solutions pour résoudre ce problème :

— Une première solution de contournement serait de ne présenter à l'utilisateur qu'une seule séance à la fois, depuis la première jusqu'à la dernière, afin qu'il choisisse lui-même quel cluster présenter à chacune des séances sans être submergé d'un excès d'information.

— Une autre solution pourrait être de revoir la sélection des étiquettes temporelles : un meilleur algorithme de vote pourrait améliorer les résultats (au lieu de l'*union*, *intersection*, ou *vote* des sections).

— Une dernière solution serait de mieux extraire les étiquettes temporelles : le découpage actuel étant très simple, il pourrait être envisagé d'extraire la structure des supports en analysant le sommaire ou les titres uniquement, mais cela implique un travail très poussé lors de la phase de reconnaissance optique (PI.2), donc avant la désambiguïsation (PI.3) et sa standardisation des termes.

Nous présentons cependant quelques opportunités offertes par cette organisation temporelle. L'analyse temporelle actuelle est spécialisée sur les documents organisés de façon séquentielle. Dans le cadre de l'enseignement supérieur et de la recherche, les articles de recherche ne sont pas du tout organisés de la même manière qu'un support de cours : il est déconseillé d'appliquer la phase d'analyse temporelle actuelle

sur des articles. Cependant, il est tout de même actuellement possible de construire des clusters avec des documents de nature variée, puis, de rechercher l'ordonnancement temporel uniquement à partir des documents organisés séquentiellement.

Nous avons vu que la méthode exige des documents dont le contenu principal est du texte, sous-entendu que de nombreux autres médias existent, afin de construire des clusters de termes. En posant l'hypothèse qu'il est possible d'adapter la phase d'analyse temporelle selon la nature de l'organisation temporelle des documents, on pourrait extraire un ordonnancement temporel depuis tous les types de documents organisés. Là où les supports de cours classiques sont séquentiels, les articles de recherche respectent eux aussi certaines règles selon les domaines (par exemple : introduction, revue de la littérature, contribution(s), expérience(s), discussion(s), conclusion). Des techniques de détection de rupture appliquées aux textes permettent typiquement d'effectuer un découpage thématique et chronologique [9][33]. Ainsi, il devient possible d'extraire les termes des introduction et contributions de plusieurs articles sur une durée donnée d'une même équipe de recherche, afin de voir l'évolution des projets et/ou les voies approfondies pour en déduire quels termes étaient porteurs à différentes époques. Cette dimension temporelle est donc particulièrement attrayante sur les possibilités qu'elle offre.

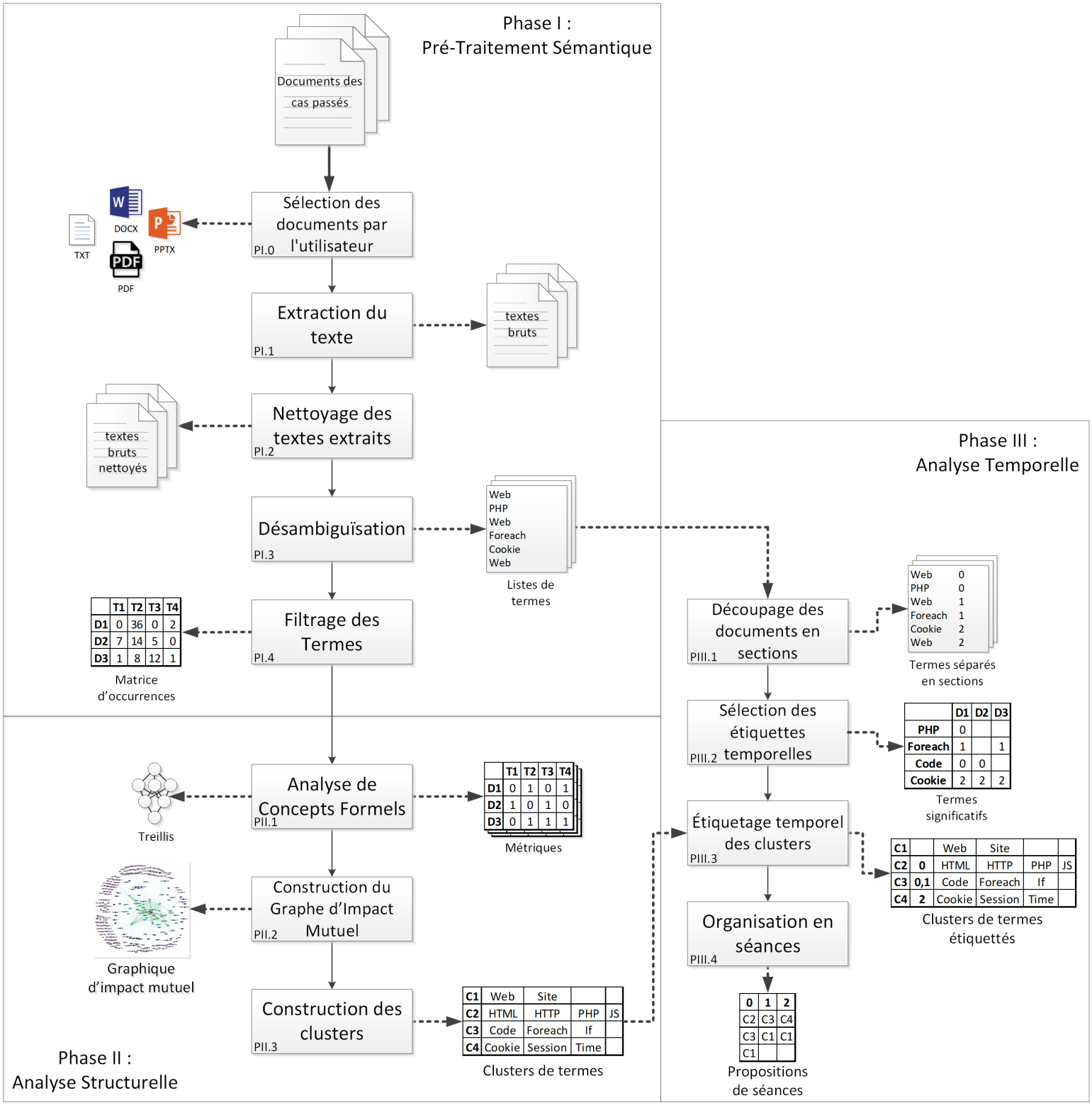


Fig. 5.15 – Détails de la méthode CREA étendue avec l'analyse temporelle

Bibliographie

[1] Andreas Abecker, Ansgar Bernardi, Knut Hinkelmann, Otto Ku, Michael Sintek, et al. Context-aware, proactive delivery of task-specific information : The knowmore project. Information Systems Frontiers, 2(3-4) :253–276, 2000.

[2] Eneko Agirre and Philip Edmonds. Word sense disambiguation : Algorithms and applications, volume 33. Springer Science & Business Media, 2007.

[3] Mian M Ajmal and Kaj U Koskinen. Knowledge transfer in project-based organizations : an organizational culture perspective. Project management journal, 39(1) :7–15, 2008.

[4] Mikhail A Babin and Sergei O Kuznetsov. Approximating concept stability. In International Conference on Formal Concept Analysis, pages 7–15. Springer, 2012.

[5] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi : an open source software for exploring and manipulating networks. In Third international AAAI conference on weblogs and social media, 2009.

[6] Radim Belohlavek. Introduction to formal concept analysis. Palacky University, Department of Computer Science, Olomouc, 47, 2008.

[7] Ilia Bider and Georgios Koutsopoulos. Introducing goal patterns for state-oriented business process modeling. In 2018 IEEE 22nd International Enterprise Distributed Object Computing Workshop (EDOCW), pages 43–47. IEEE, 2018.

[8] Fabrice Boissier, Irina Rychkova, and Bénédicte Le Grand. Challenges in knowledge intensive process management. In 2019 IEEE 23rd International Enterprise Distributed Object Computing Workshop (EDOCW), pages 65–74. IEEE, 2019.

[9] Narjès Boufaden, Guy Lapalme, and Yoshua Bengio. Découpage thématique des conversations : un outil d'aide à l'extraction. TALN 2002, 2002.

[10] Laurent Brosset. Google Livres et la numérisation : quels impacts pour les bibliothèques numériques ? Master's thesis, Université Grenoble Alpes, June 2016.

[11] Razvan Bunescu and Marius Paşca. Using encyclopedic knowledge for named entity disambiguation. In 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, April 2006. Association for Computational Linguistics.

[12] Ana Burusco Juandeaburre and Ramón Fuentes-González. The study of the l-fuzzy concept lattice. Mathware & soft computing. 1994 Vol. 1 Núm. 3 p. 209-218, 1994.

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

[14] Riccardo Cognini, Knut Hinkelmann, and Andreas Martin. A case modelling language for process variant management in case-based reasoning. In International Conference on Business Process Management, pages 30–42. Springer, 2016.

[15] Richard M Cormack. A review of classification. Journal of the Royal Statistical Society : Series A (General), 134(3) :321–353, 1971.

[16] Thomas H Davenport. Thinking for a living : how to get better performances and results from knowledge workers. Harvard Business Press, 2005.

[17] Thomas H Davenport, David W De Long, and Michael C Beers. Successful knowledge management projects. MIT Sloan Management Review, 39(2) :43, 1998.

[18] Thomas H Davenport and Nitin Nohria. Case management and the integration of labor. MIT Sloan Management Review, 35(2) :11, 1994.

[19] Guglielmo De Angelis, Alfonso Pierantonio, Andrea Polini, Barbara Re, Barbara Thönssen, and Robert Woitsch. Modeling for learning in public administrations—the learn pad approach. In Domain-Specific Conceptual Modeling, pages 575–594. Springer, 2016.

[20] Pierre Delort. Le Big Data :«Que sais-je ?» n° 4027. Que sais-je, 2018.

[21] Claudio Di Ciccio, Andrea Marrella, and Alessandro Russo. Knowledge-intensive processes : characteristics, requirements and analysis of contemporary approaches. Journal on Data Semantics, 4(1) :29–57, 2015.

[22] Claudio Di Ciccio and Massimo Mecella. Mining artful processes from knowledge workers' emails. IEEE Internet Computing, 17(5) :10–20, 2013.

[23] Claudio Di Ciccio, Massimo Mecella, and Tiziana Catarci. Representing and visualizing mined artful processes in mailofmine. In Symposium of the Austrian HCI and Usability Engineering Group, pages 83–94. Springer, 2011.

[24] Claudio Di Ciccio, Massimo Mecella, Monica Scannapieco, Diego Zardetto, and Tiziana Catarci. Mailofmine–analyzing mail messages for mining artful collaborative processes. In International Symposium on Data-Driven Process Discovery and Analysis, pages 55–81. Springer, 2011.

[25] Edwin Diday. Une représentation visuelle des classes empiétantes : les pyramides. PhD thesis, INRIA, 1984.

[26] P Dulbecco, MC Beer, J Delpech de Saint-Guilhem, S Dubourg-Lavroff, and E Pimmel. Les innovations pédagogiques numériques et la transformation des établissements d'enseignement supérieur. Les rapports de l'IGAENR, 49 :2018, 2018.

[27] Philippe Dulbecco. De l'expérimentation des innovations pédagogiques numériques à leur généralisation en france. Revue internationale d'éducation de Sèvres, 80 :103–114, 2019.

[28] Hanna Eberle, Frank Leymann, Daniel Schleicher, David Schumm, and Tobias Unger. Process fragment composition operations. In 2010 IEEE Asia-Pacific Services Computing Conference, pages 157–163. IEEE, 2010.

[29] Hanna Eberle, Tobias Unger, and Frank Leymann. Process fragments. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", pages 398–405. Springer, 2009.

[30] Amal Elgammal, Oktay Turetken, Willem-Jan van den Heuvel, and Mike Papazoglou. Formalizing and appling compliance patterns for business process compliance. Software & Systems Modeling, 15(1) :119–146, 2016.

[31] Vladimir Estivill-Castro and Jianhua Yang. Fast and robust general purpose clustering algorithms. In Pacific Rim International Conference on Artificial Intelligence, pages 208–218. Springer, 2000.

[32] Brian Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. Cluster analysis. Wiley, 5th edition, 2011.

[33] Bernard Fallery and Florence Rodhain. Quatre approches pour l'analyse de données textuelles : lexicale, linguistique, cognitive, thématique. In XVI ème Conférence de l'Association Internationale de Management Stratégique AIMS, pages pp–1. AIMS, 2007.

[34] Robert Feldt and Ana Magazinius. Validity threats in empirical software engineering research-an initial survey. In Seke, pages 374–379, 2010.

[35] Martin Fowler, Jim Highsmith, et al. The agile manifesto. Software Development, 9(8) :28–35, 2001.

[36] Chris Fraley and Adrian E Raftery. How many clusters ? which clustering method ? answers via model-based cluster analysis. The computer journal, 41(8) :578–588, 1998.

[37] Juliana BS França, Fernanda A Baião, and Flavia M Santoro. Towards characterizing knowledge intensive processes. In Computer Supported Cooperative Work in Design (CSCWD), 2012 IEEE 16th International Conference on, pages 497–504. IEEE, 2012.

[38] Gephi.org. Gephi - Tutorial Layout, June 2011.

[39] Gephi.org. Gephi - Spatialisation (FR), 2016.

[40] Corrado Gini. Measurement of inequality of incomes. The economic journal, 31(121) :124–126, 1921.

[41] André Goosse and Maurice Grevisse. Le bon usage. De Boeck Superieur, 2016.

[42] ADSE Gordon. Classification. Chapman and Hall/CRC, 1999.

[43] Alexander Gromoff, Yulia Bilinkis, and Nikolay Kazantsev. Business architecture flexibility as a result of knowledge-intensive process management. Global Journal of Flexible Systems Management, 18(1) :73–86, 2017.

[44] David Harel. Statecharts : A visual formalism for complex systems. Science of computer programming, 8(3) :231–274, 1987.

[45] Verne Harnish and Jim Collins. The greatest business decisions of all time : How Apple, Ford, IBM, Zappos, and others made radical choices that changed the course of business. Fortune Books, 2012.

[46] Mariam Ben Hassen, Mohamed Turki, and Faïez Gargouri. Choosing a sensitive business process modeling formalism for knowledge identification. Procedia Computer Science, 100 :1002–1015, 2016.

[47] Christian Herrmann and Matthias Kurz. Adaptive case management : supporting knowledge intensive processes with it systems. In International Conference on Subject-Oriented Business Process Management, pages 80–97. Springer, 2011.

[48] Alan R Hevner. A three cycle view of design science research. Scandinavian journal of information systems, 19(2) :4, 2007.

[49] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. MIS quarterly, pages 75–105, 2004.

[50] Charles Hill, Robert Yates, Carol Jones, and Sandra L Kogan. Beyond predictable workflows : Enhancing productivity in artful business processes. IBM Systems journal, 45(4) :663–682, 2006.

[51] Donald Hislop, Peter A Murray, Anup Shrestha, Jawad Syed, and Yusra Mouzughi. Knowledge management :(potential) future research directions. In The Palgrave Handbook of Knowledge Management, pages 691–703. Springer, 2018.

[52] Sebastian Huber, Peter Schott, and Matthias Lederer. Adaptive open innovation : solution approach and tool support. In Proceedings of the 7th International Conference on Subject-Oriented Business Process Management, page 12. ACM, 2015.

[53] Lawrence Hubert and Phipps Arabie. Comparing partitions. Journal of classification, 2(1) :193–218, 1985.

[54] INSEE. Indice de gini / coefficient de gini. https://www.insee.fr/fr/metadonnees/definition/c1551. [En ligne ; publication 13/01/2020].

[55] Öykü Işik, Willem Mertens, and Joachim Van den Bergh. Practices of knowledge intensive process management : quantitative insights. Business Process Management Journal, 19(3) :515–534, 2013.

[56] ISO 25964-1 : Information et documentation – Thésaurus et interopérabilité avec d'autres vocabulaires – Partie 1 : Thésaurus pour la recherche documentaire. Standard international, International Organization for Standardization, 2011.

[57] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PloS one, 9(6) :e98679, 2014.

[58] Ali Jaffal. Aide à l'utilisation et à l'exploitation de l'Analyse de Concepts Formels pour des non-spécialistes de l'analyse des données. PhD thesis, Université Panthéon - Sorbonne - Paris I, 2019.

[59] Ali Jaffal, Bénédicte Le Grand, and Manuele Kirsch-Pinheiro. Refinement strategies for correlating context and user behavior in pervasive information systems. Procedia Computer Science, 52 :1040–1046, 2015.

[60] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering : a review. ACM computing surveys (CSUR), 31(3) :264–323, 1999.

[61] Maurice G Kendall and Alan Stuart. The advanced theory of statistics. Vol. 1 : Distribution theory. Charles Griffin & Company, 1963.

[62] Janet L Kolodner. An introduction to case-based reasoning. Artificial intelligence review, 6(1) :3–34, 1992.

[63] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29(1) :1–27, 1964.

[64] Joseph B Kruskal. Multidimensional scaling. Number 11 in Quantitative Applications in the Social Sciences. Sage, 1978.

[65] Elena Kushnareva, Irina Rychkova, Rébecca Deneckére, and Bénédicte Le Grand. Modeling crisis management process from goals to scenarios. In International Conference on Business Process Management, pages 55–64. Springer, 2016.

[66] Elena Kushnareva, Irina Rychkova, and Bénédicte Le Grand. Modeling and animation of crisis management process with statecharts. In International Conference on Business Informatics Research, pages 145–160. Springer, 2015.

[67] Anton Manfreda, Brina Buh, and Mojca Indihar Štemberger. Knowledge-intensive process management : a case study from the public sector. Baltic Journal of Management, 10(4) :456–477, 2015.

[68] Lynne M Markus. Toward a theory of knowledge reuse : Types of knowledge reuse situations and factors in reuse success. Journal of management information systems, 18(1) :57–93, 2001.

[69] Andreas Martin, Sandro Emmenegger, and Gwendolin Wilke. Integrating an enterprise architecture ontology in a case-based reasoning approach for project knowledge. In Proceedings of the First International Conference on Enterprise Systems : ES 2013, pages 1–12. IEEE, 2013.

[70] Louise Merzeau, Valérie Schafer, Lionel Barbe, et al. Wikipédia, objet scientifique non identifié. Presses universitaires de Paris Nanterre, 2015.

[71] Nizar Messai. Analyse de concepts formels guidée par des connaissances de domaine : Application à la découverte de ressources génomiques sur le Web. PhD thesis, Université Henri Poincaré - Nancy I, 2009.

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

[73] George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. A semantic concordance. In HUMAN LANGUAGE TECHNOLOGY : Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993, 1993.

[74] Dilip Mookherjee and Anthony Shorrocks. A decomposition analysis of the trend in uk income inequality. The Economic Journal, 92(368) :886–902, 1982.

[75] Pauline Mornet, Stéphane Mussard, Françoise Seyte, and Michel Terraza. La décomposition de l'indicateur de gini en sous-groupes. Revue française d'économie, 29(2) :179–243, 2014.

[76] Andrea Moro, Francesco Cecconi, and Roberto Navigli. Multilingual word sense disambiguation and entity linking for everybody. In International Semantic Web Conference (Posters & Demos), pages 25–28, 2014.

[77] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation : a unified approach. Transactions of the Association for Computational Linguistics, 2 :231–244, 2014.

[78] Hamid R Motahari-Nezhad and Keith D Swenson. Adaptive case management : Overview and research challenges. In 2013 IEEE 15th Conference on Business Informatics, pages 264–269. IEEE, 2013.

[79] Ednilson Veloso Moura, Flávia Maria Santoro, and Fernanda Araujo Baião. Collaboration support for knowledge-intensive processes through a service-based approach. In Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on, pages 319–324. IEEE, 2013.

[80] Stéphane Mussard, Françoise Seyte, and Michel Terraza. La décomposition de l'indicateur de gini en sous-groupes : une revue de la littérature. Cahier de recherche/Working Paper, 6 :11, 2006.

[81] Roberto Navigli and Simone Paolo Ponzetto. Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193 :217–250, 2012.

[82] Ikujirō Nonaka and Hirotaka Takeuchi. The knowledge-creating company. Harvard business review, 85(7/8) :162, 2007.

[83] Ikujiro Nonaka, Ryoko Toyama, and Noboru Konno. Seci, ba and leadership : a unified model of dynamic knowledge creation. Long range planning, 33(1) :5–34, 2000.

[84] Klaus North and Gita Kumta. Knowledge management : Value creation through organizational learning. Springer, 2018.

[85] Amandine Pascal. L'approche du design science au cœur du débat rigueur/pertinence. In 16ème édition du Colloque de l'Association Information et Management (AIM 2011), page x, 2011.

[86] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. Journal of Machine Learning Research, 12 :2825–2830, 2011.

[87] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. Journal of management information systems, 24(3) :45–77, 2007.

[88] Stacie Petter and Adriane B Randolph. Developing soft skills to manage user expectations in it projects : Knowledge reuse among it project managers. Project Management Journal, 40(4) :45–59, 2009.

[89] Jonas Poelmans, Dmitry I Ignatov, Sergei O Kuznetsov, and Guido Dedene. Fuzzy and rough formal concept analysis : a survey. International Journal of General Systems, 43(2) :105–134, 2014.

[90] Georg Prohaska. Categorization and comparison of datasets across open data portals. Master's thesis, Department of Information Systems and Operations, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria, 2017.

[91] Graham Pyatt. On the interpretation and disaggregation of gini coefficients. The Economic Journal, 86(342) :243–255, 1976.

[92] Romain Quéré. Quelques propositions pour la comparaison de partitions non strictes. Theses, Université de La Rochelle, December 2012.

[93] Ricco Rakotomalala. Pratique des méthodes factorielles avec python. http://eric.univ-lyon2.fr/~ricco/cours/cours/Pratique_Methodes_Factorielles.pdf, 2020. p.208-209 [En Ligne ; accès Novembre 2020].

[94] William M Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336) :846–850, 1971.

[95] Delip Rao, Paul McNamee, and Mark Dredze. Entity linking : Finding extracted entities in a knowledge base. In Multi-source, multilingual information extraction and summarization, pages 93–115. Springer, 2013.

[96] Uwe V Riss, Alan Rickayzen, Heiko Maus, and Wil MP van der Aalst. Challenges for business process and task management. Journal of Universal Knowledge Management, 2 :77–100, 2005.

[97] Lior Rokach and Oded Maimon. Clustering methods. In Data mining and knowledge discovery handbook, pages 321–352. Springer, 2005.

[98] Ronald G. Ross. The business rules manifesto. http://www.businessrulesgroup.org/brmanifesto.htm, 2003. [Online ; accessed May-2019].

[99] Gilbert Saporta. Probabilités, analyse des données et statistique. Editions Technip, 2006.

[100] Silvia Schacht and Alexander Maedche. A methodology for systematic project knowledge reuse. In Innovations in Knowledge Management, pages 19–44. Springer, 2016.

[101] Helmut Schmid. Probabilistic part-ofispeech tagging using decision trees. In New methods in language processing, page 154, 1994.

[102] Helmut Schmid. Improvements in part-of-speech tagging with an application to german. In In Proceedings of the ACL SIGDAT-Workshop, pages 47–50, 1995.

[103] Robert F Simmons et al. Semantic networks : Their computation and use for understanding English sentences. Department of Computer Sciences and Computer-Assisted Instruction Laboratory . . ., 1972.

[104] Stephen Slade. Case-based reasoning : A research paradigm. AI magazine, 12(1) :42–42, 1991.

[105] Douglas Steinley. Properties of the hubert-arable adjusted rand index. Psychological methods, 9(3) :386, 2004.

[106] Keith D Swenson. White paper : State of the art in case management. Fujitsu America, Inc. March, pages 5–6, 2013.

[107] Keith D Swenson, Nathaniel Palmer, et al. Mastering the unpredictable : how adaptive case management will revolutionize the way that knowledge workers get things done, volume 1. Meghan-Kiffer Press Tampa, 2010.

[108] Jawad Syed, Peter A Murray, Donald Hislop, and Yusra Mouzughi. The Palgrave handbook of knowledge management. Springer, 2018.

[109] Hélène Zysman Sylvie Dalbin, Nathalie Yakovleff. ISO 25964-1 - Thésaurus pour la recherche documentairee. Livre blanc, AFNOR, janvier 2013.

[110] My Thao Tang. An Interactive and Iterative Knowledge Extraction Process Using Formal Concept Analysis. PhD thesis, Université de Lorraine, 2016.

[111] Nadia Tebourbi. L'apprentissage organisationnel : Penser l'organisation comme processus de gestion des connaissances et de développement des théories d'usage. Technical report, Chaire de recherche du Canada sur les enjeux socio-organisationnels de l'économie du savoir, 2000.

[112] Johannes Tenschert and Richard Lenz. Supporting knowledge work by speech-act based templates for micro processes. In International conference on business process management, pages 78–89. Springer, 2016.

[113] Johannes Tenschert and Richard Lenz. Towards speech-act-based adaptive case management. In 2016 IEEE 20th International Enterprise Distributed Object Computing Workshop (EDOCW), pages 1–8. IEEE, 2016.

[114] Thanh Tran, Erhard Weiss, Christoph Ruhsam, Christoph Czepa, Huy Tran, and Uwe Zdun. Embracing process compliance and flexibility through behavioral consistency checking in acm : A repair service management case. In Business Process Management Workshops. BPM 2016., 2015.

[115] Roman Vaculin, Richard Hull, Terry Heath, Craig Cochran, Anil Nigam, and Piyawadee Sukaviriya. Declarative business artifact centric modeling of decision and knowledge intensive business processes. In Enterprise Distributed Object Computing Conference (EDOC), 2011 15th IEEE International, pages 151–160. IEEE, 2011.

[116] Wil MP Van der Aalst, Mathias Weske, and Dolf Grünbauer. Case handling : a new paradigm for business process support. Data & Knowledge Engineering, 53(2) :129–162, 2005.

[117] Tommaso Venturini, Mathieu Jacomy, and Pablo Jensen. What do we see when we look at networks : Visual network analysis, relational ambiguity, and force-directed layouts. Big Data & Society, 8(1) :20539517211018488, 2021.

[118] MP Veyssieres and Richard E Plant. Identification of vegetation state and transition domains in california's hardwood rangelands. University of California, 101, 1998.

[119] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen,

E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0 : Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17 :261–272, 2020.

[120] Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic wikipedia. In Proceedings of the 15th international conference on World Wide Web, pages 585–594. ACM, 2006.

[121] Mark Von Rosing, Henrik Von Scheel, and August-Wilhelm Scheer. The Complete Business Process Handbook : Body of Knowledge from Process Modeling to BPM, volume 1. Morgan Kaufmann, 2014. "Phase 1 : Process Concept Evolution".

[122] Atro Voutilainen. Part-of-speech tagging. The Oxford handbook of computational linguistics, pages 219–232, 2003.

[123] WA Wagenaar and Peter Padmos. Quantitative interpretation of stress in kruskal's multidimensional scaling technique. British Journal of Mathematical and Statistical Psychology, 24(1) :101–110, 1971.

[124] Mathias Weske. Business Process Management : Concepts, Languages, Architectures. Springer, 2007.

[125] Rudolf Wille. Restructuring lattice theory : An approach based on hierarchies of concepts. In Ivan Rival, editor, Ordered Sets, volume 83 of NATO Advanced Study Institutes Series, pages 445–470. Springer Netherlands, 1982.

[126] Rudolf Wille. Formal concept analysis as mathematical theory of concepts and concept hierarchies. In Formal concept analysis, pages 1–33. Springer, 2005.

[127] Bastian Wormuth and Peter Becker. Introduction to formal concept analysis. In 2nd International Conference of Formal Concept Analysis February, volume 23, 2004.

[128] Rui Xu and Don Wunsch. Clustering, volume 10. John Wiley & Sons, 2008.

[129] Forrest W Young and David F Harris. Multidimensional scaling. LL Thurstone Psychometric Laboratory, University of North Carolina, 1983.

[130] Michael H Zack. Managing codified knowledge. Sloan management review, 40(4) :45–58, 1999.

[131] Haïfa Zargayouna, Catherine Roussey, and Jean-Pierre Chevallet. Recherche d'information sémantique : état des lieux. Traitement Automatique des Langues, 56(3) :49–73, 2015.

[132] Andrea Zasada. A box of bricks for modelling domain-specific compliance pattern. In 2018 IEEE 22nd International Enterprise Distributed Object Computing Workshop (EDOCW), pages 37–42. IEEE, 2018.

[133] André Zensen and Jochen Küster. A comparison of flexible bpmn and cmmn in practice : A case study on component release processes. In 2018 IEEE 22nd International Enterprise Distributed Object Computing Conference (EDOC), pages 105–114. IEEE, 2018.

[134] Michael Zur Muehlen, Marta Indulska, and Gerrit Kamp. Business process and business rule modeling languages for compliance management : a representational analysis. In Tutorials, posters, panels and industrial contributions at the 26th international conference on Conceptual modeling-Volume 83, pages 127–132. Australian Computer Society, Inc., 2007.

CREA : Méthode d'analyse, d'adaptation et de réutilisation des processus à forte intensité de connaissances

cas d'utilisation dans l'enseignement supérieur en informatique

La crise sanitaire du COVID-19 a particulièrement accéléré le mouvement de numérisation pourtant déjà initié depuis quelques décennies dans l'enseignement supérieur. De nombreuses activités ont dû être adaptées dans l'urgence, tout particulièrement les réunions entre enseignants, l'évaluation des étudiants et les enseignements. Ces activités sont des exemples de processus « *à forte intensité de connaissances* » (ou « *knowledge intensive processes* » en anglais) qui partagent des caractéristiques rendant difficile l'intégration du numérique, telles que :

— l'abondance de connaissances mobilisables, autant de la part des étudiants lors de leurs travaux que de la part des enseignants évaluant ou adaptant leurs cours,

— la collaboration entre toutes les parties prenantes du monde de l'enseignement supérieur,

— la créativité requise pour s'adapter au contexte incertain.

Ce besoin rapide de déployer de nouveaux processus à forte intensité de connaissances, ou d'adapter ceux qui existent, se confronte à de nombreux défis connus de ce domaine de recherche spécifique. La question est de savoir comment réutiliser des connaissances existantes, par exemple des connaissances entreposées en ligne dont l'abondance rend difficile la sélection des plus adaptées aux besoins des enseignants.

Dans cette thèse, nous proposons la méthode CREA réutilisant des cas passés dans le domaine de l'enseignement supérieur, en particulier pour la construction de cours. La méthode CREA permet de réutiliser des supports de cours existants pour tout d'abord représenter visuellement l'écart entre eux, mais également de proposer des séances de cours présentées sous forme de regroupements de sujets majeurs à aborder. D'autres types de documents peuvent également être intégrés parmi les supports de cours (des pages webs, ou des articles de recherche), afin de proposer des regroupements adaptés à un public particulier, voire de proposer des regroupements à l'état de l'art de la recherche. Cette méthode s'appuie sur des outils de traitement automatique de la langue pour extraire les termes employés indépendamment de la langue d'origine, puis sur l'analyse de concepts formels pour calculer des métriques permettant de construire des regroupements de termes et évaluer la similarité des cours fournis en entrée. Nous proposons également des résultats préliminaires d'une méthode d'ordonnancement des séances.

Mots clés : Processus à forte intensité de connaissances, Adaptive case management, Gestion de cas, Réutilisation de connaissances, Extraction de connaissances, Analyse de concepts formels, Traitement automatique du langage, Numérisation de l'enseignement, Enseignement

CREA: Method for knowledge intensive process analysis, adaptation and reuse

use case in postgraduate computer science studies

Similarly to business domains, digitalisation and virtualisation of processes in higher education started a few decades ago. During COVID-19 pandemics, the interest in solutions for developing and delivering classes on-line grew substantially. Nevertheless, solutions allowing for efficient adaptation and reuse of the existing courses and teaching materials in the new circumstances are still lagging. Transformation of a traditional course to a virtual one, developing a new course adapted for the audience are some examples of " *knowledge intensive processes* ". These processes share common properties making digitalization hard, to cite some:

— they depend on the extensive knowledge and experience of teachers analysing the context and adapting class accordingly,

— they involve an intense collaboration between all stakeholders in the higher education environment,

— they require creativity for adapting to uncertain context.

Deploying new knowledge intensive processes, or adapting existing ones in this unexpected situation, faced multiple challenges already known from this specific research domain. The main issue is: how to best reuse existing knowledge, including online stored knowledge, where the abundance of sources and data makes it difficult to select the most suited to teachers' requirements.

In this thesis we propose the CREA method that enables a reuse of past cases in the higher education domain, particularly in courses preparation. The CREA method supports the reuse of existing courses materials: (i) it presents graphically the gap (i.e. semantic difference) between the courses and (ii) it summarises the class sessions in a form of clusters of main terms to discuss. Other types of documents can also be integrated within the input courses materials (including web pages, or research articles) in order to propose adapted clusters for specific audiences, or even state-of-the-art materials. This method relies on natural language processing tools in order to extract the terms regardless of the input language, then it uses formal concept analysis for computing metrics to build clusters of terms and assess the similarity of input materials. We also propose some preliminary results of a session scheduling method.

Keywords: Knowledge intensive process, Adaptive case management, Case Management, Knowledge Reuse, Knowledge Extraction, Formal concept analysis, Natural language processing, Digitalisation of education, Education