Identification du profil des utilisateurs d'un hypermédia encyclopédique à l'aide de classifieurs basés sur des dissimilarités : création d'un composant d'un système expert pour Hypergéo

Firas Abou Latif

Institut National des Sciences Appliquées de Rouen
Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes

Thèse de doctorat

Discipline : Informatique

présentée et soutenue le 8 juillet 2011 par

Firas ABOU LATIF

Pour obtenir le titre de

DOCTEUR DE L'INSA DE ROUEN

Identification du profil des utilisateurs d'un hypermédia encyclopédique à l'aide de classifieurs basés sur des dissimilarités

Création d'un composant d'un système expert pour Hypergéo

soutenue devant le jury composé de :

Mme Yolaine BOURDA	(Rapporteur)	Professeur à Supélec
M. Nicolas DELESTRE	(Encadrant)	Maître de Conférences (INSA de Rouen)
M. Jean-Marc LABAT	(Rapporteur)	Professeur des Universités (Université Pierre et Marie Curie)
M. Mustapha LEBBAH	(Examinateur)	Maître de Conférences (Université Paris 13)
M. Jean-Pierre PÉCUCHET	(Directeur)	Professeur des Universités (INSA de Rouen)

أهدي بحثي هذا

إلى من علمني ان الحياة عمل و تسامح وعطاء ... إلى من علمني ان الفرح يكبر عندما نتقاسمه ... إلى من نسج لي ورب النجاح وكان مثلي الأعلى في الحياة...

أبي الغالي

إلى التي رأتي قلبها قبل عينيها... إلى من أرى جمال الحياة بنور عينها... و روعة النجاح بدموع مقلتيها... إلى الحضن الذي أحن اليه ابدا... و الحب الذي يحيطني دوما... إلى أحلى كلمة نطقها لساني ...

أمي الحنونة

إلى من اختارها قلبي لتشاركني الحياة ... إلى البسمة التي تزيل من نفسي هموم الأيام ... إلى من كان لدعمها وتشجيعها دور كبير بانجاز هذه الاطروحة...

زوجتي الغالية

إلى وروو شارلتهن حنان الأب والأم ... إلى من تقاسمت معهن أفراح الماضي و آمال المستقبل ... إلى من أحسست بينهن بالحب و الحنان ...

ميساء - هبه - خزامى - غيداء

إلى من وجدت بينهم الالفة والبساطة... إلى العائلة التي أحاطتني بالحب... إلى أسرتي الثانية...

عائلة زوجتي

إلى الضحكة التي تؤنس أيامي ... إلى من لا أفهم كلامه ولكنني أطرب لسماعه ...

رامي الحبيب

إلى الأرواح التي اختارت ورب العزة و الكرامة... إلى الدماء الطاهرة التي سالت لتروي تراب الوطن...

À Amjaad
À ma famille
À Ramy
À ma belle famille

REMERCIEMENTS

Cette thèse a été réalisée dans le Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes (LITIS - EA 4108) à l'Institut National des Sciences Appliquées de Rouen (INSA de Rouen). Je tiens à remercier vivement tous ceux qui ont contribué à la réalisation de ce travail ainsi que les personnes qui ont accepté de l'évaluer.

Je tiens tout d'abord à exprimer mes remerciements à Madame Yolaine Bourda professeur à Supélec ainsi qu'à Monsieur Jean-Marc Labat professeur des universités à l'Université Pierre et Marie Curie (Paris 6) pour avoir accepté d'être les rapporteurs de ce travail et de participer au jury de ma thèse. Je tiens tout particulièrement à remercier Monsieur Mustapha Lebbah, Maître de Conférences à l'université Paris 13, pour avoir accepté d'examiner ce mémoire et pour sa participation à ce jury.

Je tiens aussi à remercier Monsieur Bernard Elissalde, professeur des universités au département de Géographie de l'université de Rouen, qui nous a permis d'utiliser le site Hypergéo pour nos recherches.

Je tiens également à remercier mon directeur de thèse, Jean-Pierre Pécuchet, pour m'avoir accueilli dans son équipe.

J'exprime mes plus sincères remerciements à Nicolas Delestre pour son encadrement et pour la confiance qu'il m'a accordée durant toute ma thèse, pour sa sympathie, son encouragement, le suivi de mon travail et sa disponibilité malgré son emploi du temps souvent très chargé. Son regard critique m'a été très précieux pour structurer et améliorer mon travail.

J'adresse ma profonde gratitude à Nicolas Malandain pour son encadrement, pour ses conseils avisés. Son aide m'a été précieux pour corriger et enrichir ce manuscrit.

Mes remerciements les plus respectueux vont à Alain Rakotomamonjy pour les nombreuses discussions enrichissantes que nous avons eues. Je souhaite remercier Stéphane Canu, directeur du LITIS, qui m'a accueillie au sein de son laboratoire. Je souhaite remercier tous mes

collègues, thésards et ex-thésards : Iyadh Cabani, Karina Zapien, Benjamin Labbé, Rémi Flamary, Florian Yger, Yaquian Li, Xilan Tian, Aurélie Boisbunon, Carlo Abi-Chahine, je leur souhaite toute la réussite qu'ils méritent.

Je tiens aussi à remercier Brigitte Biarra, Sandra Hague et Florence Aubry. Je remercie également Sebastien Bonnegent, Jean-Francois Brulard, Elsa Planterose, et Gilles Gasso qui ont tous contribué a créer une atmosphère chaleureuse et conviviale.

Je tiens enfin à remercier l'ensemble du personnel du laboratoire LITIS et du département ASI qui contribue à nous faciliter la vie au quotidien.

Cette page de remerciements ne pourrait se conclure sans que je remercie vivement ma famille et ma belle famille. Ma thèse et mon diplôme de Doctorat leur font un si grand plaisir. Je les remercie pour leurs encouragements, ils m'ont toujours aidé et soutenu durant ces années, malgré la distance qui nous séparait et pour bien plus que je ne saurais écrire.

Enfin, je remercie ma chère épouse pour sa patience, son soutien quotidien indéfectible et son enthousiasme à l'égard de mes travaux comme de la vie en général.

Mes remerciements et ma reconnaissance vont également à mon pays, la Syrie, qui m'a permis de terminer ma formation universitaire en France, en me fournissant les ressources nécessaires. Que tous ceux qui ont contribué à faciliter cette étude, dans le cadre qui fut le mien, trouvent ici le témoignage de ma sincère gratitude.

Table des matières

Liste des figures

Liste des figures

Liste des tableaux

Liste des tableaux

Liste des algorithmes

Introduction

Introduction

Aujourd'hui le réseau Internet, et plus spécifiquement le web, est de plus en plus utilisé et de plus en plus complet. Les sites web sont nombreux et leur taille est souvent grandissante. Le risque de se perdre dans un site et de ne plus comprendre ce que l'on lit (surcharge coginitive) a donné lieu il y a un peu plus de dix ans à des recherches sur les hypermédias adaptables et adaptatifs [JAC 06].

On dit qu'un hypermédia est adaptable lorsqu'il permet à l'utilisateur de spécifier son profil, ses centres d'intérêt, ses préférences, etc. afin que le site lui proposent uniquement des informations qui l'intéressent. C'est par exemple ce que fait un utilisateur lorsqu'il utilise des portails d'agrégation d'informations (tel que *iGoogle*). Mais c'est aussi ce que l'on retrouve sur des sites plus institutionnels tel que celui de l'INSA (*Cf.* la figure 1). Dans ce dernier cas, l'utilisateur a la possibilité de spécifier s'il est un candidat (étudiant qui va postuler à l'INSA), un parent, un enseignant/chercheur, un ancien élève ou un professionnel. Les liens proposés sont alors en rapport avec le profil.

FIGURE 1 : Page principale du site web de l'INSA de Rouen.

Les hypermédias adaptatifs quant à eux sont des hypermédias qui se modifient automatiquement en fonction de l'utilisateur. Cette adaptation peut être le résultat d'une évaluation de l'utilisateur, il doit dans ce cas répondre à un ou plusieurs questionnaires pour le cerner. C'est par exemple le cas des hypermédias adaptatifs pédagogiques. L'apprenant au cours de sa navigation est régulièrement questionné quant à sa compréhension du cours. Ainsi le système peut prendre des décisions concernant les informations qu'il va afficher ou des liens qu'il va propo-

ser. Mais cette adaptation peut aussi être issue d'une analyse de son comportement, qui peut être sa navigation dans le site ou des actions qu'il effectue. Par exemple aujourd'hui de nombreux sites d'achats en ligne vous proposent de nouveaux articles à acheter en fonction de ce que vous avez déjà achetés et de ce qu'ont achetés d'autres utilisateurs qui vous « ressemblent ».

Quelque soit le système utilisé, l'adaptation se fait aussi bien au niveau du contenu des pages qu'au niveau des liens. L'adaptation de navigation est l'adaptation la plus courante. Elle passe par la modification des liens de la page qui est destinée à l'utilisateur avant qu'elle lui soit envoyée. Plusieurs techniques sont utilisées [Bru 98], comme le guidage direct (l'ajout d'un bouton « suivant »), l'ordonnancement des liens (afin d'organiser les liens hypertextes de la page en fonction de leur importance pour l'utilisateur), le masquage des liens (qui enlève les liens hypertextes dont la destination ne peut être comprise ou n'est pas intéressante pour l'utilisateur), ou encore l'annotation des liens (qui permet d'indiquer à l'utilisateur l'objectif du document cible). Attention toutefois à ne pas assimiler adaptation au fait de modifier les liens hypertextes d'un site web dynamiquement. Par exemple, le site web que nous avons utilisé pour notre étude, c'est-à-dire le site web Hypergeo, génère des liens hypertextes au sein des pages avant de les envoyer à l'utilisateur. Toutefois, l'ajout de ces liens est indépendant de l'utilisateur qui utilise le site, et n'est fonction que des articles [1] présents dans le système, afin uniquement d'améliorer l'interconnectivité entre ces derniers. L'adaptation de contenu quant à elle, consiste à modifier les informations présentées par chaque page. Cette adaptation est plus complexe car elle requiert de modéliser les informations que le site présente, de posséder plusieurs médias pour présenter une notion et de vérifier si l'agencement des médias sélectionnés est cohérent [Del 00].

Toutefois, quelque soit l'adaptation réalisée et que le système soit adaptatif ou adaptable, le problème principal est de définir le profil de l'utilisateur. Cet objectif de recherche, comme nous allons le voir plus en détail dans ce mémoire, se dénomme le *Web Usage Mining*. Les techniques y afférentes sont principalement basées sur l'étude de la navigation des utilisateurs. Les traces de navigation sont composées d'éléments (documents parcourus, transitions, temps, etc.). Alors que la littérature montre qu'elles fonctionnent plutôt bien sur des sites possédant de nombreuses traces de navigation (comme par exemple [Mob 07a] et [Sri 00]), l'expérience montre qu'il n'en est pas de même sur des sites possédant nettement moins de traces. Ce dernier point est important car à part quelques sites web mondialement reconnus qui possèdent un fort taux de visites par jours (jusqu'à plusieurs millions, comme c'est le cas par exemple pour Wikipédia), beaucoup n'ont que quelques dizaines voire centaines de visites par jour. L'hypothèse que nous faisons est que ces techniques ne fonctionnent pas dans ce cas car elles font abstraction du site

1. Terme qui sera précisé au chapitre 3 de ce mémoire.

en lui même et du contenu des pages que visitent les utilisateurs. Nous pensons en effet que si nous prenons en compte l'organisation du site web et le contenu des pages visitées nous serons plus à même de caractériser un utilisateur, même si sa visite est très courte et si au préalable nous avons peu de traces en amont permettant de faire cette étude. Ce changement de point de vue amène aussi des changements quant aux techniques permettant de *classer* un nouvel utilisateur en fonction de son parcours. Ces techniques doivent utiliser une *similarité* entre traces, elle même utilisant une *similarité* entre les pages visitées. Encore faut il dans ce cas que les pages soient comparables, que le contenu du site soit bien structuré et qu'elles se rapportent à des domaines connexes, comme c'est le cas par exemple pour des encyclopédies.

Ceci définit donc les objectifs de ce mémoire. Après un chapitre sur les systèmes experts et les réseaux bayésiens, nous allons commencer par faire un état de l'art des techniques permettant de classer des données, que ces données soient au préalable étiquetées ou pas. On s'intéressera aux algorithmes de classification supervisée et non-supervisée. Nous nous focaliserons alors sur les algorithmes qui nécessitent uniquement une simililarité [2] entre éléments pour pouvoir les étudier. Nous nous attarderons aussi sur des méthodes de projection qui permettent d'évaluer en amont cette similarité et les méthodes statistiques permettant d'évaluer un algorithme de classification.

Dans le chapitre 3, après une description détaillée du *Web Usage Mining* et de l'un de ces algorithmes permettant d'identifier un utilisateur à partir de son parcours, nous présenterons le site Hypergéo et les données expérimentales que nous avons recueillies. Nous montrerons alors que dans ce contexte, il nous est impossible d'utiliser cet algorithme pour identifier le profil d'un nouvel utilisateur.

Nous présenterons dans le chapitre 4 deux propositions de dissimilarité entre traces. Nous les comparerons à l'algorithme du chapitre précédant à l'aide de données synthétiques [3] afin d'estimer dans quel cadre ces dissimilarités peuvent être utilisées. À chaque fois nous effectuerons des évaluations visuelles et statistiques à l'aide des algorithmes de projection et de classification.

Sachant que la similarité entre traces choisie au chapitre précédent utilisera une similarité entre éléments de ces dernières, nous étudierons dans le chapitre 5 deux similarités entre éléments de traces adaptées à notre contexte expérimental.

Enfin, avant une conclusion et quelques perspectives, l'analyse des résultats précédents nous amènera à ré-étiqueter nos données et à montrer que dans ce contexte les résultats obtenus avec notre proposition sont alors nettement satisfaisants.

2. Ce terme, ainsi que celui de distance ou de dissimilarité, seront alors précisés.
3. Dans ce manuscrit, nous utilisons ce terme pour les données générées artificiellement, c'est la traduction de de l'expression anglaise *synthetic data*.

CHAPITRE 1

Systèmes experts et réseaux bayésiens

Sommaire

1.1 Introduction

La construction de systèmes capables d'apprendre, d'analyser et de réagir comme l'être humain est un rêve de l'homme depuis qu'il a inventé la machine. Nous allons nous intéresser aux systèmes experts qui ont donné beaucoup d'espoir et ont donc été l'objet d'une production scientifique importante.

Dans ce chapitre nous les présenterons, en commençant par un petit positionnement historique, suivi par le cycle de développement de tels systèmes. Nous nous attarderons alors un peu plus sur leur architecture en présentant les deux éléments principaux que sont la base de connaissances et le moteur d'inférence. Pour ce dernier, deux types de mécanisme seront expliqués. Nous mettrons alors en évidence les limites de ces systèmes. Ceci nous amènera à présenter les réseaux bayésiens qui peuvent être considérés comme une évolution des systèmes experts avec un apport probabiliste.

1.2 Systèmes experts

Un système à base de connaissances est un système informatique utilisant des connaissances déjà acquises pour résoudre des problèmes ou pour aider l'homme à prendre des décisions dans un domaine donné [Ake 10]. La connaissance est définie par « les faits, les informations, ou les compétences que la personne acquière par l'expérience ou l'éducation » [1].

Un système expert est un cas particulier de systèmes à base de connaissances où cette base est construite à partir de l'expertise humaine. L'objectif d'un système expert est donc de modéliser les connaissances et le raisonnement de l'expert. Autrement dit, son but est de reproduire les compétences d'un expert pour résoudre un problème. Ce système doit donc connaître des faits dans le domaine d'expertise, être capable de les traiter et de déterminer la réponse la plus adaptée au problème. Historiquement l'un des premiers systèmes experts est MYCIN [Buc 84] qui a été créé en 1974. Il a été construit pour diagnostiquer les infections bactériennes. Un deuxième exemple est PROSPECTOR (1978) qui a pour objectif d'aider à évaluer la qualité d'un site géologique en vue d'une exploitation minière [McC 96].

1.2.1 Historique

Dès la fin des années 60 et le début des années 70, la recherche autour des systèmes experts a été très active et a donné lieu à de nombreuses publications et à l'implantation de nombreux

1. Traduction de la définition du dictionnaire *Oxford*.

systèmes. Une des raisons de cet engouement était le fait que les algorithmes classiques en intelligence artificielle (tels que les algorithmes de parcours de solutions de type $min - max$, $A*$, etc.) ne permettaient pas de résoudre certains problèmes liés à l'expertise, au diagnostique, etc. La communauté scientifique pensait qu'avec ces nouveaux systèmes nous pourrions aider, conseiller, voire remplacer les experts, comme les médecins, les géologues, ou encore les enseignants lorsque l'on a voulu adosser à ces systèmes des systèmes d'apprentissage. Par exemple les systèmes d'enseignement GUIDON [Cla 83] puis GUIDON2 [Cla 86] utilisaient les systèmes expert MYCIN et NEOMYCIN [2]. Parallèlement à des logiciels spécifiques, plusieurs langages de programmation logique ont vu le jour, comme par exemple la famille des langages Prolog, dont la première version a été créé en 1973 par Colmerauer et al. [Col 73]. Jusqu'au milieu des années 80, les systèmes experts représentaient le futur de l'informatique, ils ont même fait l'objet d'un programme du gouvernement japonais en tant que fondement théorique pour la création des ordinateurs de cinquième génération.

1.2.2 Cycle de développement d'un système expert

La création d'un système expert commence par le recueil de l'expertise d'un ou plusieurs experts. Les techniques et les méthodes afférentes à cette étape sont issues de l'ingénierie des connaissances. Les professionnels de ce domaine sont des informaticiens, spécialistes en intelligence artificielle. Ils discutent avec les experts et les observent durant leur résolution de problème. Cette expertise doit ensuite être formalisée. On obtient alors des informations ou des relations qui sont toujours vraies dans le domaine étudié, c'est ce que l'on nomme les faits, et des compétences de déduction que l'on nomme les règles. Quelques fois, certaines connaissances « heuristiques », des bonnes pratiques, peuvent aussi être formalisées.

Par la suite, en fonction de ce que l'on demandera au système, l'une des deux techniques de résolution de problèmes que sont le « chaînage avant » ou le « chaînage arrière » (et quelque fois un mixte de ces deux techniques) est choisie.

Enfin, une fois développés, les systèmes experts doivent être validés. Cette validation peut être effectuée « à la main », à partir d'exemples concrets. Mais à l'image des outils de tests unitaires en génie logiciel, des *framework* de tests de systèmes experts ont été développés dans les années 80 [OLe 88].

2. Cela a représenté l'age d'or de l'Enseignement Intelligent Assisté par Ordinateur.

1.2.3 Architecture d'un système expert

La figure 1-1 représente la structure générale d'un système expert, les blocs représentent ses composants principaux (utilisateur, expert, base de connaissances, moteur d'inférence). Comme nous l'avons vu, les connaissances sont acquises à partir de(s) expert(s) du domaine. Ce sont des connaissances théoriques, académiques et/ou des connaissances heuristiques obtenues par la pratique.

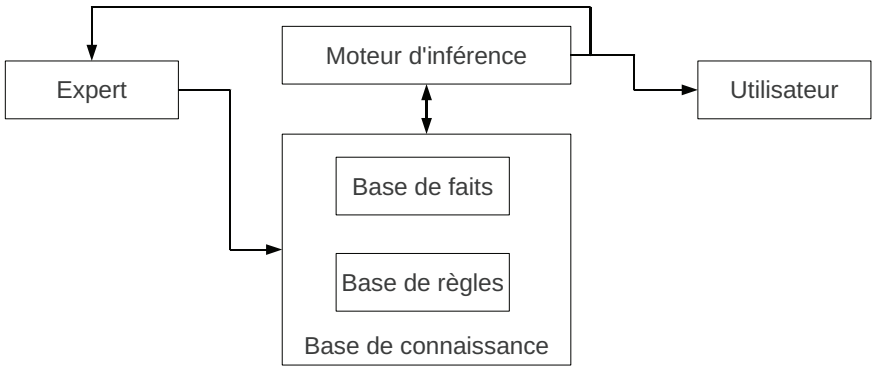


FIGURE 1-1 : L'architecture d'un système expert.

Nous allons détailler les deux éléments principaux que sont la base de connaissances et le moteur d'inférence.

1.2.3.1 Base de connaissance

La construction de la base de connaissance est une étape essentielle et difficile à réaliser. En effet, le savoir-faire de l'expertise est souvent acquis au terme d'une grande expérience, et il devient très difficile de le formaliser. Ensuite l'expert n'a souvent pas conscience de son raisonnement, il n'arrive pas à l'expliquer. Par conséquent, il est très difficile de mettre en œuvre des représentations de ces connaissances liant l'expertise humaine et le langage de programmation.

Dans la figure 1-1, nous voyons que la base de connaissances est constituée de deux parties : la base de règles et la base de faits.

Base de règles

La base de règles est l'ensemble des savoirs-faires du système. Elle est totalement formalisée par l'ingénieur des connaissances. Le système n'est pas capable de générer de nouveaux

savoirs-faires. Ces savoirs-faires sont représentés par des règles logiques (des clauses de Horn), de la forme « SI prémisses ALORS conclusion », comme par exemple « Si l'animal a une colonne vertébrale alors il est vertébré ».

Base de faits

Cette base représente la mémoire du système expert. Elle contient ce qui est vrai dans le domaine, les observations fournies par l'utilisateur et les résultats temporaires du moteur d'inférence. Au début, elle est uniquement constituée de faits et des informations fournies par l'utilisateur sur son problème. Au fur et à mesure, elle est complétée par le moteur d'inférence.

1.2.3.2 Moteur d'inférence

Le moteur d'inférence utilise la base de règles pour répondre à la demande de l'utilisateur. Il existe deux principaux types de moteurs d'inférence, le moteur guidé par le but (le chaînage arrière) et celui guidé par les données (le chaînage avant).

Chaînage arrière

Dans le cas où l'utilisateur cherche une explication ou une validation de sa demande, comme par exemple dans le cas du diagnostique d'une panne de voiture, le système expert utilise le chaînage arrière. Tout d'abord, il met la demande dans la base de faits. Puis il cherche toutes les règles ayant cette demande comme conclusion. Ses prémisses seront les nouveaux buts à résoudre. Le système travaille récursivement jusqu'à tous les buts de la base de faits aient été vérifiés. L'algorithme 1 de Lamare [3] formalise cette technique. Il est à noter que la plupart des moteurs prolog fonctionne de cette manière.

Chaînage avant

Le principe du chaînage avant est de partir de la base de faits et d'inférer de nouveaux faits à partir des règles, qui seront ajoutés à cette base. Lorsque l'utilisateur ne donne pas d'objectif de validation, le système itère jusqu'à ce qu'aucun nouveau fait ne puisse être engendré par les règles. Par contre si un but est donné, le processus itère jusqu'à ce que le but appartienne à la base des faits, ou jusqu'à ce qu'il ne puisse plus générer de nouveaux faits. Ceci peut être formalisé par l'algorithme 2 [3]

3. Inspiré du cours « Intelligence artificielle » de Philippe Lamare de l'UFR de Nantes

Algorithme 1 Algorithme du chaînage arrière $backward\_chaining(BR, BF, But)$

Entrée: BR : la base de règles,
BF : la base de faits,
But : l'ensemble des buts.
Sortie: res : Booléen permettant de savoir si le but à été atteint.

début
  si $But = \phi$ **alors**
    $res \leftarrow$ VRAI
  sinon
    $elem \leftarrow$ le premier élément dans But
    $rest \leftarrow But \setminus \{elem\}$
    si $demBut(BR, BF, elem)$ **alors**
      $res \leftarrow backward\_chaining(BR, BF, rest)$
    sinon
      $res \leftarrow$ FAUX
    finsi
  finsi
fin

---

fonction $demBut(BR, BF, curBut)$

Entrée: BR : la base de règles,
BF : la base de faits,
$curBut$: la conclusion à valider.
Sortie: res : Booléen indiquant la validation de la conclusion.

début
  si $curBut \in BF$ **alors**
    $res \leftarrow$ VRAI
  sinon
    $curBR \leftarrow BR$
    $res \leftarrow$ FAUX
    tant que ($curBR \neq \phi$) et (non res) **faire**
      $tmpR \leftarrow r \in curBR$
      $curRB \leftarrow curBR \setminus \{tmpR\}$
      si $conclusion(tmpR) = curBut$ **alors**
        $res \leftarrow backward\_chaining(BR, BF, premisses(tmpR))$
      finsi
    fintantque
  finsi
fin

Algorithme 2 Algorithme du chaînage avant $forward\_chaining(BR, BF, But)$

Entrée: BR : la base de règles,
BF : la base de faits,
But : l'ensemble des buts.
Sortie: res : Booléen indiquant la validité des buts.

début
  si $But \subset BF$ **alors**
    $res \leftarrow$ VRAI
  sinon
    $RnonDecl \leftarrow BR$
    $RaCons \leftarrow BR$
    $res \leftarrow$ FAUX
    tant que ($RaCons \neq \phi$) et (non res) **faire**
      $tmpR \leftarrow r \in RaCons$
      $RaCons \leftarrow RaCons \setminus \{tmpR\}$
      si $\forall p \in premisses(tmpR), p \in BF$ **alors**
        $BF \leftarrow BF \cup \{conclusion(tmpR)\}$
        $RnonDecl \leftarrow RnonDecl \setminus \{tmpR\}$
        $RaCons \leftarrow RnonDecl$
        si $conclusion(tmpR) = But$ **alors**
          $res \leftarrow$ VRAI
        finsi
      finsi
    fintantque
  finsi
fin

1.2.4 Limites des systèmes experts

Aujourd'hui, à par quelques exemples célèbres comme le système MEPRA qui permet d'aider à la prévention des risques d'avalanche, les systèmes experts ne sont plus beaucoup étudiés et utilisés. En effet, créer un système expert est un processus coûteux et est difficilement évolutif. Coûteux à développer car formaliser l'expertise d'un domaine est un problème complexe qui peut prendre beaucoup de temps. Difficilement évolutifs car la modification, la suppression ou l'ajout de règles peut entraîner des dysfonctionnements. Or il existe peu de domaines qui n'évoluent pas.

Mais surtout, dans le monde réel, il est rare que l'on puisse toujours répondre aux questions sans aucun doute ou de représenter les faits de manière exacte. Dès lors, l'incertitude et l'approximation ont besoin d'être représentées aussi bien au niveau des faits qu'au niveau des règles. Ceci ne peut pas être représenté avec la logique utilisé par les systèmes experts classiques, il faut donc d'autres mécanismes de représentation de ces informations. Plusieurs techniques ont été proposées comme par exemple la logique floue, les facteurs de certitudes, les réseaux bayésiens, etc. Nous allons présenter et utiliser cette dernière technique dans la suite de ce mémoire.

1.3 Réseaux bayésiens

Les réseaux bayésiens mixent la théorie des graphes avec celle des probabilités [Pea 88]. Un réseau bayésien est composé d'un graphe orienté acyclique. Les liens entre les nœuds représentent les dépendances entre eux (la probabilité conditionnelle). Plus formellement, le réseau bayésien est défini par [Nai 04] :

- un graphe orienté acyclique $G = (X, E)$, où X est l'ensemble des nœuds du graphe et E l'ensemble des arcs entre les nœuds,
- un espace probabiliste fini (Ω, Z, p),
- un ensemble de variables aléatoires définies sur (Ω, Z, p) et associées aux nœuds du graphe, tel que :

$$P(x_1, x_2, ..., x_n) = \prod_{i}^{n} P(x_i | parent(x_i))$$

De par leur représentation graphique et leurs informations associées à des probabilités, les réseaux bayésiens sont aussi dénommés « systèmes experts probabilistes » [Ler 06].

1.3.1 Structure du réseau bayésien

À l'image des règles dans les systèmes experts, la structure du réseau bayésien a pour objectif de traduire les notions de causalité entre les faits. En effet, dans un système expert, les règles sont composées de prémisses et d'une conclusion. Ces prémisses et conclusion se traduisent alors par les nœuds du graphe, ceux correspondant aux prémisses sont alors les sources des arcs vers le nœud conclusion.

Tout comme les règles, la structure d'un réseau bayésien est issue du recueil d'expertise. Il est généralement réalisé par des spécialistes en ingénierie des connaissances. Toutefois, depuis quelques années les chercheurs du domaine essaient de construire des systèmes qui seraient capable de structurer les réseaux bayésiens en fonction de données, de corpus, etc. [Oli 06]

La structure d'un réseau peut aussi dépendre de l'utilisation finale que nous allons faire du réseau. Il existe des structures types comme les réseaux bayésiens naïfs, réseaux bayésiens naïfs augmentés, réseaux bayésiens multi-nets, etc [Che 99, Che 01]. Par exemple, les réseaux bayésiens naïfs, que nous verrons en détail par la suite, permettent de choisir la classe d'une nouvelle donnée. C'est typiquement ce type réseau qui est aujourd'hui utilisé dans nos lecteurs de mails pour savoir si un message est du *spam* ou non.

1.3.2 Définition de la probabilité conditionnelle

Bien que dans un système expert les faits et les observables (informations fournies par l'utilisateur) sont représentés par la base de faits indépendemment de la base de règles, dans les réseaux bayésiens, ces informations sont intrinsèques aux graphes. Pour les renseigner, il suffit d'affecter des probabilités à certains nœuds. C'est grâce à ce mécanisme que l'on peut représenter l'incertain.

Un des avantages des réseaux bayésiens est de pouvoir représenter des expertises différentes sur un domaine. Lorsque les experts ont des points de vue différents, certains réseaux bayésiens sont capables de fusionner leurs avis. La figure 1-2, issue de [Ler 06], représente le résolution d'un problème de diagnostique de panne avec quatre points de vue différents (par exemple celui d'un électricien, d'un mécanicien, d'un énergéticien, etc.). $Etat(t)$ représente une partie du réseau bayésien contenant les observables. $Etat(t+1)$ est celle contenant la décision.

1.3.3 Inférence

L'inférence dans les réseaux bayésiens revient à calculer la probabilité conditionnelle d'un ou plusieurs nœuds sachant que certains nœuds sont des observables. Pearl et Jensen présentent

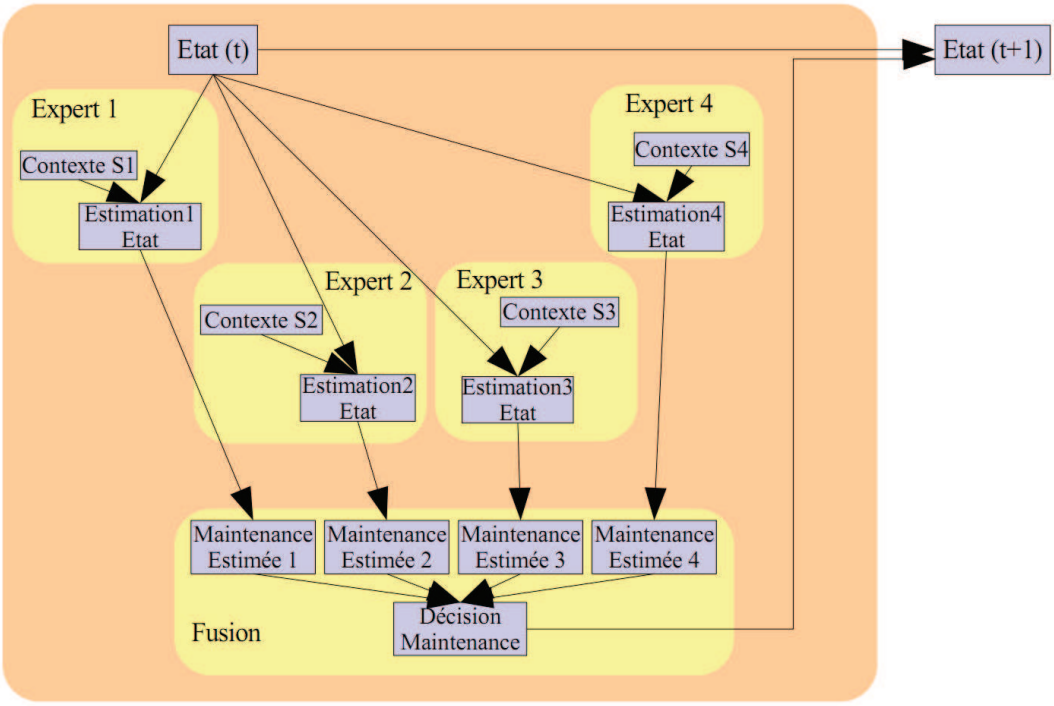


FIGURE 1-2 : Un exemple de réseau bayésien : application à la modélisation de la maintenance d'un système à partir de plusieurs expertises.

plusieurs techniques d'inférence [Pea 88] [Jen 96]. Mais toutes ces techniques utilisent la théorie de Bayes qui permet de calculer la probabilité conditionnelle d'une variable aléatoire à partir de la probabilité d'autres variables.

1.3.4 Exemple

Afin de bien comprendre ce que nous venons de voir, prenons l'exemple présenté par la figure 1-3 [4]. L'objectif ici est de déterminer quelles sont les causes du déclenchement d'une alarme d'une bijouterie. L'expertise a montré qu'il y a deux causes principales : un cambriolage et un tremblement de terre. Dans ce dernier, les radios diffuseraient alors des messages qui seraient repris par les télévisions.

Les faits sont que :

– la probabilité qu'il y ait un cambriolage est de $0{,}001$ ou de $0{,}999$;

– la probabilité qu'il y ait un séisme est de $0{,}0001$ ou de $0{,}9999$;

4. Exemple issue du cours d'introduction au réseau bayésien de P. Leray à l'INSA de Rouen.

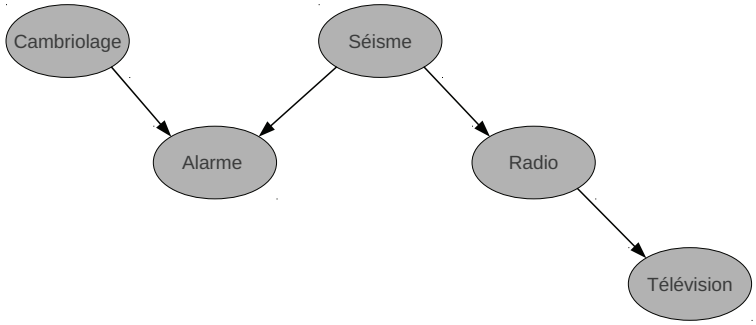


FIGURE 1-3 : Un exemple jouet de réseau bayesien.

On peut alors calculer la probabilité que l'alarme se déclenche ou pas, c'est à dire $P(Alarme|Cambriolage, Séisme)$:

	(Cambriolage, Séisme)			
	(Oui,Oui)	(Oui,Non)	(Non,Oui)	(Non,Non)
Alarme = Oui	0,75	0,10	0,99	0,10
Alarme = Non	0,25	0,90	0,01	0,90

TABLEAU 1-1 : Calcul de $P(Alarme|Cambriolage, Sisme)$.

On peut aussi calculer la probabilité qu'il y ait un cambriolage en fonction du fait que l'alarme sonne.

1.4 Conclusion

La construction d'un système expert est une idée ancienne, mise en œuvre avec le premier système créé au début des années 70. Dans ce chapitre nous nous sommes attardés sur leurs caractéristiques et sur la façon de les concevoir en explicitant plus en détail le mécanisme d'inférence. Nous avons mis en évidence leurs limites ce qui nous a amenés à étudier les réseaux bayésiens.

Les systèmes experts comme les réseaux bayésiens sont des systèmes d'aide à la décision qui demandent une expertise poussée en amont. Nous allons voir dans le chapitre suivant, d'autres techniques capables de construire leur fonction de décision sans l'expertise humaine, mais directement à partir des données.

CHAPITRE 2

Apprentissage supervisé et Apprentissage non supervisé

Sommaire

2.1 Introduction

Les méthodes d'apprentissage cherchent à déterminer des classifieurs à partir d'un ensemble de données. Chaque donnée est présentée par le couple (x, y). $x \in X$ est une donnée observée, $y \in Y$ est un label à prédire.

L'ensemble des données peut être soit linéairement séparable, soit non linéairement séparable ou pas du tout séparable. De même, ces données peuvent être représentées à l'aide d'attributs ou à l'aide de modèles plus complexes comme par exemple les graphes. Pour les données ayant des attributs, ces attributs peuvent être [Wit 05] :

- des attributs qualitatifs (attributs nominaux) comme par exemple un énuméré représentant des couleurs. Nous ne pouvons pas effectuer de calcul sur ces attributs ;
- des attributs quantitatifs (attributs ordinaux) comme par exemple un réel représentant une température ;
- des attributs composés de sous-attributs comme par exemple une date.

Le label y est une valeur discrète (souvent symbolique), ou continue (forcément numérique).

Le problème de l'apprentissage est donc de trouver la fonction f, $f : X \to Y$, qui lie chaque x à un certain y. Il y a deux types principaux d'apprentissage : l'apprentissage supervisé (la classification) et l'apprentissage non supervisé (le *clustering*).

Avant d'expliciter ces deux types d'apprentissage, nous allons présenter la notion de distance qui prend une place importante dans ces méthodes.

2.2 Distance, dissimilarité et similarité

La distance sur l'ensemble X est la fonction d définie par :

$$d : X \times X \to \mathbb{R}^+$$

La distance vérifie les propriétés suivantes :

1. $\forall x_1, x_2 \in X$; $d(x_1, x_2) = d(x_2, x_1)$ (la symétrie)
2. $\forall x_1, x_2 \in X$; $d(x_1, x_2) = 0 \Leftrightarrow x_1 = x_2$ (l'identité)
3. $\forall x_1, x_2, x_3 \in X$; $d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$ (l'inégalité triangulaire)

Selon le cas, l'algorithme d'apprentissage utilise une distance d ou une similarité Sim. Ces deux notions se différencient par le fait que :

– une similarité est bornée (comprise entre 0 et 1), alors que la distance ne l'est pas,

– la valeur 1 pour une similarité est équivalente à la valeur de 0 pour la distance (deux objets égaux auront une valeur de similarité de 1),

– une similarité ne valide pas obligatoirement l'inégalité triangulaire, nous avons juste :

1. $\forall x_1, x_2 \in S\,;\, Sim(x_1, x_2) = Sim(x_2, x_1)$
2. $\forall x_1, x_2 \in S\,;\, Sim(x_1, x_2) = 1 \Leftrightarrow x_1 = x_2$

Il est possible de créer une similarité à partir d'une distance. Toutefois, l'inverse n'est pas forcément vrai, à cause de la propriété d'inégalité triangulaire. Les transformations les plus courantes sont :

– $Sim(x_1, x_2) = \frac{1}{1+d(x_1,x_2)}$

– $Sim(x_1, x_2) = e^{-d(x_1,x_2)}$

– $Sim(x_1, x_2) = e^{-\frac{d(x-1,x_2)^2}{\sigma}}$

Il existe enfin la notion de dissimilarité. Cette mesure est une restriction de la distance car elle ne valide pas l'inégalité triangulaire. De plus elle est souvent normalisée. Dans la littérature, le terme de distance non métrique est aussi utilisé comme synonyme de la dissimilarité.

2.3 Apprentissage non supervisé (*clustering*)

Dans l'apprentissage non supervisé, les informations sur les labels de données ne sont pas disponibles. L'objectif est de regrouper les données selon leurs similarités. Il existe plusieurs méthodes d'apprentissage non supervisé comme l'ACP (l'Analyse en Composantes Principales, *Principal Component Analysis*) qui fait l'apprentissage en utilisant un nombre d'attributs de données moins grand que le nombre d'attributs initiaux. Il y a aussi l'algorithme des *Kmeans* et les cartes de Kohonen que nous allons présenter en détail.

2.3.1 *Kmeans*

Le but de l'algorithme *Kmeans* est de regrouper les données en k partitions différentes. Le nombre de groupes k est défini par l'utilisateur. L'algorithme *Kmeans* est ancien, puisque publié en 1967 [Mac 67], mais il est encore largement utilisé aujourd'hui, grâce à sa rapidité d'exécution et sa facilité d'utilisation.

2.3. Apprentissage non supervisé (*clustering*)

Toutefois, contrairement à d'autres algorithmes de *clustering* qui permettent à une donnée d'appartenir à plus d'un groupe, dans l'algorithme *Kmeans* une donnée ne peut appartenir qu'à un seul groupe.

Pour l'initialisation de l'algorithme *Kmeans*, k moyennes ou centroïdes sont choisis. Ensuite les distances entre ces centroïdes et toutes les données sont calculées et les données sont associées au centroïde le plus proche. Puis, le centroïde de chaque groupe est recalculé. Ces étapes sont répétées jusqu'à ce que les groupes ne changent plus.

La distance euclidienne [1] est souvent utilisée dans l'algorithme des *Kmeans*, mais d'autres distances peuvent être appliquées comme *cityblock* [2], cosinus [3], etc.

L'algorithme 3 représente les étapes du *Kmeans* [Pre 09].

Algorithme 3 *Kmeans*

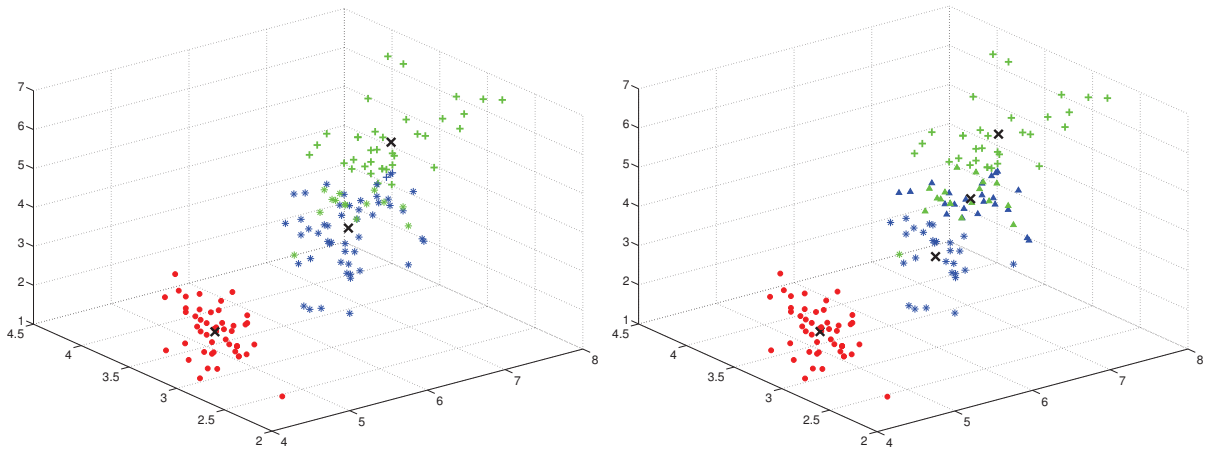
Entrée: X un ensemble de N exemples, k.
Sortie: X dégroupé en $G_{j\in\{1,...,k\}}$ groupes.
début
  Prendre k centroïdes $c_{j\in\{1,...,k\}}$ au hasard
  $G_{j\in\{1,...,k\}} \leftarrow \emptyset$
  répéter
    $G^*_{j\in\{1,...,k\}} \leftarrow \emptyset$
    pour $i \in \{1, ..., N\}$ **faire**
      $m \leftarrow \underset{m\in\{1,...,k\}}{\arg\min}\, d(x_i, c_m)$
      $G^*_m \leftarrow G^*_m \bigcup \{x_i\}$
    finpour
    pour $j \in \{1, ..., k\}$ **faire**
      $G_j \leftarrow G^*_j$
      $c_j \leftarrow$ le nouveau centroïde de G_j
    finpour
  jusqu'à ce que $G_{j\in\{1,...,k\}}$ ne changent pas
fin

L'algorithme *Kmeans* dépend de la sélection des centroïdes dans la phase d'initialisation. Si ces centroïdes sont mals sélectionnés, le *Kmeans* peut donner un mauvais résultat. C'est pourquoi, il est conseillé de répéter l'algorithme *Kmeans* plusieurs fois avec différentes initialisations de centroïdes.

La figure 2-1 présente les résultats du *Kmeans* sur l'ensemble de données Iris [Fis 88] [4]. Cette base de données, très souvent utilisée dans le domaine de l'apprentissage, est compo-

1. La distance euclidienne : $d(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$
2. La distance *cityblock* : $d(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|}$
3. La distance cosinus : $d(x,y) = 1 - \frac{x.y}{||x||.||y||}$
4. Les résultats sont visualisés dans l'espace 3D en utilisant un algorithme de projection.

sée de 150 exemples appartenant à trois classes *Setosa, Versicolor, Virginica* (50 exemples par classe). Chaque exemple est une fleur décrite à l'aide de quatre attributs numériques : la longueur et la largeur des sépales, ainsi que la longueur et la largeur des pétales. Dans cette figure nous pouvons voir l'influence de la valeur de k ($k = 3$ et $k = 4$). Nous avons utilisé les couleurs pour distinguer les exemples, *Setosa* en rouge, *Versicolor* en bleu, et *Virginica* en vert.



(a) $k = 3$ (b) $k = 4$

FIGURE 2-1 : L'application de *Kmeans* sur l'ensemble de données Iris.

Pour $k = 3$ (la figure 2-1(a)), le groupe 1 (les ronds) est constitué de 50 exemples de *Setosa*, le groupe 2 (les étoiles) de 48 exemples de *Versicolor* et 14 exemples de *Virginica*, et le groupe 3 (les croix) de 2 exemples de *Versicolor* et 36 exemples de *Virginica*.

Pour $k = 4$ (la figure 2-1(b)), les exemples sont regroupés en :

Groupe 1 (les ronds) : 50 exemples de *Setosa*.
Groupe 2 (les étoiles) : 27 exemples de *Versicolor* et 1 exemple de *Virginica*.
Groupe 3 (les triangles) : 23 exemples *Versicolor* et 17 exemples de *Virginica*.
Groupe 4 (les croix) : 32 exemples de*Virginica*.

Le *Kmeans* est un algorithme simple, rapide mais sensible au choix de la valeur k. Nous verrons dans les chapitres suivants, que nous utiliserons des algorithmes de projections de données pour nous aider à sélectionner la bonne valeur de k.

2.3.2 Carte de Kohonen

La carte de Kohonen est aussi appelée SOM (*Self-organizing map*). C'est un type de réseau de neurones qui est à la fois un mécanisme de projection de données et un algorithme d'apprentissage non supervisé.

2.3. Apprentissage non supervisé (*clustering*)

Une carte de Kohonen se compose de K unités (ou neurones) et P entrées. Chaque unité i est représentée par un vecteur de poids w_i de la même dimension que les données d'entrées P. Chaque unité i reçoit les P entrées par une connexion de poids w_{ie} où e indique l'une des P entrées [Koh 01] (*Cf.* figure 2-2, chaque cercle représente une unité).

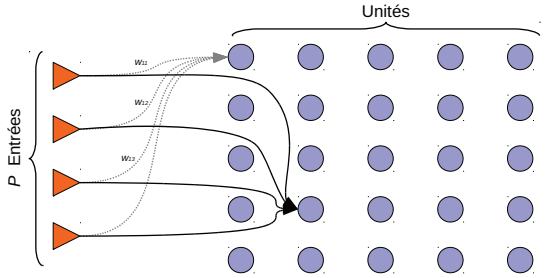


FIGURE 2-2 : La carte de Kohonen.

Le nombre d'unités utilisées dans la carte de Kohonen varie selon l'objectif. Pour projeter les exemples, le nombre d'unités choisi doit être assez grand pour obtenir une représentation qui projette au mieux les exemples. L'utilisation d'une carte de Kohonen pour faire de la classification non supervisée demande une unité pour chaque classe.

Avant de pouvoir utiliser une carte de Kohonen, pour l'apprentissage non supervisé, ou la projection de donnée, il faut l'initialiser à l'aide d'une phase d'apprentissage. Dans chaque étape de la phase d'apprentissage, le vecteur x représente une donnée. Les distances entre ce vecteur et toutes les unités k (tous les vecteurs de poids w) sont calculées. L'unité gagnante m_g est celle qui est la plus proche de x. Elle s'appelle le *BMU* (*Best-Matching Unit*) [Ves 00]

$$m_g = \arg\min_j d(x, w_j) \quad , \quad j \in P \tag{2-1}$$

Ensuite, les vecteurs w de l'unité gagnante et de tous ses voisins sont mis à jour par l'équation suivante :

$$w_i(t+1) = w_i(t) + \alpha(t) h_{gi}(t)(x - w_i(t)) \tag{2-2}$$

où t représente une étape de la phase d'apprentissage, et α est le taux d'apprentissage qui est une fonction décroissante avec t. Elle peut être une fonction linéaire [5], puissance [6], inversée [7], etc. h_{gi} est la fonction de voisinage. C'est une fonction décroissante ou stable comme par exemple la fonction gaussienne.

5. $\alpha(t) = \alpha_0(1 - \frac{t}{T})$, où T est la durée d'apprentissage, α_0 est le initial taux d'apprentissage
6. $\alpha(t) = \alpha_0(c \times \alpha_0)^{\frac{t}{T}} \quad , \quad c < 1$
7. $\alpha(t) = \frac{\alpha_0}{1+\frac{t}{T}}$

Ce processus d'apprentissage est répété plusieurs fois jusqu'à avoir une faible modification des poids (*Cf.* 2-3). Ceci est formalisé par l'algorithme 4 issu de [Pre 09].

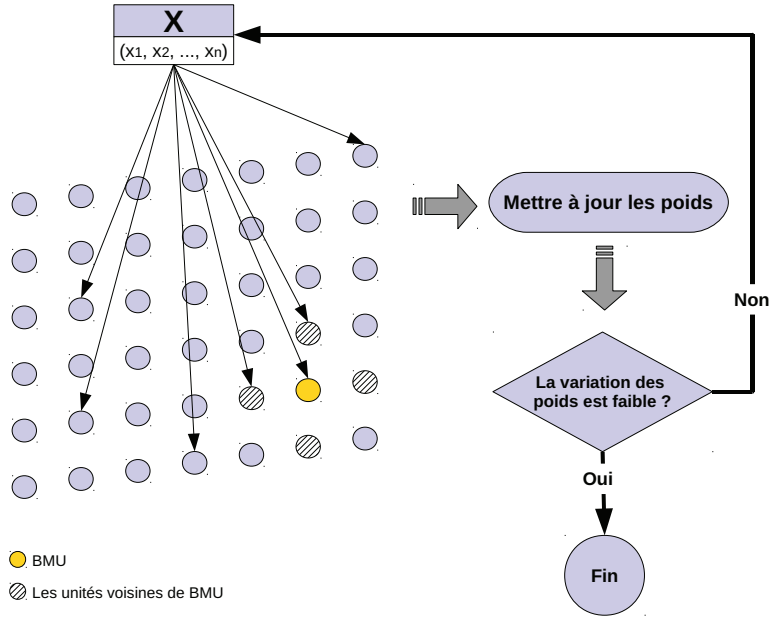


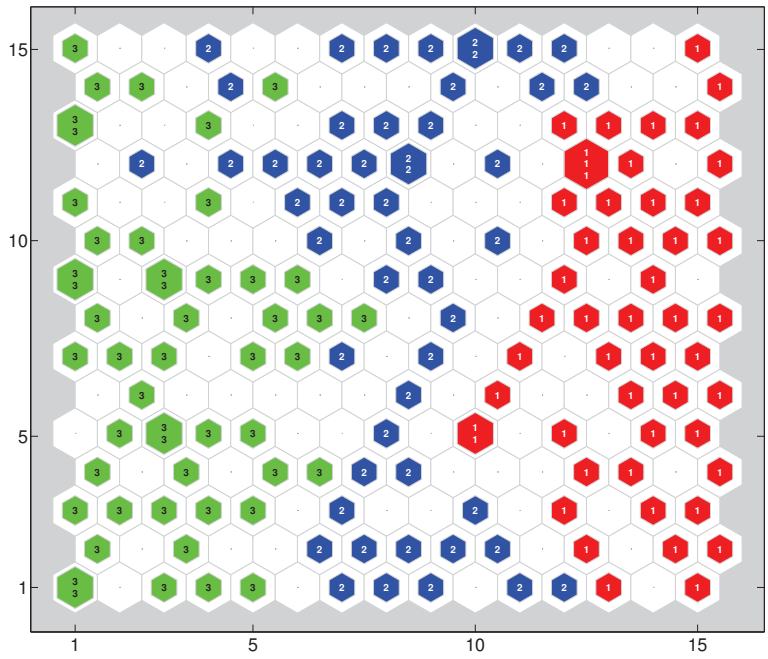
FIGURE 2-3 : La phase d'apprentissage de la carte de Kohonen.

La figure 2-4 présente l'utilisation de cartes de Kohonen à partir de l'ensemble de données Iris [Fis 88].

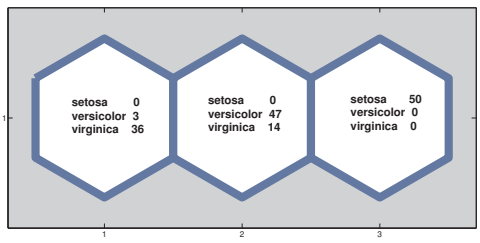
Dans la figure 2-4(a), nous utilisons une grille de 15×15 unités, pour montrer l'utilisation des cartes de Kohonen en projection. Les couleurs et le numéro de la classe sont utilisés pour distinguer les unités entre elles (*Setosa* en rouge et de valeur 1, *Versicolor* en bleu et de valeur 2, et *Virginica* en vert et de valeur 3). Cette figure nous permet aussi de voir les exemples liés à chaque unité. À titre d'exemple, 2 exemples de *Virginica* sont liés à l'unité $(1, 1)$, et 1 exemple de *Versicolor* est lié à l'unité $(7, 1)$.

Dans la figure 2-4(b) nous utilisons une grille de 1×3 sur ces mêmes données. Cette figure nous permet de voir l'utilisation de carte de Kohonen pour faire de l'apprentissage non supervisé. Dans la figure 2-4(b) est indiqué le nombre des exemples de chaque classe qui est lié à chaque unité. Par exemple, il y a 3 exemples de *Versicolor* et 36 exemples de *Virginica* liés à l'unité $(1, 1)$.

À la différence de l'algorithme du *Kmeans*, les cartes de Kohonen peuvent être utilisées pour deux objectifs différents. Elles peuvent être utilisées pour déterminer le nombre de classes



(a) 15×15 unités



(b) 1×3 unités

FIGURE 2-4 : Nombre d'exemples liés à chaque unité.

Algorithme 4 L'étape d'apprentissage de la carte de Kohonen.

Entrée: X un ensemble de N exemples.
Sortie: X grouper selon la carte de Kohonen.
début
  mélanger les exemples
  répéter
    pour $x_i \in X$ **faire**
      placer x_i en entrée
      déterminer l'unité la plus activée (unité gagnante)
      déterminer les unités voisines de l'unité vainqueur
      mettre à jour les poids des unités
    finpour
  jusqu'à ce que la variation des poids est faible
fin

intrinsèques aux données (c'est le cas lorsqu'on les utilise en tant qu'algorithme de projection). De plus, elles peuvent être utilisées dans le but de faire de l'apprentissage non supervisé.

2.4 Apprentissage supervisé (classification)

Dans l'apprentissage supervisé contrairement au non supervisé, les classes des données sont connues à l'avance. Le problème de la classification supervisée peut donc être défini de la façon suivante : soit un ensemble de données $X = \{x_1, x_2, ..., x_n\}$ et un ensemble de classes (ou étiquettes) $Y = \{y_1, y_2, ..., y_k\}$. L'objectif de la classification est de trouver la fonction de décision $f : X \rightarrow Y$. Cette fonction, appelée le classifieur, regroupe les données X dans des ensembles Y tel que $y_j = \{x_i, \quad f(x_i) = y_j \quad ; \quad x_i \in X\}$.

Cette définition caractérise le problème de la classification supervisée comme étant une projection des données dans un espace de classes. L'obtention de la fonction f se fait à partir des données d'apprentissage préalablement étiquetées. Par la suite, cette fonction est utilisée sur toute nouvelle donnée pour déterminer sa classe [FAO 07].

Il existe plusieurs types de fonctions de décision. Certaines utilisent toutes les données de la base d'apprentissage pour prendre la décision, on dit que ce sont des algorithmes non parcimonieux. Au contraire, les algorithmes utilisant uniquement une partie de la base d'apprentissage pour décider sont appelés algorithmes parcimonieux. Les algorithmes parcimonieux seront donc à privilégier avec de grandes bases d'apprentissage. De plus certains algorithmes ne font pas d'hypothèse sur le modèle que suivent les données, on les appelle des classifieurs non paramétriques. Un exemple d'algorithme paramétrique est le réseau bayésien naïf, que nous verrons dans le chapitre suivant. Celui-ci fait l'hypothèse qu'il n'y a pas de corrélations

entre les attributs des données.

Il y a plusieurs algorithmes qui permettent de résoudre le problème de l'apprentissage supervisé comme par exemple la régression linéaire, la régression logistique, les arbres de décision, le $k-ppv$ ou encore les SVM. Nous allons présenter plus en détail ces deux derniers puisque le premier est l'algorithme non parcimonieux le plus connu, et que le deuxième est un algorithme parcimonieux très utilisé.

2.4.1 Algorithme des k plus proches voisins

L'algorithme $k-ppv$ (k plus proches voisins, ou en anglais *k-nearest neighbor KNN*) est une méthode de classification supervisée non paramétrique et non parcimonieuse.

Elle suppose que seulement les k plus proches voisins influent sur la décision de la classification d'une nouvelle donnée [Tac 06]. Dès lors, la classe de la nouvelle donnée est la classe la plus représentée parmi les k plus proches exemples [Vin 03].

Le principe de l'algorithme du $k-ppv$ est :

- de trouver les k exemples les plus proches de la nouvelle donnée,
- d'attribuer la classe de la nouvelle donnée à celle de la classe de majorité des k plus proches exemples.

Ceci peut être formalisé par l'algorithme 5 (inspiré de l'algorithme proposé par P. Preux [Pre 09]), qui utilise la notion de la similarité pour identifier les k plus proches voisins.

Algorithme 5 $k-ppv$

X un ensemble de N exemples
Entrée: Une donnée x_{new}
$k \in \mathbb{N}^+$
Sortie: La classe de x_{new}
début
  pour tous $x_i \in X$ **faire**
    Calculer la similarité entre x_i et x_{new} : $Sim(x_i, x_{new})$
  finpour
  pour j $\leftarrow$1 **à** k **faire**
    $kppv[j] \leftarrow \arg\max_{i\in\{1,\ldots,N\}} Sim(x_i, x_{new})$
    $Sim(x_i, x_{new}) \leftarrow 0$
  finpour
  Déterminer la classe de x_{new} à partir de la classe des exemples dont le numéro est stocké dans le tableau $kppv$
fin

La figure 2-5 montre l'influence de la valeur k dans la décision de la classification. Lorsque $k = 5$ la nouvelle donnée (hexagone rouge) est classée comme un triangle vert, alors qu'avec $k = 7$ elle est classée comme une étoile bleue.

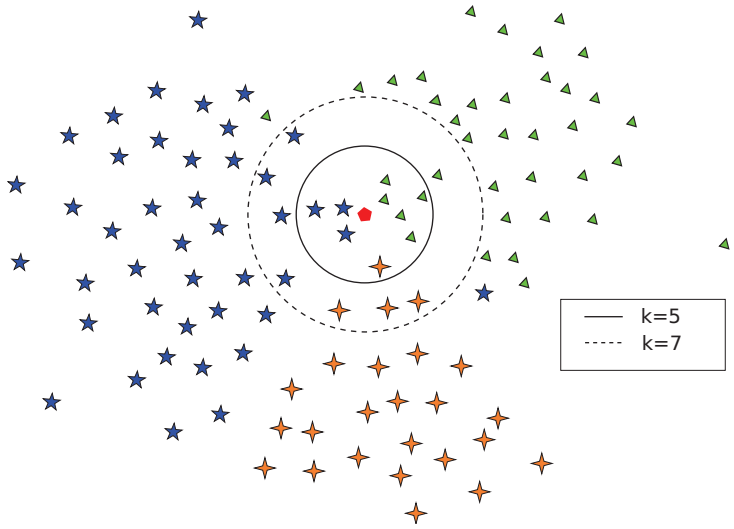


FIGURE 2-5 : L'influence de k sur le classifieur $k - ppv$.

L'avantage du $k - ppv$ est qu'il est capable de traiter des données avec ou sans attributs (parce qu'il nécessite uniquement une distance ou une similarité entre les données). Il est aussi capable de traiter des problèmes non linéaires.

Toutefois, cet algorithme a aussi l'inconvénient d'être non parcimonieux, il n'est donc pas bien adapté à des bases de données volumineuses. De plus, la grande difficulté dans l'utilisation de cet algorithme réside principalement dans le choix de la valeur k. En fait, il n'y a pas une méthode universelle. Plusieurs valeurs de k doivent être testées sur la base d'apprentissage pour trouver la meilleure valeur de k.

2.4.2 SVM, le séparateur à vaste marge (*Support Vector Machine*)

Le SVM est une méthode de classification supervisée non paramétrique et parcimonieuse. Elle repose sur l'existence d'un classifieur linaire au sein des données. Elle calcule alors cette fonction de décision par combinaison linéaire des observations.

Initialement le SVM a été conçu pour séparer les données en deux classes (nommées exemples positifs et exemples négatifs). Mais il est possible d'utiliser le SVM dans les cas où il y a plusieurs classes (en combinant alors plusieurs SVM suivant différentes stratégies, telles que « un contre tous », « un contre un » [Sch 95, Has 98])

Le SVM est aussi défini comme étant un type de machine à noyaux. C'est-à-dire que le classifieur f, qui associe à chaque donnée x_i un label y_j, utilise un noyau $K(x, x_i)$. Le noyau $K(x, x_i)$ est une mesure de ressemblance entre deux observations x et x_i. Par exemple, dans l'espace $\mathbb{R}^n$, le noyau $K(\vec{x_1}, \vec{x_2})$ peut-être le produit scalaire $\langle \vec{x_1}, \vec{x_2} \rangle$.

Toutefois, pour que le SVM fonctionne correctement, il est préférable que le noyau K soit défini semi-positif. Une matrice M symétrique est définie semi-positive si et seulement si, $\forall v \in \mathbb{R}^n, v^t M v \geq 0$. Pratiquement, il faut en fait calculer les valeurs propres de la matrice M et vérifier que celles-ci soient bien toutes positives ou nulles (*Cf.* [Sha 04]).

2.4.2.1 Problème linéairement séparable

Dans ce contexte, les données appartiennent à deux classes linéairement séparables. Il existe donc un hyperplan H qui les sépare. $H+$ est l'hyperplan parallèle à H qui contient les données positives les plus proches de H ($H-$ est l'hyperplan parallèle à H qui contient les données négatives les plus proches de H).

Le principe est de trouver la fonction de décision qui maximisera la marge m (la distance entre $H+$ (ou $H-$) et H) comme le montre la figure 2-6.

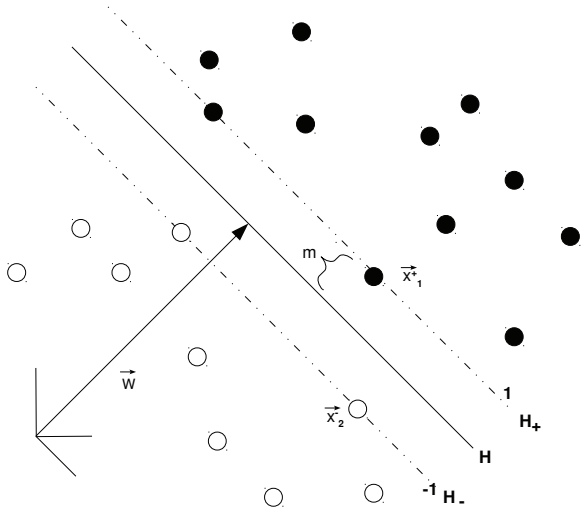


FIGURE 2-6 : SVM pour les données issues de deux classes. Les données x_1^+, x_2^- sur les marges $H+, H-$ sont appelées vecteurs supports (*support vectors*).

H a pour équation :

$$\langle \vec{w}, \vec{x} \rangle + b = 0 \tag{2-3}$$

où

– $\vec{x}$ est une donnée d'apprentissage,

– $\langle \vec{w}, \vec{x} \rangle$ dénote le produit scalaire entre les vecteurs $\vec{w}$ et $\vec{x}$.

Donc

$$\begin{cases} \langle \vec{w}, \vec{x} \rangle + b \geq 1 & \text{si } y = 1 \\ \langle \vec{w}, \vec{x} \rangle + b \leq -1 & \text{si } y = -1 \end{cases}$$

C'est-à-dire $y(\langle \vec{w}, \vec{x} \rangle + b) - 1 \geq 0$

Il faut alors trouver l'hyperplan H qui maximise la marge m [Sch 99]. Il faut donc trouver les paramètres qui maximisent la formule suivante :

$$\max_{w,b} \quad \min \| \vec{x} - \vec{x_i} \| : x \in \mathbb{R}^N, \langle \vec{w}, \vec{x} \rangle + b = 0 (x \in H), i = 1, ..., \text{nb des données} \quad (2\text{-}4)$$

Soit $\vec{x_1^+} \in H_+$ et $\vec{x_2^-} \in H_-$. De par la définition de H_+ et H_-, on a :

$$\langle \vec{w}, \vec{x_1^+} \rangle + b = 1$$
$$\langle \vec{w}, \vec{x_2^-} \rangle + b = -1$$

Donc

$$\langle \vec{w}, (\vec{x_1^+} - \vec{x_2^-}) \rangle = 2$$

$$\langle \frac{\vec{w}}{\| \vec{w} \|}, (\vec{x_1^+} - \vec{x_2^-}) \rangle = \frac{2}{\| \vec{w} \|} \quad (2\text{-}5)$$

Dès lors, maximiser la marge m revient en fait à minimiser la norme de $\vec{w}$ ($\| \vec{w} \|$) ou minimiser $\| \vec{w} \|^2$.

Minimiser $\| \vec{w} \|^2$ est un problème d'optimisation non linaire avec contraintes. Ce problème est résolu par la méthode de Lagrange. Cette méthode transforme un problème d'optimisation de fonction avec contraintes en un problème d'optimisation de fonction sans contraintes. On obtient donc :

$$L_p = \frac{1}{2} \| \vec{w} \|^2 - \sum_{i=1}^{N} \alpha_i y_i (\langle \vec{w}, \vec{x} \rangle + b) + \sum_{i=1}^{N} \alpha_i \quad (2\text{-}6)$$

où $\alpha_i \in \mathbb{R}$ sont les multiplicateurs de Lagrange.

Lp (la formule première de Lagrange) doit être minimisée par rapport à $\vec{w}$ et b. Il faut donc que les dérivées par rapport aux α_i soient nulles.

$$\left.\begin{array}{l} \frac{\partial L_p}{\partial w} = 0 \\ \\ \frac{\partial L_p}{\partial b} = 0 \end{array}\right\} \Rightarrow \begin{array}{l} \vec{w} = \sum_{i=1}^{N} \alpha_i y_i \vec{x_i} \\ \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{array}$$

Ceci peut être résolu à l'aide de la formule duale de Lagrange, c'est-à-dire :

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \vec{x_i}, \vec{x_j} \rangle \tag{2-7}$$

où $\alpha_i \geq 0$.

Les vecteurs supports $(\vec{x_i^+}, \vec{x_j^-})$ sont les données de la base d'apprentissage dont les multiplicateurs de Lagrange associés sont non nuls ($\alpha_i \neq 0$).

Dès lors, l'équation suivante peut être utilisée pour classer une nouvelle donnée [Pre 09] :

$$\begin{aligned} sign(\langle \vec{w}, \vec{x} \rangle + b) &= sign(\sum_{i=1}^{N} \alpha_i y_i (\langle \vec{x}_i, \vec{x} \rangle + b)) \\ &= sign(\sum_{j=1}^{N_s} \alpha_j y(s_j) (\langle \vec{s}_j, \vec{x} \rangle + b)) \end{aligned} \tag{2-8}$$

où $\vec{x}$ est l'instance à classer, les $\vec{x_i}$ sont les données de la base d'apprentissage, les $\vec{s_j}$ sont les vecteurs supports et N_s le nombre de vecteurs supports.

2.4.2.2 Cas non linéairement séparable

Les cas non linéairement séparables signifient qu'il n'existe pas d'hyperplan qui sépare les données positives des données négatives. Mais l'utilisation d'un noyau K peut permettre de représenter ces même données dans un autre espace où elles deviennent alors linéairement séparables [Sha 04]. Alors, le SVM trouve l'hyperplan linéaire dans la nouvelle dimension et nous pouvons donc projeter l'hyperplan dans la dimension originale (*Cf.* figure 2-7).

Dans ce cas, la formule de Lagrange devient :

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2-9}$$

Il existe plusieurs propositions pour la fonction noyau K comme :

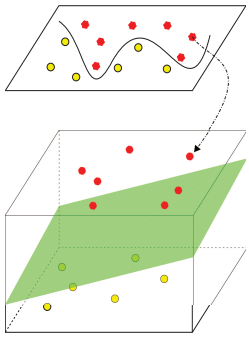


FIGURE 2-7 : Le changement de l'espace de présentation des données par la fonction à noyau K.

- le noyau gaussien : $K(x_1, x_2) = e^{\frac{\| \vec{x_1} - \vec{x_2} \|^2}{2\sigma^2}}$,
- le noyau sigmoïde : $K(x_1, x_2) = tanh(K\langle \vec{x_1}, \vec{x_2} \rangle - \delta)$ où $K > 0$ et $\delta > 0$,
- le noyau polynomial : $K(x_1, x_2) = (\langle \vec{x_1}, \vec{x_2} \rangle + c)^m, m \in \mathbb{N}$ et $c > 0$.

Il existe plusieurs algorithmes pour implanter la méthode SVM comme *incremental SVM* [Cau 01], *DirectSVM* [Roo 02] et *SimpleSVM* [Vis 02]. Dans nos travaux, nous utilisons « *SVM and Kernel Methods Matlab Toolbox* [Can 05] ». Cette Toolbox utilise l'algorithme *SimpleSVM* formalisé par l'algorithme 6.

Algorithme 6 SimpleSVM

Entrée: X un ensemble de N exemples pour l'apprentissage
Sortie: Un ensemble de *Support Vectors* SV_{set}
début

$$SV_{set} \leftarrow \{p_1, p_2\} \quad : \begin{cases} p_1 \in classe1 \\ p_2 \in classe2 \\ \arg\min \quad dis(p_1, p_2) \end{cases}$$

tant qu'il existe des points de violation **faire**
$c_{new} \leftarrow$ un violateur
$SV_{set} \leftarrow SV_{set} \bigcup \{c_{new}\}$
$SV_{set} \leftarrow SV_{set} \setminus \{p_i\} \quad : \alpha_{p_i} < 0$
fintantque
fin

Pour finir, le SVM est donc un classifieur non linéaire, parcimonieux et capable de classer des données avec ou sans attributs grâce à l'utilisation de noyaux.

2.5 Évaluation de la qualité de la mesure d'une similarité et d'une classification

Dans les sections précédentes, nous avons présenté des algorithmes d'apprentissage. On remarque que souvent ces algorithmes utilisent une mesure de similarité entre les données. La qualité de la classification sera donc fonction de cette mesure de similarité et de l'algorithme choisi. Il nous faut donc pouvoir évaluer ces deux paramètres.

2.5.1 Évaluation de la mesure d'une similarité

L'évaluation de la qualité d'une mesure de similarité [8] est un tâche difficile. Nous allons présenter deux méthodes pour évaluer cette mesure : la matrice de Gram et les méthodes de visualisation.

2.5.1.1 Matrice de Gram

La matrice de Gram G (en anglais *Gramain matrix*) est constituée de $N \times N$ éléments, tel que N est le nombre de données.

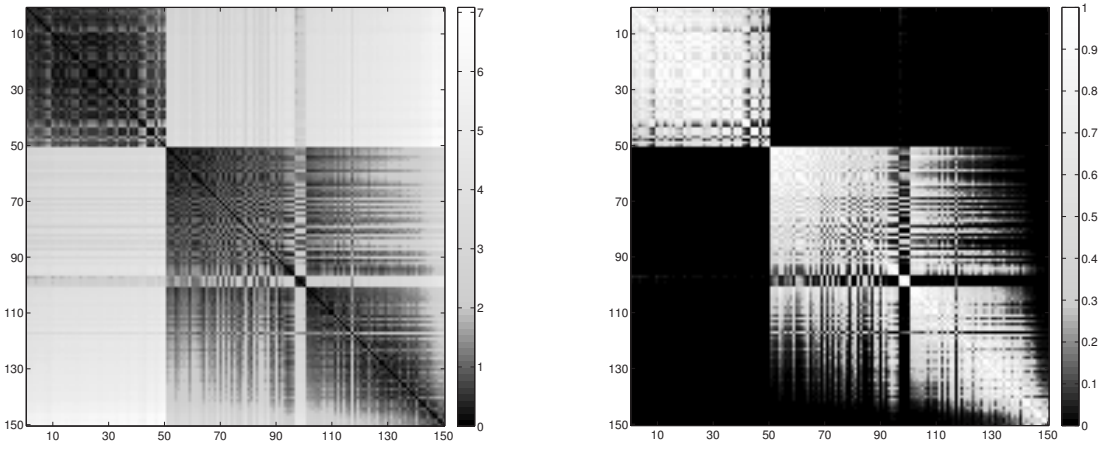
Chaque élément G_{ij} représente la similarité entre l'exemple i et l'exemple j, généralement $G_{ij} = < \vec{i}, \vec{j} >$. Dès lors, puisque la similarité est un réel, la matrice de Gram peut être représentée graphiquement (les similarités entre éléments étant représentées par une couleur ou un niveau de gris).

Dans l'apprentissage supervisé, où les exemples sont étiquetés, nous pouvons alors ordonner les données en fonction de leur classe. La similarité entre données de même classe devant être plus faible qu'entre données de classes différentes, la représentation graphique de la matrice de Gram [9] permet de visualiser la qualité de cette similarité. En effet, lorsque la similarité est bonne, la représentation graphique est alors composée de rectangles de couleurs homogènes identifiant les classes.

La figure 2-8(a) représente la matrice de Gram de l'ensemble de données Iris [Fis 88] en utilisant la distance euclidienne. La figure 2-8(b) quant à elle représente la similarité entre ces même exemples ($sim = \exp^{-dis(x_1,x_2)^2}$). Nous pouvons noter que la mesure de cette similarité identifie bien la première classe, mais pour les deux autres classes il existe quelques erreurs.

8. Par simplification, dans cette partie nous utiliserons indistinctement les termes « mesure de similarité », « similarité » et « distance ».

9. Par abus de la langage cette représentation est aussi appelée matrice de Gram.



(a) Distance (b) Similarité

FIGURE 2-8 : La matrice de Gram sur la base de données Iris.

Par la suite le critère d'alignement peut être calculé. Il correspond au produit scalaire entre la matrice de Gram et la matrice parfaite issue des labels, généralement lorsque les labels sont représentés par un vecteur y, cette matrice est égale à $y.y^t$. Lorsque cette valeur est proche de 0 la similarité est mauvaise, et lorsqu'elle est proche de 1, elle est considérée comme idéale (attention toutefois dans ce cas à ne pas faire du sur-apprentissage).

L'avantage de l'utilisation de la matrice de Gram est qu'elle est visuelle et rapide à obtenir. Toutefois, elle exige que les données soient étiquetées, ce qui n'est pas toujours le cas. Ceci nous amène donc à étudier des méthodes génériques de visualisation de données.

2.5.1.2 Méthodes de visualisation de données

De manière générale, les méthodes de visualisation de données permettent de mieux interpréter les relations qui les unissent. Cependant ces données sont souvent représentées à l'aide d'un grand nombre d'attributs. Elles sont donc humainement non interprétables. Dès lors les méthodes de visualisation utilisent principalement des algorithmes de réductions de dimensions, que l'on nomme aussi algorithme de projection, afin de pouvoir les représenter en deux ou trois dimensions.

Il existe plusieurs algorithmes de projection comme par exemple :

– Le *Multidimensional Scaling* (MDS) calcule la matrice de similarité (sim_x) entre les exemples x_i. Elle calcule aussi la similarité (sim_p) entre les projection p_i. Ensuite, elle modifie les positions des points p_i de telle sorte que $sim_p(p_i, p_j) \approx sim_x(x_i, x_j)$.

– Le *Locally Linear Embedding* (LLE) identifie les plus proches voisins v_{x_i} de chaque donnée x_i dans l'espace d'origine. Il calcule ensuite les coefficients α_{x_i} tel que les coordonnées de chaque x_i soient une combinaison linéaire des coordonnées de ses voisins v_{x_i} ($x_i = \sum_j \alpha_{ij} v_{x_i}$). Enfin, cet algorithme représente les projections p_i dans l'espace de projection comme combinaison linéaire des v_{p_i} (projection des v_{x_i}) en utilisant les mêmes coefficients α_{x_i}. [Row 00],

– Les Cartes de Kohonen (*self-organizing map*) que nous avons présentées dans la section 2.3.2 (page 41),

– Le *Stochastic Neighbor Embedding* (SNE) et (*t-SNE*).

C'est cette dernière méthode que nous allons étudier plus en détail, car il a été montré que sa projection était très souvent de meilleur qualité [Maa 09].

SNE

De manière globale, le principe de l'algorithme du SNE est de déplacer les projections p_i des x_i de manière à ce que la matrice a_{ij} (qui représente la probabilité conditionnelle que x_j soit voisin de x_i) et la matrice b_{ij} (qui représente la probabilité conditionnelle que p_j soit voisin de p_i) soient les plus semblables possibles [Hin 03].

Ceci est obtenu à l'aide de la méthode suivante :

– Soit x_i les données dans l'espace de départ et p_i leurs projections dans l'espace d'arrivée. Initialement les p_i sont positionnés aléatoirement autour de 0.

– Soient les matrices A et B, telles que A est la matrice de probabilité de voisinage sur les x_i et B est la matrice de probabilité de voisinage sur les p_i

  A est calculée de la manière suivante :

$$a_{ij} = \frac{\exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma^2)} \tag{2-10}$$

  tel que σ est la valeur qui représente l'entropie de la distribution sur les voisins, et qui est égale à $\log(k)$(k est le nombre effectif des voisins locaux ou *perplexity*).

  De la même manière, nous calculons la matrice B, telle que :

$$b_{ij} = \frac{\exp(-||p_i - p_j||^2)}{\sum_{k \neq i} \exp(-||p_i - p_k||^2)} \tag{2-11}$$

avec $a_{ii} = 0$ et $b_{ii} = 0$.

– Nous pouvons calculer un coût C (dit de *Kullback-Leibler*) entre ces deux matrices, tel que :

$$C = \sum_i \sum_j b_{ij} \log \frac{a_{ij}}{b_{ij}} \tag{2-12}$$

– Dès lors en utilisant la méthode de descente de gradient, nous pouvons minimiser la fonction du coût C et donc obtenir les p_i optimaux (ou localement optimaux).

La figure 2-9 montre l'utilisation du SNE pour projeter 100 points de $[0-225]^3$ vers $\mathbb{R}^2$. Les couleurs RGB (*Red*, *Green*, *Blue*) des points de cette projection représentent les coordonnées de ces même points dans l'espace de départ. Par exemple, la projection du point de coordonnée $(0, 0, 0)$ sera représentée par un point noir, celle du point $(255, 255, 255)$ sera représenté par un point blanc, celle du point $(255, 0, 0)$ par un point rouge, etc. Nous pouvons remarquer que deux points de couleurs proches (c'est-à-dire proche dans l'espace de départ) sont aussi proches dans l'espace d'arrivée.

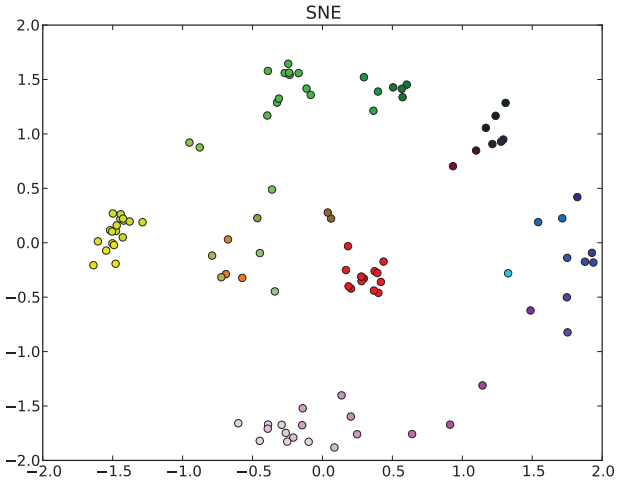


FIGURE 2-9 : La projection de SNE de 100 points de $[0 - 225]^3$.

Outre la qualité de sa projection, il est à noter que cet algorithme, à la différence des autres listés précédemment dans cette section, n'a besoin que d'une matrice de similarités entre les données x_i pour effectuer cette projection. De plus, à la différence de la matrice de Gram, il n'est pas nécessaire d'avoir des données étiquetées.

t-SNE

Le t-SNE est un algorithme s'inspirant du SNE. Il propose deux modifications [Maa 08]. Tout d'abord, il utilise une version symétrisée de la fonction de coût (si x_i est voisin de x_j, alors x_j est voisin de x_i, ce qui n'est forcément le cas dans le SNE). Ensuite il utilise la loi de Student (*Student's t-distribution*) plutôt qu'une loi Gaussienne pour calculer la probablité de voisinage entre deux données dans l'espace de projection.

Ainsi, dans le t-SNE :

$$q_{ij} = \frac{(1 + ||p_i - p_j||^2)^{-1}}{\sum_{k \neq i}(1 + ||p_i - p_k||^2)^{-1}} \tag{2-13}$$

La figure 2-10 montre la projection à l'aide du t-SNE de la base de données Iris. Les Setosa sont représentés à l'aide des ronds rouges, les Vesicolor à l'aide des étoiles bleues et enfin les Virginica à l'aide des croix vertes. Nous pouvons remarquer que la projection respecte bien ces trois classes.

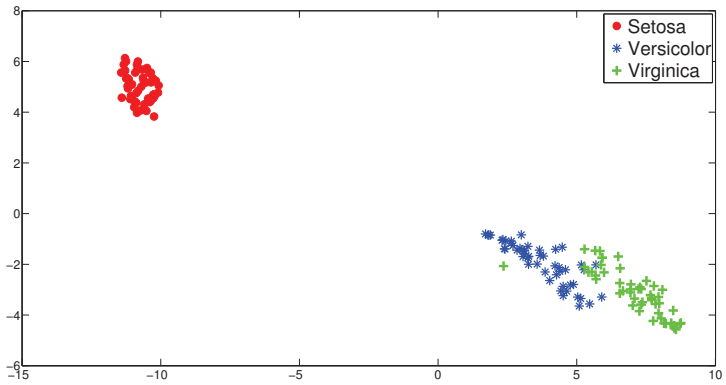


FIGURE 2-10 : La projection de t-SNE de la base de données Iris.

Comme l'indique [Maa 08], l'avantage du t-SNE est de préserver les proximités locales tout en permettant d'identifier plus facilement des structures plus lointaines (mettant ainsi en avant l'apparition de clusters). De plus dans les faits, cet algorithme est nettement plus rapide que le SNE.

2.5.2 Évaluation d'une classification

Outre l'évaluation de la mesure de la similarité, nous devons évaluer la qualité de la classification. C'est généralement la matrice de confusion qui est utilisée pour le faire.

Le tableau 2-1 présente la matrice de confusion pour deux classes (P pour positive et N pour négative). On retrouve au niveau des colonnes, les classes des éléments testés. Au niveau

des lignes, on retrouve les décisions du classifieur.

		Classe	
		P	N
Classification	P'	vrai positif (TP)	faux positif (FP)
	N'	faux négatif (FN)	vrai négatif (TN)

vrai positif (TP) : nombre des données positives qui ont été classées positives (*hit*).
faux positif (FP) : nombre des données négatives qui ont été classées positives (*false alarme*).
faux négatif (FN) : nombre des données positives qui ont été classées négatives (*miss*).
vrai négatif (TN) : nombre des données négatives qui ont été classées négatives (*correct rejection*).

P, N, P', N' sont défini de la façon suivante :

$$P = TP + FN \qquad N = FP + TN$$
$$P' = TP + FP \qquad N' = TN + FN$$

TABLEAU 2-1 : La matrice de confusion.

À partir de cette matrice, plusieurs mesures peuvent être calculées. Le tableau 2-2 en donne quelques exemples [Faw 06]. Chaque mesure prend une valeur entre 0 et 1. Suivant l'objectif quant à l'utilisation du classifieur, on en utilise habituellement une ou plusieures. Par exemple, la mesure Acc est utilisée quand nous cherchons un classifieur donnant une bonne reconnaissance. Alors que le facteur FPR permet de sélectionner des classifieurs ayant peu de fausses alarmes.

$TPR = \frac{TP}{P}$	Taux de vrai positif ou la sensibilité (*recall*)
$FPR = \frac{FP}{N}$	Taux de faux positif
$Acc = \frac{TP+TN}{P+N}$	La bonne reconnaissance (*accuracy Acc*)
$PPV = \frac{TP}{TP+FP}$	La valeur prédictive positive - la précision (*positive predictive value*)
$NPV = \frac{TN}{TN+FN}$	La valeur prédictive négative (*nigative predictive value*)
$F_1 = \frac{2}{\frac{1}{\text{précision}}+\frac{1}{\text{sensibilité}}} = \frac{2*TP}{P+P'}$	*F1-score*

TABLEAU 2-2 : Exemples de mesures de la qualité de classification calculées à partir de la matrice de confusion.

Outre ces évaluations statistiques, il existe d'autres évaluations graphiques telles que la ROC (*Receiver Operating Characteristics*) et l'AUC (*Area Under the ROC Curve*).

2.5. Évaluation de la qualité de la similarité et de la classification

Dans l'espace de ROC ($[0-1]^2$), les ROC sont des graphiques en deux dimensions dans lesquels le taux de vrais positifs TPR est représenté sur l'axe Y et le taux de faux positifs FPR est représenté sur l'axe X. Une ROC représente un compromis relatif entre les avantages (vrais positifs) et les coûts (faux positifs) [Bra 97].

Un classifieur discret produit seulement une étiquette de classe. Chaque classifieur discret produit un TPR et un FPR. Donc, un classifieur discret correspond à un seul point dans l'espace de ROC.

Par exemple, le tableau 2-3 donne un exemple avec quatre classifieurs. Vingt données issus de deux classes (P et N) ont été classées par les classifieurs A,B,C et D (attribution de la classe p ou n).

	Classe	Classification			
		A	B	C	D
1	P	p	p	p	n
2	P	p	p	p	n
3	P	p	p	p	n
4	P	p	p	p	n
5	P	p	p	p	n
6	P	p	p	p	n
7	P	p	n	p	n
8	P	n	n	n	p
9	P	n	n	n	p
10	P	n	n	n	p

	Classe	Classification			
		A	B	C	D
11	N	p	p	p	n
12	N	p	p	p	n
13	N	n	p	p	n
14	N	n	p	n	p
15	N	n	n	n	p
16	N	n	n	n	p
17	N	n	n	n	p
18	N	n	n	n	p
19	N	n	n	n	p
20	N	n	n	n	p

TABLEAU 2-3 : Exemple de quatre classifieurs.

Le tableau 2-4 donne les matrices de confusion de ces classifieurs. La figure 2-11 quant à elle, présente ces classifieurs dans l'espace ROC.

À la lecture de ce graphique, on remarque que A est le meilleur classifieur puisqu'il a un taux de vrai positif (TPR) assez élevé et un taux de faux positif (FPR) assez bas (graphiquement il est proche du haut $(0,1)$). Le classifieur D est quant à lui *a priori* le moins bon classifieur, puisqu'il a un TPR bas et un FPR haut. Toutefois, ce dernier résultat n'est pas aussi mauvais qu'il en a l'air, parce que nous pouvons prendre le contraire de sa décision, et dans ce cas nous obtenons un résultat identique au classifieur C (D est graphiquement le symétrique de C par rapport à la droite $y = x$). Notons qu'au niveau des matrices de confusion, cela se traduit par le fait que la matrice de confusion du classifieur D est égale à la matrice transposée de celle du C. Nous pouvons en déduire que le classifieur D n'a pas été utilisé correctement [Faw 06].

Contrairement aux classifieurs discrets qui ne donnent qu'un point dans l'espace ROC, il existe d'autres classifieurs donnant une valeur continue pour chaque donnée. Ce type de

(a) classifieur A

7	2
3	8

$TPR = 0.7$

$FPR = 0.2$

(b) classifieur B

6	6
4	4

$TPR = 0.6$

$FPR = 0.6$

(c) classifieur C

7	3
3	7

$TPR = 0.7$

$FPR = 0.3$

(d) classifieur D

3	7
7	3

$TPR = 0.3$

$FPR = 0.7$

TABLEAU 2-4 : Matrices de confusion des quatre classifieurs du tableau 2-3.

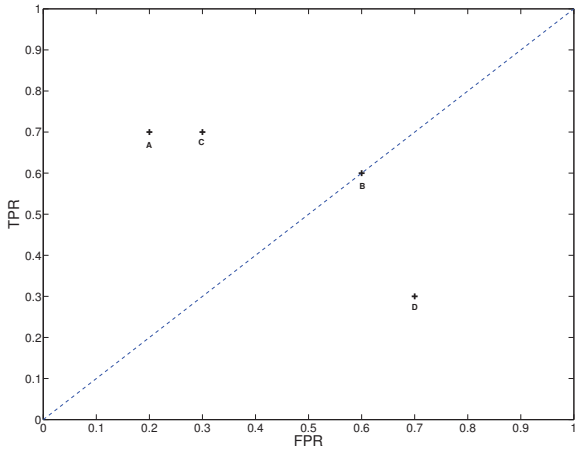


FIGURE 2-11 : Les quatre classifieurs dans l'espace ROC.

classifieur donne alors des courbes dans l'espace de ROC. Dans ce cas, pour comparer des classifieurs, la surface sous la courbe ROC (AUC *Area Under the* ROC *Curve*) est calculée pour les comparer. Le classifieur le plus performant est celui qui possède l'AUC la plus élevée [Han 09].

2.6 Conclusion

Dans ce chapitre, nous avons présenté deux catégories d'apprentissage, l'apprentissage supervisé et non supervisé. Nous nous sommes attardés sur quelques algorithmes de chaque type, l'algorithme du $Kmeans$ et les cartes de Kohonen pour l'apprentissage non supervisé et l'algorithme du $k-ppv$ et du SVM pour l'apprentissage supervisé. Ensuite, nous avons étudié des méthodes et techniques permettant d'évaluer la qualité d'une similarité et d'une classification supervisée.

Par la suite, nous allons voir les utilisations possibles des algorithmes d'apprentissage supervisé et non supervisé afin de identifier les profils des utilisateurs. Cette identification de profil représente une information essentielle pour une majorité des systèmes d'adaptation d'hypermédia. Lorsque nous aurons besoin d'un algorithme d'apprentissage, il faudra le choisir en fonction de certains critères, comme par exemple la taille de la base d'apprentissage (choix d'un algorithme parcimonieux ou pas) ou le fait que nous ayons des *a priori* concernant nos données (choix d'un algorithme paramétrique ou pas). De la même manière, nous devrons évaluer ce choix sur nos données, sachant que dans cette partie uniquement des problèmes à deux classes ont été pris en compte. Ceci nous amènera certainement à adapter l'évaluation de la classification.

CHAPITRE 3

Web Usage Mining et son utilisation sur les traces d'Hypergéo

Sommaire

3.1 Introduction

L'analyse des comportements d'utilisations d'un site web aide son responsable à connaître les besoins et les intérêts des utilisateurs. Ces connaissances sont très utiles pour adapter le site et pour mieux répondre aux attentes des visiteurs. Le *Web Usage Mining* (WUM) est un domaine de recherche qui adapte les techniques du *knowledge data discovery* (KDD) aux données d'utilisation du web.

Dans ce chapitre nous présenterons le WUM et les étapes essentielles de son processus (la collecte et le prétraitement des données, la découverte et l'analyse de modèles). Après nous montrons l'emploi du WUM pour adapter des hypermédias. Puis nous expliquerons en détail l'algorithme des motifs caractéristiques que nous utilisons pour la découverte de modèle. Ensuite, nous présenterons notre contexte de travail (le site Hypergéo). Enfin, avant une conclusion, nous appliquerons l'algorithme des motifs caractéristiques à nos données expérimentales.

3.2 WUM : *Web Usage Mining*

La fouille du web (*Web mining*) utilise les outils et les techniques de fouille de données pour extraire des connaissances. Elle est composée de trois sous-domaines [Gui 04] :

- le *web content mining*, qui traite les données, les informations et les connaissances présentes dans le contenu des pages du web ;
- le *web structure mining*, qui analyse la structure et les hyperliens entre les pages du site web ;
- le *web usage mining* (WUM), qui étudie le comportement de navigation des utilisateurs afin de les caractériser, notamment par l'émergence de modèles utilisateurs [Sri 00, Wan 06].

Le WUM est utilisé pour révéler les informations potentiellement utiles et non connues à l'avance à partir des navigations des utilisateurs [Kos 00]. Le but est de capturer, de modéliser, et d'analyser les modèles et les profils des utilisateurs d'un site Web. Les modèles découverts sont habituellement représentés comme des collections de pages, de transactions, d'objets, ou de ressources qui sont accédés fréquemment par un ou des groupes d'utilisateurs ayant des intérêts communs [Mob 07b].

Le résultat du WUM peut aider à améliorer les stratégies de e-marketing, à optimiser la fonctionnalité des applications web, et à fournir un contenu plus personnalisé aux utilisateurs.

Le processus du WUM peut être divisé en trois étapes interdépendantes : **la collecte et le prétraitement des données**, **la découverte de modèle**, et **l'analyse de modèle** [Coo 00]. Nous allons étudier uniquement les deux premières étapes, car la troisième étape dépend trop fortement du contexte applicatif.

3.2.1 Collecte et prétraitement des données

L'objectif de cette étape est de construire une base de données sur laquelle des techniques de fouille de données peuvent être appliquées. Ce processus est souvent l'étape la plus longue et requiert de nombreux calculs. La qualité de la collecte et du prétraitement des données est un critère essentiel pour l'extraction des modèles. Les données collectées peuvent être classées suivant quatre groupes [Sri 00].

Le premier groupe représente **les données d'utilisation** sauvegardées dans les fichiers de *log* des serveurs. Ces données sont la source primaire de données dans le WUM. Chaque requête vers le serveur, correspondant à une requête HTTP, produit une entrée simple dans les fichiers de log du serveur. Cette entrée contient des informations telles que la date et l'heure de la demande, l'adresse IP de l'utilisateur, la ressource demandée, la méthode de HTTP employée (Get, Post, etc.), et d'autres informations [W3C 10]. Ces données sont interprétées différemment selon les objectifs de l'analyse. En général, la consultation d'une page (*pageview*) est la forme la plus fréquente dans le WUM. Une ***pageview*** est la représentation d'une collection d'objets du web affichée sur le navigateur suite à une seule action de l'utilisateur (le texte ainsi que les images, sons, etc.). « Au niveau de l'utilisateur, le niveau le plus fondamental de l'abstraction comportementale est celui d'une **session**. Une session est un ensemble délimité de cliques d'un seul utilisateur sur un ou plusieurs serveurs web » [W3C 99] , [W3C 10].

Le deuxième groupe de données représente **les données de contenu**. Ce sont les données qui sont transmises à l'utilisateur comme des textes, des images et des vidéos.

Le troisième groupe est constitué **des données de structure**, qui représentent la vue du concepteur sur l'organisation du contenu du site web. Cette organisation peut prendre plusieurs formes, par exemple linéaire, hiérarchique ou en graphe.

Enfin, le dernier groupe représente les **données des utilisateurs** qui sont la base de données opérationnelle pour le site. Elles peuvent inclure des informations additionnelles sur les utilisateurs comme, par exemple, les informations contenues au sein des *cookies*.

3.2.2 Découverte de modèle

Après la phase de prétraitement, les techniques de fouille de données peuvent être appliquées. Les techniques les plus fréquemment utilisées sont [Tan 05] :

- le *cluster mining*,
- l'*association rule mining*,
- le *sequential pattern mining*.

3.2.2.1 *Cluster mining*

Le *cluster mining* est utilisé pour regrouper une collection d'objets non étiquetés en groupes significatifs [Jai 88]. C'est donc des techniques d'apprentissage non supervisé qui sont utilisées, suivies d'une étape d'interprétation des ensembles identifiées.

Dans le domaine du web, les objets sont les documents du web, les références à ces documents, ou les visites des utilisateurs. Le regroupement des documents du web est généralement basé sur le contenu des données et vise à déterminer les documents à contenus similaires. Le *clustering* des références de documents du web regroupe des documents qui sont destinés à un usage similaire, tandis que le *clustering* des visites des utilisateurs révèle des modèles de navigation des utilisateurs semblables [Kou 05].

Les résultats du *clustering* varient selon les méthodes utilisées pour regrouper les données. Nous avons principalement les méthodes de partition et les méthodes hiérarchiques [Han 06]. La méthode de partition divise les données en k groupes représentant les classes comme la méthode *Kmeans*, que nous avons vu dans le chapitre précédent. La méthode hiérarchique, quant à elle, construit un arbre en commençant soit par la racine et divise alors les données, soit par les feuilles et agglomère alors les données similaires.

Dans le *clustering*, un élément peut appartenir soit à un et seul un *cluster*, soit à plusieurs *clusters*, soit à un *cluster* avec une certaine probabilité.

Perkowitz et Etzioni présentent l'algorithme *PageGather* pour découvrir les *clusters* significatifs à partir de fichiers de *log*. Ils construisent une matrice de similarité entre les pages, plus les pages visitées apparaissent ensemble dans les logs, plus elles sont similaires. Ensuite, ils construisent un graphe à partir de cette matrice pour identifier les *clusters* [Per 98, Per 00].

Shahabi et al. [Sha 97] ont proposé une approche de *clustering* sur les navigations d'utilisateurs. Ils représentent les chemins de navigation des utilisateurs par des séquences de paires de documents. Ensuite, leur approche est réalisée en utilisant l'algorithme *Kmeans* [Que 67]. Par contre, Nasraoui et al. font du *clustering* en utilisant la logique flou [Nas 00].

Enfin, Soriano-Asensi et al. [Sor 08] utilisent l'algorithme *Growing Hierarchical SOM* (GHSOM), une amélioration du SOM (*Self-Organizing Map* ou carte de Kohonen [Koh 89]), pour analyser l'utilisation du site gouvernemental de la région de Valence en Espagne.

3.2.2.2 *Association rule mining*

La fouille des règles d'association (*Association rule mining*) tente de trouver des groupes d'éléments apparaissant ensemble dans les transactions utilisateurs. Elle identifie toutes les associations entre les éléments de sorte que la présence d'un sous-ensemble de ces éléments dans une transaction implique la présence d'autres aussi [Mob 96].

Une règle d'association est de la forme $A \Rightarrow B$, où A et B sont des sous-ensembles disjoints d'un ensemble d'éléments. La règle est accompagnée de deux mesures significatives, la confiance c et le support s.

Le support s de la règle d'association $A \Rightarrow B$ mesure le pourcentage de transactions qui contiennent A et B dans toutes les transactions. Donc le support $s(A \Rightarrow B)$ est [Han 06] :

$$s(A \Rightarrow B) = p(A \; et \; B) \tag{3-1}$$

La confiance c mesure le pourcentage de transactions contenant A, qui contiennent également B, donc la confiance $c(A \Rightarrow B)$ est :

$$\begin{aligned} c(A \Rightarrow B) = & \quad p(A|B) \\ = & \quad \frac{p(A \; et \; B)}{p(A)} \\ = & \quad \frac{s(A \; et \; B)}{s(A)} \end{aligned} \tag{3-2}$$

Agrawal et al. proposent dans [Agr 93, Agr 94] l'algorithme Apriori comme un algorithme efficace pour générer toutes les règles d'association significatives pour l'analyse d'une grande masse de données. Pour chaque transaction, cet algorithme trouve les ensembles d'éléments qui ont au moins une fréquence minimum égale à C.

Mobasher et al. utilisent les règles d'association pour construire un système du *web usage mining* (*WEBMINER*) [Mob 96]. *WEBMINER* regroupe (*clustering*) les entrées datées du fichier de *log* d'accès au serveur, il utilise ensuite l'algorithme Apriori pour la fouille les règles d'association.

Borges et Levene [Bor 98] proposent l'extraction de règles d'association à partir de la structure de données et des hyperliens du site. Ils proposent de trouver des modèles de comportement des utilisateurs à partir de ces règles.

La méthode proposée par Wong et al. est basée sur l'approche du raisonnement par cas. Ils trouvent le modèle d'accès de l'utilisateur du site en appliquant les règles d'association floues [Won 01].

3.2.2.3 *Sequential pattern mining*

Contrairement aux règles d'association, la méthode des motifs séquentiels (*Sequential pattern mining*) tient compte de l'ordre d'accès aux pages des sites web. Ainsi, le motif $m_1 = \{s_1, s_2\}$ est différent du motif $m_2 = \{s_2, s_1\}$.

Pour l'ensemble des séquences $S = (s_1, s_2, s_3, s_4)$, deux types de motifs séquentiels sont proposés [Mob 07a] :

- **Les séquences fermées**, dans ce cas, nous pouvons dire que $s_1, s_2 \Rightarrow s_3$, mais nous ne pouvons pas dire $s_1, s_2 \Rightarrow s_4$.
- **Les séquences ouvertes**, les deux séquences $s_1, s_2 \Rightarrow s_3$ et $s_1, s_2 \Rightarrow s_4$ sont acceptées.

Comme les règles d'association, le motif séquentiel a deux paramètres : le support s et la confiance c.

Il existe plusieurs algorithmes pour obtenir ces motifs séquentiels. Par exemple Chen et al. [Che 98] introduisent l'algorithme *maximal forward references*. Cet algorithme convertit les séquences originales du fichier de *log* en ensembles de motifs qui représentent les chemins suivis par l'utilisateur. Chaque motif est construit à partir du premier document du site visité par l'utilisateur jusqu'à ce qu'il rebrousse chemin dans sa navigation. Dans le *WAP-tree mining* algorithme (*Web Access Patterns tree mining*), Pei, et al. construisent un arbre de séquences à partir du fichier *log*. Ensuite, ils proposent un algorithme d'analyse de cet arbre [Pei 00].

L'inconvénient de cette méthode est que le taille de l'ensemble des motifs séquentiels est souvent faible. En effet les utilisateurs aujourd'hui utilisent de plus en plus souvent les moteurs de recherche externes (type google). Dès lors ils entrent et ressortent constamment des sites web. Ils n'ont plus un parcours linéaire au sein de ce dernier.

Le tableau 3-1 résume les algorithmes présentés.

En comparant la fouille des règles d'association avec le *clustering*, nous constatons que les deux méthodes visent à identifier les motifs à partir des documents visités du site web. Mais aucun d'eux prend en compte les informations concernant l'ordre de visite des documents. Toutefois, les règles d'association fournissent des informations supplémentaires sur l'antécédent et sur les valeurs de la confiance et du support.

Source	Type de WUM	Algorithme utilisé
[Agr 93, Agr 94]	*Association rule mining*	l'algorithme Apriori
[Mob 96]	*Association rule mining*	*clustering* + l'algorithme Apriori
[Sha 97]	*Clustering*	*Kmeans*
[Bor 98]	*Association rule mining*	fouille des règles d'association à partir de la structure de données du site
[Che 98]	*Sequential pattern mining*	*maximal forward references*
[Per 98, Per 00]	*Clustering*	PageGather
[Nas 00]	*Clustering*	la logique flou
[Pei 00]	*Sequential pattern mining*	*WAP-tree mining (Web Access Patterns tree mining)*
[Sor 08]	*Clustering*	GHSOM (carte de Kohonen)

TABLEAU 3-1 : Les algorithmes du WUM présentés.

Le choix de l'algorithme du WUM, lors de la découverte du modèle, dépend de trois paramètres : les classes des données qui sont ou non prédéfinies, le volume des données à traiter et le but final quant à l'utilisation de cet algorithme. Nous allons présenter maintenant les différentes manières d'utiliser le WUM pour adapter un hypermédia.

3.3 Utilisations du WUM pour l'adaptation des hypermédias

Dans [Kou 05], Kourtri et *al.* proposent une état de l'art sur l'utilisation du WUM pour l'adaptation des hypermédias. Dans ce cadre, ils identifient trois techniques d'adaptation. La première est basée sur des systèmes de recommandations. La deuxième est la modification dynamique du contenu et/ou des liens de l'hypermédia en fonction de l'utilisateur. Enfin la troisième est la modification statique de l'hypermédia *a posteriori*.

Comme nous l'avons vu, le WUM propose plusieurs outils pour découvrir les modèles qui peuvent être utilisés pour ces trois types d'adaptation, mais chacun a un champ d'application qui lui est propre.

L'utilisation des résultats des algorithmes de *clustering* varient en fonction des données prises en compte. Par exemple le *clustering* de documents est fortement utilisé dans l'adaptation statique. Le responsable de l'hypermédia peut ainsi améliorer le système en ajoutant par exemple un lien entre deux documents similaires. Le *clustering* des parcours utilisateurs quant à lui est principalement utilisé dans les systèmes de recommandation. Deux utilisateurs ayant un comportement similaire se verront proposés les même recommandations.

L'*association rule mining* s'intéresse aux motifs qui représentent les documents visités par

les utilisateurs sans prendre en compte l'ordre chronologique. Cela permet de mettre en évidence des interconnexions entre documents lorsque ces derniers apparaissent dans de nombreux parcours, sans qu'il y ait obligatoirement de liens hypertextes entre eux. Dès lors l'*association rule mining* est souvent utilisé pour faire de l'adaptation dynamique ou statique.

Par contre, le *sequential pattern mining* prend en compte l'ordre de navigation. Les motifs sont donc nettement moins fréquents (la probabilité que deux personnes fassent exactement le même parcours est plus faible). Dès lors il est difficile de l'utiliser pour faire de l'adaptation, mais elle est souvent employée pour faire de la recommandation.

Comme nous allons le voir par la suite, dans notre contexte, nous connaissons *a priori* les classes des utilisateurs de notre site web. Sachant de plus que les données recueillies seront peu nombreuses, nous avons donc choisi d'utiliser les règles d'association et plus particulièrement l'algorithme des motifs caractéristiques que nous allons étudier plus en détail maintenant.

3.4 Motifs caractéristiques

Un motif peut être défini comme étant un ou plusieurs éléments qui se retrouvent dans plusieurs ensembles d'éléments. Dans le cadre du WUM, les motifs sont des documents ou des transitions qui appartiennent aux traces des utilisateurs du site. Ces motifs peuvent être utilisés pour différents objectifs.

Par exemple, le motif commun, qui est le motif le plus représenté dans les transactions, sert à identifier les pages les plus visitées du site. Cette information permet d'améliorer l'organisation du site.

Un autre exemple, celui qui nous intéresse plus particulièrement dans le cadre de cette thèse, est l'identification d'utilisateur. Gao et Sheng [Gao 08] ont proposé une méthode d'identification des utilisateurs qui suit les étapes suivantes :

- identification et fouille des motifs caractéristiques à l'aide de la méthode *FP-tree* et *FP-growth*,
- construction d'une fonction de décision (identification des utilisateurs) à partir de ces motifs en utilisant les réseaux bayesiens naïfs multi-nets.

Avant que nous commencions à détailler la méthode des motifs caractéristiques, nous définissons les termes utilisés dans ce chapitre :

La transaction HTTP	La communication entre le navigateur de l'utilisateur et le serveur hébergeant le site web en utilisant le protocole HTTP (*Cf.* [W3C 10]).
La transition	Le fait d'aller d'un document à un autre. Elle est composée de plusieurs transactions HTTP.
La transaction	L'ensemble de transitions réalisées par un utilisateur pendant une période donnée.
U	L'ensemble des utilisateurs.
T	L'ensemble des transactions de tous les utilisateurs.
T_i (la trace)	L'ensemble des transactions de l'utilisateur u_i.
T_{ik}	L'ensemble des transactions de l'utilisateur u_i qui contient le motif m_k.
TM_k	L'ensemble des transactions contenant le motif m_k.
T^c	L'ensemble des motifs caractéristiques de tous les utilisateurs.

3.4.1 Identification et fouille des motifs caractéristiques

L'objectif de cette étape est de trouver tous les motifs caractéristiques pour chaque utilisateur.

Pour chaque utilisateur $u_i \in U$ et pour chaque motif m_k constitué d'un ou plusieurs éléments (pages, transitions, ...), il y a deux paramètres qui caractérisent m_k : le support $s(m_k, u_i)$ et la confiance $c(m_k, u_i)$.

Le support $s(m_k, u_i)$ est la probabilité conditionnelle que le motif m_k existe dans les transactions de l'utilisateur u_i, donc $s(m_k, u_i) = P(m = m_k | u = u_i)$.

$s(m_k, u_i)$ peut être estimé à partir de l'ensemble des transactions T :

$$s(m_k, u_i) = \frac{|T_{ik}|}{|T_i|} \tag{3-3}$$

La confiance $c(m_k, u_i)$ est la probabilité conditionnelle qu'une transaction contenant le motif m_k soit une transaction de l'utilisateur u_i, c'est-à-dire $c(m_k, u_i) = P(u = u_i | m = m_k)$

Ainsi, $c(m_k, u_i)$ peut être estimé par la formule suivante :

$$c(m_k, u_i) = \frac{|T_{ik}|}{|TM_k|} \tag{3-4}$$

On définit alors les **motifs caractéristiques** comme étant les motifs m_k qui satisfont les

deux conditions suivantes :

$$s(m_k, u_i) \geqslant s_{min} \tag{3-5}$$

$$c(m_k, u_i) \geqslant c_{min} \tag{3-6}$$

où le s_{min} et le c_{min} sont prédéfinis pour tous les utilisateurs.

L'équation 3-5, que l'on nomme « critère minimum du support d'utilisateur » (ou critère s_{min}), assure que le motif caractéristique m_k a assez d'occurrences dans les transactions de l'utilisateur u_i. Par conséquent il reflète l'un des comportements de l'utilisateur u_i.

L'équation 3-6, que l'on nomme « critère de la confiance d'utilisateur » (ou critère c_{min}), garantit que le motif caractéristique m_k d'un utilisateur particulier u_i est assez spécial et que peu d'autres utilisateurs possèdent le même motif.

Le processus d'identification des motifs caractéristiques comporte deux étapes :

– Découvrir des motifs fréquents qui valident le critère s_{min} (l'équation 3-5) pour chaque utilisateur [Han 00, Han 04]. Cette étape est composée de deux sous étapes :

  1. des arbres représentant les éléments fréquents de chaque utilisateur sont construits en utilisant l'algorithme *FP-tree*,
  2. l'algorithme *FP-tree growth* est appliqué sur ces arbres pour trouver les motifs fréquents.

– Parmi ces motifs fréquents, trouver les motifs caractéristiques validant le critère c_{min} (l'équation 3-6).

3.4.1.1 Méthode des arbres des motifs fréquents (*FP-tree*).

Initialement l'obtention de l'ensemble des motifs fréquents part du principe qu'un motif de longueur $k+1$ est fréquent si et seulement si il est composé d'un motif fréquent de longueur k (algorithme *a priori*). Dès lors, il suffit d'identifier les motifs fréquents de longueur $k=1$ et itérativement d'identifier les motifs fréquents de longueur $k+1$. Toutefois cette méthode bien que simple possède l'inconvénient d'être d'une forte complexité algorithmique. C'est pourquoi, Han et al. dans [Han 00] ont proposé une nouvelle méthode basée sur l'utilisation d'arbre de motifs fréquents.

L'arbre des motifs fréquents (en anglais, *FP-tree* pour *frequent-pattern tree*) permet de représenter les éléments fréquents issus des transactions des utilisateurs en regroupant les motifs qui ont les même préfixes au sein d'une même branche.

Afin d'être plus facilement compréhensible, nous allons décrire cette méthode à partir d'un exemple. Le tableau 3-2 donne, pour trois utilisateurs, un ensemble de transactions T ordonnées lexicographiquement (chaque transaction est composée d'un ou plusieurs éléments représentés ici par des lettres).

ID de utilisateurs	ID de transaction	Transactions
1	1	*aef*
1	2	*act*
1	3	*bgkt*
1	4	*bdt*
2	5	*abcgr*
2	6	*acgrt*
2	7	*efgt*

ID de utilisateurs	ID de transaction	Transactions
2	8	*adefgk*
2	9	*aefk*
2	10	*abcefgm*
2	11	*acgr*
3	12	*efgh*
3	13	*cfhw*
3	14	*ach*

TABLEAU 3-2 : Les transactions T ordonnées lexicographiquement.

Nous allons identifier les éléments fréquents. Supposons que nous prenions $s_{min} = 0{,}4$. Dans ce cas, il faut que $s(m_k, u_i) \geqslant 0{,}4$ (*Cf.* équation 3-5). Donc pour l'utilisateur 2, il est nécessaire que $\frac{|T_{2k}|}{|T_2|} \geqslant 0{,}4$, c'est-à-dire :

$$T_{2k} \geqslant 0{,}4 \times 7 \Rightarrow T_{2k} \geqslant 2{,}8$$

Il faut donc que la fréquence d'un élément soit supérieure ou égale à $2{,}8$ pour être un « élément fréquent » de l'utilisateur 2. De la même manière, pour les utilisateurs 1 et respectivement 3, la fréquence d'un élément fréquent doit être supérieure ou égale à $1{,}6$ et respectivement $1{,}2$.

De ce fait, pour l'utilisateur 2, les éléments *a*, *c*, *e*, *f*, *g* et *r* sont considérés comme des éléments fréquents. Et les éléments *b*, *d*, *k*, *t* et *m* ne sont pas considérés comme des éléments fréquents. Le tableau 3-3 donne les éléments fréquents pour chaque utilisateur avec $s_{min} = 0{,}4$.

La méthode du *FP-tree* est appliquée à chaque utilisateur pour construire son arbre. La structure (élément, occurrences) identifie le nom des éléments et leur nombre d'occurrences. La liste des éléments fréquents de l'utilisateur 2 selon l'ordre d'occurrence des éléments est $\{(a, 6), (g, 6), (c, 4), (e, 4), (f, 4), (r, 3)\}$.

Les étapes suivantes montrent comment construire le *FP-tree* de l'utilisateur 2 :

– La racine d'un arbre est créée et marquée comme *Racine*.

ID de utilisateurs	ID de transaction	Éléments fréquents	Éléments fréquents triés selon leurs occurrences
1	1	*a*	*a*
1	2	*at*	*ta*
1	3	*bt*	*tb*
1	4	*bt*	*tb*
2	5	*acgr*	*agcr*
2	6	*acgr*	*agcr*
2	7	*efg*	*gef*

ID de utilisateurs	ID de transaction	Éléments fréquents	Éléments fréquents triés selon leurs occurrences
2	8	*aefg*	*agef*
2	9	*aef*	*aef*
2	10	*acefg*	*agcef*
2	11	*acgr*	*agcr*
3	12	*fh*	*hf*
3	13	*cfh*	*hcf*
3	14	*ch*	*hc*

TABLEAU 3-3 : Les éléments fréquents avec $s_{min} = 0{,}4$.

- Les motifs fréquents de la première transaction de l'utilisateur 2, $(agcr)$, mènent à la construction de la première branche de l'arbre : $< (a,1), (g,1), (c,1), (r,1) >$.
- Les motifs fréquents de la deuxième transaction sont encore $(agcr)$. Ce chemin existe déjà, l'occurrence de chaque nœud dans ce chemin est alors incrémentée de 1.
- Les motifs fréquents de la troisième transaction, (gef), ne partagent aucun préfixe commun avec le chemin existant. Ainsi, une deuxième branche de l'arbre est construite : $< (g,1), (e,1), (f,1) >$.
- Les motifs fréquents de la quatrième transaction sont $(agef)$. Dans l'arbre il existe une branche associée à une partie de ces motifs, (ag) . Dès lors, l'occurrence des nœuds (a,g) dans ce chemin est incrémentée de 1. Un chemin $< (e,1), (f,1) >$ est construit comme un fils du chemin $< (a,3), (g,3) >$.
- Nous continuons de la même manière pour toutes les transactions.

Pour parcourir l'arbre des motifs fréquents facilement, un tableau de tête des éléments fréquents est construit en commençant par l'élément le plus fréquent jusqu'au moins fréquent (de *a* jusqu'à *r* pour l'utilisateur 2). Chaque élément référence sa première occurrence insérée dans l'arbre. Les nœuds de l'arbre de même nom sont liés en séquence. La figure 3-1 représente le *FP-tree* pour l'utilisateur 2.

L'algorithme 7 est proposé par Han et al. [Han 00] pour construire l'arbre des motifs fréquents (*frequent-pattern tree, FP-tree*).

Algorithme 7 Algorithme de construction de l'arbre des motifs fréquents (*FP-tree)*

Entrée: T : Ensemble de transactions, s_{min} : Le minimum du support.
Sortie: R : Arbre de motifs fréquents
début
  $F \leftarrow \phi$
  pour chaque $t \in T$ **faire**
    $t_{new} \leftarrow \phi$
    pour chaque $i \in t$ et $frequence(i) \geqslant s_{min}$ **faire**
      $t_{new} \leftarrow t_{new} \cup i$
    finpour
    $F \leftarrow F \cup t_{new}$
  finpour
  $F' \leftarrow t', t' \in F$ et t' est ordonnée selon les fréquences décroissantes de ses éléments.
  $R \leftarrow creerNoeud(" Racine")$, la racine du *FT-tree*
  pour chaque élément t, $t \in F'$ **faire**
    $ajout\_page(R, t)$.
  finpour
fin

---

procédure $ajout\_page$(**Entree/Sortie** arb : Arbre des motifs fréquents, **Entree** $trans$: Ensemble des transactions)
début
  Soit $trans = (p.P')$, p le premier élément de $trans$ et P' les éléments suivants
  si arb a un fils n tel que $nom(n) = p$ **alors**
    $incrementerOccurence(n)$
  sinon
    $n \leftarrow creerNoeudAvecOccurence(p, 1)$
    $ajouterFils(arb, n)$
  finsi
  si $P' \neq \varnothing$ **alors**
    $ajout\_page(n, P')$.
  finsi
fin

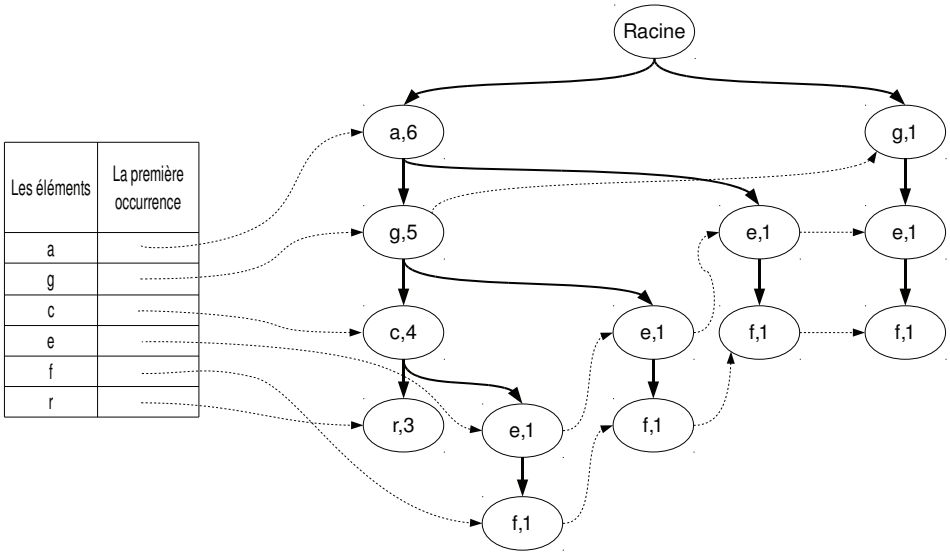


FIGURE 3-1 : Le *FP-tree* pour l'utilisateur 2.

3.4.1.2 Analyse du *FP-tree* (la méthode *FP-tree growth*)

L'objectif de l'algorithme *FP-tree growth* est de trouver tous les motifs fréquents existant dans le *FP-tree*. À partir du Tableau de Tête des Éléments Fréquents ($TTEF$), nous pouvons obtenir tous les motifs fréquents contenant chaque élément du tableau [Han 04]. Comme précédemment, nous allons décrire cet algorithme à l'aide du même exemple.

Pour rappel, les éléments du $TTEF$ sont des motifs fréquents. Il reste à déterminer les motifs fréquents incluant ces motifs. On effectue ceci en commençant par le dernier élément du $TTEF$ (*Cf.* figure 3-1).

Le nœud r a un chemin dans le *FP-tree* qui est $< (a,6), (g,5), (c,4), (r,3) >$. Comme le nœud r de cette branche [1] possède une fréquence de 3, cela signifie que le motif $agcr$ apparaît trois fois dans les transactions de l'utilisateur 2. On associe alors au motif $agcr$ cette fréquence de la manière suivante : $(agcr, 3)$. De la même manière, le motif agc apparaît quatre fois dans les transactions de cet utilisateur, on a donc $(agc, 4)$, $(ag, 5)$ et $(a, 6)$. Toutefois, le motif agc n'apparait que trois fois lorsqu'il est associé à r (en effet il apparaît aussi une fois lorsqu'il est associé à ef). On dit donc que la fréquence du motif agc est de 3 sachant r, ce qui se note $(agc, 3)|r$. Pour identifier les autres motifs fréquents contenant r, on construit un nouvel *FP-tree* à partir de ce motif, c'est ce que l'on nomme un *FP-tree* conditionnel d'un motif, dans notre cas de r. La figure 3-2 présente tous les *FP-trees* conditionnels de r, de cr, de

1. Il pourrait très bien y avoir un autre nœud r dans une autre branche, comme c'est le cas par exemple pour g.

gr, etc. Par conséquent, l'ensemble de motifs fréquents contenant r est $\{(r,3), (ar,3), (gr,3), (cr,3), (agr,3), (acr,3), (gcr,3), (agcr,3)\}$. On peut remarquer que l'identification des motifs fréquents d'un *FP-tree* n'ayant qu'un seul chemin produit toutes les combinaisons des éléments de ce chemin.

Le nœud f quant à lui a quatre chemins dans le *FP-tree* : $\{(a,6), (g,5), (c,4), (e,1), (f,1)\}$, $\{(a,6), (g,5), (e,1), (f,1)\}$, $\{(a,6), (e,1), (f,1)\}$, et $\{(g,1), (e,1), (f,1)\}$. Donc, les chemins contenants f sont $\{(agce,1), (age,1), (ae,1), (ge,1)\}$. La figure 3-3 représente tous les *FP-trees* conditionnels de f. Dès lors, l'ensemble des motifs fréquents contenant f est $\{(f,4), (ef,4), (af,3), (gf,3), (eaf,3), (egf,3)\}$.

On fait ensuite de même avec les motifs fréquents e, c, g, puis a. L'union de ces ensembles est l'ensemble des motifs fréquents de l'utilisateur 2.

Le tableau 3-4 représente les motifs fréquents pour tous les utilisateurs avec $s \geqslant s_{min}$ c'est-à-dire $s \geqslant 0{,}4$ (les équations 3-3, 3-5).

Utilisateur	Motifs fréquents
1	$\{b\}$, $\{bt\}$, $\{a\}$, $\{t\}$
2	$\{r\}$, $\{ar\}$, $\{gr\}$, $\{cr\}$, $\{agr\}$, $\{acr\}$, $\{gcr\}$, $\{agcr\}$, $\{f\}$, $\{ef\}$, $\{af\}$, $\{gf\}$, $\{eaf\}$, $\{egf\}$, $\{e\}$, $\{eg\}$, $\{ea\}$, $\{c\}$, $\{cg\}$, $\{ca\}$, $\{cga\}$, $\{g\}$, $\{ga\}$, $\{a\}$
3	$\{f\}$, $\{fh\}$, $\{c\}$, $\{ch\}$, $\{h\}$

TABLEAU 3-4 : Les motifs fréquents pour tous les utilisateurs avec $s_{min} = 0{,}4$.

L'algorithme 8 formalise la méthode *FP-tree growth*. Pour le premier appel à cet algorithme, il faut utiliser les arguments $(FP - tree, s_{min}, \phi)$.

Après avoir identifié tous les motifs fréquents, nous appliquons la condition de la confiance $c \geqslant c_{min}$ (les équations 3-4, 3-6) pour trouver les motifs caractéristiques. À la fin de cette étape, la phase d'apprentissage est terminée.

3.4.2 Fonction d'identification à base de réseaux bayésiens

Les utilisateurs qui naviguent sur un site web, parcours séquentiellement des documents (d_1, d_2, …, d_n). Certains regroupements de ces documents (indépendemment de leur ordre) peuvent (ou pas) représenter des motifs caractéristiques calculés précédemment. De ce fait pour une trace donnée et l'ensemble des motifs caractéristiques T^c, nous avons besoin d'une fonction qui calcule l'ensemble des motifs caractéristiques contenus dans cette trace (cet ensemble peut

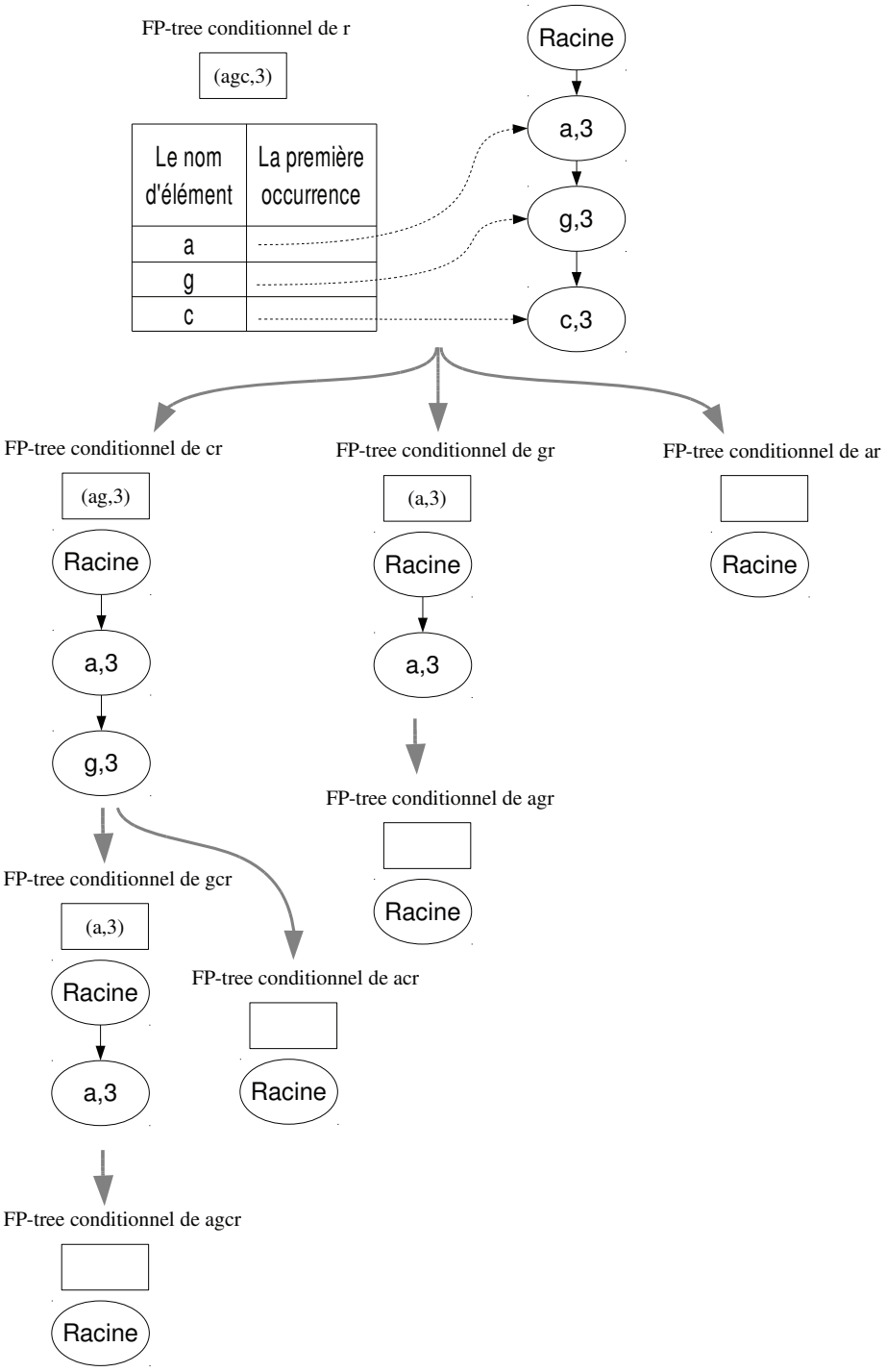


FIGURE 3-2 : Les *FP-trees* conditionnels de r.

Algorithme 8 Algorithme *FP-growth*(arb, s_{min}, α)

Entrée: *FP-tree*, le minimum du support s_{min}, l'ensemble des motifs fréquents précédent α.
Sortie: l'ensemble complet des motifs fréquents.
début
  si arb contient un seul chemin préfix **alors**
    $P \leftarrow$ le seul chemin préfixe de arb
    $Q \leftarrow$ l'ensemble des arbres issus de la première séparation de arb
    $\beta \leftarrow$ l'ensemble de toutes les combinaisons des nœuds de P
    // On génére l'ensemble de motifs fréquents de P (freq-pat-set(P)).
    $freq\_pat\_set\_P \leftarrow \alpha \cup \bigcup_{\beta_i \in \beta} \beta_i, support(\beta_i) \geqslant s_{min}$
  sinon
    $Q \leftarrow arb$
  finsi
  pour chaque élément $a_i \in Q$ **faire**
    support(α) $\leftarrow$ support(a_i)
    $\beta \leftarrow \{a_i\} \cup \alpha$
    // On génére la base de motifs conditionnels de β, et le *FP-tree* conditionnel de β (*FP-tree*$_\beta$)
    si *FP-tree*$_\beta \neq \phi$ **alors**
      $E \leftarrow \beta \cup$ *FP-growth*(*FP-tree*$_\beta, s_{min}, \beta$)
    sinon
      $E \leftarrow \beta$
    finsi
    $freq\_pat\_set\_Q \leftarrow freq\_pat\_set\_Q \cup E$
  finpour
  retourner { $freq\_pat\_set\_P \cup freq\_pat\_set\_Q \cup \{freq\_pat\_set\_P \times freq\_pat\_set\_Q\}$ }
fin

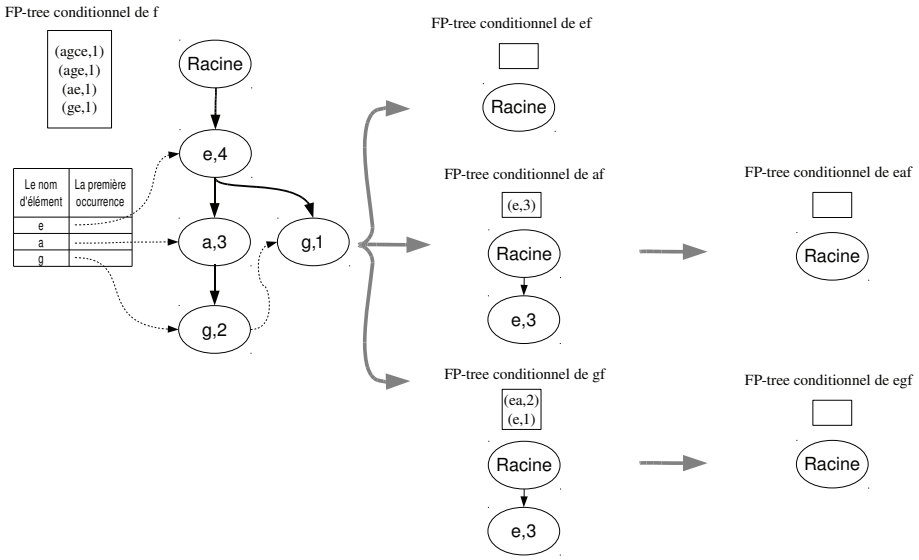


FIGURE 3-3 : Les *FP-trees* conditionnels de f.

être vide). Ceci est formalisé par la fonction suivante :

$$motifsCaractéristiques : Trace \times T^c \rightarrow M$$

C'est à partir de ces ensembles de motifs caractéristiques que nous devons définir une fonction d'identification, de décision. Nous avons étudié dans la section 1.3 les réseaux bayésiens. L'algorithme des motifs caractéristiques utilise, comme fonction d'identification, un certain type de réseau bayésien, les réseaux bayésiens naïfs.

3.4.2.1 Réseaux bayésiens naïfs et réseaux bayésiens naïfs multi-net

Un réseau bayésien naïf est un réseau bayésien simple, qui permet de prendre une décision. Comme présenté dans la figure 3-4, sa structure est composée d'un nœud représentant la classe C et des nœuds fils représentant les caractéristiques A_i. Aucune autre connexion n'est autorisée dans cette structure. Le réseau bayésien naïf suppose donc que toutes les caractéristiques soient indépendantes les unes des autres. Un réseau bayésien naïf représente une classe [2].

Les réseaux bayésiens multi-net (*Cf.* figure 3-5) permettent de représenter plusieurs classes (un réseau bayésien naïf par classe). Dès lors la topologie est fixée pour tous les réseaux, les

2. Lorsqu'il y a uniquement deux classes, un seul réseau bayésien naïf suffit, car la probabilité d'appartenance à la seconde classe est égale à un moins la probabilité d'appartenance à la première.

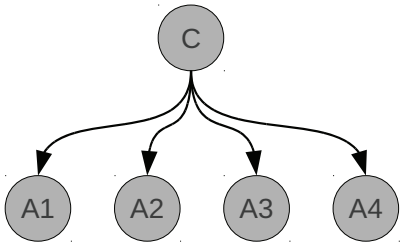


FIGURE 3-4 : Le réseau bayésien naïf.

liens de causalités sont donc les même. Par contre les probabilités associées à chaque nœud caractéristique A_i sont propres à chaque réseau et donc à chaque classe. L'appartenance d'une donnée à une classe est obtenue en sélectionnant le réseau bayésien naïf dont la probabilité calculée au niveau du nœud C_i est la plus grande.

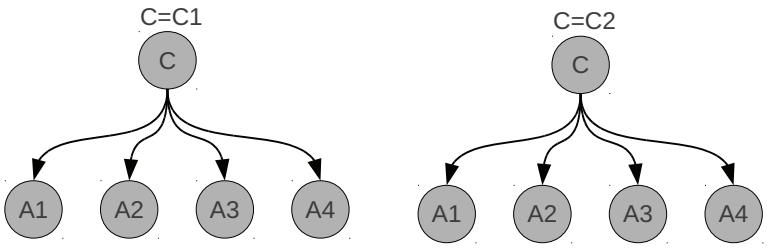


FIGURE 3-5 : Le réseau bayésien multi-net.

3.4.2.2 Construction des réseaux bayésiens pour les motifs caractéristiques

Dans notre contexte, les profils d'utilisateurs représentent les classes de notre problème. Comme le nombre de classes *a priori* est supérieur à deux, nous devons dans la fonction de décision utiliser un réseau bayésien naïf multi-nets. Chaque profil étant alors représenté par un réseau bayésien naïf.

Tous les réseaux bayésiens ayant la même topologie, ils possèdent autant de nœuds fils que de motifs caractéristiques de tous les profils T^c définis dans la partie précédente.

Il reste alors à initialiser le réseau bayésien multi-nets, c'est-à-dire pour chaque réseau bayésien associé au profil u_i de calculer pour chaque motif caractéristique, la probabilité d'appartenir à ce profil. m_k prend la valeur *oui* ou *non* (c'est-à-dire, m_k existe ou non dans la transaction). Cette probabilité est $P(m_k = oui | u = u_i)$ où m_k et u sont des variables aléatoires.

D'après l'équation 3-3, nous avons donc :

$$P(m_k|u = u_i) = s(m_k, u_i) = \frac{|T_{ik}|}{|T_i|}$$

Ceci peut être représenté par la figure 3-6.

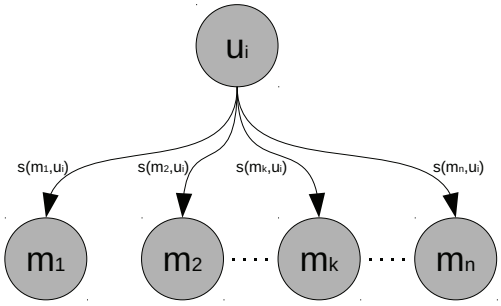


FIGURE 3-6 : Réseau bayésien naïf pour le profil u_i.

3.4.2.3 Identification d'un nouvel utilisateur

La fonction de décision revient alors à utiliser le réseau bayésien naïf multi-nets en calculant $P(u = u_i|m_1, m_2, ..., m_n)$.

Selon le théorème de Bayes :

$$P(B_i|A) = \frac{p(A|B_i)p(B_i)}{p(A)} \quad \Rightarrow \quad P(A|B_i) = \frac{P(B_i|A)p(A)}{p(B_i)}$$

Donc :

$$P(u = u_i|m_1, m_2, ..., m_n) = \frac{P(m_1, m_2, ..., m_n|u = u_i)P(u = u_i)}{P(m_1, m_2, ..., m_n)} \tag{3-7}$$

Pour une transaction t, $P(m_1, m_2, ..., m_n)$ est le même pour tous les utilisateurs. Ainsi le problème d'identification de l'utilisateur revient à trouver u_i qui maximise le numérateur de l'équation 3-7, c'est-à-dire :

$$P(m_1, m_2, ..., m_n|u = u_i)P(u = u_i)$$

Selon l'algorithme proposé par [Gao 08], l'identification du nouvel utilisateur revient à

trouver u_{new} tel que :

$$\begin{aligned} u_{new} = & \quad \arg\max_{u_i}\{P(m_1, m_2, ..., m_n|u = u_i)P(u = u_i)\} \\ = & \quad \arg\max_{u_i}\{\prod_{k1} P(m_{k1} = oui|u = u_i) \times \prod_{k2} P(m_{k2} = non|u = u_i)\} \end{aligned} \tag{3-8}$$

Où m_{k1} sont les motifs caractéristiques se trouvant dans les transactions du nouvel utilisateur, et m_{k2} sont tous les motifs caractéristiques qui n'existent pas dans les transactions de cet utilisateur.

Nous venons d'étudier le WUM en général et l'algorithme des motifs caractéristiques en particulier. Notons toutefois que cet algorithme est habituellement utilisé pour identifier un utilisateur qui aurait déjà visité le site. On se propose de tester ce dernier sur nos données expérimentales. Mais dans notre cas, plutôt qu'identifier un utilisateur, on se propose de voir s'il est possible d'identifier le profil d'un utilisateur parmi un certain nombre de profils. Avant cela, nous allons décrire notre contexte de travail, un site web encyclopédique thématique.

3.5 Contexte : le site web Hypergéo

Hypergéo est un site web représentant une encyclopédie géographique ouverte à tout public (*Cf.* figure 3-7) . Ce site [3] est né dans le cadre du Groupe de Recherche Libergéo. Il est disponible en trois langues (français, anglais et espagnol). Le projet a débuté en 2001, son objectif est d'expliquer les principaux concepts et théories en géographie [Eli 07].

3.5.1 Organisation

L'organisation d'Hypergéo est hiérarchique. Pour chaque langue les informations sont ordonnées sous une des quatre rubriques principales : « Géographie », « Régions et Territoires », « Relations Sociétés/Environnement » et « Spatialité des sociétés ». Ensuite, chaque rubrique est divisée en rubriques ou articles. Par exemple la figure 3-7 présente l'article « La région comme système » qui est inclus dans la hiérarchie de rubriques « Géographie/Concepts/Autres conceptions de la région ». La figure 3-8 présente cette organisation, les rubriques et les articles sont respectivement représentés à l'aide de losanges et d'ovales. En 2008, le site Hypergéo contenait 358 articles et 76 rubriques rédigés par 46 experts en géographie.

3. www.hypergeo.eu

3.5. Contexte : le site web Hypergéo

FIGURE 3-7 : Le site web hypergéo.

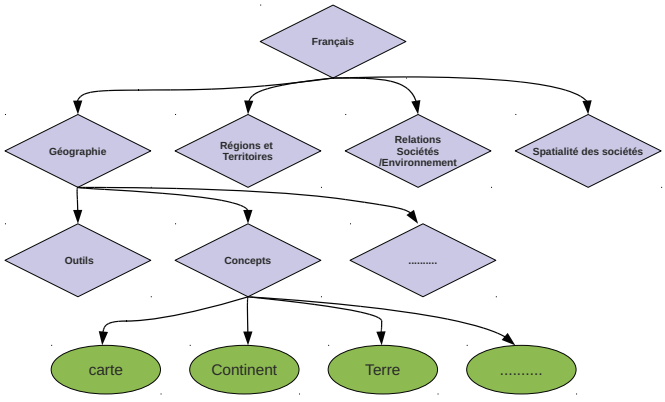


FIGURE 3-8 : Extrait de l'organisation d'Hypergéo.

3.5.2 Système de publication SPIP

Hypergéo utilise le *Content Managment System* SPIP[4] afin de publier les articles. Comme beaucoup d'autres CMS, il permet la publication multi-lingues. Il s'agit d'un logiciel libre sous licence GNU/GPL, écrit en PHP et utilisant la base de données MySQL. Les contenus des pages sont sauvegardés dans cette base de données. Quand une page est demandée, SPIP la génère à partir de la base de données et il peut alors y ajouter des informations supplémentaires, comme par exemple des liens internes qui n'auraient pas été explicités lors de la création de la page.

Grâce à cette propriété de SPIP, dès que nous aurons identifié le profil d'un utilisateur, nous pourrons ajouter, à l'aide de scripts, des informations en fonction de son profil. Hypergéo sera alors un hypermédia adaptatif comme proposé dans l'introduction de ce mémoire.

3.5.3 Recueil des données

Afin d'étudier les comportements des utilisateurs, nous avons recueilli leurs navigations, c'est-à-dire leurs transactions HTTP pendant environ quatre mois (du 14 février 2008 au 30 mai 2008).

Pour ne pas perturber le fonctionnement du site Hypergéo, notamment son indexation par des moteurs de recherche, nous l'avons dupliqué. Seules les requêtes des utilisateurs étaient redirigées vers cette copie en toute transparence, l'annexe A explique cette opération. L'ensemble de ces transitions datées était stocké dans une base de données. Nous avons décidé de ne pas garder l'adresse IP du navigateur Web, car cette information n'est pas fiable (plusieurs utilisateurs peuvent partager une même adresse IP).

Durant ce recueil, quand un utilisateur visitait le site pour la première fois, une page d'accueil lui demandait de sélectionner l'un des sept profils déterminés par un expert du domaine. Les profils proposés sont :

1. étudiant en géographie,
2. étudiant dans autre domaine,
3. enseignant en géographie,
4. enseignant dans autre domaine,
5. enseignant/chercheur en géographie,
6. enseignant/chercheur dans autre domaine,
7. autre.

À la fin de la période de recueil des données, nous avons obtenu les informations de navigation de 17792 utilisateurs qui ont effectué 89787 transitions. Le tableau 3-5 présente le nombre des utilisateurs et des transitions effectuées selon les profils.

4. http://www.spip.net/

Profil	Nombre des utilisateurs	Nombre des transitions	Moyenne $\overline{X}$	Écart type empirique $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}$
Étudiant en géographie	10249	55993	5,4633	13,5058
Étudiant dans autre domaine	3678	12403	3,3722	9,1197
Enseignant en géographie	539	4471	8,2950	17,9761
Enseignant dans autre domaine	291	1034	3,5533	7,34561
Enseignant/chercheur en géographie	673	6392	9,4978	18,5104
Enseignant/chercheur dans autre domaine	464	1954	4,2112	9,7401
Autre	1898	7540	3,9726	9,7359

TABLEAU 3-5 : Les utilisateurs et leurs transitions selon les profils.

Parmi les utilisateurs, il y a 8397 utilisateurs qui ont effectué 52637 transitions sur des documents français. Le tableau 3-6 présente quelques statistiques quant à leurs navigations. Nous pouvons constater qu'à chaque fois l'écart type est assez grand. Ce qui indique que la longueur des traces entre utilisateurs de même profil est très variable. Par exemple pour le profil étudiant en géographie francophone (ayant visité au moins un document en français), la trace la plus courte comporte 1 transition et la trace la plus longue comporte 354 transitions.

Profil	Nombre des utilisateurs	Nombre des transitions	Moyenne $\overline{X}$	Écart type empirique $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}$
Étudiant en géographie	4833	33306	6,8914	15,1537
Étudiant dans autre domaine	1549	6638	4,2853	11,9304
Enseignant en géographie	341	2837	8,3196	12,6498
Enseignant dans autre domaine	150	581	3,8733	7,1641
Enseignant/Chercheur en géographie	441	4257	9,6531	18,2863
Enseignant/Chercheur dans autre domaine	203	968	4,7685	10,6457
Autre	880	4050	4,6023	10,6348

TABLEAU 3-6 : Les utilisateurs et leurs transitions sur des documents français selon les profils.

Appliquons maintenant l'algorithme des motifs caractéristiques à ces données pour identi-

fier les profils des utilisateurs du site Hypergéo.

3.6 Applications des motifs caractéristiques à Hypergéo

Après la préparation des données de navigation sur Hypergéo, nous allons chercher à classer un nouvel utilisateur parmi les sept profils prédéfinis :

1. étudiant en géographie (**ETG**) ;
2. étudiant dans autre domaine (**ETAD**) ;
3. enseignant en géographie (**EG**) ;
4. enseignant dans autre domaine (**EAD**) ;
5. enseignant/chercheur en géographie (**ECG**) ;
6. enseignant/chercheur dans autre domaine (**ECAD**) ;
7. autre (**Autre**).

Toutefois au vue des statistiques précédentes, il apparaît que pour certains profils le nombre d'utilisateurs est faible. Nous allons donc aussi essayer de classer un nouvel utilisateur parmi trois profils qui identifie *a priori* leur domaine de compétences, c'est-à-dire :

1. spécialiste en géographie (**G**),
2. spécialiste dans autre domaine (**AD**)
3. autre (**Autre**)

Dans nos expériences, nous prenons en compte uniquement les utilisateurs qui consultent les documents français. Comme nous l'avons présenté dans la section 3.5.3 page 84, il y a 8397 utilisateurs effectuant des transitions sur les documents français.

Le tableau 3-7 représente la distribution des utilisateurs francophones selon le nombre de transitions effectuées.

NT (Nombre de transitions effectuées)	Nombre d'utilisateurs
$NT < 3$	4775
$2 < NT < 11$	2397
$10 < NT < 16$	447
$15 < NT$	778

TABLEAU 3-7 : La distribution des utilisateurs francophones selon le nombre de transitions effectuées.

La figure 3-9 représente le nombre d'utilisateurs francophones du site Hypergéo selon leurs nombres de transitions effectuées.

Malheureusement, le tableau 3-7 et la figure 3-9 montrent que la plupart des utilisateurs font une ou deux transitions. Ces utilisateurs n'apportent pas beaucoup d'informations. Nous

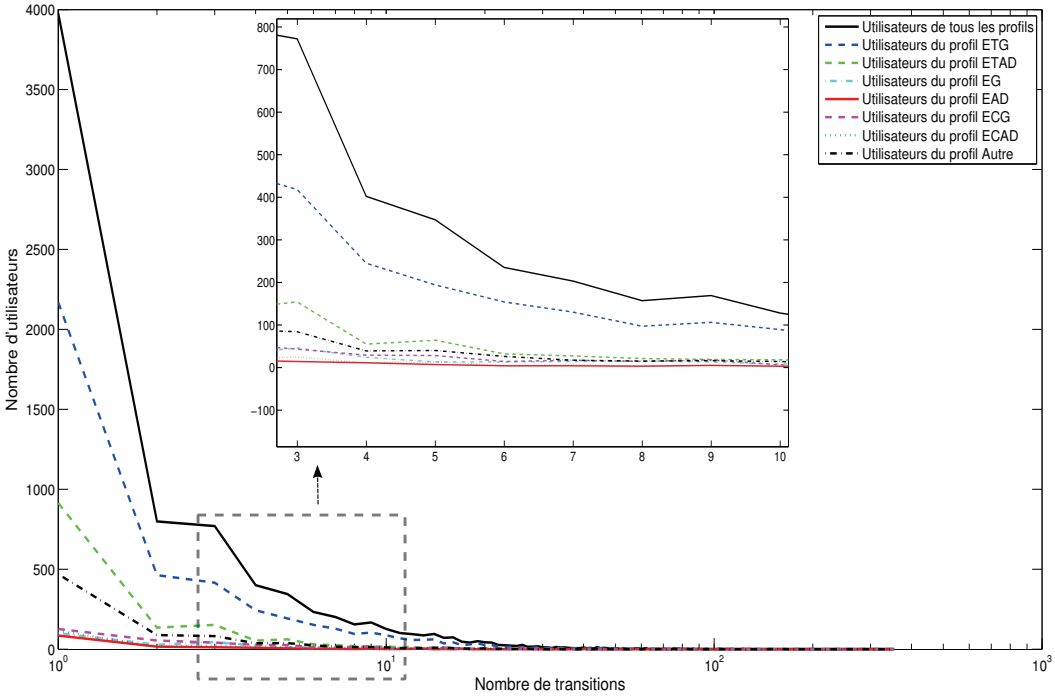


FIGURE 3-9 : Le nombre d'utilisateurs francophones du site Hypergéo selon leurs nombres de transitions effectuées.

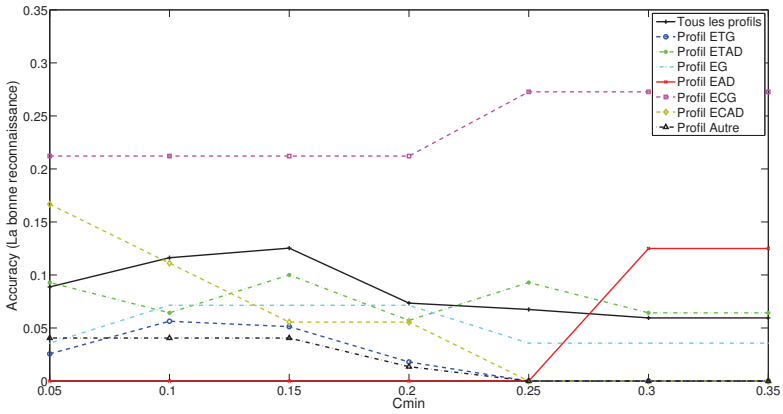
pouvons penser qu'ils visitent le site hypergéo un peu au hasard. Nous voyons aussi qu'il y a peu d'utilisateurs qui effectuent plus de 10 transitions, la grande majorité est issue des profils **ETG** et **ETAD**. Les prendre en compte risqueraient de fausser nos algorithmes d'apprentissage au dépend des autres profils. Dès lors dans nos expérimentations, nous allons seulement utiliser ceux qui effectuent entre trois et dix transitions. Cela représente 2397 utilisateurs (*Cf.* le tableau 3-7). Ils constituent les bases de données pour l'apprentissage et pour les tests. Nous avons pris 60% d'utilisateurs pour l'apprentissage et 40% pour les tests.

Le tableau 3-8 donne la distribution d'utilisateurs selon leur profil. Dans ce tableau, nous pouvons voir les déséquilibres qu'il y a entre les nombres d'utilisateurs de chaque profil. À titre d'exemple, plus de la moitié des utilisateurs appartiennent au profil **ETG**, alors qu'il y a peu d'utilisateurs du profil **EAD**.

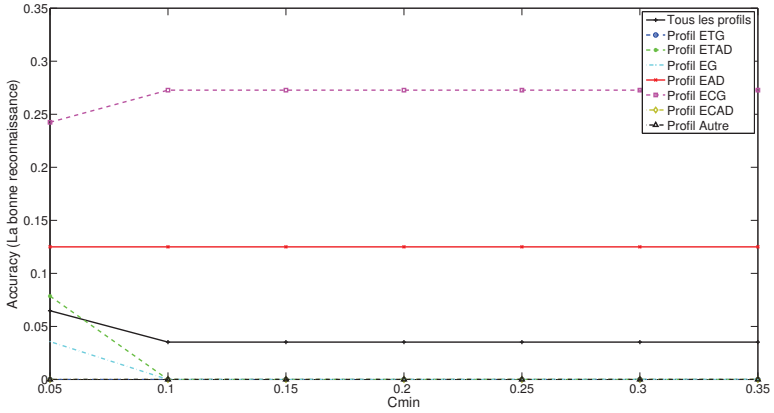
Après la présentation des résultats de nos manipulations, dans la conclusion, nous expliquerons l'influence de ces déséquilibres.

Tout d'abord, nous essayons de classer les utilisateurs de test parmi sept profils (**ETG**, **ETAD**, **EG**, **EAD**, **ECG**, **ECAD**, **Autre**).

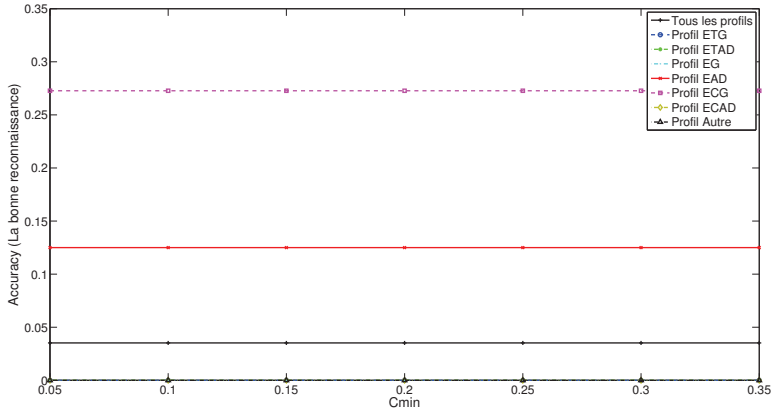
La figure 3-10 représente les résultats de l'utilisation de la méthode des motifs caractéristiques pour plusieurs valeurs de s_{min} et c_{min}.



(a) $Smin = 0{,}1$



(b) $Smin = 0{,}2$



(c) $Smin = 0{,}3$

FIGURE 3-10 : Les résultats de l'application des motifs caractéristiques sur les données des utilisateurs ayant effectués entre 3 et 10 transitions pour 7 profils.

Profil	Nombre d'utilisateurs
1. Étudiant en géographie (**ETG**)	1417
2. Étudiant dans autre domaine (**ETAD**)	374
3. Enseignement en géographie (**EG**)	130
4. Enseignement dans autre domaine (**EAD**)	35
5. Professeur en géographie (**ECG**)	153
6. Professeur dans autre domaine (**ECAD**)	53
7. Autre (**Autre**)	235

TABLEAU 3-8 : La distribution des utilisateurs francophones qui exécutent entre trois et dix transitions selon leur profil.

D'après la littérature, nous devons prendre des valeurs assez grandes pour s_{min} ou c_{min}, c'est-à-dire généralement supérieures à $0{,}3$. Cependant pour nos données, il nous est impossible de choisir de telles valeurs. En effet lorsque que l'on essaye d'augmenter la valeur de s_{min}, à partir de $s_{min} = 0{,}2$ (et $c_{min} = 0{,}1$), il n'y a plus de motifs caractéristiques pour les profils **EG**, **EAD**, **ECG**, et **ECAD**. Dans ce cadre, nous obtenons un seul motif caractéristique pour le profil **ETAD**, deux pour le profil **Autre** et trois pour les profil **ETG**.

De la même manière, lorsque nous essayons de prendre une valeur de c_{min} plus grande, dès la valeur de $0{,}25$ (avec $s_{min} = 0{,}1$), nous obtenons des motifs caractéristiques uniquement pour les profils **ETG** et pour **EAD**.

La figure 3-11 propose le nombre de traces des utilisateurs ayant des motifs caractéristiques selon différentes valeurs de s_{min} et c_{min}. Nous notons bien que pour les petites valeurs de s_{min} et c_{min} ($s_{min} = 0{,}2$ et $c_{min} = 0{,}1$), seulement 284 traces parmi 691, soit 44%, ont des motifs caractéristiques. Même pour $s_{min} = 0{,}1$ et $c_{min} = 0{,}1$, il existe uniquement 387 traces ayant des motifs caractéristiques, soit 56% des traces initiales. Dès lors, même dans ce cas extrême, 44% des traces de la base de tests ne possèdent pas de motifs caractéristiques, et ne peuvent donc pas être classées.

Les figures 3-10(a), 3-10(b), 3-10(c) évaluent la classification selon le taux de reconnaissance (*Accuracy*). Nous trouvons que la méthode des motifs caractéristiques ne donne pas de bons résultats sur nos données pour classer les utilisateurs parmi les sept profils.

Par exemple, pour $s_{min} = 0{,}1$ (*Cf.* la figure 3-10(a)) le taux de la reconnaissance pour tous les profils n'atteind pas $0{,}15$ ($0{,}12$ pour $c_{min} = 0{,}15$, la ligne noire). Dans la même figure, le meilleur résultat est $0{,}27$ pour le profil **ECG** (la ligne mauve pointillée avec $s_{min} = 0{,}1$, $c_{min} = 0{,}25$). Des résultats similaires sont trouvés pour $s_{min} = 0{,}2$ (*Cf.* la figure 3-10(b)) et $s_{min} = 0{,}3$ (*Cf.* la figure 3-10(c)). Dans ces figures (3-10(a), 3-10(b), 3-10(c)), plus la valeur de s_{min} est grande plus les résultats (taux de reconnaissance) sont mauvais (la meilleure valeur

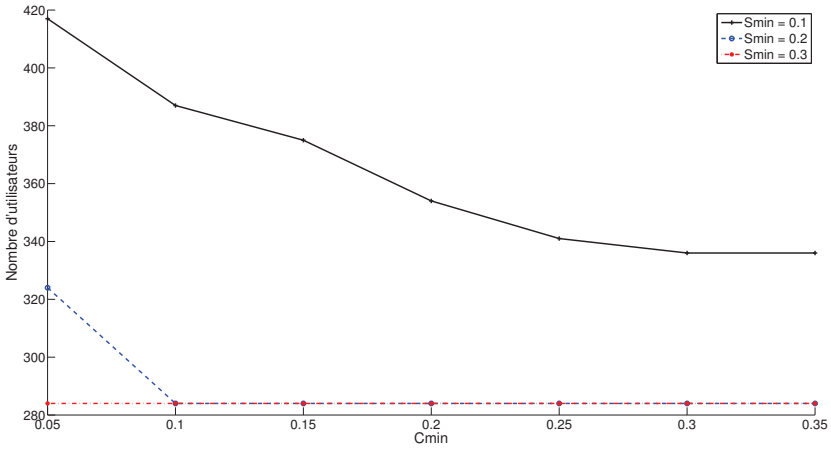


FIGURE 3-11 : Nombre d'utilisateurs qui ont des motifs caractéristiques parmi 691 utilisateurs de test.

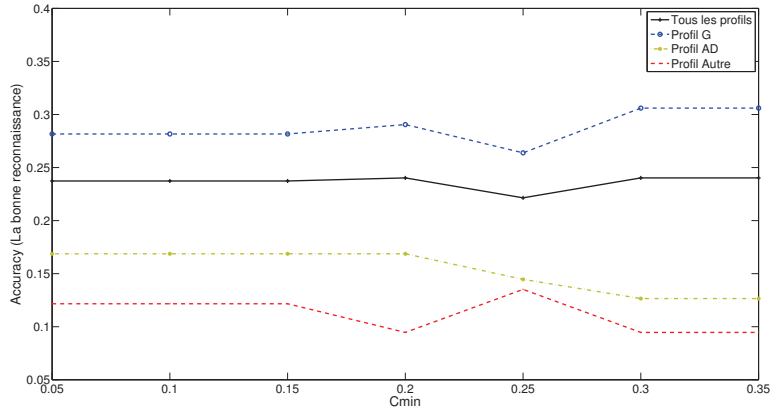
pour tous les profils est de $0{,}12$ pour $s_{min} = 0{,}1$, $0{,}06$ pour $s_{min} = 0{,}2$ et $0{,}03$ pour $s_{min} = 0{,}3$). Ces résultats peuvent être expliqués par le fait que plus la valeur de s_{min} est grande, plus le nombre d'utilisateurs et le nombre de profils qui n'ont pas de motifs caractéristiques sont grands. Dans ces figures nous remarquons aussi que s_{min} joue un rôle plus important que c_{min} sur nos données, parce que les courbes ne changent pas beaucoup quand la valeur de c_{min} est modifiée.

Dans la deuxième expérimentation, nous essayons de classer les utilisateurs parmi trois profils (**G**, **AD**, **Autre**). La figure 3-12 présente ces résultats. Même si les résultats sont supérieurs à ceux calculés pour les sept profils, ils restent insuffisants. En effet dans le meilleur cas, le taux de reconnaissance ne dépasse pas $0{,}3$ pour tous les profils (la ligne noire dans la figure 3-12(c)).

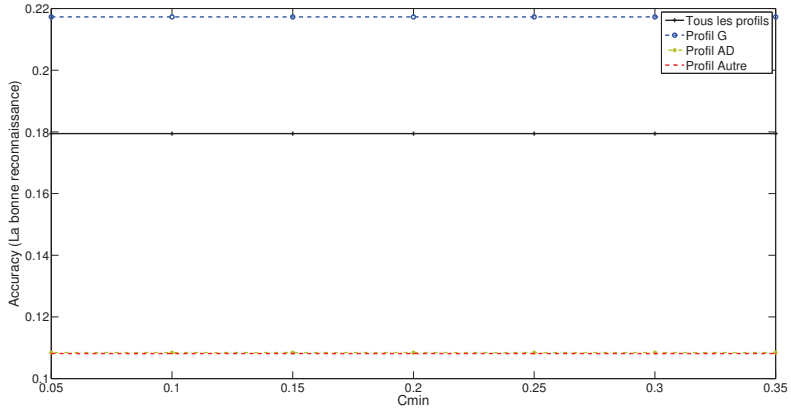
3.7 Conclusion

Le *web usage mining* concerne l'application des techniques de fouille de données sur les données brutes d'accès à un site pour trouver les motifs significatifs. Les motifs séquentiels sont difficiles à mettre en œuvre et nécessitent des bases de données riches afin de refléter les comportements des utilisateurs. Nous les avons donc écartés au profit de la méthode des motifs caractéristiques. Nous avons présenté en détail cette dernière. Elle est conjointement utilisée avec les réseaux bayésiens pour identifier un nouvel utilisateur.

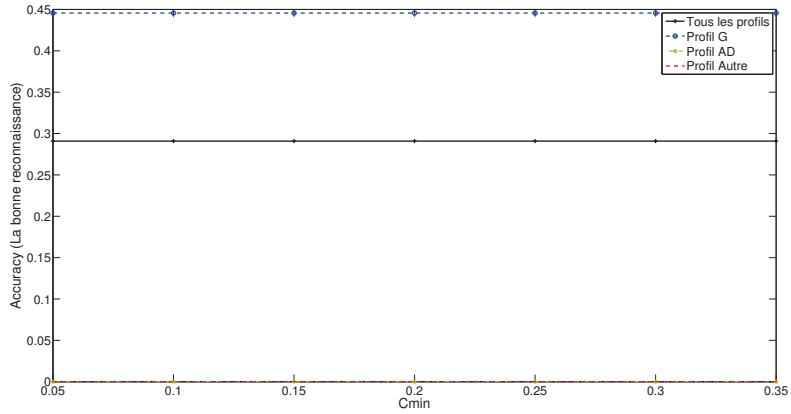
La méthode des motifs caractéristiques a été appliquée sur nos données représentant les traces des utilisateurs francophones d'Hypergéo afin de classer des nouveaux utilisateurs soit



(a) $Smin = 0{,}1$



(b) $Smin = 0{,}2$



(c) $Smin = 0{,}3$

FIGURE 3-12 : Les résultats de l'application des motifs caractéristiques sur les données des utilisateurs ayant effectués entre 3 et 10 transitions pour 3 profils.

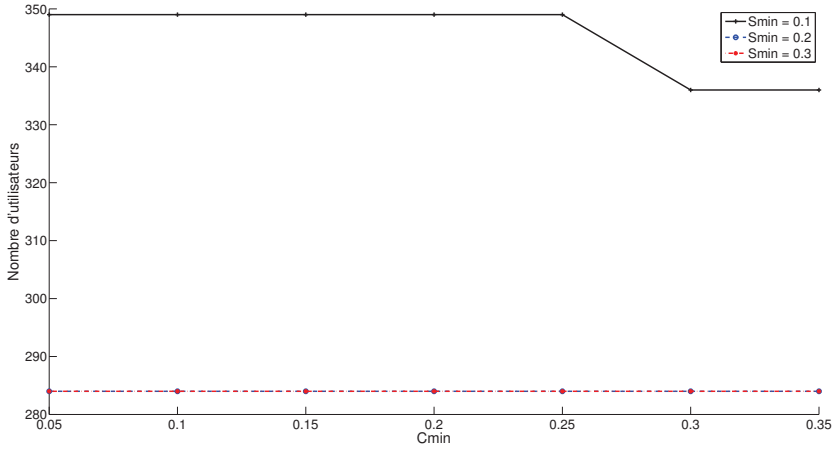


FIGURE 3-13 : Nombre d'utilisateurs qui ont des motifs caractéristiques parmi 691 utilisateurs de test.

parmi sept profils, soit parmi trois profils. Malheureusement, les résultats ne sont pas probants et nous n'arrivons pas à identifier correctement ces nouveaux utilisateurs.

La méthode des motifs caractéristiques ne fonctionne pas correctement sur nos données pour deux raisons :

1. Il y a beaucoup de données inutiles, par exemple parmi les 8397 utilisateurs, car il y a 4775 utilisateurs qui ont effectué seulement une ou deux transitions. Même pour les données utiles, il y a un grand déséquilibre entre le nombre d'utilisateurs pour chaque profil. Par exemple pour les utilisateurs ayant entre trois et dix transitions, il y a 1417 étudiants en géographie, mais seulement 35 enseignants d'un autre domaine. À cause de ces déséquilibres, nous ne pouvons pas trouver les motifs caractéristiques. En effet lorsqu'il y a des motifs qui réalisent le critère de s_{min} (équation (3-5)), il est difficile de réaliser le critère de c_{min} (équation (3-6)).
2. Nous n'avons pas de motifs caractéristiques qui ont s_{min} ou c_{min} assez grand pour déterminer les profils.

Dans les chapitres suivants, nous allons présenter une nouvelle approche utilisant des algorithmes à noyau, c'est-à-dire des algorithmes qui nécessite une similarité entre nos traces.

CHAPITRE 4

Dissimilarités entre traces basées sur la similarité entre éléments

Sommaire

4.1 Introduction

Dans le chapitre 3, nous avons montré que la méthode des motifs caractéristiques ne permet pas de résoudre le problème d'identification des profils sur nos données. Une des raisons qui empêche l'application de cette méthode est le fait que pour trouver des motifs caractéristiques il faut choisir des petites valeurs pour s_{min} et c_{min}, et donc les quelques motifs caractéristiques trouvés ne représentent pas assez les profils des utilisateurs. Cela est dû au fait que le nombre de traces pour chaque profil est assez faible au regard du nombre de documents que peuvent visualiser les utilisateurs.

Une des façons de résoudre ce problème est de réduire le nombre de documents au sein des traces, c'est-à-dire considérer que deux documents différents dans une trace, mais assez proches, représentent un seul document. Cette approche comporte toutefois deux inconvénients. Tout d'abord il faut déterminer à partir de quel seuil deux documents similaires sont considérées comme représentant un seul document. Ensuite, il faut appliquer à chaque trace une fonction qui la transforme en remplaçant les documents considérés.

Une autre approche, celle que nous appliquons ici, est de calculer une similarité entre les traces, utilisant une similarité entre des documents ou des transitions (comme nous allons le voir par la suite), et appliquer alors des algorithmes de classification à noyaux.

Dans ce chapitre nous allons donc étudier les navigations des utilisateurs du site Hypergéo. Ensuite, nous allons proposer plusieurs mesures de similarité entre traces qui seront validées à l'aide de données synthétiques.

4.2 Traces des utilisateurs d'Hypergeo

Nous prenons quelques exemples des navigations d'utilisateurs du site Hypergéo pour mieux comprendre leurs comportements. Le tableau 4-1 représente les documents visités par 4 utilisateurs :

ID	Titres des documents visités
80	'Statistique spatiale', 'Carte choroplèthe', 'carte'
322	'CARTOGRAPHIE THÉMATIQUE', 'carte', 'Figuration cartographique'
377	'La dynamique des systèmes selon J.W. Forrester', 'Localisation', 'Système spatial', 'Théories de l'Analyse Spatiale'
597	'Auto-organisation', 'Résilience', 'Localisation selon F.Durand-Dastés', 'Qu'est-ce qu'un système-expert ?'

TABLEAU 4-1 : Les titres des documents visités par quatre utilisateurs du site Hypergéo.

Selon la méthode des motifs caractéristiques, il n'y a qu'un motif commun entre les utilisateurs 80 et 322, et il n'existe aucun motif commun entre les utilisateurs 377 et 597. Mais, il semble que les utilisateurs 80 et 322 ont des intérêts similaires : les documents visités concernent tous le concept de « carte ». De la même manière, les utilisateurs 377 et 597 ont visité des documents qui concernent la localisation et le concept de système.

Lors de l'identification d'un nouvel utilisateur, il faudrait pouvoir prendre en compte ces similarités entre documents au sein des traces. Il faut donc définir une similarité entre traces et utiliser des algorithmes qui puissent en tirer partie.

Cette constatation nous amène à proposer une nouvelle approche basée sur la méthode de classification à noyau pour identifier les utilisateurs. Ces méthodes de classification ont besoin d'une mesure de similarité entre les informations recueillies lors des navigations des utilisateurs. La structure de la trace d'un utilisateur du site Hypergéo est comparable à un graphe chronologique orienté où les nœuds du graphe représentent les pages de l'hypermédia, et les arcs représentent les transitions (*Cf.* la figure 4-1).

Puisqu'il y a un ordre temporel dans ce graphe, une trace peut aussi être représentée par une liste de transitions sans perte d'information. Toutefois, si nous acceptons une perte d'information, nous pouvons aussi représenter une trace par un ensemble de transitions (on perd alors l'information temporelle et le fait qu'un utilisateur ait effectué plusieurs fois la même transition). De même nous pouvons représenter une trace par un ensemble de documents visités. Dans ce cas, les transitions ne sont plus représentées, on ne sait donc plus si un document est source ou destination d'une transition.

En résumé, les traces peuvent être appréhendées sous plusieurs aspects [Abo 10] :

- un ensemble de nœuds (documents),
- un ensemble de transitions (nœud$X \rightarrow$ nœud$Y$, ...),
- une liste de transitions (dans un ordre chronologique).

Nous avons vu qu'il n'y a pas beaucoup des motifs fréquents dans les traces des utilisateurs du site Hypergéo. Dès lors conserver l'information liée à la chronologie des transitions diminue d'autant la découverte de motifs. Par conséquent, nous choisissons de considérer les traces comme des ensembles d'éléments. Ces éléments seront alors soit des documents du site, soit des transitions entre documents du site.

À titre d'exemple, la figure 4-2 représente les traces de deux utilisateurs U_a et U_b. L'ensemble des documents de la trace U_a est $A_{doc} = \{2, 3, 8, 6\}$. Pour U_b, c'est $B_{doc} = \{1, 2, 5, 8\}$. L'ensemble des transitions de la trace U_a est $A_{trans} = \{(2, 3), (3, 8), (8, 3), (3, 6), (6, 2)\}$. Pour U_b, c'est $B_{trans} = \{(1, 2), (2, 5), (5, 8)\}$. Nous notons que la transition $(2, 3)$ apparaît deux fois

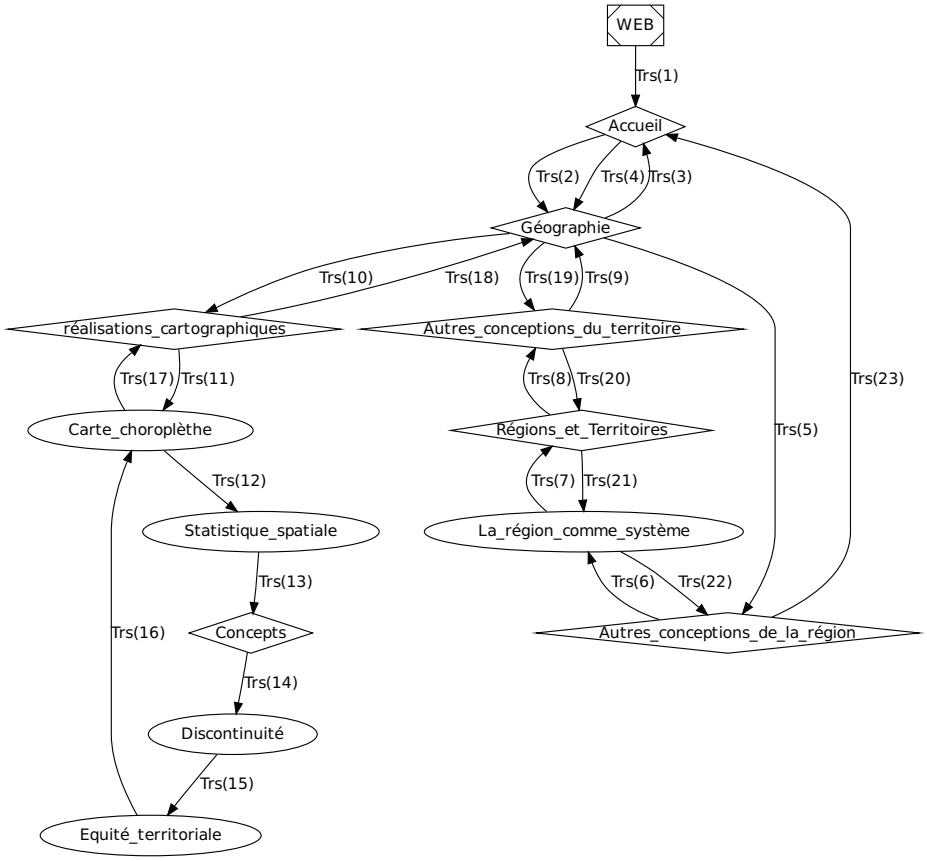


FIGURE 4-1 : Une trace d'un utilisateur du site Hypergéo.

dans la trace U_a, mais une seule fois dans l'ensemble A (parce qu'un ensemble ne peut pas posséder plusieurs fois le même élément).

Pour savoir si deux utilisateurs du site appartiennent au même profil, nous avons besoin de définir une similarité entre leurs traces, et donc de définir une similarité entre les éléments de leur trace. Une similarité possible entre la trace de U_a et la trace de U_b, basée sur une dissimilarité entre ces traces $D_{trace}(U_a, U_b)$ peut être définie par :

$$Sim_{trace}(U_a, U_b) = \exp^{(-\frac{(D_{trace}(U_a,U_b))^2}{\sigma})} \quad (4\text{-}1)$$

Avant d'étudier plusieurs propositions de dissimilarités D_{trace}, nous devons définir leurs propriétés et voir comment nous allons évaluer ces propositions.

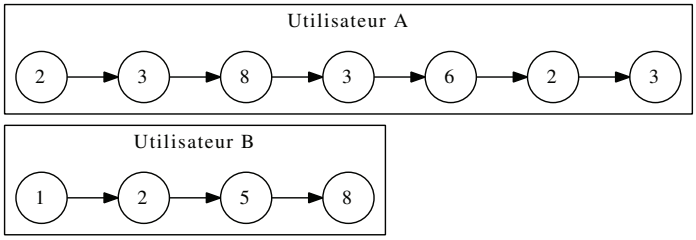


FIGURE 4-2 : Un exemple des traces des utilisateurs.

4.3 Propriétés de la dissimilarité entre traces

Dans notre contexte, les éléments en commun caractérisent un intérêt similaire des utilisateurs. Nous voulons donc qu'une bonne dissimilarité satisfasse, en plus des deux propriétés intrinsèques au concept de dissimilarité, la contrainte suivante : le nombre des éléments en commun ou les plus proches doit prévaloir face à la longueur des traces.

Ces propriétés vont nous guider dans la construction d'une dissimilarité entre traces. Mais plutôt que de les confronter directement à nos données « réelles », nous avons décidé de les valider à l'aide de données synthétiques. Ceci nous permettra de faire varier les caractéristiques de ces données et ainsi définir dans quel contexte ces mesures fonctionnent convenablement.

4.3.1 Description des données synthétiques

Nous rappelons qu'une « trace » est l'ensemble des données recueillies lors de la navigation d'un utilisateur. Dans notre cas, nous avons extrait ces données à partir des informations enregistrées sur le serveur (c'est-à-dire que nous perdons quelques informations comme par exemple l'utilisation de la fonction « page précédente » des navigateurs).

Pour valider notre proposition, plusieurs bases de données synthétiques sont construites à l'image des traces des utilisateurs issues du site Hypergéo, c'est-à-dire :

- le même nombre de classes (sept ou trois),
- la même proportion du nombre de traces dans chacune des classes (la première classe est prépondérante par rapport aux six autres classes, *Cf.* le tableau 3-8 page 88),
- la longueur des traces entre 10 et 15 éléments (pour povoir faire varier le paramètre $DiscNB$, défini ci dessous, de 0,1 à 0,6) .

Pour la construction des bases de données synthétiques, nous allons faire varier deux paramètres $DiscNB$ et $SimDisc$.

1. $DiscNB$ est le pourcentage des éléments discriminants dans une trace (l'élément discriminant d'une classe est l'élément qui apparaît dans cette classe de manière beaucoup plus fréquente que dans les autres classes),

$$p(element_i \in TransDisc_i | profil_i) \geqslant DiscNB \tag{4-2}$$

2. $SimDisc$ est la similarité minimale entre les éléments discriminants d'un même profil. Dés lors, nous avons construit une matrice de similarité, telle que deux éléments de la même classe ont une similarité aléatoire supérieure à $SimDisc$, alors que deux éléments de classes différentes ont une similarité aléatoire strictement inférieure à $SimDisc$.

4.3.2 Cas d'un noyau non semi-défini positif

Comme l'indique [Haa 04], le fait qu'un « noyau ne soit pas semi-défini positif apparaît lorsque la distance [utilisée] n'est pas métrique [1], c'est-à-dire qu'elle viole l'inégalité triangulaire [2] ». De ce fait si les dissimilarités que nous allons définir ne sont pas des distances, il y a un risque pour que le noyau calculé ne soit pas semi-défini positif. Or, lors de la phase d'apprentissage du SVM, l'utilisation d'un noyau semi-défini positif peut empêcher de trouver les vecteurs supports, et donc que l'on ne puisse pas utiliser cette méthode.

On se propose alors de représenter nos données dans un autre espace qui utilisera la distance euclidienne. Comme le propose [Pek 02], il est possible de représenter les données x de l'espace de départ dans un nouvel espace, on note alors cette représentation $\overline{x}$, telle que les caractéristiques (coordonnées) de $\overline{x}$, soient égales à la distance entre x et toutes les données de l'espace de départ. Si le nombre de données est n, cet espace d'arrivée est donc de dimension n et les données $\overline{x} \in [0,1]^n$. Nous allons nommer cet espace, l'espace de voisinage.

Cet espace de voisinage conserve la similarité entre nos données. En effet si deux points x et y, de l'espace de départ sont proches, alors leur distance vis à vis des autres points doivent être équivalentes. Dès lors puisque ces distances vis à vis de tous les autres points représentent les coordonnées de leurs images $\overline{x}$ et $\overline{y}$ dans l'espace d'arrivée, alors $\overline{x}$ et $\overline{y}$ seront proches (distance euclidienne proche de 0) dans l'espace de voisinage.

Formellement, soit les $x^1 \dots x^n$ les points de notre espace de départ. On considère que x^a

1. Ce que nous appelons la dissimilarité
2. Traduction de *It allows to conclude missing pd-ness of DS-kernels that arise from distances which are non-metric, e.g. violate the triangle inequality*

et x^b sont proches (au sens d'une dissimilarité d_d) dans l'espace de départ si :

$$d_d(x^a, x^b) \simeq 0 \tag{4-3}$$

$$\forall x^i, d_d(x^a, x^i) \simeq d_d(x^b, x^i) \tag{4-4}$$

Dans l'espace de voisinage, on a $\overline{x^i}$ qui a pour coordonnées $(d_d(x^i, x^1), d_d(x^i, x^2), \ldots, d_d(x^i, x^n))$.

Dans ce cas la distance euclidienne entre $\overline{x^a}$ et $\overline{x^b}$ est :

$$\begin{aligned} d_a(\overline{x^a}, \overline{x^b}) &= \sqrt{\sum_{i=1}^{n} (\overline{x_i^a} - \overline{x_i^b})^2} \\ &= \sqrt{\sum_{i=1}^{n} (d_d(x^a, x^i) - d_d(x^b, x^i))^2} \end{aligned} \tag{4-5}$$

Or d'après l'équation 4-4, cette somme de différences est proche de 0, donc :

$$d_a(\overline{x^a}, \overline{x^b}) \simeq 0 \tag{4-6}$$

Donc $\overline{x^a}$ et $\overline{x^b}$ sont proches dans l'espace de voisinage. De la même manière, on peut montrer que deux points éloignés dans l'espace de voisinage, auront comme image deux points éloignés dans l'espace d'arrivée.

Nous allons vérifier ce constat à l'aide d'une méthode de projection de données. Nous utilisons la méthode $t - SNE$ pour visualiser que les relations de similarité entre les données sont conservées. La figure 4-3 représente la projection $t - SNE$ pour les éléments de la base de données synthétiques quand $SimDisc = 0{,}1$. Nous voyons que les éléments discriminants pour chaque classe sont regroupés.

La figure 4-4 quant à elle, présente la projection $t - SNE$ de ces mêmes données dans l'espace de voisinage. Nous notons bien que la représentation des données dans cet espace conserve les relations de proximité.

En conclusion, nous choisissons de généraliser cette approche, et donc d'utiliser le noyau K suivant pour la classification SVM :

$$K(x^a, x^b) = \exp^{-\frac{||\overline{x^a} - \overline{x^b}||^2}{\sigma}} \tag{4-7}$$

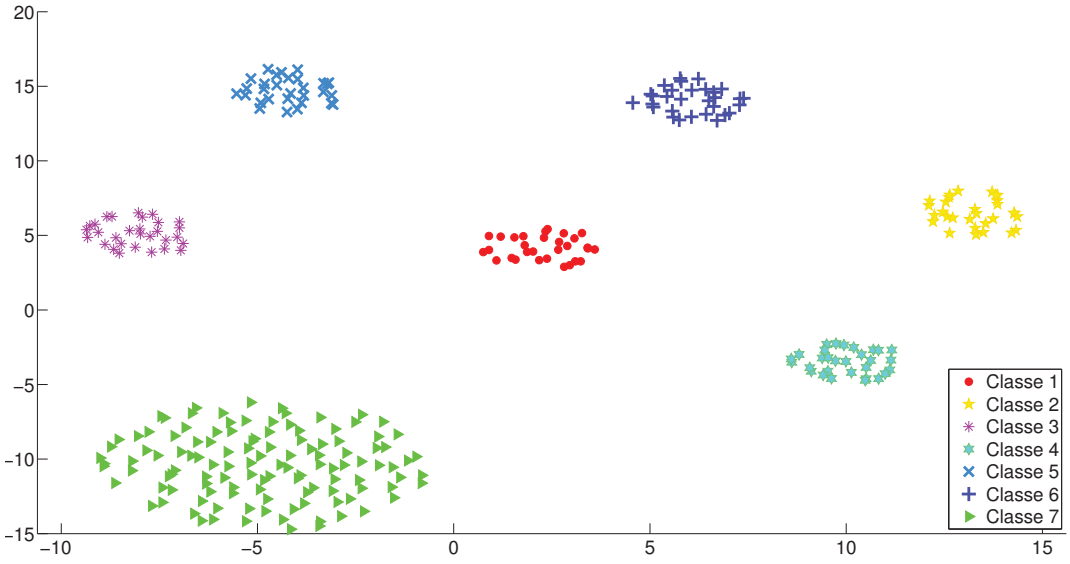


FIGURE 4-3 : La projection $t - SNE$ pour les données synthétiques avec $SimDisc = 0{,}1$.

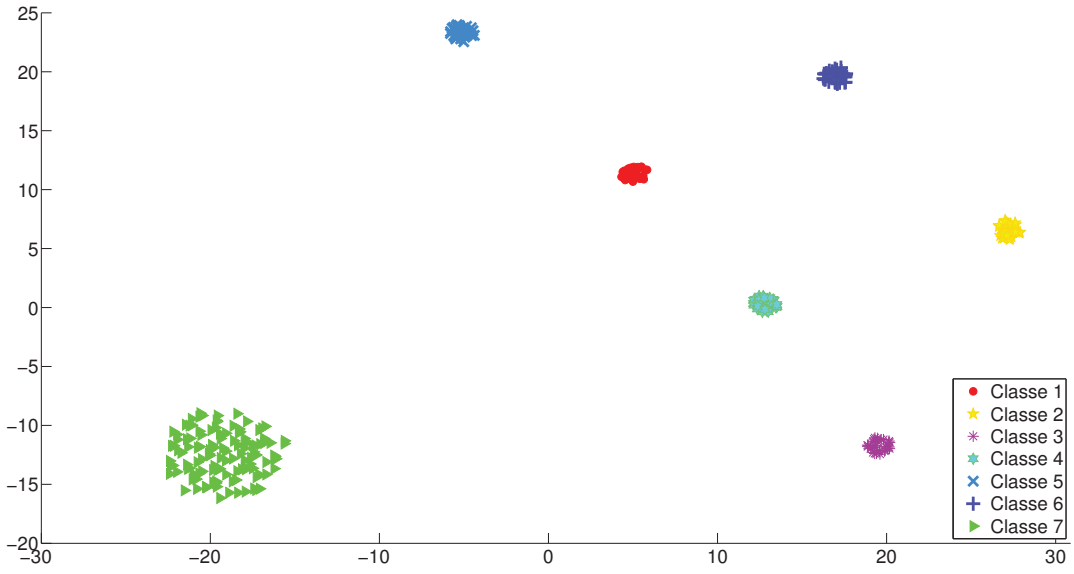


FIGURE 4-4 : La projection $t - SNE$ pour les données synthétiques avec $SimDisc = 0{,}1$ dans l'espace de voisinage.

4.3.3 Mesure de la performance

Nous avons vu dans 2.5.2 (page 56) que l'on peut évaluer les performances d'un classifieur à l'aide de plusieurs mesures, dont la mesure *Accuracy*. Nous avons à chaque fois montré que ces mesures fonctionnent pour des problèmes à deux classes et aucune information n'était précisée concernant la taille de chacune des classes.

Or dans notre cas, nous avons un problème avec 7 classes de tailles très différentes, l'une est même largement prédominante sur les 6 autres (près de 50% du nombre total d'éléments appartiennent à la première classe). De ce fait, lorsque nous testerons la performance de nos dissimilarités, les données de la première classe seront présentes en plus grand nombre. Et donc, le fait de bien classer un élément de cette classe augmentera d'autant la mesure globale de performance. À l'extrême, un classifieur, qui attribuerait la première classe à toute donnée de test, obtiendrait dans ce cas un bon taux de classification.

Il nous faut donc adapter la mesure de performances proposée. À notre connaissance, il n'y a pas dans la littérature de mesure reconnue pour ce type de problème (multi-classes avec déséquilibre entre classes). Nous proposons donc une nouvelle mesure, inspirée de la mesure *Accuracy*, qui prend en compte le résultat de classification de chacune des classes et qui en fait la moyenne.

$$\text{taux de reconnaissance} = \frac{1}{\text{Nb de classes}} \sum_i \frac{a_{ii}}{\sum_j a_{ij}} \quad ; \quad (i, j = 1, 2, \ldots, \text{Nb de classes}) \tag{4-8}$$

4.3.4 Résultats de référence pour l'évaluation

Afin d'évaluer nos propositions de dissimilarité entre traces, nous proposons de les comparer aux mesures obtenues par la méthode des motifs caractéristiques sur le même jeu de données synthétiques.

Bien entendu, le paramètre $SimDisc$ n'ayant pas de sens dans cette méthode, seuls les résultats de classification en fonction du paramètre $DiscNB$ sont présentés. La figure 4-5 présente les meilleurs résultats (lorsque la valeur de S_{min} vaut $0{,}1$). On remarque que c'est la courbe noire ($C_{min} = 0{,}1$) qui obtient globalement le meilleur résultat. C'est donc par rapport à cette courbe que l'on comparera nos propositions.

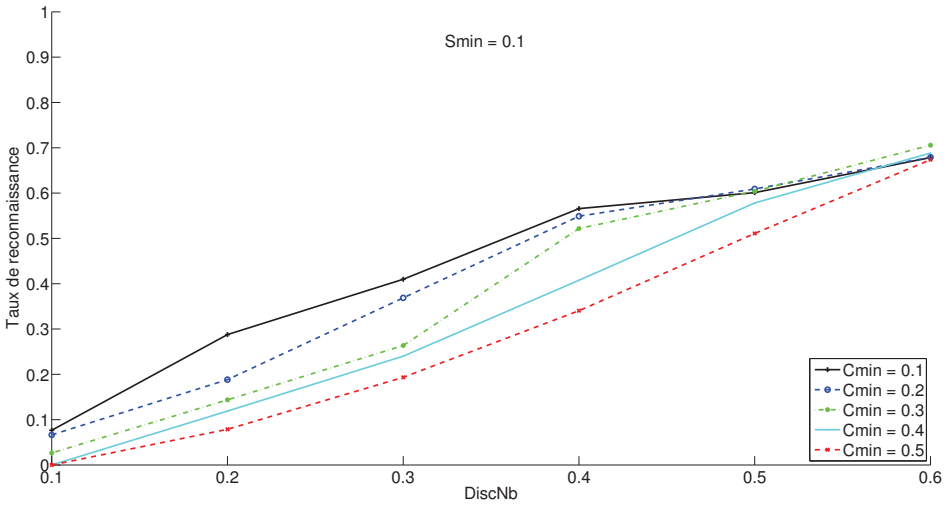


FIGURE 4-5 : Résultats des motifs caractéristiques sur les données synthétiques.

4.4 Dissimilarité comme moyenne des distances

Nous avons vu qu'une trace est un graphe chronologique. Nous pouvons alors utiliser des mesures de distance entre graphes, comme la mesure proposée par Suard et Rakotomamonjy [Sua 07]. Ils proposent de calculer la distance entre deux graphes G_a et G_b comme la moyenne des distances entre toutes les transitions de G_a et G_b. Ainsi, la formule pour calculer la distance entre deux graphes est :

$$D(G_a, G_b) = \frac{1}{|A||B|} \sum_{i:t_i \in A} \sum_{j:t_j \in B} D_{element}(t_i, t_j) \tag{4-9}$$

Le problème est que cette proposition, dans certains cas, ne satisfait pas la propriété que l'on s'est fixé. Par exemple soit $T_1 = \{1, 2, 3, 4, 5\}$, $T_2 = \{1, 2\}$ et $T_3 = \{7, 8, 9\}$ trois traces, tel que :

- les dissimilarités entre les éléments 1 et 2 vis à vis des éléments 3,4 et 5 soient de 1
- les dissimilarités entre les éléments 1 et 2 vis à vis des éléments 7, 8, 9 soient de 0,5.

D'après la propriété, que l'on s'est fixée dans la partie 4.3, nous voulons que la dissimilarité entre T_1 et T_2 soit plus petite que la dissimilarité entre T_2 et T_3, et ce quelque soient les dissimilarités entre élément sauf si 1 et 2 sont très proches de 7, 8 ou 9 (ce qui n'est pas le cas ici). Or dans notre exemple, la dissimilarité entre T_1 et T_2 est de 0,6 et la dissimilarité entre T_2 et T_3 est de 0,5.

Tout en s'inspirant de cette dernière, nous proposons une autre dissimilarité entre traces définie de la manière suivante :

- Soit A l'ensemble des éléments (documents ou transitions) de l'utilisateur U_a, B celui de l'utilisateur U_b (*Cf.* la figure 4-6),
- Soit $C = A \bigcap B$,
- Soit $D = A \setminus C$,
- Soit $E = B \setminus C$.

Nous définissons la dissimilarité $D_{trace\_1}(U_a, U_b)$ entre traces U_a et U_b comme :

$$\begin{aligned} D_{trace\_1}(U_a, U_a) = & \frac{|C||D|}{fac} \sum_{i:t_i \in C} \sum_{j:t_j \in D} D_{element}(t_i, t_j) + \\ & \frac{|C||E|}{fac} \sum_{i:t_i \in C} \sum_{j:t_j \in E} D_{element}(t_i, t_j) + \\ & \frac{|D||E|}{fac} \sum_{i:t_i \in D} \sum_{j:t_j \in E} D_{element}(t_i, t_j) \end{aligned} \tag{4-10}$$

où $fac = (|C||D| + |C||E| + |D||E|)$.

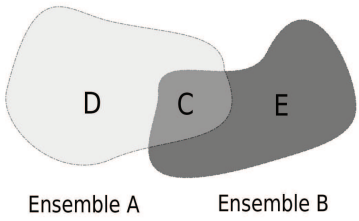


FIGURE 4-6 : Les ensembles A et B.

Cette dissimilarité valide le critère précédent. En effet, si $U_a = U_b$, alors D et E sont des ensembles vides ($|D| = |E| = 0$) et $D_{trace\_1}(U_a, U_b) = 0$. Si nous calculons la dissimilarité entre une trace et sa sous-trace, E sera l'ensemble vide, et donc seule la première partie de l'équation 4-10 n'est pas nulle. Enfin, pour deux traces différentes, si elles ne partagent pas d'éléments, nous revenons à l'équation 4-9, sinon, nous appliquons l'équation 4-10 avec ses trois parties.

Nous allons maintenant évaluer la qualité de cette dissimilarité en fonction des deux paramètres ($DiscNB$ et $SimDisc$) à l'aide de l'algorithme de projection de données t-SNE puis à l'aide de l'algorithme de classification SVM.

4.4.1 Validation visuelle

Pour étudier l'influence de $DiscNB$ et $SimDisc$, nous appliquons le protocole suivant :

- pour chaque base de données synthétiques, nous calculons la matrice de similarité entre les traces,
- puis, nous appliquons l'algorithme du $t - SNE$ sur ces matrices en prenant des valeurs différentes pour $DiscNB$ et $SimDisc$.

Les figures 4-7 et 4-8 représentent les résultats $t - SNE$ pour $DiscNB = 0{,}2|0{,}4$ et $SimDisc = 0{,}2$. Nous voyons que plus la valeur de $DiscNB$ est grande, meilleur est la projection $t - SNE$.

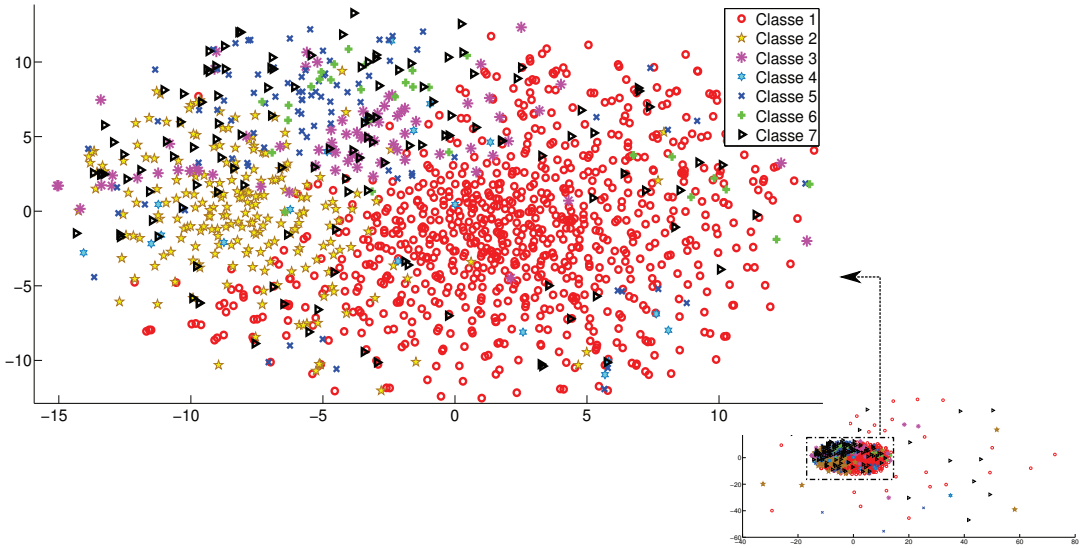


FIGURE 4-7 : La projection $t - SNE$ pour $DiscNB = 0{,}2$ et $SimDisc = 0{,}2$.

Maintenant, nous prenons différentes valeurs de $SimDisc$ pour $DiscNB = 0{,}2$. La figure 4-9 représente la projection $t - SNE$ pour $DiscNB = 0{,}2$ et $SimDisc = 0{,}4$.

Dans cette figure l'influence de $SimDisc$ sur la projection $t - SNE$ n'est pas très probante. Ces résultats nous amènent donc à étudier plus profondément la relation entre le taux de reconnaissance et les différentes valeurs de $DiscNB$ et $SimDisc$. Nous allons donc utiliser l'algorithme de classification SVM.

4.4.2 Validation statistique

Nous utilisons l'algorithme du SVM pour étudier le rôle de $DiscNB$ et $SimDisc$ dans l'identification du profil. Lors de la phase d'apprentissage du SVM, nous multiplions la valeur

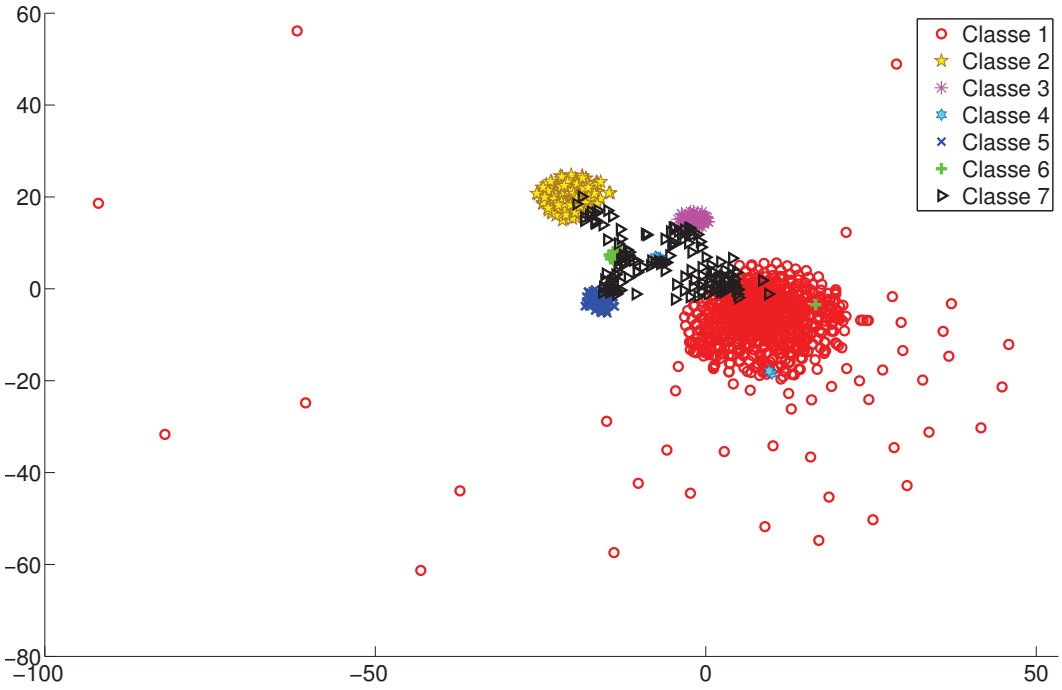


FIGURE 4-8 : La projection $t - SNE$ pour $DiscNB = 0{,}4$ et $SimDisc = 0{,}2$.

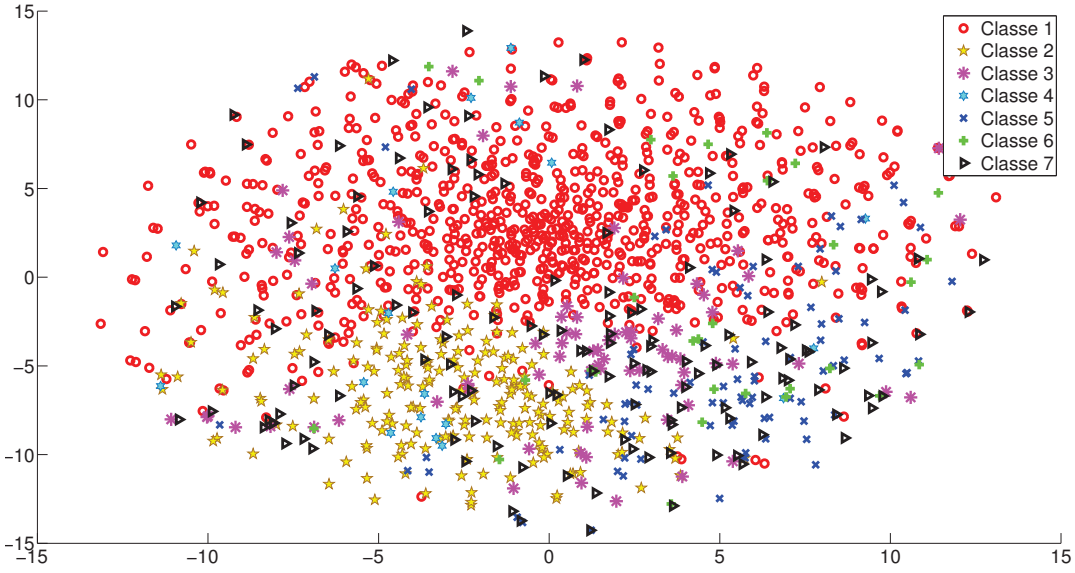


FIGURE 4-9 : La projection $t - SNE$ pour $DiscNB = 0{,}2$ et $SimDisc = 0{,}4$.

de coût C (le coût d'erreur du classifieur) par un cofficient co_i pour chaque profil i.

$$co_i = \frac{\text{la taille du profil le plus grand}}{\text{la taille du profil}_i} \tag{4-11}$$

Donc, une erreur de positionnement, par rapport à l'hyperplan, pour une donnée appartenant à un profil de petite taille aura un coût plus important que pour une donnée d'un profil de grande taille.

Pour rappel, pour chaque base de données synthétiques, nous prenons 60% des données pour la phase d'apprentissage et le reste pour la phase de test. $DiscNB$ varie de $0{,}1$ à $0{,}6$ par pas de $0{,}1$. Nous faisons la même chose pour $SimDisc$.

La figure 4-10 représente la matrice de Gram pour $DiscNB = 0{,}4$, $SimDisc = 0{,}4$ et $\sigma = 0{,}1$. Dans cette matrice, nous pouvons bien distinguer les sept classes et donc *a priori* nous devrions obtenir de bon résultat avec le SVM.

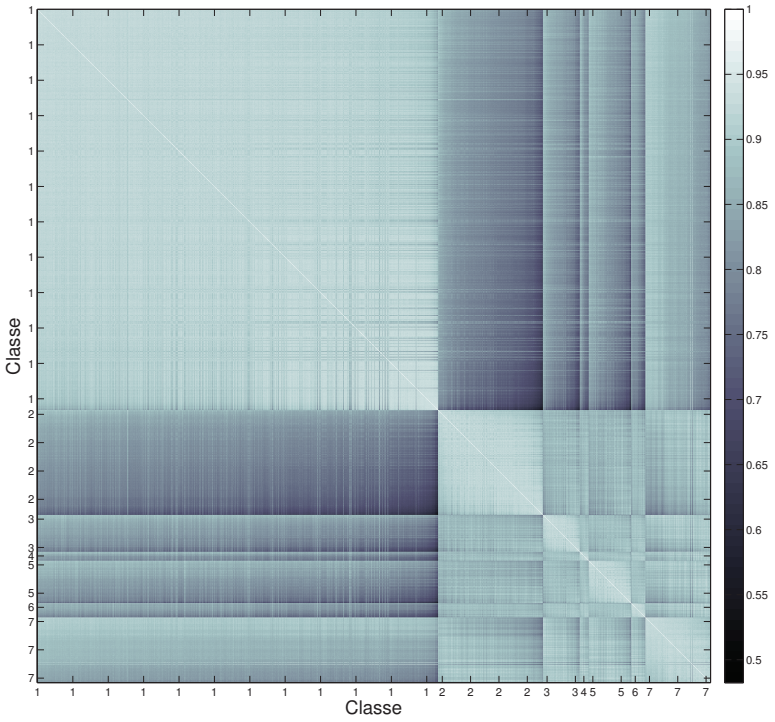


FIGURE 4-10 : La matrice de Gram du noyau gaussien, $DiscNB = 0{,}4$, $SimDisc = 0{,}4$.

Pour chaque base de données synthétiques, une matrice de confusion est générée pour étudier la qualité de classification. Pour rappel, la matrice de confusion (*Cf.* 2.5.2, page 56) est représentée par un tableau 4-2, où a_{ij} est le nombre de traces qui appartiennent au profil i et classées comme appartenant au profil j.

		Décision		
		1	i	k
Profil	1	a_{11}	a_{1i}	a_{1k}
	i	a_{i1}	a_{ii}	a_{ik}
	k	a_{k1}	a_{ki}	a_{kk}

TABLEAU 4-2 : Un exemple de la matrice de confusion.

Le tableau 4-3 présente la matrice de confusion en appliquant les SVM avec $DiscNB = 0{,}2$ et $SimDisc = 0{,}4$.

		Décision						
		1	2	3	4	5	6	7
Profil	1	438	43	14	25	14	16	18
	2	3	125	6	7	6	1	4
	3	6	1	32	2	4	3	4
	4	1	1	2	8	1	3	0
	5	12	7	3	0	38	3	0
	6	3	2	2	2	2	11	1
	7	17	10	12	14	13	19	11

TABLEAU 4-3 : La matrice de confusion avec $DiscNB = 0{,}2$ et $SimDisc = 0{,}4$.

Notons bien que dans ce tableau, nous obtenons de bons résultats de classification pour les classes 1, 2 (77%, 82%). Mais, nous pouvons identifier seulement 47% (respectivement 11%) des profils 6 (respectivement profil 7). Ces résultats sont compatibles avec la projection obtenue par le $t - SNE$ dans la figure 4-9, car les profils 6 et 7 dans cette projection sont très mélangés avec les autres profils.

La figure 4-11 représente les résultats de l'utilisation du SVM pour plusieurs valeurs de $DiscNB$ et $SimDisc$. Chaque ligne représente les résultats de la classification pour une valeur de $SimDisc$ fixe en faisant varier $DiscNB$.

Dans la figure 4-12, les lignes représentent les résultats de classification pour une valeur de $DiscNB$ fixe en faisant varier $SimDisc$. La variation du taux de reconnaissance est faible lorsque l'on modifie les valeurs de $SimDisc$. En plus, il y a une forte influence de la valeur de $DiscNB$ sur ce taux de reconnaissance. Alors que dans la figure 4-11, les lignes qui représentent les $SimDisc$ sont proches. C'est-à-dire que l'influence de la $SimDisc$ sur le taux de reconnaissance est faible.

Pour conclure sur cette proposition, nous pouvons remarquer que les résultats de classification sont supérieurs aux résultats de référence (celui des motifs caractéristiques). Quelque soit la valeur de $DiscNB$, même avec un $SimDisc$ de 0,1, la qualité de la classification est supé-

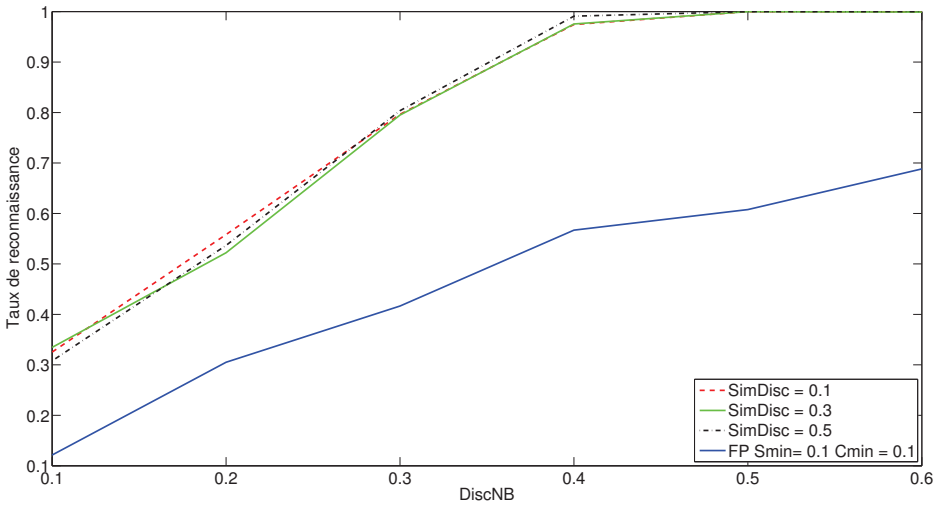


FIGURE 4-11 : Les résultats de l'application du SVM sur les données synthétiques en faisant varier $DiscNB$.

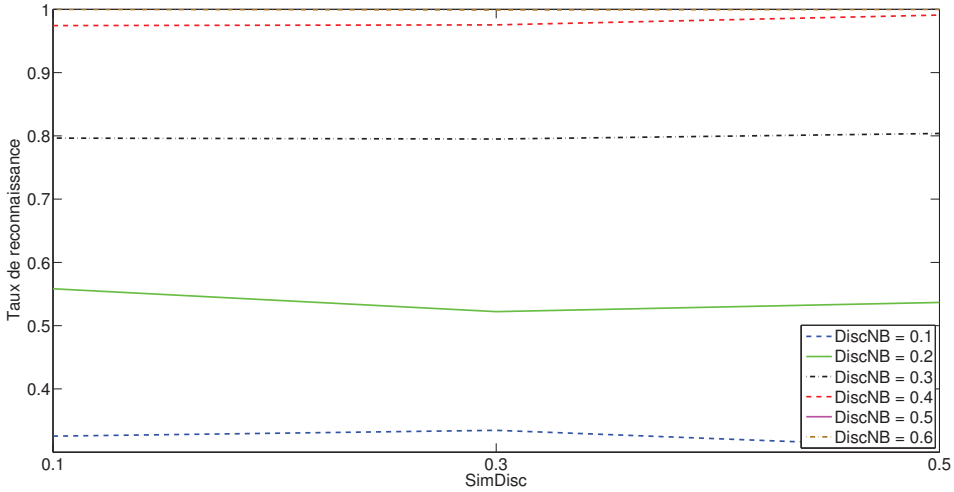


FIGURE 4-12 : Les résultats d'application du SVM sur les données synthétiques en faisant varier $SimDisc$.

rieur. Ainsi pour un $DiscNB$ de $0{,}2$ la classification est supérieure de 20%. Pour une valeur de $0{,}3$ elle est de plus 35%. Enfin, pour $0{,}4$ elle est de plus de 40%.

Par contre nous pouvons aussi remarquer que l'augmentation de la valeur du paramètre $SimDisc$ influence beaucoup moins la qualité de la classification que l'augmentation du paramètre $DiscNB$. Nous avons pour des valeurs de $DiscNB$ faibles, une augmentation de 3% de la qualité de la classification entre les valeurs du paramètre $SimDisc$ la plus faible et la plus forte. Ceci nous amène à étudier une autre dissimilarité.

4.5 Dissimilarité comme extension des vecteurs caractéristiques

Cette approche s'inspire de la dissimilarité entre les cartes conceptuelles proposée par Delestre et Malandain [Del 07]. En effet, ils proposent une dissimilarité entre ensembles possédant des éléments sur lesquels il existe aussi une dissimilarité (ce qui est notre cas). Ils sont parties de la proposition de Besson dans [Bes 73], qui propose une distance entre deux ensembles A et B définie par l'équation 4-12.

$$d_2(A,B) = \frac{(A \setminus B) \cup (B \setminus A)}{|A \cup B|} \tag{4-12}$$

L'idée de Delestre et Malandain est alors de prendre en compte, dans le numérateur, la similarité entre éléments des ensembles A et B. Ce calcul de numérateur est formalisé par l'algorithme 9. Dans leur contexte, la dissimilarité entre éléments est normalisée (comprise entre 0 et 1), et dans le cas extrême où les éléments des deux ensembles sont totalement différents deux à deux (dissimilarités égalent à 1) ils obtiennent une dissimilarité équivalente à celle proposée par Besson.

Dans notre contexte, pour calculer la dissimilarité entre deux traces d'utilisateur U_a et U_b avec A l'ensemble des éléments (documents ou transitions) de l'utilisateur U_a, B celui de l'utilisateur U_b, nous proposons la formule suivante :

$$D_{trace\_2}(U_a,U_b) = \frac{D_{ensemble}(A,B)}{|A \cup B|} \tag{4-13}$$

Le problème de cet algorithme est la grande influence des différences de longueur entre les deux traces dans le calcul de cette dissimilarité. Afin d'être plus clair, prenons un exemple. Soit trois traces représentées par trois ensembles $A = \{b, e\}$, $B = \{a, d\}$ et $C = \{a, b, d\}$. La figure 4-13 présente entre autres ces trois traces, tel que la trace A est représentée par une ligne verte en tirets courts et longs, la trace B par une ligne violette avec des grands tirets longs et la trace C par une polyligne rouge en trait plein.

Algorithme 9 Algorithme de $D_{ensemble}(A, B)$

Entrée: Deux ensembles A et B, $D_{element}$ une dissimilarité entre éléments des ensembles A et B
Sortie: d_{AB}
début
  $C \leftarrow A \cap B$
  $D \leftarrow A \setminus B$
  $E \leftarrow B \setminus A$
  si $D = \phi$ **alors**
    $d_{AB} \leftarrow |E|$
  sinon
    si $E = \phi$ **alors**
      $d_{AB} \leftarrow |D|$
    sinon
      $(e, d) \leftarrow \arg\min_{e \in E, d \in D} D_{element}(e, d)$
      $d_{AB} \leftarrow 2 * D_{element}(e, d) + D_{ensemble}(E \setminus \{e\}, D \setminus \{d\})$
    finsi
  finsi
fin

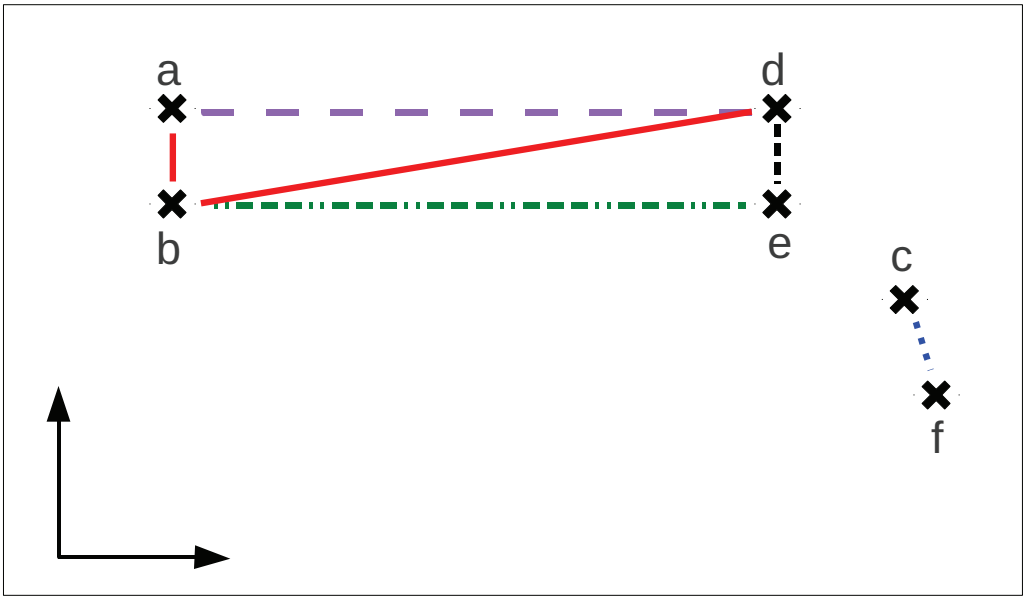


FIGURE 4-13 : Exemple des traces.

La dissimilarité entre les éléments de ces traces est donnée par leur distance euclidienne entre eux dans le plan. L'ensemble de ces distances est donné par le tableau 4-4.

	a	b	c	d	e	f
a	0	0,1	0,85	0,6	0,7	0,95
b	0,1	0	0,85	0,7	0,8	0,9
c	0,85	0,85	0	0,3	0,2	0,1
d	0,6	0,7	0,3	0	0,1	0,5
e	0,7	0,8	0,2	0,1	0	0,3
f	0,95	0,9	0,1	0,5	0,3	0

TABLEAU 4-4 : Dissimilarité entre les éléments des ensembles des traces de la figure 4-13.

Les traces A, B et C se ressemblent. Les trois dissimilarités entre elles devraient donc être faibles et se rassembler. Qui plus est, la trace C a des points en commun avec les traces A et B (ce qui n'est pas le cas pour les traces A et B entre elles). Dès lors il faudrait que les dissimilarités entre la trace C et les traces A et B soient plus petites que la dissimilarité entre les traces A et B. Or avec la précédente dissimilarité, nous obtenons $D_{trace\_2}(U_a, U_b) = 0{,}1$, $D_{trace\_2}(U_a, U_c) = 0{,}3$ et $D_{trace\_2}(U_b, U_c) = 0{,}33$. Ceci ne correspond pas à nos attentes.

Comme nous le disions précédemment, ceci est dû au fait que la différence de taille des deux traces a une grande influence. En effet, dans l'algorithme 9, la dissimilarité entre chaque élément restant dans la trace la plus longue, après avoir enlevé les éléments les plus proches, est égale à 1, même si ces éléments supplémentaires sont proches des éléments de l'autre trace. Donc, pour améliorer cet algorithme, nous proposons de calculer la dissimilarité entre ces éléments restant et les éléments de l'autre trace qui leurs sont plus proches. Cette amélioration est formalisée par l'algorithme 10. Cet algorithme prend deux paramètres supplémentaires que sont A_{ref} et B_{ref} qui correspondent aux deux ensembles initiaux, alors que A et B peuvent être privés d'un élément à chaque appel récursif.

Finalement, comme précédemment afin de normaliser cette dissimilarité, nous divisons $D_{ensemble}(A, B, A, B)$ par la taille de la trace la plus longue, On obtient alors la dissimilarité D_{trace_am} dont la formule est proposée par l'équation 4-14.

$$D_{trace\_am}(U_a, U_b) = \frac{D_{ensemble}(A, B, A, B)}{\max(|A|, |B|)} \tag{4-14}$$

Pour les mêmes ensembles A, B et C de l'exemple précédent, nous obtenons alors les dissimilarités $D_{trace\_am}(U_a, U_b) = 0{,}1$, $D_{trace\_am}(U_a, U_c) = 0{,}06$ et $D_{trace\_am}(U_b, U_c) = 0{,}03$. Ces résultats sont alors compatibles avec nos attentes.

Notons toutefois que l'algorithme 10 fonctionne bien malgré le fait que dans quelque cas

Algorithme 10 Algorithme amélioré de $D_{ensemble}(A, B, A_{ref}, B_{ref})$

Entrée: Quatre ensembles A,B,A_{ref} et B_{ref} , $D_{element}$ une dissimilarité entre éléments des ensembles
Sortie: d_{AB}
début
  si $A = \phi$ et $B = \phi$ **alors**
    $d_{AB} \leftarrow 0$
  sinon
    si $A = \phi$ et $B \neq \phi$ **alors**
      $a, b \leftarrow \underset{\substack{a \in A_{ref} \\ b \in B}}{\text{argmin}} D_{element}(a, b)$
      $d_{AB} \leftarrow D_{element}(a, b) + D_{ensemble}(A, B \setminus \{b\}, A_{ref}, B_{ref})$
    sinon
      si $A \neq \phi$ et $B = \phi$ **alors**
        $a, b \leftarrow \arg\min_{\substack{a \in A \\ b \in B_{ref}}} D_{element}(a, b)$
        $d_{AB} \leftarrow D_{element}(a, b) + D_{ensemble}(A \setminus \{a\}, B, A_{ref}, B_{ref})$
      sinon
        $a, b \leftarrow \arg\min_{\substack{a \in A \\ b \in B}} D_{element}(a, b)$
        $d_{AB} \leftarrow D_{element}(a, b) + D_{ensemble}(A \setminus \{a\}, B \setminus \{b\}, A_{ref}, B_{ref})$
      finsi
    finsi
  finsi
fin

particuliers il ne donne pas la dissimilarité optimale. Par exemple, soient les traces $D = \{e, d\}$, $E = \{c, f\}$ proposées aussi dans la figure 4-13.

Selon la formule 4-14, la dissimilarité entre ces deux traces est :

$$D_{trace\_am}(U_e, U_f) = \frac{D_{element}(e, c) + D_{element}(d, f)}{2} = \frac{0{,}2 + 0{,}5}{2} = 0{,}35$$

Alors que la dissimilarité optimale est :

$$\frac{D_{element}(e, f) + D_{element}(d, c)}{2} = \frac{0{,}3 + 0{,}3}{2} = 0{,}3$$

Ceci est dû au fait que nous choisissons un compromis entre la performance de l'algorithme et la qualité du résultat obtenu. En effet, rechercher la dissimilarité optimale est en complexité en temps plus grande que celle de notre algorithme.

4.5.1 Validation visuelle

Tout comme la proposition de dissimilarité précédente, nous avons effectué une projection $t - SNE$ de nos données synthétiques en faisant varier les deux paramètres $DiscNB$ et $SimDisc$. Les figures 4-14, 4-15 et 4-16 présentent trois exemples de paramétrage.

Tout comme avec la première proposition lorsque $DiscNB$ vaut $0{,}2$ et $SimDisc$ vaut $0{,}2$ (*Cf.* figure 4-14), les classes 1 et 2 sont homogènes. Par contre, la classe 3 est un peu plus dispersée par rapport à la première proposition. Et les autres classes sont totalement dispersées. Donc a priori, dans ce contexte, les résultats semblent moins bon.

Par contre, lorsque $DiscNB$ vaut $0{,}4$ et $SimDisc$ vaut $0{,}2$, le résultat de cette projection semble permettre de mieux discriminer les classes, tout particulièrement la septième.

Enfin, la figure 4-16 qui montre la projection des données lorsque que $DiscNB$ vaut $0{,}2$ et $SimDisc$ vaut $0{,}4$, à l'exception de la classe 7, semble de meilleure qualité comparée à la première proposition de dissimilarité avec ces même paramètres. En effet, les classes 1 et 2 sont bien identifiables. Et les autres classes forment des petits clusters, ce qui n'était pas le cas précédemment.

Donc *a priori*, plus l'un des deux paramètres augmente, meilleur devrait être la classification. Nous allons maintenant le vérifier de manière statistique.

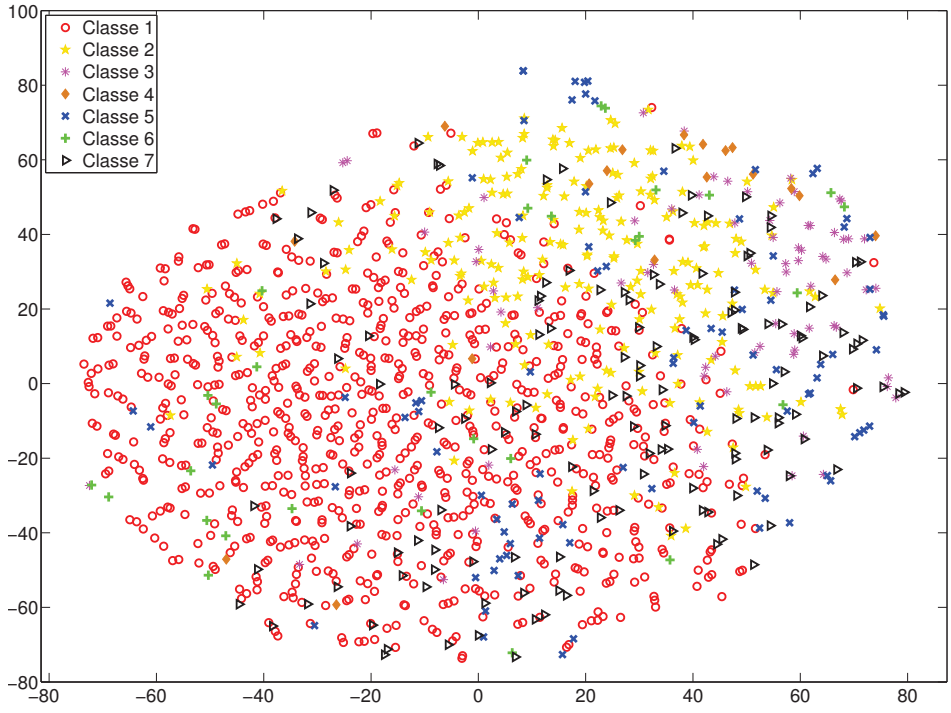


FIGURE 4-14 : La projection $t - SNE$ pour $DiscNB = 0{,}2$ et $SimDisc = 0{,}2$.

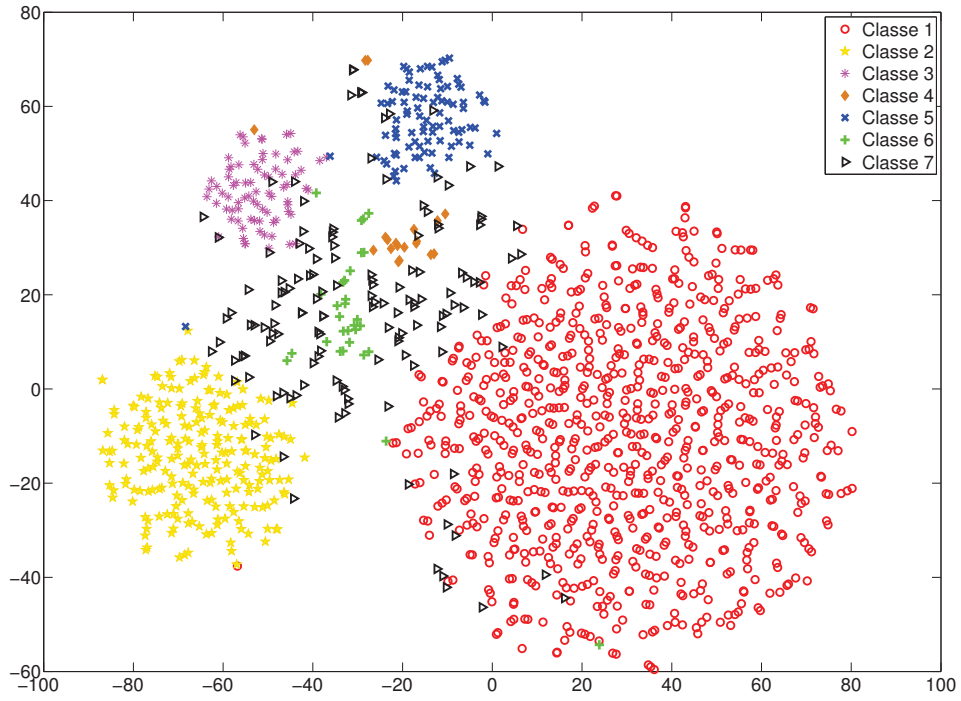


FIGURE 4-15 : La projection $t - SNE$ pour $DiscNB = 0{,}4$ et $SimDisc = 0{,}2$.

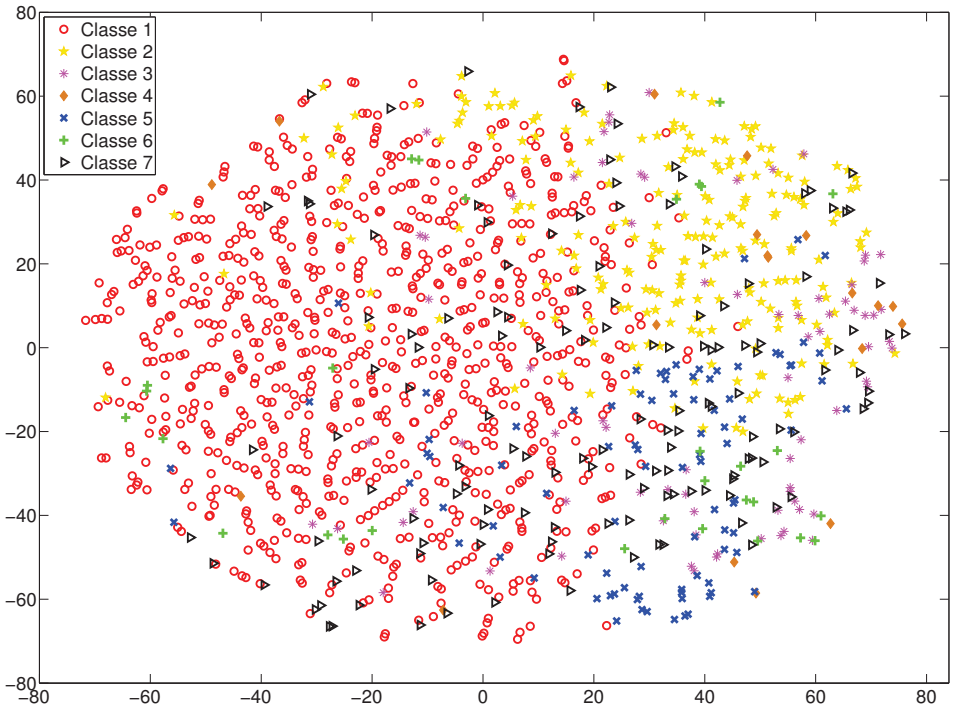


FIGURE 4-16 : La projection $t - SNE$ pour $DiscNB = 0{,}2$ et $SimDisc = 0{,}4$.

4.5.2 Validation statistique

Les résultats de l'utilisation du SVM avec cette dissimilarité sur les données synthétiques sont représentés dans les figures 4-17 et 4-18.

Les résultats de cette approche sont supérieurs à ceux de la première approche, surtout pour les petites valeurs de $DiscNB$. Nous voyons aussi que l'augmentation de la valeur de $SimDisc$ permet d'améliorer légèrement la qualité de classification pour des valeurs faibles de $DiscNB$. Par exemple pour $DiscNB = 0{,}2$, le taux de reconnaissance est supérieur de 5% quand la valeur de $SimDisc$ passe de $0{,}3$ à $0{,}5$.

4.6 Conclusion

Dans ce chapitre nous sommes partis du constat que des utilisateurs du site Hypergéo pouvaient avoir des intérêts communs même s'ils ne visitaient pas des documents identiques. Nous avons alors proposé d'utiliser des algorithmes de classification à noyau en lieu et place de l'algorithme des motifs caractéristiques. Nous avons proposé deux mesures de dissimilarité entre traces. Nous les avons testées à l'aide de données synthétiques en les validant visuellement

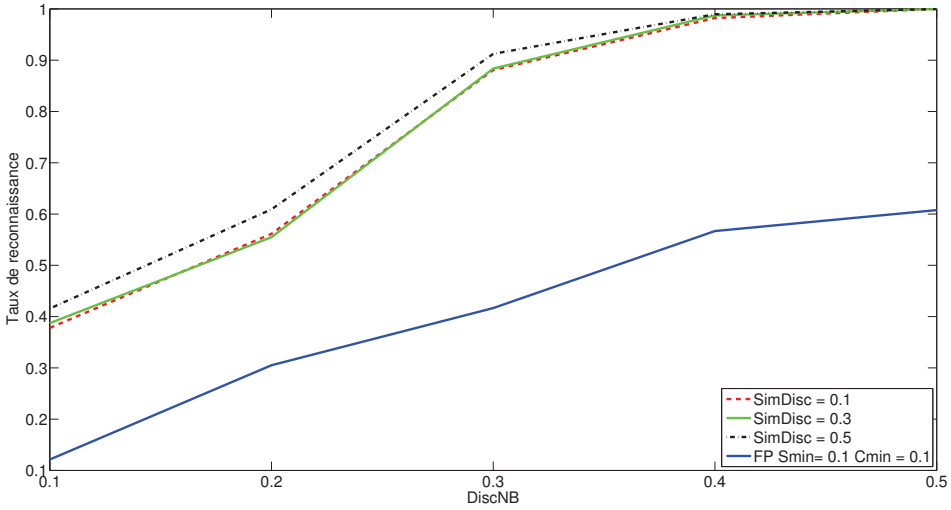


FIGURE 4-17 : Les résultats de l'application du SVM avec l'approche de D_{trace_am} sur des données synthétiques en fonction de $DiscNB$.

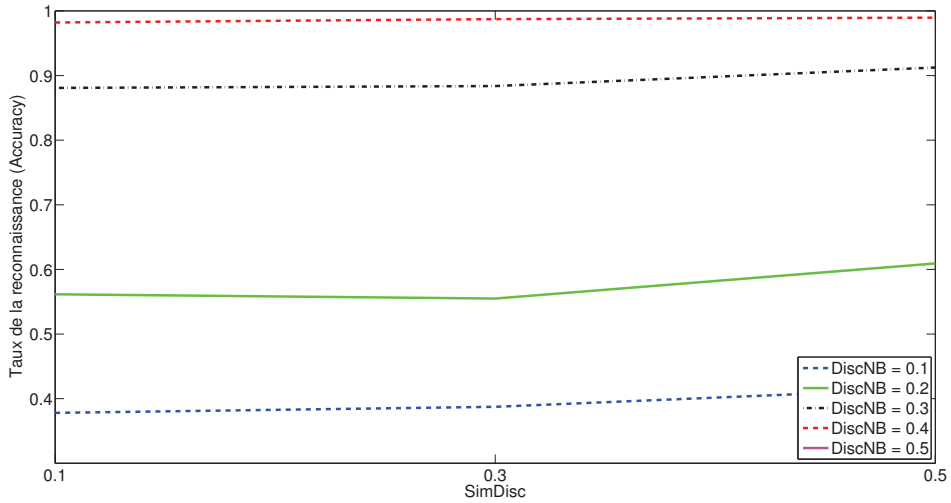


FIGURE 4-18 : Les résultats de l'application du SVM avec l'approche de D_{trace_am} sur des données synthétiques en fonction de $SimDisc$.

à l'aide du $t - SNE$ et statistiquement à l'aide la classification SVM. Ces données synthétiques ont été générées en faisant varier deux paramètres $DiscNB$ et $SimDisc$. Dans le cas de la validation statistique, nous avons confronté ces résultats aux résultats obtenus par l'utilisation des motifs caractéristiques sur ces mêmes données synthétiques. Au vue des résultats, nos propositions donnent des résultats supérieurs à ceux obtenus avec l'algorithme des motifs caractéristiques.

Il nous faut donc maintenant déterminer ce que représente les éléments de nos traces (les documents visités ou les transitions effectuées). Il faudra alors définir une dissimilarité entre ces éléments, et valider l'ensemble sur nos données réelles.

CHAPITRE 5

Dissimilarités entre éléments pour les traces d'Hypergéo

Sommaire

5.1 Introduction

Nous avons proposé deux mesures pour calculer la similarité entre les traces des utilisateurs du site Hypergéo, au sein desquelles, nous avons utilisé la dissimilarité entre les éléments des traces. Dans le contexte d'Hypergéo, les éléments de trace sont des documents ou des transitions. Nous avons défini la transition comme le fait d'aller d'un document à l'autre. Une dissimilarité entre deux transitions peut alors être calculée comme la moyenne des dissimilarités D_{doc} entre les documents de départs doc_{start} et les documents d'arrivés doc_{end}, c'est-à-dire :

$$D_{trans}(t_1, t_2) = \frac{D_{doc}(doc_{start_1}, doc_{start_2}) + D_{doc}(doc_{end_1}, doc_{end_2})}{2} \quad (5\text{-}1)$$

Quelque soit le point vue abordé, il faut définir une distance entre documents. Il est à noter que les traces que nous avons récupérées sont de longueur assez modeste (au maximum une dizaine de transitions). De plus, comme nous allons le voir, nous sommes obligés dans certains cas de leur appliquer des filtres, qui les raccourcissent d'autant. Il est donc préférable d'utiliser la représentation qui maximise leurs caractéristiques communes. Ainsi nous avons donc décidé de considérer qu'une trace est un ensemble de documents (les éléments d'une trace sont donc des documents du site Hypergéo). Notre problème revient alors à trouver une dissimilarité entre les documents du site Hypergéo.

Dans ce chapitre, nous allons présenter deux mesures des dissimilarité entre les documents d'Hypergéo (la dissimilarité hiérarchique et la dissimilarité *tf-idf*). Pour chaque dissimilarité, nous calculerons une projection *t-SNE* des documents, une projection *t-SNE* des traces et nous appliquerons une classification à l'aide du SVM, que nous comparerons aux résultats obtenus dans la chapitre 3. Lors de cette classification, nous utiliserons soit les classes initiales, soit deux regroupements de trois classes, comme précédemment l'un axé sur le domaine de compétence des utilisateurs (avec les profils "Spécialiste en géographie" *G*, "Spécialiste dans autre domaine" *AD* et *Autre*) et en plus l'autre axé sur le niveau de connaissance ("Étudiant" *Et*, "Enseignant" *En* et *Autre*).

5.2 Première proposition : une dissimilarité hiérarchique

Le site hypergéo est une encyclopédie sur la géographie. Les experts du domaine, concepteurs du site, l'ont donc organisé en thèmes et sous-thèmes. Ils ont créé une organisation hiérarchique de l'information. Des utilisateurs intéressés par la même problématique devraient alors visiter les mêmes thèmes et donc les mêmes parties du site web.

Après avoir étudié plus en détail cette organisation, nous proposerons une distance entre

éléments des traces basée sur cette organisation.

5.2.1 Organisation hiérarchique d'Hypergéo

De manière générale, les sites web peuvent être représentés à l'aide d'un graphe orienté, où les nœuds du graphes sont les pages HTML et les arcs représentent les liens hypertextes. Le site hypergéo est une encyclopédie sur la géographie, sa structure est donc plus proche d'un arbre. Les nœuds de cet arbre sont les documents du site (les articles ou les rubriques). Les arcs de l'arbre sont les liens hypertextes entre rubriques et leurs sous-rubriques (et inversement), ainsi qu'entre rubriques et leurs articles (et inversement). Ceci toutefois est une simplification de la réalité, car il existe quelques liens transversaux entre articles et articles ou rubriques, comme par exemple les liens de la partie « Mots-clés », ou les liens au sein des articles générés dynamiquement (*Cf.* figure 3-7 page 82). La figure 5-1 présente une partie de cet arbre. Les nœuds sont représentés par des ovales pour les articles et des losanges pour les rubriques. Les arcs orientés noirs représentent les liens hypertextes hiérarchiques que nous utilisons, et les arcs bleus en pointillé représentent des liens non hiérarchiques et donc non pris en compte dans nos travaux.

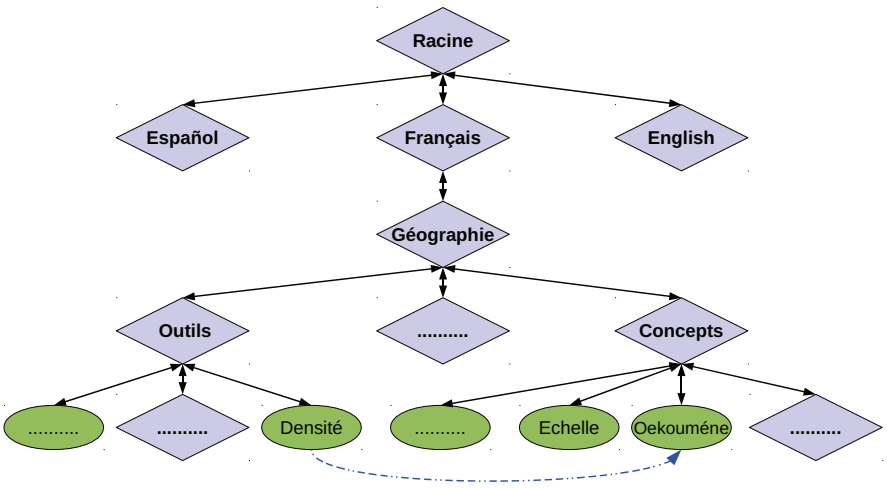


FIGURE 5-1 : Extrait de l'organisation du site Hypergéo.

À partir de cette représentation, il est possible de définir une dissimilarité hiérarchique entre documents (rubriques ou articles) du site Hypergéo. Elle peut être définie comme la normalisation de la distance entre les nœuds de l'arbre définie comme étant le nombre d'arcs qui relient deux nœuds.

La figure 5-2 représente la projection *t-SNE* des documents d'Hypergéo en appliquant cette dissimilarité. Les documents d'Hypergéo sont bien identifiés selon leurs langues. Dans ces figures, les documents sont composés des petits groupes selon la dissimilarité entre eux. Quelquefois, ils sont regroupés en cercles concentriques. Les cercles représentent les rubriques et ses sous rubriques et ensuite leurs articles.

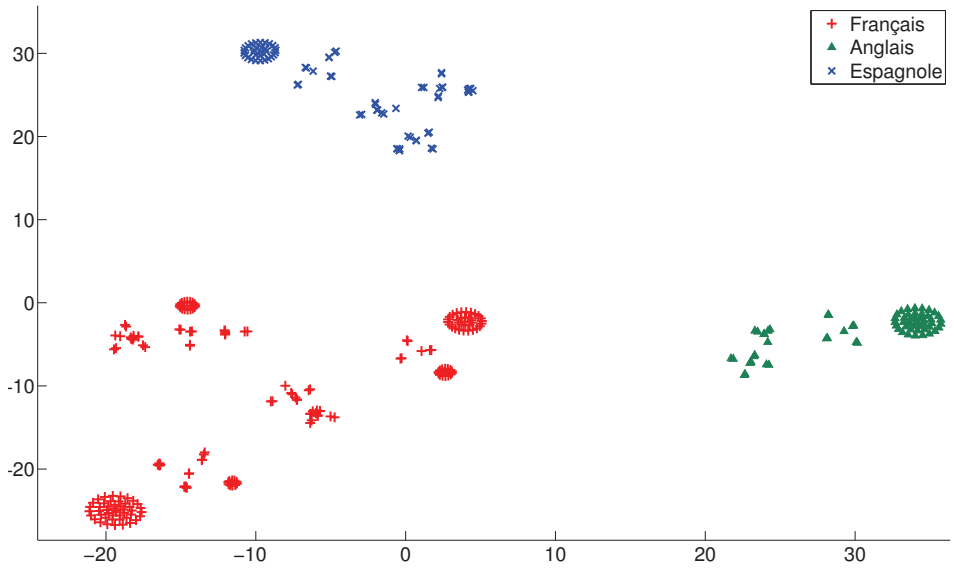


FIGURE 5-2 : La projection *t-SNE* des documents d'Hypergéo en appliquant la dissimilarité hiérarchique.

Nous allons appliquer la dissimilarité hiérarchique entre les éléments.

5.2.2 Application de la dissimilarité hiérarchique aux données de Hypergéo

Comme pour le chapitre 3, nous allons appliquer cette proposition aux utilisateurs francophones. L'évaluation des performances de cette dissimilarité sera tout d'abord visuelle puis statistique.

5.2.2.1 Résultats visuels en utilisant *t-SNE*

La figure 5-3 représente la projection de *t-SNE* des utilisateurs pour les sept profils initiaux. Aucun profil ne se démarque sur cette projection. Il en est de même sur les figures 5-4 et 5-5 qui représentent les projections des ces traces pour respectivement les profils identifiant le domaine de compétences et le niveau de connaissance. Nous pouvons donc émettre l'hypothèse que les utilisateurs, quelque soit leur profil, ne se focalisent pas sur une partie du site.

De plus, aucun *cluster* ne se détache réellement dans cette projection. Ce qui signifie que les traces sont toutes plus ou moins à la même distance les unes des autres. Cela est dû au fait que la façon de calculer cette dissimilarité discrétise fortement l'intervalle $[0..1]$. En effet pour normaliser la dissimilarité, nous divisons la distance entre deux documents par la dissimilarité entre les deux documents les plus éloignés d'Hypergéo, c'est-à-dire 10. De plus le numérateur prend des valeurs de 0 jusqu'à 10 par pas de 1. Dès lors, il existe uniquement dix valeurs de dissimilarité possibles.

Nous allons confirmer cela à l'aide d'une évaluation statistique.

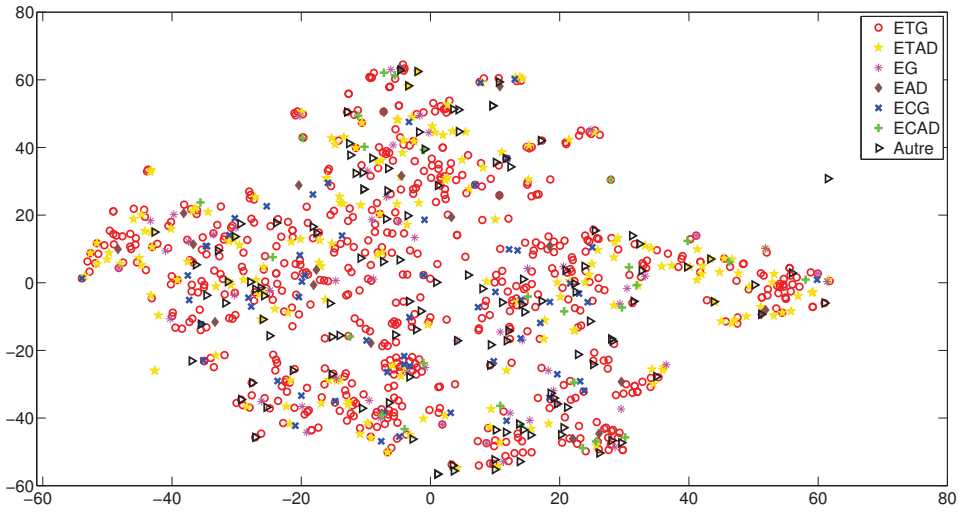


FIGURE 5-3 : La projection *t-SNE* des utilisateurs français d'Hypergéo classés dans sept profils en appliquant la dissimilarité hiérarchique.

5.2.2.2 Résultats statistiques en utilisant le SVM

Nous appliquons la méthode SVM sur les traces pour identifier les sept profils initiaux. Le tableau 5-1 donne la matrice de confusion en utilisant des valeurs de coût C différentes selon les profils (*Cf.* équation 4-11 page 106).

Nous pouvons comparer ces résultats à ceux obtenus dans le chapitre 3 avec l'algorithme des motifs caractéristiques (*Cf.* figure 3-10). Ainsi, pour chaque profil, nous remarquons que le taux de reconnaissance (*Accuracy*) est souvent supérieur à ceux obtenus précédemment, c'est le cas par exemple pour les profils ETG ou ETAD. Toutefois, lorsque nous calculons l'indicateur global pondéré par l'influence de la taille des classes (*Cf.* l'équation (4-8)), nous obtenons une valeur de $0{,}16$, qui est proche au hasard $(1/7)$. Nous pouvons alors en déduire que le taux de reconnaissance dans sa globalité n'est toujours pas acceptable mais souvent supérieur au niveau des profils à l'algorithme des motifs caractéristiques.

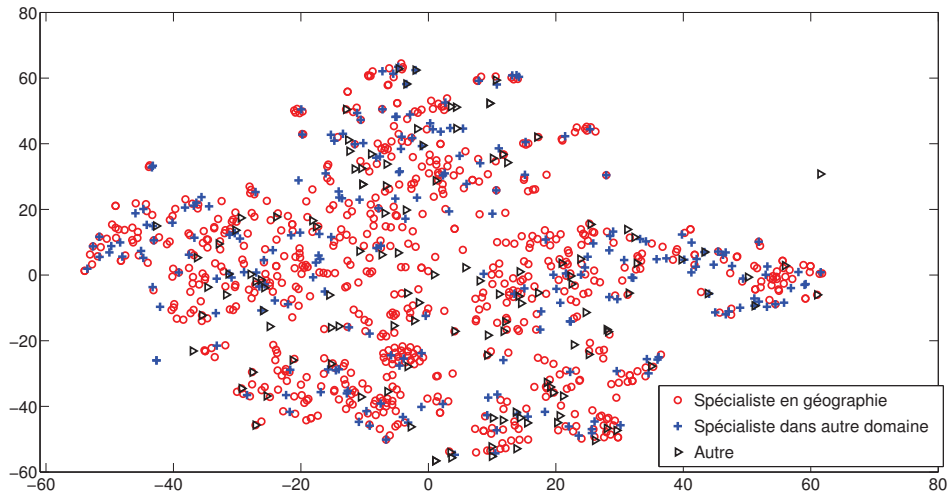


FIGURE 5-4 : Représentation de la figure 5-3 en fonction du domaine de compétences des utilisateurs.

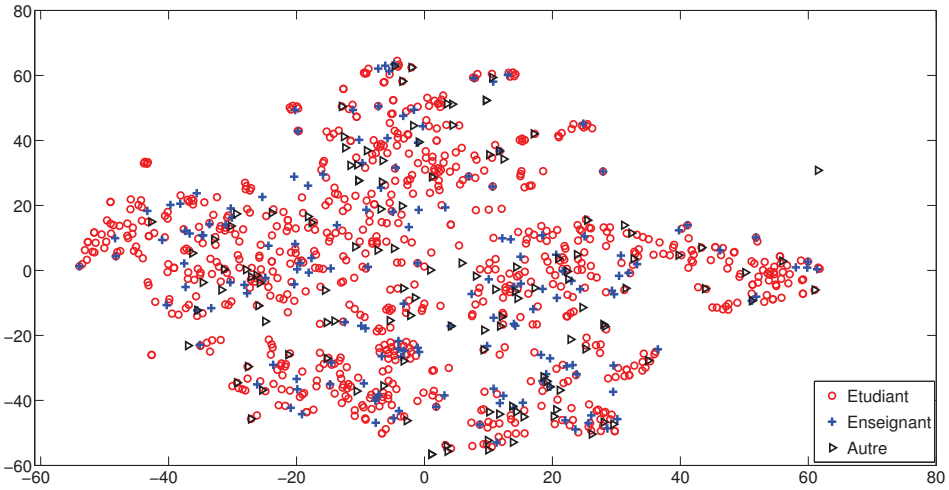


FIGURE 5-5 : Représentation de la figure 5-3 en fonction du niveau de connaissance des utilisateurs.

		Décision							
		ETG	ETAD	EG	EAD	ECG	ECAD	Autre	*Accuracy*
Profil	ETG	193	85	5	1	19	72	15	0,49
	ETAD	56	39	1	3	12	22	7	0,28
	EG	16	5	0	0	2	3	2	0
	EAD	2	3	0	0	3	0	0	0
	ECG	19	4	2	0	5	3	0	0,15
	ECAD	13	2	1	0	0	1	1	0,06
	Autre	37	18	0	0	1	10	8	0,1
		Indicateur global							0,16

TABLEAU 5-1 : La matrice de confusion pour les sept profils initiaux en utilisant des valeurs différentes de C pour chaque profil et en utilisant la dissimilarité hiérarchique.

Il en est globalement de même lorsque l'on utilise les deux regroupements de profils. Les tableaux 5-2, 5-3 présentent respectivement les matrices de confusion selon les profils axés sur les domaines de compétences des utilisateurs, et sur le niveau de connaissances de ces derniers. Les indicateurs globaux (respectivement $0{,}37$ et $0{,}35$) sont alors un peu supérieurs au hasard ($1/3$). Ils ne permettent pas de faire une bonne classification. Toutefois au niveau des profils les résultats sont supérieurs à ceux obtenus par la méthode des motifs caractéristiques (par exemple le taux de bonne reconnaissance pour le profil « Étudiant » est de pratiquement de 50%).

		Décision			
		Spécialiste en géographie	Spécialiste dans autre domaine	Autre	*Accuracy*
Profil	Spécialiste en géographie	124	177	150	0,27
	Spécialiste dans autre domaine	33	68	65	0,41
	Autre	10	33	31	0,42
				Indicateur global	0,37

TABLEAU 5-2 : La matrice de confusion pour trois profils en fonction du domaine de compétences des utilisateurs en utilisant la dissimilarité hiérarchique.

En utilisant cette dissimilarité, nous ne pouvons pas identifier les profils. Mais nous remarquons que les traces peuvent être regroupées en quatre ou cinq groupes (*Cf.* figure 5-3). L'inconvénient de la dissimilarité hiérarchique est qu'elle ne prend pas en compte les dissimilarités sémantiques entre documents. Par exemple la dissimilarité entre les deux articles « carte », qui sont en réalité le même article mais positionnés à deux endroits différents dans la hiérarchie, est égale à $0{,}4$. Nous proposons donc une dissimilarité sémantiques pour prendre en compte la

		Décision			
		Étudiant	Enseignant	Autre	*Accuracy*
Profil	Étudiant	262	138	130	0,49
	Enseignant	41	26	20	0,30
	Autre	36	20	18	0,24
				Indicateur global	0,35

TABLEAU 5-3 : La matrice de confusion pour trois profils en fonction du niveau de connaissance des utilisateurs en appliquant la dissimilarité hiérarchique.

ressemblance entre les contenus des documents.

5.3 Deuxième proposition : une dissimilarité sémantique

Cette proposition a pour objectif de prendre en compte le contenu des documents visités. Nous commencerons par présenter une méthode pour calculer la proximité sémantique entre deux documents. Puis, nous adapterons cette dernière aux caractéristiques de nos données. Enfin nous l'utiliserons au sein de notre similarité entre traces que nous testerons en utilisant le *t-SNE* et le SVM.

5.3.1 Dissimilarité *tf-idf*

La dissimilarité *tf-idf*[1] est souvent utilisée dans le domaine de fouille de textes. Dans cette méthode deux facteurs sont calculés, tf (*Term Frequency*) et idf (*Inverse Document Frequency*).

Le facteur $tf_{i,d}$ représente le nombre d'occurrences du terme i dans le document d. $tf_{i,d}$ peut être normalisé en divisant le nombre d'occurrences du terme i sur la somme du nombre d'occurrences de tous les termes dans le document d.

$$tf_{i,d} = \frac{n_{i,d}}{\sum_k n_{k,d}} \tag{5-2}$$

avec $n_{i,d}$ qui est le nombre d'occurrences du terme i dans le document d, et $n_{k,d}$ qui est le nombre d'occurrences du terme k dans le document d, $k \in \{1, ..., \text{ le nombre des termes dans } d\ \}$.

Le facteur idf quant à lui donne une indication sur l'occurrence du terme i dans tous les documents :

$$idf_i = \log \frac{|D|}{|d_j|} \tag{5-3}$$

1. *Term Frequency-Inverse Document Frequency*

où $|D|$ est le nombre de tous documents, et $|d_j|$ est le nombre de documents ayant le terme i.

Pour chaque terme i et document d, on définit la valeur $w_{i,d}$ tel que :

$$w_{i,d} = tf_{i,d} \times idf_i \tag{5-4}$$

Une fois que les $w_{i,d}$ sont calculées pour tous les termes et tous les documents, nous obtenons une matrice de distance de $|D| \times |T|$, tel que $|D|$ est le nombre de tous documents et $|T|$ est le nombre de termes. À partir de cette matrice, nous pouvons calculer une similarité entre les documents (par exemple, en utilisant la distance cosinus ou la distance euclidienne).

Nous avons appliqué cette méthode sur les articles français du site Hypergéo en utilisant le distance cosinus. La figure 5-6 représente la projection *t-SNE* de ces articles. Bien que plus

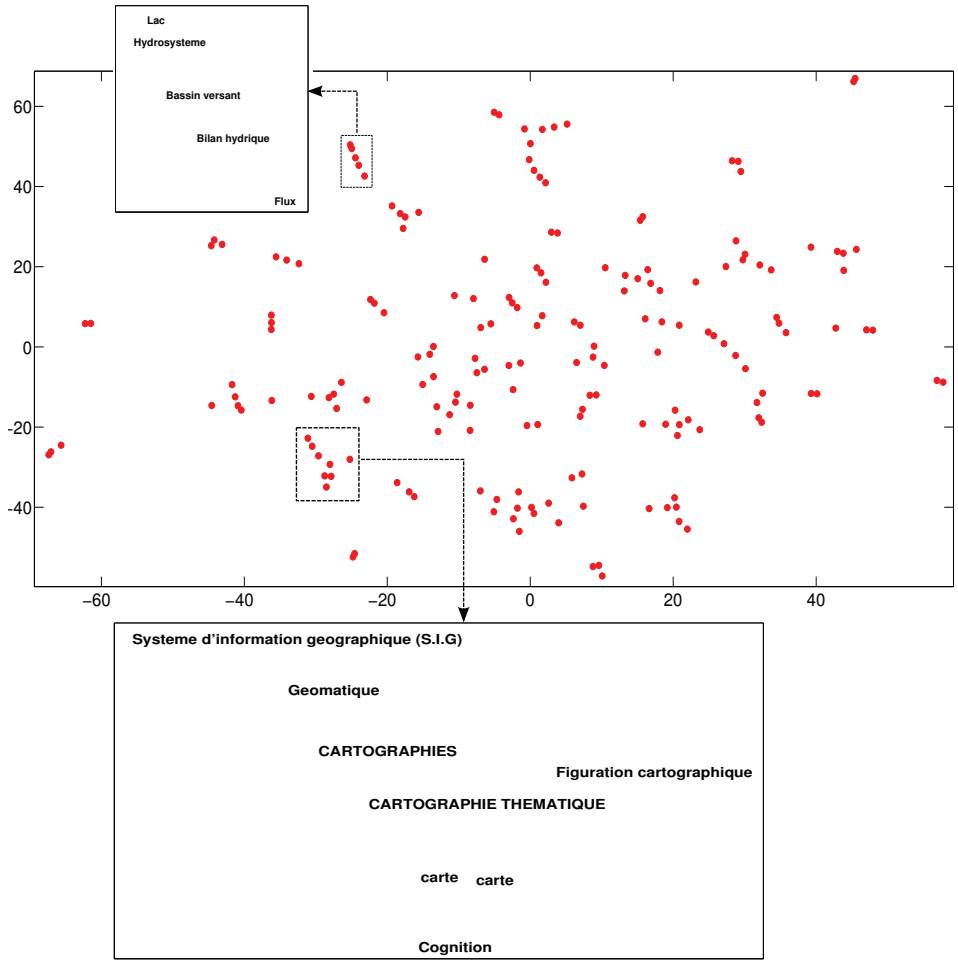


FIGURE 5-6 : La projection *t-SNE* de la similarité entre les articles français selon la méthode *tf-idf*.

diffus que les données synthétiques que nous avons générés dans le chapitre précédent (*Cf.*

la figure 4-3 page 100), les projections d'articles forment régulièrement de petits ensembles. L'analyse de ces deniers permet de vérifier que deux articles proches dans cette projection sont aussi proches sémantiquement. À titre d'exemple (*Cf.* la figure 5-6), il y a un ensemble qui est composé des articles { *'Systeme d'information geographique (S.I.G)'* , *'Geomatique'*, *'CARTOGRAPHIES'*, *'Figuration cartographique'*, *'CARTOGRAPHIE THEMATIQUE'*, *'carte'*, *'carte'*, *'Cognition'*}.

Rappelons que nous avons dans le tableau 4-1 (page 94) des articles de cet ensemble qui sont apparus dans les navigations des utilisateurs 80 et 322. Nous voyons aussi dans la figure 5-6 qu'il y a deux articles « carte ». Il s'agit en fait du même article qui apparaît à deux endroits différents dans le site Hypergéo (article 266 et article 381). L'article 266 se trouve dans le chemin *Français* > *Géographie* > *Outils* > *carte*, et l'article 381 se trouve dans le chemin *Français* > *Géographie* > *Concepts* > *carte*. Les articles en double risquent de diminuer la qualité de l'identification des utilisateurs surtout quand les méthodes appliquées ne prennent pas en compte les contenus des articles comme la méthode des motifs caractéristiques.

Pour confirmer l'intuition que nous avons eu lors de l'analyse du tableau 4-1 (page 94), nous avons représenté ces quatre traces sur cette projection. Nous obtenons alors la figure 5-7 qui montre que les utilisateurs 80 et 322 ont bien des intérêts communs parce que les articles visités par ces utilisateurs sont des articles similaires, donc proches au niveau de leurs projections. Nous notons que c'est le même cas pour les utilisateurs 377 et 597.

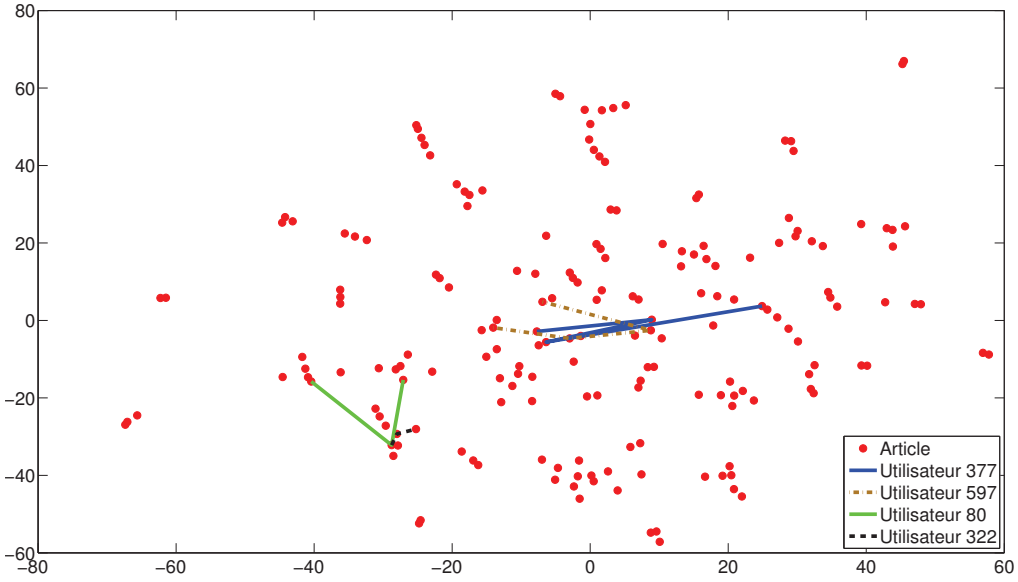


FIGURE 5-7 : Les navigations des utilisateurs 377, 597, 80 et 322 sur la projection *t-SNE* de la similarité entre les articles français selon la méthode *tf-idf*.

Nous allons donc généraliser cette approche en appliquant la dissimilarité *tf-idf* entre les

documents pour calculer la similarité entre les traces.

5.3.2 Application du *tf-idf* aux données de Hypergéo

Grâce à la méthode *tf-idf* nous sommes capables de trouver une similarité entre les traces selon la sémantique des documents visités. Cet avantage devrait nous permettre d'identifier des utilisateurs visitant des documents qui sont non identiques, mais qui sont proches sémantiquement. Toutefois, nous rencontrons deux problèmes de perte d'informations.

Tout d'abord il y a le problème des rubriques. Nous ne pouvons pas prendre en compte les rubriques visitées, parce qu'elles n'ont pas de contenus. Nous ne pouvons donc pas calculer le *tf-idf* entre deux rubriques ou entre une rubrique et un article. Nous sommes donc obligés d'enlever les rubriques des traces. Par exemple, voici deux traces (a est l'abréviation d'article, et r est celle de rubrique) :

- $trace1 = \{(r0 > r6); (r6 > r76); (r76 > r6); (r6 > r7); (r7 > r1); (r1 > r11); (r11 > r90)\}$,
- $trace2 = \{(r6 > r74); (r74 > r7); (r7 > a407); (a407 > r7); (r7 > r74); (r74 > r6)\}$

Pour pouvoir appliquer la méthode *tf-idf*, nous sommes obligés de supprimer la première trace de notre base de données et la deuxième ne contiendra finalement qu'un seul article.

Ensuite il y a le problème des articles étrangers. Comme Hypergéo est disponible en trois langues, il existe des traces qui ont des articles rédigés dans des langues différentes. Comme la méthode *tf-idf* est dépendante de la langue, nous ne pouvons pas l'appliquer sur deux articles rédigés dans deux langues différentes. Pour résoudre ce problème, trois propositions sont possibles :

1. Les articles en langues étrangères sont enlevés, on garde uniquement les articles en français dans les traces.
2. Une grande matrice de similarité S entre tous les articles est construite, tel que la dissimilarité entre deux articles d'une même langue est calculée à l'aide du *tf-idf*, et celle calculée entre deux articles dans deux langues différentes est égale à 0, soit :

$$S = \begin{pmatrix} \overbrace{(S_{fr})}^{\text{articles français}} & 0 & 0 \\ 0 & \overbrace{(S_{en})}^{\text{articles anglais}} & 0 \\ 0 & 0 & \overbrace{(S_{es})}^{\text{articles espagnols}} \end{pmatrix}$$

Sachant que S_{fr}, S_{en} et S_{es} sont respectivement les matrices de similarité entre les articles français, entre les articles anglais et entre les articles en espagnol.

3. Une grande matrice de similarité S entre toutes les articles est construite après avoir traduit tous les articles étrangers en français.

Dans notre contexte, nous utilisons l'outil *TreeTagger* [Sch] pour faire de l'analyse syntaxique en langue naturelle. Nous avons rencontré beaucoup de problèmes avec ce dernier pour analyser des textes en espagnol. Nous avons donc dû écarter la deuxième solution. Nous avons aussi écarté la troisième solution, car nous n'avons pas la possibilité de faire traduire ces textes par des experts en géographie anglophones et hispanophones. Nous avons aussi écarté la traduction automatique générique (tel que *google translator*), car les articles d'Hypergéo utilisent un vocabulaire très spécialisé, et cela risquerait d'apporter des erreurs.

Dès lors, après avoir filtré les traces pour enlever les rubriques et les articles en langues étrangères, nous pouvons calculer une dissimilarité entre elles.

5.3.2.1 Résultats visuels de *t-SNE*

La figure 5-8 représente la projection *t-SNE* des traces qui seront utilisées comme base d'apprentissage pour le SVM. Nous pouvons constater que toutes ces traces sont mélangées et qu'il sera très difficile d'avoir un bon taux de reconnaissance.

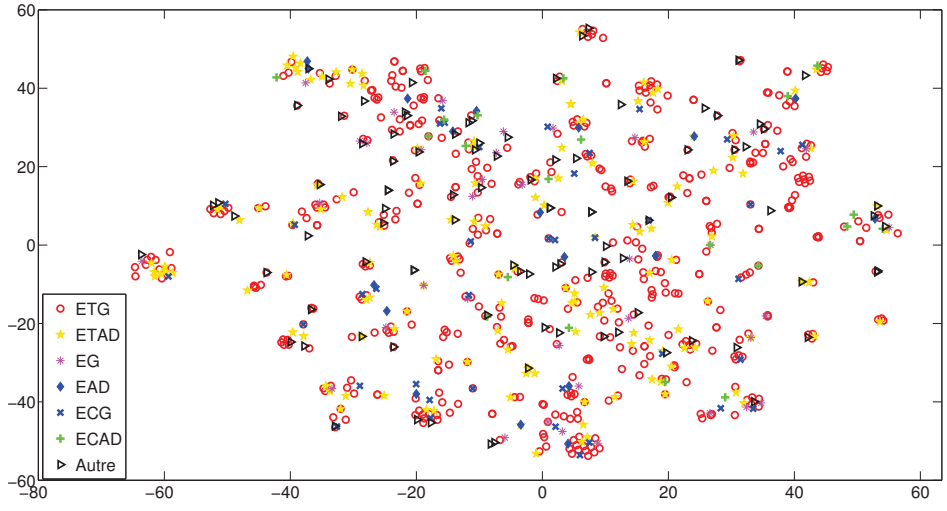


FIGURE 5-8 : La projection *t-SNE* des utilisateurs francophones d'Hypergéo de la base d'apprentissage classés dans sept profils après l'application de la dissimilarité *tf-idf*.

Quand nous examinons la matrice de dissimilarité entre les articles français d'Hypergéo, nous remarquons que la plupart (98%, $\frac{34294}{34969}$) des valeurs sont supérieures à $0{,}8$, et les autres

sont proches de 0. Pour rappel, la dissimilarité entre deux traces est une somme de la dissimilarité entre les éléments de ces traces (*Cf.* l'algorithme 9 page 110). Sachant que la plage des dissimilarités entre articles est faible, la répartition des dissimilarités entre traces l'est aussi. C'est pourquoi le *t-SNE* représente l'ensemble de ces traces sous forme d'une boule, il n'arrive pas à distinguer le voisinage de chacune des traces. Or ceci est contraire à l'intuition que nous avions en regardant la figure 5-7.

Nous proposons donc d'utiliser une matrice de dissimilarité basée sur la projection *tf-idf* des articles en français (la figure 5-6 page 127). Nous utilisons cette matrice à la place de la matrice de dissimilarité *tf-idf*. Le résultat de projection *t-SNE* des traces de la base d'apprentissage en utilisant cette nouvelle matrice est proposé par figure 5-9.

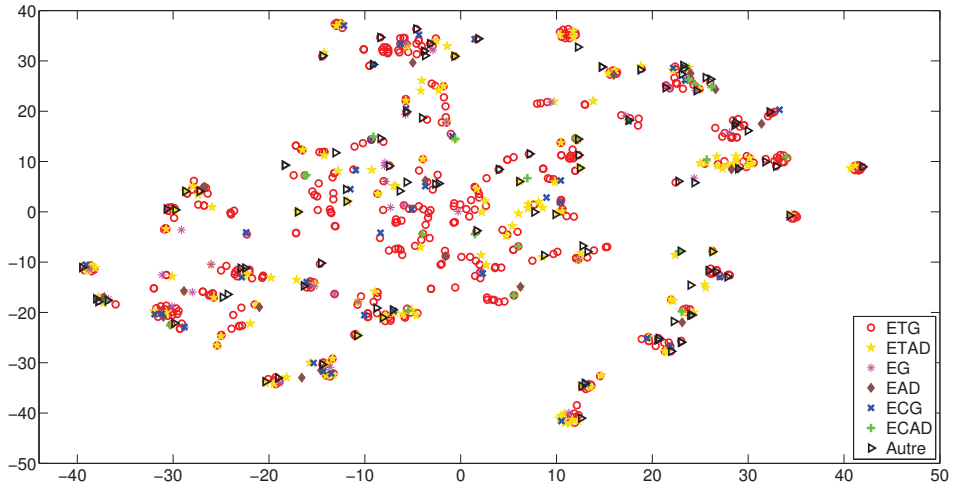


FIGURE 5-9 : La projection *t-SNE* des utilisateurs francophones de la base d'apprentissage utilisant la dissimilarité entre traces basée sur la projection *t-SNE* des articles (*Cf.* figure 5-6).

Les sept profils sont encore mélangés, mais nous voyons que les traces sont divisées en plusieurs groupes. Nous remarquons aussi qu'il y quelques traces ayant des positions identiques. Ceci est dû à la perte d'informations dans le filtrage des traces que nous effectuons (après le filtrage, il existe des traces qui ont un seul article identique).

Les figures 5-10 et 5-11 représentent respectivement la projection *t-SNE* pour les profils axés sur les domaines de compétences et pour les profils axés sur le niveau de connaissance. Nous remarquons que les trois groupes qui se forment ne permettent pas d'identifier ces différents profils.

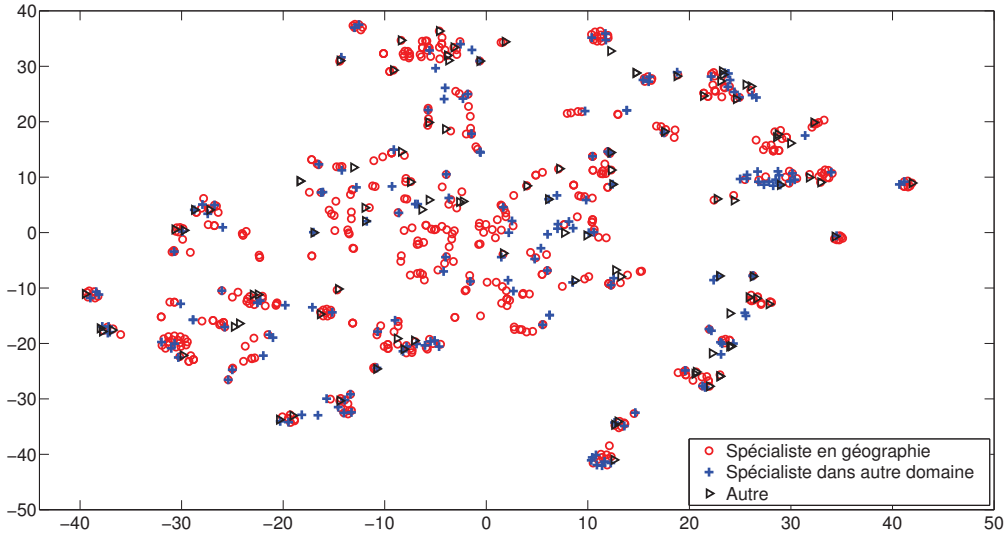


FIGURE 5-10 : Représentation de la figure 5-9 en fonction du domaine de compétences des utilisateurs.

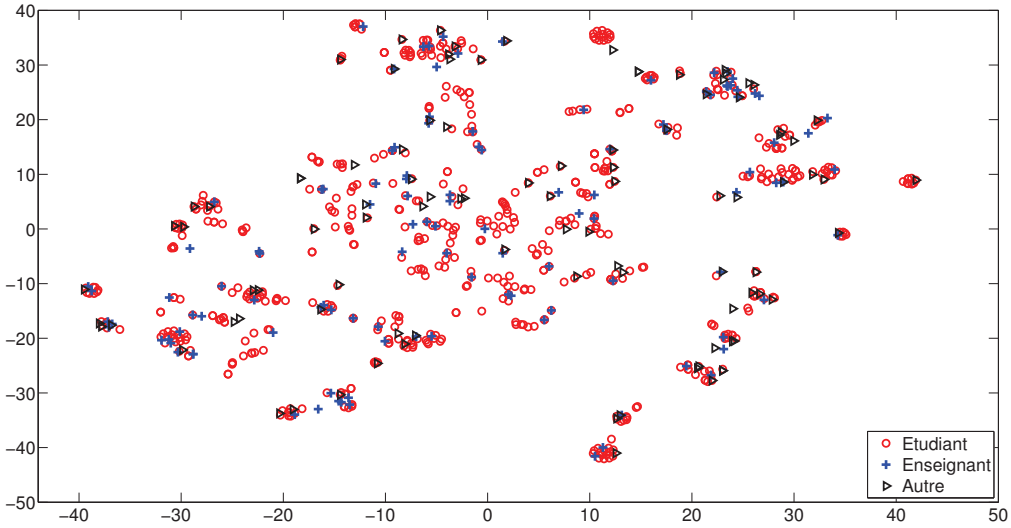


FIGURE 5-11 : Représentation de la figure 5-9 en fonction du niveau de connaissance des utilisateurs.

5.3.2.2 Résultats statistiques de SVM

Nous appliquons la méthode SVM sur les traces pour identifier les sept profils initiaux. Le tableau 5-4 donne la matrice de confusion en utilisant des valeurs de coût C différentes selon le profil (*Cf.* équation 4-11 page 106). Nous pouvons comparer ces résultats à ceux obtenus dans le chapitre 3 avec l'algorithme des motifs caractéristiques (*Cf.* figure 3-10). Ainsi, pour chaque profil, nous remarquons que le taux de reconnaissance (*Accuracy*) est souvent supérieur à ceux obtenus précédemment, c'est le cas par exemple pour les profils ETG, ETAD ou Autre. Toutefois, lorsque nous calculons l'indicateur global pondéré par l'influence de la taille des classes (*Cf.* (4-8)), nous obtenons une valeur $0{,}14$ qui est proche au hasard $(1/7)$. Nous pouvons alors en déduire que le taux de reconnaissance dans sa globalité n'est toujours pas acceptable mais souvent supérieur au niveau des profils à l'algorithme des motifs caractéristiques.

		Décision							
		ETG	ETAD	EG	EAD	ECG	ECAD	Autre	*Accuracy*
Profil	ETG	63	46	32	32	44	37	62	0,20
	ETAD	24	17	20	13	21	11	18	0,14
	EG	10	0	1	3	5	4	2	0,04
	EAD	2	0	0	0	2	0	2	0
	ECG	5	4	2	4	4	1	3	0,17
	ECAD	1	3	3	0	1	2	2	0,17
	Autre	17	3	1	7	10	6	16	0,27
								Indicateur global	0,14

TABLEAU 5-4 : La matrice de confusion pour les sept profils initiaux en utilisant des valeurs différentes de C pour chaque profil (la dissimilarité basée sur la projection *t-SNE*).

		Décision			
		Spécialiste en géographie	Spécialiste dans autre domaine	Autre	*Accuracy*
Profil	Spécialiste en géographie	144	131	89	0,40
	Spécialiste dans autre domaine	55	56	31	0,39
	Autre	18	21	21	0,35
				Indicateur global	0,38

TABLEAU 5-5 : La matrice de confusion pour trois profils en fonction du domaine de compétences des utilisateurs (la dissimilarité basée sur la projection *t-SNE*).

Il en est globalement de même lorsque l'on utilise les deux regroupements de profils. Les tableaux 5-5, 5-6 présentent respectivement les matrices de confusion selon les profils axés sur

les domaines de compétences des utilisateurs, et sur le niveau de connaissances de ces derniers. Les indicateurs globaux (respectivement 0,38 et 0,33) ne permettent pas de faire une bonne classification. Toutefois au niveau des profils, les résultats sont supérieurs à ceux obtenus par la méthode des motifs caractéristiques (par exemple le taux de bonne reconnaissance pour le profil « spécialistes en géographie » est de pratiquement 40% alors qu'il était inférieur à 30% avec la précédente méthode (*Cf.* figure 3-12(a)).

		Décision			
		Étudiant	Enseignant	Autre	*Accuracy*
Profil	Étudiant	260	103	77	0,59
	Enseignant	46	15	5	0,23
	Autre	32	17	11	0,18
				Indicateur global	0,33

TABLEAU 5-6 : La matrice de confusion pour trois profils en fonction du niveau de connaissance des utilisateurs (la dissimilarité basée sur la projection *t-SNE*).

Ces résultats peuvent s'expliquer par le fait que nous soyons obligés de filtrer les traces de façon à ne garder que les articles en français. Dès lors, un grand nombre d'éléments de ces traces sont supprimés, il y a donc un perte d'information importante. Enfin, à la lecture de la figure 5-9, cinq groupements de traces apparaissent (comme nous avons vu aussi dans la figure 5-3 page 123).

5.4 Conclusion

Pour trouver les profils des utilisateurs d'Hypergéo, nous avons défini dans le chapitre 4 une dissimilarité entre leurs traces. Cette dissimilarité a reposé sur la dissimilarité entre les éléments de ces traces. À cause des longueurs courtes de la majorité des traces, nous avons pris en compte les documents visités pour présenter les éléments de traces.

Deux méthodes pour calculer la dissimilarité entre documents ont été proposées, une dissimilarité hiérarchique et une dissimilarité sémantique. La dissimilarité hiérarchique dépend de la structure du site. Autrement dit, elle reflète le point du vue du responsable d'Hypergéo. Nous avons montré les inconvénients de cette méthode (par exemple, les documents en double). Nous avons ensuite expliqué la méthode *tf-idf*. Cette méthode nous a permis de calculer une dissimilarité sémantique entre les documents. Comme elle dépend de la langue et du contenu du document, nous avons été obligés d'enlever les rubriques et les articles en langue étrangère des traces. Beaucoup d'informations ont été perdues à cause de ce filtrage.

Les résultats du SVM appliqués à notre dissimilarité sont un peu supérieurs aux résultats des motifs caractéristiques, mais ils ne nous permettent pas de bien identifier les profils prédéfinis. Nous remarquons toutefois dans les projections *t-SNE* (*Cf.* figures 5-3, 5-8) que les traces peuvent être regroupées dans quatre ou cinq groupes. Nous allons étudier dans le chapitre suivant la signification de ces groupes et la possibilité qu'ils puissent présenter de nouveaux profils.

CHAPITRE 6

Identification des classes a posteriori

Sommaire

6.1 Introduction

Dans le chapitre précédent, nous avons essayé d'identifier des utilisateurs d'Hypergéo selon sept profils prédéfinis, ou selon deux regroupements de ces profils. La méthode des motifs caractéristiques n'arrive pas à résoudre ce problème. Même si notre proposition donne des résultats supérieurs aux motifs caractéristiques, elle ne nous permet pas d'attendre notre objectif. Nous pensons qu'il peut y avoir trois hypothèses qui expliquent ce fait.

Tout d'abord, il se peut que les traces des utilisateurs, avec les informations que l'on utilise, ne soient pas séparables. En effet, si deux utilisateurs, ayant des profils définis *a priori* comme différents et qu'ils effectuent le même parcours, quelque soit la distance utilisée, quelque soit la méthode de classification utilisée, il sera impossible de les distinguer.

La deuxième hypothèse concerne le manque d'informations complémentaires, informations que nous ne possédons pas. En effet, les informations que nous avons recueillies concernent uniquement les documents visités par les utilisateurs. Il se peut qu'elles ne suffisent pas à les identifier correctement. Par exemple, il existe d'autres informations, comme les temps de visualisation des documents, les mots clés utilisés pour arriver aux documents, le pays du visiteur, etc. qui pourraient aider à mieux les caractériser.

Enfin, la troisième hypothèse est la possibilité que les profils prédéfinis ne représentent pas bien les utilisateurs d'Hypergéo. Dans ce cas, il existe peut-être un autre étiquetage qui permettrait de mieux les caractériser.

Comme nous avons uniquement les documents visités par les utilisateurs pour représenter les traces, dans ce chapitre, nous allons étudier plus en plus en détail les causes possibles soulevées par la troisième hypothèse. Ainsi, à partir de l'analyse de la projection *t-SNE* des traces d'apprentissage, nous allons essayer d'identifier le nombre de profils des utilisateurs. Ensuite, nous ré-étiquetterons les données d'apprentissage automatiquement en utilisant l'algorithme *Kmeans*. Nous étudierons alors ces classes pour vérifier qu'elles ont un sens. Puis, pour vérifier que cet étiquetage est généralisable, nous projetterons ensemble les données d'apprentissage et les données tests. Enfin après un ré-étiquetage manuel des données de test, nous comparerons les résultats de l'algorithme des motifs caractéristiques et de l'utilisation du SVM.

6.2 Étude des traces

Comme nous avons commencé à le supposer dans le chapitre précédent, à partir de l'étude de la projection des traces des utilisateurs de la base d'apprentissage utilisant la dissimilarité sémantique entre les documents (*Cf.* la figure 5-9 page 131), il semblerait que ces traces puissent

être regroupées en cinq groupes assez distincts.

Notre objectif est de vérifier que ces regroupements de traces puissent être identifiés automatiquement. Ensuite il faut donner sens à ces regroupements, c'est-à-dire interpréter le contenu de ces traces, et vérifier que les traces d'un même groupe sont sémantiquement proches.

6.2.1 *Clustering* des traces

Comme nous venons de le voir, à partir de la figure 5-9, nous sommes en mesure d'identifier visuellement principalement cinq groupements : un premier en bas à gauche, un second central, un troisième tout en haut, un quatrième plutôt en haut à droite et enfin un cinquième en bas à droite. Il faut maintenant vérifier que l'on peut identifier ces groupements automatiquement. Ceci peut être réalisé avec un algorithme de classification non supervisée.

Nous avons donc utilisé l'algorithme *Kmeans* sur cette base d'apprentissage avec k qui est égal à 5. Toutefois, l'algorithme Kmeans a besoin, en entrée, des données de $\mathbb{R}^n$ pour fonctionner. Comme nous l'avons fait pour le SVM (*Cf.* la deuxième et troisième paragraphe de la section 4.3.2), nous avons utilisé la dissimilarité entre les traces comme données d'entrée. Nous obtenons donc pour chaque trace un label ainsi que les cordonnées des centroïdes de chaque classe.

Pour s'assurer que les classes déterminées par le *Kmeans* représentent bien les groupements que nous avons identifiés, nous utilisons la même figure mais avec les nouveaux labels obtenus par le *Kmeans*. Nous obtenons donc la figure 6-1 ou chacune des cinq classes est représentée par une couleur. Nous constatons que les classes déterminées par l'algorithme du *Kmeans* correspondent globalement à celles que nous avions précédemment supposées.

Toutefois, nous avons besoin de vérifier que ces classes représentent toute la base de données et pas seulement la base d'apprentissage. Autrement dit, est-ce que les traces d'apprentissage représentent bien toutes les traces ? Ainsi il faut vérifier que l'ajout de nouvelles traces ne remet pas en cause le *clustering* que nous avons effectué. Cela implique de vérifier tout d'abord qu'il n'y aura pas de nouveaux *clusters* qui apparaîtront. Ensuite il faut aussi vérifier que les futures traces appartiendront bien aux cinq classes que le $Kmeans$ à identifier.

On se propose de vérifier cela en utilisant encore une fois la projection *t-SNE*. Tout d'abord, nous avons calculé la dissimilarité entre toutes les traces (de la base d'apprentissage et de la base de test). Ensuite nous les avons projetées en utilisant les cinq couleurs précédentes pour les traces de la base d'apprentissage et les cercles noirs pour les traces de test. Nous obtenons la figure 6-2.

Dans cette figure, les traces de test sont présentes dans toutes les classes et elles ne forment aucun nouveau *cluster*. Dès lors, ce résultat nous assure que les traces d'apprentissage repré-

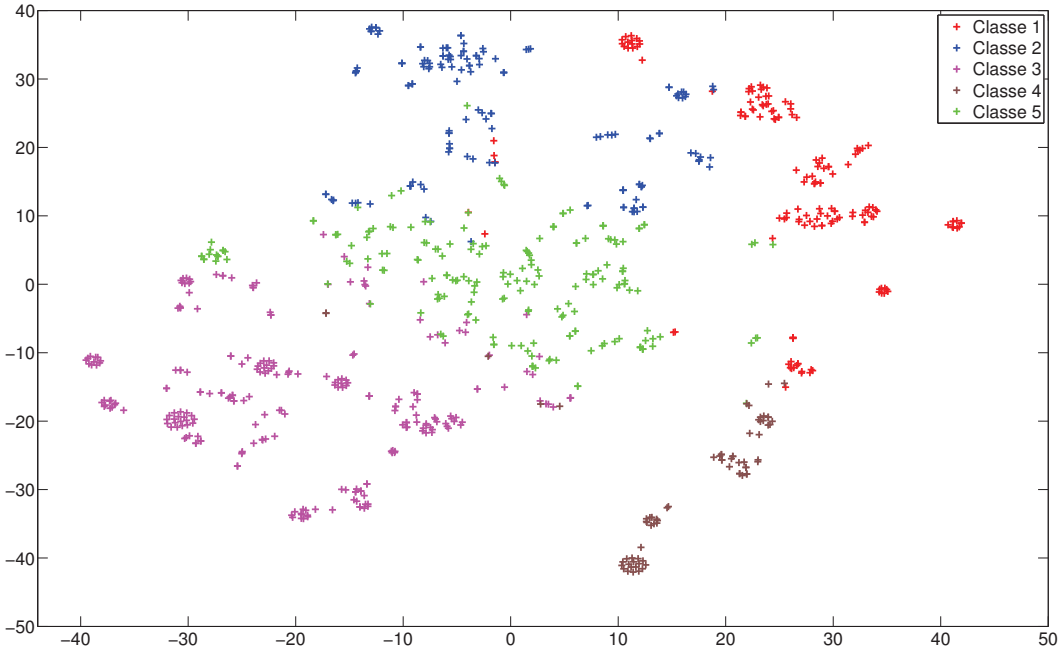


FIGURE 6-1 : La projection *t-SNE* de la base d'apprentissage avec les labels issus du *Kmeans*.

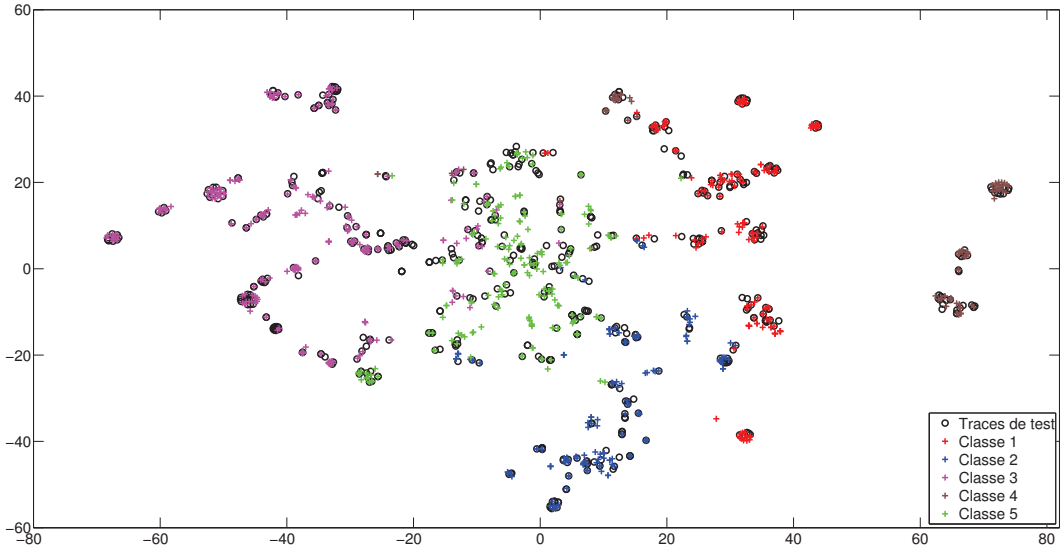


FIGURE 6-2 : La projection *t-SNE* de toutes les données avec les labels issus du *Kmeans* des données d'apprentissage.

sentent bien toutes les traces et que les classes du *Kmeans* sont caractéristiques de toutes les traces. Il nous faut maintenant identifier les sens de ces classes.

6.2.2 Identification du sens des classes

Pour interpréter le sens de ces classes identifiées par le *Kmeans*, nous avons choisi quelques traces de chaque classe. Le tableau 6-2 présente les titres des documents appartenant à certaines de ces traces.

À première vue, sans être expert du domaine les documents de ces traces sont de manière générale assez différents. Mais surtout pour les quatre premières classes, les documents des traces de chaque classe semblent appartenir aux mêmes thèmes.

Une analyse plus détaillée, nous a permis d'identifier le ou les thèmes de chaque classe. Par exemple, pour la classe 1, les thèmes émergeant sont la « cartographie » et « l'utilisation des mathématiques en géographie ». Pour la classe 2, c'est plutôt la notion de « paysage » qui semble se dégager. Quant à la classe 3, les documents font référence au concept de « mondialisation » et des problèmes qui lui sont associés (par exemple l'équité). Pour la classe 4, ce sont les concepts de « discontinuité » et tout ce qui est lié aux domaines aquatiques qui semble apparaître. Par contre, nous n'arrivons pas à trouver des concepts discriminants pour la classe 5. En effet, les traces appartenant à cette classe partagent de nombreux documents avec les traces des quatre autres classes. Cette hypothèse est confirmée par la position des traces de cette classe dans la projection *t-SNE*. En effet, elle se positionne au centre de la projection et est assez mélangée avec les autres classes (*Cf.* la figure 6-1). Le tableau 6-2 synthétise cette analyse.

Ceci nous amène donc à appréhender nos traces d'une nouvelle façon. Dans les chapitres précédents, nous avons essayé de déterminer le profil des utilisateurs en fonction de leur domaine de compétence ou en fonction de leur niveau de connaissance. Nous allons maintenant essayer d'identifier leur profil en fonction de leur(s) centre(s) d'intérêt.

6.3 Classification des traces en utilisant les nouvelles classes

Nous avons montré que les traces d'apprentissage peuvent être regroupées en cinq catégories représentant les centres d'intérêt des utilisateurs. Il nous faut maintenant vérifier que cette classification est généralisable, c'est-à-dire confronter les données de test à l'aide d'un algorithme de classification supervisée. Pour étiqueter les données d'apprentissage nous avons utilisé l'algorithme d'apprentissage non supervisé *Kmeans* avec $k = 5$. Il n'est pas possible d'utiliser ce même algorithme pour les traces de test, car étiqueter ces traces avec l'algorithme de *Kmeans* qui utilise la même distance que l'algorithme de classification supervisé, nous as-

	Documents visités
Classe 1	Variables qualitatives, Analyse de données, Mesure, Discrétisation
	CARTOGRAPHIE THÉMATIQUE, carte ,Figuration cartographique
	carte, Carte choroplèthe, Statistique spatiale
	Topographie (historique de la notion), CARTOGRAPHIES
	Topographie, carte
	Analyse de données, Modelés statistiques, Statistique spatiale
	Carte choroplethe, Discretisation, Variables quantitatives
Classe 2	Historique du paysage , Paysage , Statut spatial du paysage , Paysage selon le laboratoire THEMA
	Paysage selon le laboratoire THEMA , Paysage
	Télédétection , Paysage
	La trajection paysagère , Utilisations des paysages , Paysage selon le laboratoire THEMA ,Production du paysage
	Perception des paysages, Statut spatial du paysage
	Paysage, Constructivisme
Classe 3	Centre Périphérie, Équité territoriale, Mondialisation, Centralite
	Aire d'influence, Centralite, Polarisation
	Mondialisation
	Organisation de l'espace , Orient
	Nord , Mondialisation
Classe 4	Insularité, Île
	Seuil, Flux
	Insularité, Discontinuité
	Insularité, Littoral, Littoral, Anthropisation
	Bilan hydrique, Densité
Classe 5	Interaction spatiale, Mondialisation, Périphérie, Bifurcation, Organisation de l'espace
	Evolution des entités géographiques, Auto-organisation, Bifurcation
	Système spatial, Géographie, Polarisation
	Organisation de l'espace, Auto-organisation
	Analyse spatiale, Configuration, Organisation de l'espace
	Auto organisation, Complexité
	Échelle, Théories de l'Analyse Spatiale, Localisation

TABLEAU 6-1 : Titres des documents des exemples des traces classées par *Kmeans*.

Classe Nb	Concepts principaux
Classe 1	Cartographie, Outils mathématiques
Classe 2	Paysage
Classe 3	Mondialisation, Polarisation
Classe 4	Eau, Discontinuité
Classe 5	Mélange entre les tous concepts (*autre*)

TABLEAU 6-2 : Les concepts principaux des classes de *Kmeans*.

surerait artificiellement un excellent résultat. Nous allons donc classer les traces de la base de test manuellement.

6.3.1 Étiquetage des traces de test

Pour étiqueter les traces de la base de test, nous avons analysé les documents visités afin d'extraire les centres d'intérêt de l'utilisateur correspondant. Ensuite, ces centres d'intérêt ont été comparés avec ceux que nous avons définis précédemment. Nous avons affecté alors à chaque trace, le label de la classe dont les centres d'intérêt sont les plus proches. Toutefois dans certains cas, il peut y avoir ambiguïté. Ainsi, une deuxième étude de ces traces a été faite pour confirmer ou non la décision. Le tableau 6-3 présente le nombre de traces pour chaque classe que nous avons étiquetées, et le nombre de labels confirmés par la deuxième évaluation.

Classe Nb	Étiquetage	Confirmation d'étiquetage	Pourcentage
Classe 1	99	69	70%
Classe 2	37	17	50%
Classe 3	62	35	56%
Classe 4	46	30	65%
Classe 5	26	6	23%

TABLEAU 6-3 : La décision d'étiquetage et sa confiance.

On remarque dans ce tableau que le pourcentage de confirmation pour classe 2 est de 50%. Nous pouvons expliquer cela par le fait que pour cette classe un seul concept principal a été identifié (la notion de « Paysage »). Or souvent ces traces comportent aussi des documents faisant référence à d'autre concept. C'est pourquoi il y a, pour la moitié d'entre elles, des doutes quant à leur étiquetage. Nous voyons aussi que pour la classe 5, la confiance quant à la classification manuelle est très basse. En effet, comme nous l'avons déjà vu, cette classe n'a pas été clairement définie.

Nous allons maintenant évaluer notre dissimilarité en visualisant la projection des traces d'apprentissage et des traces de test avec ces nouveaux labels. Ensuite, notre approche sera validée statistiquement.

6.3.2 Validation visuelle

Notre dissimilarité est satisfaisante si, pour une projection, des traces d'apprentissage et de test du même profil sont proches, sachant que les profils des traces d'apprentissage ont été fixés par le $Kmeans$ et ceux des traces de tests ont été fixés manuellement.

La figure 6-3 représente cette projection. Les couleurs représentent les classes. Les croix et les cercles représentent respectivement les traces d'apprentissage et celles de test.

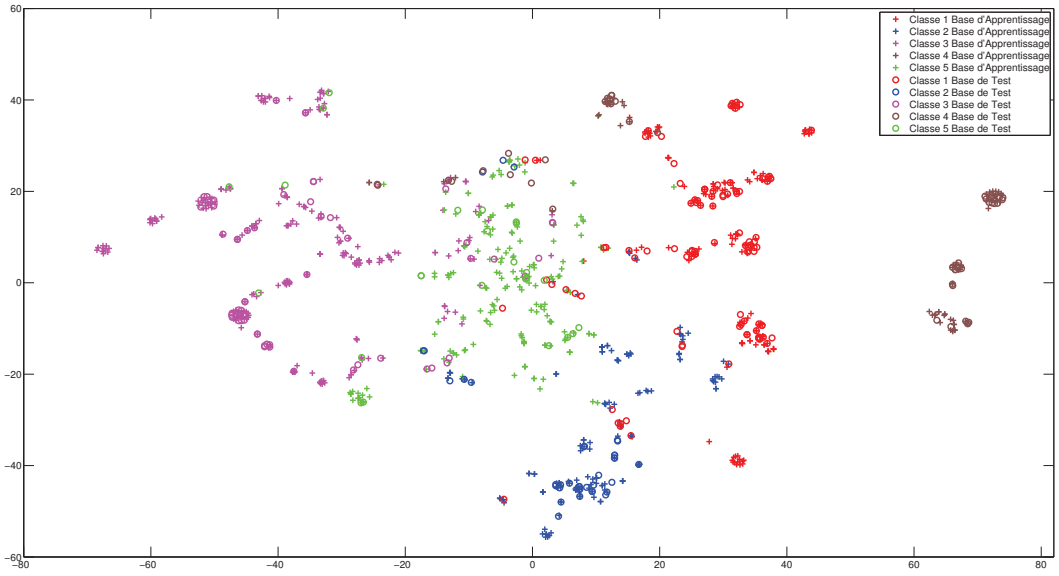


FIGURE 6-3 : La projection *t-SNE* des traces d'apprentissage et de test avec les labels de *Kmeans* des données d'apprentissage.

Pour la même classe, nous voyons bien que les positions des traces d'apprentissage sont assez proches de celles de test. Bien sûr, il existe quelques traces de test qui sont mal positionnées. Par exemple quelques traces de test de la classe 1 sont mélangées avec celles des classes 2 et 5. Mais en règle générale, nous arrivons à classer correctement les nouvelles traces.

Ces résultats nous indiquent que notre dissimilarité est bien capable de répartir les traces en plusieurs classes et d'identifier la classe d'une nouvelle trace. Nous allons maintenant valider ceci statistiquement à l'aide de l'algorithme du SVM. Puis ces résultats seront comparés avec ceux des motifs caractéristiques.

6.3.3 Validation statistique et comparaison avec l'algorithme des motifs caractéristiques

Le tableau 6-4 représente la matrice de la confusion des résultats de SVM pour la classification des traces de test. Dans ce tableau l'indicateur global, calculé à partir de l'équation 4-8 page 101, est de $0{,}82$, ce qui est un bon résultat. Il en est de même pour le taux de reconnaissance pour chaque classe (les valeurs sont comprises entre $0{,}77$ et $0{,}87$). Nous pouvons donc affirmer que nous sommes en mesure de bien identifier toute nouvelle trace.

		Décision					
		1	2	3	4	5	*Accuracy*
Classe	1	78	14	0	7	0	0,79
	2	0	31	1	5	0	0,84
	3	0	0	54	8	0	0,87
	4	2	0	2	38	4	0,83
	5	0	0	6	0	20	0,77
		Indicateur global					0,82

TABLEAU 6-4 : La matrice de confusion du SVM en appliquant les classes de *Kmeans*.

Nous avons décidé de comparer ces résultats avec ceux qui peuvent être obtenus avec l'algorithme des motifs caractéristiques. Ce résultat est représenté dans le tableau 6-5. Même si les motifs caractéristiques arrivent dans le cas de la classe 4 à convenablement classer une nouvelle trace (taux de reconnaissance de $0{,}67$), son indicateur global $0{,}32$ n'est pas acceptable, il est assez proche du hasard $(0{,}2)$.

		Décision						
		1	2	3	4	5	Non classé	*Accuracy*
Profil	1	60	6	0	1	1	31	0,61
	2	0	2	0	0	0	35	0,05
	3	0	0	16	0	0	46	0,26
	4	0	0	1	31	1	13	0,67
	5	0	0	1	1	0	24	0
		Indicateur global						0,32

TABLEAU 6-5 : La matrice de confusion des motifs caractéristiques en utilisant l'étiquetage issue du *Kmeans*, $s_{min} = 0{,}1$ et $c_{min} = 0{,}1$.

Nous pouvons aussi remarquer que l'algorithme des motifs caractéristiques fonctionne d'autant mieux que les traces étiquetées manuellement le sont de manière certaine. En effet, les meilleurs résultats ont été obtenus pour les traces de la classe 1 et 4. Or ce sont les deux

classes qui ont été étiquetées avec une confiance élevée (*Cf.* le tableau 6-3). Nous pouvons expliquer ceci par le fait que les personnes qui ont étiquetées les traces ont d'autant plus confiance dans leur étiquetage lorsqu'ils ont déjà rencontré les mêmes documents au sein des traces, c'est-à-dire qu'un document devient caractéristique de cette classe.

6.4 Conclusion

La projection *t-SNE* des traces d'apprentissage (en utilisant notre dissimilarité entre traces et la dissimilarité sémantique entre les documents) montre que ces traces peuvent être regroupées en cinq *clusters*. Dans ce chapitre, l'algorithme d'apprentissage non supervisé *Kmeans* a été utilisé pour donner des nouveaux labels à ces traces. Nous avons vérifié que les nouvelles classes identifiées par le *Kmeans* représentent bien toutes les traces.

Pour valider notre approche, l'utilisation de cette méthode d'étiquetage des traces de test n'est pas envisageable. Nous avons alors analysé des exemples de traces d'apprentissage en étudiant les documents visités. Les centres d'intérêt de chaque classe ont alors été extraits. Pour étiqueter les traces de test, nous les avons analysées afin de trouver la classe dont les centres d'intérêt sont les plus proches.

La validation visuelle de notre approche a été faite. Les résultats ont montré que les traces d'apprentissage et de test de même classe apparaissent ensemble avec peu d'erreurs. L'utilisation du SVM nous a permis de valider statistiquement notre proposition. L'indicateur global du taux de reconnaissance est égal à $0{,}82$. Par contre, cet indicateur est égal à $0{,}32$ pour les motifs caractéristiques.

À partir de ces résultats, nous pouvons conclure par le fait que notre méthode fonctionne bien dans les cas où l'algorithmique des motifs caractéristiques ne fonctionne pas du tout, comme par exemple pour les classes 2 et 5. Et elle donne aussi de meilleurs résultats dans le cas ou l'algorithme des motifs caractéristiques fonctionne, comme c'est le cas par exemple pour les classes 1 et 4.

Conclusion et perspectives

Conclusion

Les hypermédias d'aujourd'hui, de par la richesse de leurs contenus, posent le problème de perte d'attention de leurs utilisateurs. Les hypermédias adaptatifs sont une des solutions à ce problème, mais ils nécessitent l'identification du profil des utilisateurs. Les travaux de cette thèse s'inscrivent dans cette démarche.

Nous avons commencé par présenter la structure des systèmes experts, leurs domaines d'application et leurs limitations. Deux types de moteurs d'inférence ont été exposés (le raisonnement guidé par le but et le raisonnement guidé par les données). Puis, nous avons présenté les réseaux bayésiens qui sont des systèmes experts probabilistes.

Ensuite, nous avons dressé un panorama des méthodes d'apprentissage supervisé et non supervisé. Les notions de distance, de dissimilarité et de similarité ont été définies. Nous nous sommes attardés sur les algorithmes *Kmeans* et les cartes de Kohonen pour les classifications non supervisées, et les algorithmes *kppv* et SVM pour les supervisées. Ces deux dernières méthodes d'apprentissage sont des méthodes à noyaux, c'est-à-dire qu'elles ont uniquement besoin d'une mesure de similarité pour fonctionner. Nous avons aussi présenté des techniques de projection de données, entre autres l'algorithme *t-SNE* qui nous a permis d'évaluer visuellement la qualité des similarités.

Puis, nous nous sommes attardés sur les trois méthodes principales utilisées dans *web usage mining* (WUM) pour découvrir le modèle des utilisateurs d'hypermédia. Nous avons décrit le site Hypergéo qui est l'hypermédia sur lequel nous avons travaillé. C'est une encyclopédie géographique en trois langues ouverte à tout public. La taille des traces que nous avons recueillies n'étant pas grande, nous avons dû exclure l'utilisation de l'algorithme des motifs séquentiels. De plus, puisque nous les avions préalablement pré-étiquetées, nous n'avons pas utilisé les méthodes de *clustering*. Nous avons choisi un algorithme de la famille de la fouille des règles d'association : l'algorithme des motifs caractéristiques. En appliquant cet algorithme sur nos données expérimentales, nous nous sommes confrontés à une de ses limites : cet algorithme exige un grand nombre de traces pour bien fonctionner. De plus les éléments de ces traces doivent être assez caractéristiques de chaque profil. C'est la raison pour laquelle, ces techniques sont plus adaptées à identifier un utilisateur plutôt qu'à identifier leurs profils.

Nous avons donc proposé d'utiliser les algorithmes de classification à la place des motifs caractéristiques. Ceci nous a amenés à définir une dissimilarité entre les traces des utilisateurs. Deux dissimilarités entre traces ont donc été définies et testées successivement en utilisant une dissimilarité entre les éléments des traces. Nous avons construit des bases de traces synthétiques pour connaître les conditions d'utilisation optimale de ces dissimilarités. Dans notre contexte expérimentale, comme nous l'avons dit, les traces sont courtes. Dès lors nous ne pouvions

pas représenter les éléments de ses traces par des transitions, ce qui les auraient raccourcis d'autant. Nous avons donc dû les représenter à l'aide des documents visités. Encore une fois, deux dissimilarités entre documents ont été adaptées et testées : la dissimilarité hiérarchique et la dissimilarité sémantique. Bien que les résultats de l'application de nos propositions soient supérieurs à ceux des motifs caractéristiques, ils sont tout de même insatisfaisants.

En utilisant notre proposition de dissimilarité entre traces, nous avons remarqués que leurs projections ont formé des groupes assez distincts. Nous les avons étudiés pour leur donner un sens. Nous avons alors remarqué que pour chaque groupe, il existait des centres d'intérêt communs. Ces groupes peuvent donc former une nouvelle façon de dresser des profils utilisateurs. Nous avons alors utilisé l'algorithme *Kmeans* pour ré-étiqueter les données d'apprentissage. Les données de test quant à elles ont été ré-étiquetées manuellement. Les résultats du SVM pour identifier les classes des nouvelles traces à partir de leurs centres d'intérêt sont très bons, nettement supérieurs à ceux obtenus avec l'algorithme des motifs caractéristiques.

Il est à noter que notre démarche scientifique est basée sur l'itération de plusieurs cycles « hypothèse, expérimentation et validation ». En effet à chaque fois que nous avons formalisé une idée, nous nous sommes attachés à la valider expérimentalement à l'aide de données synthétiques ou de données réelles. Les résultats de ces expérimentations nous ont alors permis de confirmer ou d'infirmer les hypothèses proposées.

Contributions

Dans ces travaux nous avons mis en évidence certains cas où les techniques de *Web Usage Mining* ne fonctionnent pas. Une nouvelle approche basée sur l'utilisation conjointe d'algorithme de classification à noyau et d'une dissimilarité entre les traces utilisateurs a été proposée. Cette dissimilarité utilise les dissimilarités entre éléments des traces, mais elle est générique et indépendante des types de ses éléments (transitions, documents, etc.).

Nous avons aussi utilisé régulièrement la projection *t-SNE*. Tout d'abord elle nous a servi à tester la qualité des dissimilarités avant d'appliquer les algorithmes de classification. Cette projection, qui est plus rapide que la classification, nous a permis de faire des suppositions quant aux futurs résultats des classifications. Ensuite elle nous a aidé à interpréter des données. Par exemple à partir des projections, nous avons validé l'utilisation du *tf-idf* pour mesurer la distance entre les articles d'Hypergéo, et nous avons pu identifier le nombre de groupes que formaient les traces Hypergéo selon notre dissimilarité sémantique.

Nous avons aussi identifié les utilisateurs d'Hypergéo selon leurs centres d'intérêt. Cela a été possible en prenant en compte en plus des documents visités leurs contenus et la structure de l'hypermédia.

Limitations

Nous avons proposé une dissimilarité entre traces indépendante de celle entre éléments. Mais la nature de cette dernière (hiérarchique, sémantique, etc.) influence la dissimilarité entre traces. La dissimilarité entre éléments que nous avons choisie permet de mettre en évidence le thème de la recherche des utilisateurs, mais pas le niveau de connaissance ou le domaine de compétence. C'est pourquoi notre dissimilarité entre traces ne fonctionne pas bien sur les profils prédéfinis. Si nous avions défini une dissimilarité entre éléments basée sur le niveau de connaissance ou de compétence, peut-être aurions nous pu identifier les traces avec les profils définis *a priori*.

Une autre limite concerne l'utilisation de la dissimilarité sémantique. Dans cette dissimilarité nous sommes obligés de supprimer les rubriques et les documents des langues étrangères. Par conséquent, nous avons perdu beaucoup d'informations qui auraient pu être utile pour mieux classer les traces.

Enfin, nous avons montré que notre solution fonctionne mieux que les techniques du WUM car nous prenons en compte le contenu des éléments des traces. Cet apport est coûteux en temps mais permet de travailler sur des volumes plus faibles, ce que le WUM ne sait pas faire. Se pose alors la question du seuil concernant le volume de la base d'apprentissage à partir duquel notre proposition n'est plus aussi bénéfique.

Perspectives

Différentes perspectives de travail sont alors envisageables.

Tout d'abord, des améliorations pourraient être menées concernant l'étiquetage des traces. Par exemple, l'analyse des traces n'ayant pas été effectuée par des experts en géographie, il se peut que certains thèmes n'aient pas été bien appréhendés. Un ré-étiquetage pourrait améliorer les résultats. De plus, comme nous pouvons le constater dans la projection des traces d'apprentissage (*Cf.* la figure 6-1 page 140), le choix de cinq groupes de profils de traces pourraient être affiné. En effet, l'analyse du rôle de ces traces a montré que dans chaque profil un ou deux thèmes émergeaient. Par exemple pour le premier profil, nous avons identifié les thèmes « cartographie » et « outils mathématiques ». Or ceci transparaît en partie dans la projection, puisque le groupe de traces de la première classe peut lui même être divisé en sous-groupes (au moins sept). La difficulté réside ici à ne pas trop sectoriser les classes afin d'avoir des traces toujours homogènes sémantiquement.

Ensuite, notre travail a pour objectif de définir le profil des utilisateurs afin d'adapter le site Hypergéo. Avec nos résultats, plusieurs types d'adaptations sont possibles comme par exemple

proposer des recommandations de documents à visiter ou changer les liens hypertextes au sein des documents. Dans le premier cas, nous pouvons construire des listes de liens hypertextes vers des documents dont les centres d'intérêt appartiennent au profil de l'utilisateur courant. Dans le deuxième cas, au lieu de générer les liens entre documents automatiquement quelque soit l'utilisateur d'Hypergéo [1], nous proposons ici d'ajouter un lien hypertexte uniquement lorsque le document cible est un document qui a déjà été visité par les utilisateurs du même profil. Pour évaluer la pertinence de l'adaptation et donc indirectement la qualité de notre identification de profil, deux méthodes pourraient être adoptées. Dans le premier cas, n'ayant aucun élément de comparaison, puisqu'une liste de recommandation serait ajoutée, nous validerions l'intérêt de cette liste par son utilisation. Dans le deuxième cas, les liens hypertextes entre documents sont présents par défaut. Une comparaison statistiques du nombre de fois où les liens seraient utilisés, avant et après l'adaptation, permettrait d'évaluer notre approche.

Nous pourrions aussi travailler sur notre dissimilarité entre traces. En effet, les données que nous avons pour identifier les utilisateurs sont uniquement les documents visités. Or obtenir d'autres informations sur l'utilisateur pourraient certainement nous permettre d'améliorer la capacité du système à identifier son profil. Par exemple, à l'aide des adresses IP, il serait possible d'obtenir le pays d'où provient la requête et ainsi en déduire des informations quant au niveau de connaissance du lecteur (un utilisateur lisant un texte dans une langue étrangère est certainement plus avancé scolairement). De la même manière, le temps de visualisation d'un document peut renseigner quant au comportement de l'utilisateur [BOU 11]. Tous ces indicateurs pourraient être utilisées, mais encore faudrait il déterminer comment les insérer dans la mesure de dissimilarité entre traces.

Toujours au niveau de la dissimilarité entre traces, dans ce mémoire nous avons utilisé soit la dissimilarité hiérarchique soit la dissimilarité sémantique. Bien que ce soit la dernière que nous ayons privilégié, la dissimilarité hierarchique a aussi ses avantages (elle utilise les traces dans leur ensemble, et pas uniquement les articles), et elle pourrait mieux être prise en compte. Il faudrait par exemple étudier une nouvelle dissimilarité utilisant une moyenne pondérée des deux précédentes et voir ainsi si de nouveaux profils d'utilisateur émergent.

Un travail important pourrait aussi être mené sur la mesure de dissimilarités entre articles. Tout d'abord nous avons utilisé la dissimilarité *tf-idf*, or cela peut poser quelque fois des problèmes. Par exemple le terme de « flux » est assez utilisé dans hypergéo [2] pour représenter

1. À l'heure ou nous écrivons ce mémoire, la technique permettant d'interconnecter automatiquement deux articles d'Hypergéo n'est plus active sur la version en ligne. Le principe de cette technique était, pour l'article demandée, d'ajouter des liens vers des articles dont le titre était présent dans son contenu. La détection de ces titres était réalisée à l'aide d'expressions rationnelles.

2. La requète google `flux site:hypergeo.eu` permet de voir que le terme flux est présent dans soixante trois articles.

« les flux migratoires », « les flux de véhicules », « les flux d'information », « les flux de marchandise », « les flux hydrauliques », etc. Il faudrait donc essayer de représenter les documents par des concepts à la place des termes, en utilisant par exemple le *Latent Semantic Analysis* [Lan 98]. Nous pourrions aussi améliorer cette dissimilarité en prenant en compte les articles de différentes langues. On pourrait alors s'inspirer des travaux effectués récemment sur l'*alignement mutlilingue de corpus comparables spécialisés* (*Cf.* [PRO 10]). Puis comme nous le disions précédemment, la sémantique de notre dissimilarité entre traces étant dépendante de la sémantique de la dissimilarité entre documents, en créant une dissimilarité entre documents mettant en exergue le niveau de langue, nous pourrions, peut être, être en mesure de classer les utilisateurs en fonction de leur niveau de connaissance.

Pour finir, ce travail de thèse a été validé expérimentalement sur les données synthétiques et sur les données réelles d'Hypergéo. L'application de nos propositions sur un autre hypermédia permettrait de vérifier que notre approche est généralisable.

Appendices

ANNEXE A

Duplication du site Hypergéo sur les serveurs de l'INSA de Rouen

Dans cette annexe nous présentons la démarche que nous avons suivie pour pouvoir recueillir les traces étiquetées des utilisateurs d'Hypergéo. Nous allons expliquer les procédures que nous avons suivi pour modifier le site *www.hypergeo.eu* de manière à envoyer toutes les requêtes vers le site *hypergeotraces.insa-rouen.fr* pour enregistrer les traces des utilisateurs. Ce site est une copie du site *www.hypergeo.eu*. Avant mise en production, nous avons validé notre démarche à l'aide d'un serveur de test, le site *hypergeotestredirection.insa-rouen.fr*.

A.1 Construction du site *hypergeotraces.insa-rouen.fr* et envoie des requêtes

Nous avons commencé par copier toutes les informations du site *http ://www.hypergeo.eu* sur un serveur de l'INSA, sur lequel nous avions installé à l'époque la dernière mise à jour à la version SPIP, la 1.92. À l'aide du pare-feu de l'INSA, nous avons rendu le site *hypergeotraces.insa-rouen.fr* invisible pour toutes les requêtes issues d'internet sauf celles qui provenait du site *www.hypergeo.eu*. Nous avons modifié certains fichiers de SPIP, entre autres le formulaire de recherche sur les fichiers suivants :

- `aide.html`.
- `article.html`.
- `auteur-list.html`.
- `menu.html`.
- `plan.html`.
- `recherche.html`.
- `resume.html`.
- `rubrique_bug.html`.
- `rubrique.html`.
- `sommaire.html`.

Ceci a été effectué en insérant le code suivant dans chacun de ces fichiers.

```
<!-- Formulaire de recherche -->
        <div class="formrecherche" align="right">
                <form action="spip.php?page=recherche" method="get">
                <input name='page' value='recherche' type='hidden' />
                 Rechercher <input type="text" name="recherche" size="20"
                    maxlength="50"><input type="submit" value="ok">
                </form>
        </div>
```

A.2 Modification du site *www.hypergeo.eu*

Nous avons modifié le site *www.hypergeo.eu* pour laisser passer les requêtes qui provenaient des robots des moteurs de recherches. En revanche, nous avons renvoyé toutes les autres requêtes vers le site *hypergeotraces.insa-rouen.fr* de manière transparente (l'utilisateur ne s'en apercevait pas). Pour réaliser cela, nous avons utilisé la classe HTTP_Request. Sur le site *www.hypergeo.eu* , nous avons créé un répertoire appelé `"changeproxy"` dans lequel nous avons copié plusieurs fichiers PHP permettant de réaliser un proxy web (les fichiers `(Request.php, Socket.php, URL.php, classWebPage.php, PEAR.php, config.php)`). Dès lors il nous a fallu modifier le fichier `config.php` de façon à définir la redirection :

```
$server ="http ://www.hypergeo.eu" ;
$redirectIP = "http ://hypergeotraces.insa-rouen.fr/hypergeotraces" ;
```

De façon à rediriger les requêtes vers l'INSA et à laisser passer les robots des moteurs de recherche, dans le répertoire `"changeproxy"`, nous avons créé les fichiers `(page.php, modifierlien.php)` de la manière suivante :

Le fichier `page.php`

```
<?
if (!(eregi ("Googlebot|bot|msnbot|Slurp|VoilaBot|Gigabot|Openbot|Lycos",
   $_SERVER["HTTP_USER_AGENT"])))
        {
                include ("modifierlien.php");
                include("config.php");
                require_once('Request.php');
                $newaddress=$redirectIP;
```

```
$adresspage=  $_SERVER["REQUEST_URI"];
$adresspage= ereg_replace("hypergeotestredirection/","",
    $adresspage);
$comp="/article_pdf/i";
if (!(preg_match ($comp,$adresspage)))
{
        $comp= "/spip.php?/i";
        if (!(preg_match ($comp,$adresspage))) $adresspage
            =trouve_page($adresspage);
        $newaddress .= $adresspage;
        $req = &new HTTP_Request($newaddress);

        if (isset($_COOKIE["visiteurhypergeo"])){
                $req->addCookie("visiteurhypergeo",
                    $_COOKIE["visiteurhypergeo"]);
        }
        $req->setMethod(HTTP_REQUEST_METHOD_GET);
        $req->sendRequest();
        $response = $req->getResponseBody();
        $response= ereg_replace("trace_hypergeo","hypergeo"
            ,$response);
        header('Content-Type: text/html; charset=iso-8859-1
            ');
        echo $response;
        exit();
}
else
{
$comp="/article_pdf.php3/i";
$oldadress=$_SERVER["REQUEST_URI"];
if (!(preg_match ($comp,$oldadress)))
        {
        $newaddress .= $adresspage;
        $req = &new HTTP_Request($newaddress);
        if (isset($_COOKIE["visiteurhypergeo"])){
                $req->addCookie("visiteurhypergeo",
                    $_COOKIE["visiteurhypergeo"]);
        }
        $req->setMethod(HTTP_REQUEST_METHOD_GET);
        $req->sendRequest();
        $adresspage= ereg_replace("spip.php\?page=
            article_pdf&","article_pdf.php3?",$adresspage);
```

```
                        $adresspage=$server . $adresspage;
                        header('Content-Type: text/html; charset=iso-8859-1
                            ');
                        header("Location: ".$adresspage);
                        exit();
                        }
                }

        }

?>
```

Le fichier `modifierlien.php`

```
<?
        function trouve_page($current_adresse)
        {
                $current_page="/";
                if (preg_match ("/_article_PDF/i",$current_adresse))
                        $current_page= $current_adresse;
                elseif (preg_match ("/article/i",$current_adresse))
                        $current_page= ereg_replace("/article.php3\?
                            id_article=","/spip.php?article",
                            $current_adresse);
                elseif (preg_match ("/rubrique/i",$current_adresse))
                        $current_page= ereg_replace("/rubrique.php3\?
                            id_rubrique=","/spip.php?rubrique",
                            $current_adresse);
                elseif (preg_match ("/aide/i",$current_adresse))
                        $current_page= ereg_replace("/aide.html","/spip.
                            php?page=aide",$current_adresse);
                elseif (preg_match ("/auteur-list/i",$current_adresse))
                        $current_page= ereg_replace("/auteur-list.php3","/
                            spip.php?page=auteur-list",$current_adresse);
                elseif (preg_match ("/imprimersans/i",$current_adresse))
                        $current_page= ereg_replace("/imprimersans.php3\?
                            id_article=","/spip.php?page=imprimersans&
                            id_article=",$current_adresse);
                elseif (preg_match ("/plan/i",$current_adresse))
                        $current_page= ereg_replace("/plan.php3","/spip.
                            php?page=plan",$current_adresse);
                elseif (preg_match ("/resume/i",$current_adresse))
```

```
                $current_page= ereg_replace("/resume.php3","/spip.
                    php?page=resume",$current_adresse);
            elseif (preg_match ("/mot/i",$current_adresse))
                $current_page= ereg_replace("/mot.php3\?id_mot=","
                    /spip.php?mot",$current_adresse);
            elseif (preg_match ("/sommaire/i",$current_adresse))
                $current_page= ereg_replace("/sommaire.php3","/
                    spip.php?page=sommaire",$current_adresse);
            return $current_page;
        }
?>
```

Enfin, à la racine du site *www.hypergeo.eu*, nous avons redéfinis la page d'accueil (`bienvenue.php`) et le fichiers général de SPIP (`spip.php`).

Le fichier `bienvenue.php`

```
<?php
include_once("changeproxy/config.php");
require_once('changeproxy/Request.php');
$adresspage="http://hypergeotraces.insa-rouen.fr/hypergeotraces/bienvenue.
    php";
$req = &new HTTP_Request($adresspage);
$nametype="type";
$valuetype=$_POST["type"];
$req->addPostData($nametype,$valuetype);
$namearrivee="arrivee";
$valuearrivee=$_POST["arrivee"];
$req->addPostData($namearrivee,$valuearrivee);
$req->setMethod(HTTP_REQUEST_METHOD_POST);
$req->sendRequest();
$valeur_cooki=$req->getResponseCookies ();
$valuearrivee= ereg_replace("/hypergeotraces","/hypergeotestredirection",
    $valuearrivee);
setcookie("visiteurhypergeo",$valeur_cooki[0]["value"],time()+3600);
header('Content-Type: text/html; charset=iso-8859-1');
header("Location: ".$valuearrivee);
exit();
?>
```

Le fichier `spip.php`

```
<?php
include ("inc-public.php3");
?>
```

Enfin, nous ajoutons la ligne suivante au fichier `mes_fonctions.php3`

```
include ("changeproxy/page.php");
```

A.3 Rétablissement du site *www.hypergeo.eu* dans son état original

Pour remettre le site *www.hypergeo.eu* dans son état initial :

- Dans le fichier `mes_fonctions.php3`, il a fallu supprimer la ligne

```
include ("changeproxy/page.php");
```

- De la racine du site *www.hypergeo.eu*, nous avons supprimé les fichiers `spip.php` et `bienvenue.php`.
- De la racine du site *www.hypergeo.eu*, nous avons supprimé le répertoire `"changeproxy"`.

ANNEXE B

Articles visités par des utilisateurs de site Hypergéo

TABLEAU B-1 : Exemple d'articles d'Hypergéo visités par des utilisateurs de la base d'apprentissage (après application du filtre).

$N°$	**Articles visités**
1	« Organisation de l'espace », « Spatialite », « Centre Peripherie », « Territoire »
2	« Contrainte », « La region comme systeme »
3	« Carte choroplethe »
4	« Constructivisme », « Paysage »
5	« Polarisation »
6	« Latitude », « TOPOGRAPHIE », « Orient », « Meridien », « Topographie (historique de la notion) », « EST »
7	« Echelle »
8	« Mondialisation », « Actualisme »
9	« Densite », « Oekoumene »
10	« Representation », « Region », « Pays »
11	« Centre Peripherie », « Paysage »
12	« Gradient », « Espace transfrontalier »
13	« Configuration »
14	« Interaction spatiale », « Discontinuite », « Centre Peripherie », « Puissance »
15	« Distance », « Reseau », « Equite territoriale », « Site », « Lieu », « Voisinage »
16	« Analyse spatiale »
17	« Historique du territoire »
18	« Biocenose », « Region »
19	« Historique du paysage », « Paysage selon le laboratoire THEMA », « Statut spatial du paysage », « Paysage »
20	« CARTOGRAPHIES », « carte », « Centre Peripherie », « CARTOGRAPHIE THEMATIQUE »
21	« TOPOGRAPHIE », « Densite », « Historique du territoire », « carte »
22	« Representation »

suite sur la page suivante

$N°$	**Articles visités**
23	« Constructivisme »
24	« CARTOGRAPHIES », « CARTOGRAPHIE THEMATIQUE »
25	« Aire d'influence », « Theories de l'Analyse Spatiale », « Modele gravitaire », « Centralite »
26	« Theories de l'Analyse Spatiale », « Representation », « La region espace vecu », « Region polarisee », « Territoire »
27	« Hierarchie »
28	« Espace transfrontalier », « Dispute territoriale », « Ville frontaliere », « Structure spatiale », « Geopolitique », « Polarisation », « Communautes fermees (gated communities) », « Evolution des entites geographiques », « Discontinuite »
29	« CARTOGRAPHIE THEMATIQUE », « Figuration cartographique »
30	« Equite territoriale »
31	« Carte choroplethe »
32	« La dynamique des systemes selon J.W. Forrester », « La region comme systeme »
33	« Metropolisation »
34	« Terrain », « Carte choroplethe », « CARTOGRAPHIES »
35	« Constructivisme »
36	« Organisation de l'espace », « Oekoumene »
37	« Production du paysage », « Statut spatial du paysage », « Geosysteme »
38	« Historique du paysage »
39	« Geosysteme »
40	« Segregation », « Dynamique des versants »
41	« Communaute transnationale », « Diaspora »
42	« Mondialisation »
43	« Nature et culture »
44	« Organisation de l'espace », « Oekoumene »
45	« Interaction spatiale », « Complexite », « Theories de l'Analyse Spatiale », « Analyse spatiale », « Nomothetisme », « Spatialite »
46	« Graphe », « Analyse spatiale », « Nomothetisme », « Reseau », « Modele »
47	« Paysage selon le laboratoire THEMA », « Paysage »
48	« Reseau », « Systeme », « Structure spatiale », « Lieux centraux »
49	« Organisation de l'espace », « Analyse spatiale »
50	« Nature et culture », « La region espace vecu », « Christaller (modele) »
51	« La region espace vecu », « Region »
52	« Configuration », « Theories de l'Analyse Spatiale », « Analyse spatiale », « Nomothetisme »
53	« Paysage »
54	« Interaction spatiale », « Organisation de l'espace », « Mondialisation », « Bifurcation », « Centre Periphe-rie »

suite sur la page suivante

$N°$	**Articles visités**
55	« Interaction spatiale », « Determinisme », « La region espace vecu », « Lieu »
56	« Interaction spatiale »
57	« Theories de l'Analyse Spatiale », « Mondialisation »
58	« Graphe », « Hierarchie », « Reseau », « Connectivite », « Connexite2 », « Accessibilite »
59	« Terrain », « Ethnie »
60	« Dispute territoriale », « Ville frontaliere », « Discontinuite », « Le territoire selon Claude Raffestin », « Discontinuite »
61	« Representation »
62	« Centralite », « Lieux centraux »
63	« Agregation », « Distance », « Equite territoriale »
64	« Biome »
65	« Discretisation », « Variables quantitatives », « CARTOGRAPHIE THEMATIQUE »
66	« carte »
67	« Carte choroplethe », « Statistique spatiale », « carte »
68	« TOPOGRAPHIE », « CARTOGRAPHIES », « Figuration cartographique »
69	« Systeme spatial »
70	« Connectivite », « Accessibilite », « Metropolisation »
71	« La region comme systeme », « Region naturelle »
72	« Ecotone », « Mondialisation », « Lieux centraux », « Christaller (modele) »
73	« Auto organisation », « Configuration », « Organisation de l'espace », « Spatialite », « Centre Peripherie », « Processus »
74	« Constructivisme »
75	« Interaction spatiale », « Theories de l'Analyse Spatiale », « Equite territoriale », « Diffusion spatiale », « Accessibilite »
76	« Theories de l'Analyse Spatiale », « Biocenose », « Centre Peripherie », « Diffusion spatiale », « Lieux centraux »
77	« Etagement »
78	« Theories de l'Analyse Spatiale », « Nomothetisme »
79	« Densite », « Equite territoriale »
80	« Connexite », « Graphe », « Interaction spatiale », « Connectivite », « Connexite2 », « Accessibilite »
81	« Carte choroplethe », « Vallee »
82	« Representation », « Historique du territoire », « Perception des paysages »
83	« Constructivisme »
84	« Postmodernisme », « Region naturelle », « Constructivisme »
85	« Theories de l'Analyse Spatiale », « Modele », « Polarisation », « Centre Peripherie », « Interaction »
86	« Vulnerabilite », « Cindynique »
87	« Segregation », « Communautes fermees (gated communities) »
88	« Frontiere »

suite sur la page suivante

N°	**Articles visités**
89	« Contrainte », « Bilan hydrique »
90	« Segregation »
91	« Dispute territoriale », « Geopolitique »
92	« Centralite », « Christaller (modele) », « Metropolisation »
93	« Orient », « Equite territoriale »
94	« carte »
95	« Representation », « Echelle », « Coupe (Transect) »
96	« Centre Peripherie »
97	« Historique du paysage », « Utilisations des paysages », « Geosysteme »
98	« Christaller (modele) »
99	« Modele gravitaire », « Loi »
100	« Oekoumene », « Region naturelle »
101	« Littoral »
102	« Front pionnier », « Anthropisation », « Oekoumene »
103	« Insularite », « Littoral »
104	« Theories de l'Analyse Spatiale », « Localisation », « Echelle »
105	« Geopolitique », « Equite territoriale »
106	« Puissance »
107	« Theories de l'Analyse Spatiale », « Christaller (modele) », « Lieux centraux »
108	« Mondialisation »
109	« Etagement », « Coupe (Transect) »
110	« Bilan hydrique »
111	« Anthropisation », « Geochronologie »
112	« Contingence »
113	« Historique du territoire »
114	« Littoral »
115	« Seuil »
116	« Centre Peripherie », « La dynamique des systemes selon J.W. Forrester », « Lieux centraux »
117	« Mondialisation »
118	« Fixisme Mobilisme »
119	« Longitude », « Insularite », « Ile »
120	« Centre Peripherie »
121	« Attraction Attractivite », « Communautes fermees (gated communities) »
122	« CARTOGRAPHIE THEMATIQUE », « Figuration cartographique »
123	« Constructivisme »
124	« Frontiere », « Mondialisation », « Evolution des entites geographiques »
125	« Auto organisation », « Evolution des entites geographiques »
126	« Complexite »

suite sur la page suivante

N°	**Articles visités**
127	« Analyse spatiale »
128	« Auto organisation », « Centre Peripherie », « Evolution des entites geographiques »
129	« Anthroposysteme »
130	« Nature et culture »
131	« Paysage selon le laboratoire THEMA »
132	« Front pionnier », « Frontiere », « Dispute territoriale »
133	« Frontiere », « Pays »
134	« Ethnie », « Evolution des entites geographiques »
135	« Terrain », « Echelle »
136	« Lieux centraux »
137	« Mondialisation »
138	« Evolution des entites geographiques »
139	« Interaction spatiale », « Complexite », « Theories de l'Analyse Spatiale », « Modele gravitaire »
140	« La trajection paysagere »
141	« Theories de l'Analyse Spatiale », « Geographie », « Centre Peripherie »
142	« Christaller (modele) »
143	« Mondialisation »
144	« Metropolisation »
145	« carte », « Oekoumene »
146	« Centralite »
147	« Front pionnier »
148	« Statistique spatiale », « carte », « Geographie », « Systeme d'information geographique (S.I.G) »
149	« Front pionnier », « Equite territoriale », « Lieux centraux », « Christaller (modele) »
150	« Discontinuite »
151	« Front pionnier »
152	« Geosysteme »
153	« Carte choroplethe », « Variables quantitatives », « CARTOGRAPHIE THEMATIQUE »
154	« Christaller (modele) »
155	« Mondialisation », « Processus »
156	« Organisation de l'espace », « Geographie »
157	« Etagement », « Site »
158	« Discretisation »
159	« Seuil »
160	« Mondialisation », « Savoirs vernaculaires »
161	« Seuil »
162	« Geosysteme », « Paysage »
163	« Paysage »
164	« Le territoire selon Claude Raffestin »

suite sur la page suivante

N°	**Articles visités**
165	« Contrainte »
166	« Carte choroplethe », « Discretisation », « Analyse de donnees »
167	« Geochronologie »
168	« Modele »
169	« Representation »
170	« Latitude », « Longitude »
171	« Region polarisee »
172	« Nord »
173	« Representation »
174	« Theories de l'Analyse Spatiale », « Localisation »
175	« Region polarisee »
176	« Contingence », « Constructivisme »
177	« TOPOGRAPHIE », « Topographie (historique de la notion) »
178	« Realisme »
179	« Discretisation »
180	« Mondialisation », « Geographie », « Centralite », « Lieux centraux »
181	« La region espace vecu », « Region polarisee », « Lieux centraux »
182	« Theories de l'Analyse Spatiale », « Distance », « Connectivite », « Lieux centraux »
183	« Situation », « TOPOGRAPHIE », « Longitude », « CARTOGRAPHIES », « Meridien »
184	« Echelle »
185	« Communautes fermees (gated communities) »
186	« Geographicite », « Fondements epistemologiques »
187	« Ville frontaliere »
188	« Meridien »
189	« Spatialite », « Nature et culture », « La region espace vecu »
190	« Equite territoriale »
191	« Discretisation », « CARTOGRAPHIE THEMATIQUE »
192	« Discretisation », « CARTOGRAPHIE THEMATIQUE »
193	« Distance »
194	« Discontinuite »
195	« Aire d'influence », « Christaller (modele) »
196	« Representation », « Communaute transnationale », « Diaspora »
197	« Geographie », « Localisation »
198	« Agregation », « carte », « CARTOGRAPHIE THEMATIQUE », « Figuration cartographique »
199	« carte », « Teledetection »
200	« Systeme spatial », « Organisation de l'espace », « Centre Peripherie »
201	« Region polarisee »
202	« Configuration », « Organisation de l'espace »

suite sur la page suivante

$N°$	**Articles visités**
203	« Latitude », « carte »
204	« TOPOGRAPHIE »
205	« Oekoumene »
206	« CARTOGRAPHIES », « Topographie (historique de la notion) »
207	« Constructivisme »
208	« Statistique spatiale », « Variables qualitatives »
209	« Carte choroplethe », « Discretisation »
210	« carte », « CARTOGRAPHIE THEMATIQUE »
211	« Seuil »
212	« Diffusion spatiale »
213	« Segregation », « Meridien »
214	« Historique du paysage »
215	« CARTOGRAPHIES »
216	« Frontiere »
217	« Aire d'influence », « Metropolisation »
218	« Metropolisation »
219	« Carte choroplethe », « Analyse de donnees », « Statistique spatiale »
220	« Carte choroplethe », « Mondialisation »
221	« Analyse spatiale »
222	« Terrain », « Metropolisation »
223	« Carte choroplethe », « Discretisation »
224	« Mondialisation », « Teledetection »
225	« Bilan hydrique »
226	« Autocorrelation », « Diffusion spatiale », « Voisinage »
227	« Frontiere »
228	« TOPOGRAPHIE », « Variables quantitatives », « CARTOGRAPHIE THEMATIQUE », « Figuration cartographique »
229	« Mondialisation », « Connectivite », « Metropolisation »
230	« Biotope », « Biocenose »
231	« Mondialisation »
232	« Insularite », « Ile »
233	« Aire d'influence »
234	« Historique du territoire », « Mondialisation »
235	« Geomatique », « Geographie »
236	« Coupe (Transect) »
237	« Mondialisation », « Modele gravitaire »
238	« Mondialisation »
239	« Anthroposysteme »

suite sur la page suivante

N°	**Articles visités**
240	« Mondialisation »
241	« Segregation », « Equite territoriale »
242	« carte », « CARTOGRAPHIE THEMATIQUE », « Figuration cartographique »
243	« Paysage selon le laboratoire THEMA »
244	« Carte choroplethe », « CARTOGRAPHIE THEMATIQUE », « Variables quantitatives »
245	« Geographie »
246	« Mondialisation »
247	« Seuil »
248	« Mondialisation », « Centre Peripherie »
249	« Carte choroplethe », « CARTOGRAPHIES », « carte », « Figuration cartographique »
250	« Front pionnier », « Diffusion spatiale »
251	« Front pionnier »
252	« Analyse spatiale »
253	« Mondialisation »
254	« Discretisation », « Variables quantitatives »
255	« Etagement », « Coupe (Transect) »
256	« CARTOGRAPHIE THEMATIQUE »
257	« Structure spatiale », « Mondialisation »
258	« Aire de marche », « Communautes fermees (gated communities) », « Agglomeration »
259	« Densite »
260	« Risque »
261	« Carte choroplethe », « CARTOGRAPHIE THEMATIQUE »
262	« Le territoire selon Claude Raffestin », « Mondialisation »
263	« Carte choroplethe »
264	« Historique du territoire »
265	« Latitude », « TOPOGRAPHIE », « Longitude »
266	« Perception des paysages »
267	« Auto organisation », « Organisation de l'espace », « Attraction Attractivite »
268	« Topographie (historique de la notion) »
269	« Geographie », « Christaller (modele) »
270	« Geosysteme »
271	« Organisation de l'espace »
272	« Anthropisation », « La trajection paysagere »
273	« Theories de l'Analyse Spatiale », « Christaller (modele) », « Lieux centraux »
274	« Systeme d'information geographique (S.I.G) », « CARTOGRAPHIE THEMATIQUE »
275	« Equite territoriale »
276	« Mondialisation », « Equite territoriale »
277	« Paysage »

suite sur la page suivante

$N°$	**Articles visités**
278	« Equite territoriale »
279	« Metropolisation »
280	« Dynamique des versants », « Evolution des entites geographiques »
281	« Domestication Neolithisation »
282	« Insularite », « Ile »
283	« Insularite »
284	« Nature et culture », « Actualisme »
285	« Bilan hydrique »
286	« Contrainte », « Bilan hydrique »
287	« Systeme morphogenetique »
288	« Equite territoriale », « Centre Peripherie »
289	« Configuration »
290	« Terrain », « Cognition »
291	« Aire d'influence », « Polarisation », « Region polarisee »
292	« Diaspora »
293	« Estuaire »
294	« Resilience », « Hydrosysteme »
295	« Aire d'influence », « Hierarchie », « Lieux centraux »
296	« Puissance », « Lieux centraux », « EST »
297	« Littoral », « Etagement », « Echelle »
298	« Discretisation », « Variables qualitatives »
299	« Espace transfrontalier », « Mondialisation »
300	« Frontiere », « Espace transfrontalier »
301	« Analyse spatiale », « Organisation de l'espace », « Structure spatiale », « Spatialite »
302	« Representation », « La region espace vecu »
303	« Ullman Edward L. (1912 1976) »
304	« Latitude », « TOPOGRAPHIE », « Densite »
305	« Geosysteme »
306	« Teledetection », « Echelle »
307	« Littoral », « Littoral »
308	« Organisation de l'espace »
309	« Auto organisation », « Bifurcation », « Evolution des entites geographiques »
310	« Geomatique », « Carte choroplethe », « Statistique spatiale », « CARTOGRAPHIE THEMATIQUE »
311	« Historique du paysage », « Geosysteme »
312	« Mondialisation », « Centre Peripherie »
313	« Paysage visible », « Perception des paysages »
314	« La region espace vecu »
315	« Interaction spatiale », « Analyse spatiale »

suite sur la page suivante

N°	**Articles visités**
316	« Realisme », « Constructivisme »
317	« Christaller (modele) »
318	« Discretisation », « Analyse de donnees », « Variables quantitatives », « Variables qualitatives »
319	« Aire d'influence »
320	« TOPOGRAPHIE »
321	« Montagne », « Etagement »
322	« Biome »
323	« Discretisation », « Variables quantitatives », « CARTOGRAPHIE THEMATIQUE »
324	« Etagement », « Coupe (Transect) »
325	« Carte choroplethe », « Variables quantitatives », « CARTOGRAPHIE THEMATIQUE »
326	« Vulnerabilite », « Cindynique »
327	« Region »
328	« Terrain »
329	« Haut lieu », « Anthroposysteme »
330	« Mondialisation », « Equite territoriale »
331	« Insularite »
332	« Lieux centraux »
333	« Diffusion spatiale »
334	« Centralite », « Lieux centraux »
335	« Gradient », « Mondialisation », « Evolution des entites geographiques »
336	« Diffusion spatiale », « Evolution des entites geographiques »
337	« Ullman Edward L. (1912 1976) »
338	« Insularite », « Ile »
339	« Qu'est ce qu'un systeme expert ? », « Region »
340	« Pays »
341	« Segregation », « Le territoire selon Claude Raffestin »
342	« Topographie (historique de la notion) », « Pays »
343	« Christaller (modele) »
344	« Geosysteme »
345	« Region », « Pays »
346	« Seuil », « Discontinuite »
347	« carte », « Echelle »
348	« Frontiere »
349	« Region », « Pays »
350	« Discretisation », « Geographie », « CARTOGRAPHIE THEMATIQUE »
351	« Cindynique »
352	« Organisation de l'espace », « Territoire »
353	« Interaction spatiale », « Reseau », « Connexite2 »

suite sur la page suivante

$N°$	**Articles visités**
354	« Representation », « La region espace vecu »
355	« Analyse spatiale »
356	« Determinisme »
357	« Ullman Edward L. (1912 1976) »
358	« Mondialisation », « Centre Peripherie »
359	« Realisme »
360	« Etagement »
361	« Aire d'influence », « Hierarchie », « Lieux centraux »
362	« Ecotone », « Biocenose »
363	« Nature et culture »
364	« Geochronologie », « Actualisme »
365	« Carte choroplethe », « Theories de l'Analyse Spatiale », « Analyse spatiale », « Statistique spatiale »
366	« Insularite »
367	« Theories de l'Analyse Spatiale », « Organisation de l'espace », « Lieux centraux », « Territoire »
368	« Organisation de l'espace », « carte »
369	« Connectivite », « Metropolisation »
370	« Paysage selon le laboratoire THEMA », « Production du paysage »
371	« Figuration cartographique »
372	« Vulnerabilite », « Biotope », « Biocenose »
373	« Aire d'influence »
374	« Representation »
375	« Paysage selon le laboratoire THEMA », « Production du paysage », « Geographie »
376	« La region comme systeme »
377	« Insularite », « Ile »
378	« La region espace vecu »
379	« Analyse de donnees », « Variables qualitatives », « Mesure »
380	« Bilan hydrique »
381	« Vulnerabilite », « Risque », « Cindynique »
382	« Systeme spatial », « Resilience »
383	« Orient », « Biocenose »
384	« Polarisation »
385	« Bilan hydrique »
386	« Mondialisation »
387	« Le territoire selon Claude Raffestin »
388	« Littoral », « Littoral »
389	« Paysage visible », « Paysage »
390	« Mondialisation », « Puissance »
391	« Biotope », « Biome »

suite sur la page suivante

N°	**Articles visités**
392	« Etagement », « Coupe (Transect) »
393	« Biome »
394	« Paysage selon le laboratoire THEMA », « Paysage visible »
395	« Theories de l'Analyse Spatiale », « Analyse spatiale »
396	« Le territoire selon Claude Raffestin »
397	« Postmodernisme »
398	« Voisinage »
399	« Contrainte »
400	« Region »
401	« Representation », « Oekoumene »
402	« Communaute transnationale »
403	« Connectivite », « Metropolisation »
404	« Metropolisation »
405	« Geopolitique »
406	« Communaute transnationale », « Diaspora »
407	« Aire d'influence »
408	« Complexite », « Bassin versant », « classification »
409	« Figuration cartographique »
410	« Postmodernisme », « Processus », « Constructivisme »
411	« Mondialisation », « Centre Peripherie »
412	« Geopolitique »
413	« Front pionnier »
414	« Resilience »
415	« Localisation », « Echelle », « Discontinuite »
416	« Dynamique des versants », « Etagement »
417	« Terrain »
418	« Equite territoriale »
419	« Qu'est ce qu'un systeme expert ? »
420	« Anthroposysteme », « Geosysteme »
421	« Qu'est ce qu'un systeme expert ? »
422	« Analyse de donnees »
423	« TOPOGRAPHIE », « carte »
424	« Determinisme », « La region espace vecu »
425	« Mondialisation »
426	« Segregation », « Equite territoriale »
427	« Anthroposysteme »
428	« CARTOGRAPHIES », « Figuration cartographique »
429	« Segregation »

suite sur la page suivante

$N°$	**Articles visités**
430	« Geosysteme »
431	« Historique du paysage »
432	« Representation », « La region espace vecu »
433	« Discontinuite », « Insularite », « Evolution des entites geographiques »
434	« Modeles statistiques », « Mesure »
435	« Systeme spatial », « Interaction spatiale », « Organisation de l'espace », « Polarisation »
436	« Risque »
437	« Constructivisme »
438	« Discretisation », « CARTOGRAPHIE THEMATIQUE »
439	« Analyse de donnees », « Variables quantitatives »
440	« Biotope », « Biocenose »
441	« Equite territoriale »
442	« Configuration », « Organisation de l'espace », « Reseau », « Region »
443	« Constructivisme »
444	« Littoral », « Littoral », « Etagement »
445	« Production du paysage », « Anthropisation »
446	« Mondialisation », « La region espace vecu », « Region »
447	« Ecotone »
448	« Organisation de l'espace », « Mondialisation »
449	« Orient », « Biocenose »
450	« Nature et culture »
451	« Reseau », « Evolution des entites geographiques »
452	« Segregation »
453	« Aire d'influence », « Aire de chalandise », « Polarisation », « Centralite »
454	« Centre Peripherie »
455	« Geopolitique », « Puissance »
456	« Utilisations des paysages », « Equite territoriale »
457	« Interaction spatiale », « Echelle »
458	« Systeme spatial », « Geographie », « Polarisation »
459	« Contingence »
460	« Postmodernisme »
461	« Geosysteme »
462	« Anthropisation », « Geosysteme »
463	« carte », « Etagement », « Coupe (Transect) »
464	« CARTOGRAPHIES », « Echelle », « Energie »
465	« Equite territoriale »
466	« Equite territoriale »
467	« Analyse spatiale », « Fondements epistemologiques »

suite sur la page suivante

$N°$	**Articles visités**
468	« Interaction spatiale », « Modele gravitaire », « Evolution des entites geographiques »
469	« Historique du paysage », « Energie »
470	« Friction », « Mondialisation », « Modele gravitaire »
471	« Littoral », « Littoral »
472	« Mondialisation »
473	« Statistique spatiale », « Modeles statistiques », « Variables qualitatives »
474	« Orient », « Insularite »
475	« Connexite2 »
476	« TOPOGRAPHIE », « Topographie (historique de la notion) »
477	« Nature et culture »
478	« Latitude », « Longitude »
479	« Vulnerabilite », « Cindynique »
480	« Fondements epistemologiques », « Constructivisme »
481	« Theories de l'Analyse Spatiale », « Modele gravitaire », « Flux », « Interaction »
482	« Interaction spatiale », « Mesure », « Modele gravitaire »
483	« Representation », « Spatialite », « La region espace vecu »
484	« Discretisation », « Analyse de donnees », « Statistique spatiale », « Modeles statistiques », « Variables qualitatives », « Mesure »
485	« Geographicite », « Geographie »
486	« Centralite »
487	« Geographicite », « Fondements epistemologiques », « La trajection paysagere », « Imaginaire spatial »
488	« Coupe (Transect) »
489	« carte »
490	« Geographie », « Oekoumene », « La trajection paysagere »
491	« Region polarisee »
492	« Analyse spatiale », « Modele », « Loi »
493	« Segregation », « Ethnie », « Equite territoriale »
494	« Segregation », « Risque »
495	« Nature et culture »
496	« Carte choroplethe », « Discretisation », « Analyse de donnees », « Modeles statistiques », « Variables quantitatives »
497	« Postmodernisme »
498	« Historique du paysage »
499	« Systeme spatial », « Friction », « Modele gravitaire », « carte », « Polarisation », « Flux », « La dynamique des systemes selon J.W. Forrester », « Anthroposysteme », « Interaction »
500	« Theories de l'Analyse Spatiale », « Evolution des entites geographiques », « Lieux centraux »
501	« Theories de l'Analyse Spatiale », « Mondialisation »
502	« Insularite », « Ile »

suite sur la page suivante

N°	**Articles visités**
503	« Etagement », « Coupe (Transect) »
504	« Postmodernisme », « Representation »
505	« Seuil », « Flux »
506	« Interaction spatiale », « Attraction Attractivite », « Modele gravitaire »
507	« Equite territoriale »
508	« Espace transfrontalier »
509	« Communaute transnationale », « Diaspora »
510	« Configuration », « Organisation de l'espace », « Analyse spatiale »
511	« Bassin versant », « TOPOGRAPHIE »
512	« Lieux centraux », « Christaller (modele) »
513	« Montagne », « Etagement »
514	« Littoral », « Oekoumene »
515	« Geomatique », « Geosimulation »
516	« Reseau »
517	« Constructivisme »
518	« Segregation », « Communautes fermees (gated communities) »
519	« Bilan hydrique »
520	« Representation »
521	« Antipode »
522	« CARTOGRAPHIES »
523	« Homogeneite »
524	« Attraction Attractivite », « Structure spatiale »
525	« Ethnie »
526	« carte », « Coupe (Transect) »
527	« Theories de l'Analyse Spatiale »
528	« Connectivite », « Centralite », « Lieux centraux »
529	« Littoral », « Etagement »
530	« Front pionnier »
531	« Theories de l'Analyse Spatiale », « Distance », « Spatialite »
532	« Insularite »
533	« Paysage selon le laboratoire THEMA », « Paysage visible », « Perception des paysages »
534	« Structure spatiale »
535	« Centralite », « Lieux centraux »
536	« La dynamique des systemes selon J.W. Forrester »
537	« Vulnerabilite »
538	« Representation »
539	« Mondialisation »
540	« Savoirs vernaculaires »

suite sur la page suivante

N°	**Articles visités**
541	« Statistique spatiale », « Modeles statistiques », « Mesure »
542	« Aire d'influence », « Polarisation »
543	« TOPOGRAPHIE », « Topographie (historique de la notion) »
544	« Perception des paysages », « Mondialisation »
545	« Orient », « Mondialisation »
546	« Biotope »
547	« Catastrophes (theorie des) », « Causalite »
548	« Segregation »
549	« Metropolisation »
550	« Terrain », « Coupe (Transect) »
551	« Gradient », « Seuil »
552	« Discretisation », « Analyse de donnees », « Variables qualitatives », « Mesure »
553	« Oekoumene », « Figuration cartographique »
554	« TOPOGRAPHIE », « Longitude »
555	« Topographie (historique de la notion) », « Localisation »
556	« Nature et culture », « Oekoumene »
557	« Constructivisme »
558	« Le territoire selon Claude Raffestin »
559	« Representation », « Analyse spatiale », « Ethnie »
560	« Modeles statistiques », « Variables qualitatives », « Mesure »
561	« Le territoire selon Claude Raffestin »
562	« Terrain »
563	« Modeles statistiques »
564	« Equite territoriale »
565	« Connexite », « Graphe », « Connectivite »
566	« Site », « Lieu », « Paysage »
567	« Littoral »
568	« Geographicite », « Contingence »
569	« Constructivisme »
570	« Interaction spatiale »
571	« Frontiere », « Pays »
572	« Paysage », « Pays »
573	« Analyse de donnees », « Modeles statistiques », « Mesure »
574	« carte », « Echelle », « Territoire »
575	« Ullman Edward L. (1912 1976) »
576	« Mondialisation »
577	« Interaction spatiale », « La dynamique des systemes selon J.W. Forrester », « Interaction »
578	« Autocorrelation », « Voisinage »

suite sur la page suivante

$N°$	**Articles visités**
579	« Systeme spatial », « Theories de l'Analyse Spatiale », « La region espace vecu », « Imaginaire spatial », « Territoire »
580	« Variables quantitatives »
581	« Discretisation », « Statistique spatiale », « Variables quantitatives »
582	« Lieux centraux »
583	« Analyse de donnees », « Variables quantitatives », « Variables qualitatives »
584	« La region espace vecu »
585	« Insularite »
586	« Orient »
587	« Structure spatiale », « Accessibilite »
588	« Bassin versant »
589	« Frontiere », « Ville frontaliere », « Mondialisation », « Equite territoriale »
590	« Insularite »
591	« Puissance »
592	« Littoral », « Littoral »
593	« Dispute territoriale »
594	« Analyse de donnees », « Variables quantitatives »
595	« Theories de l'Analyse Spatiale », « Christaller (modele) »
596	« Postmodernisme », « Representation »
597	« Insularite »
598	« Biocenose »
599	« Christaller (modele) »
600	« Segregation »
601	« Insularite »
602	« Theories de l'Analyse Spatiale », « Metropolisation »
603	« Terrain »
604	« Front pionnier »
605	« Hierarchie », « Organisation de l'espace »
606	« Risque »
607	« Insularite »
608	« Mondialisation », « Puissance », « Lieux centraux »
609	« Theories de l'Analyse Spatiale », « Analyse spatiale », « Flux »
610	« Orient », « Region »
611	« carte »
612	« Frontiere »
613	« Anthropisation », « Bilan hydrique »
614	« Christaller (modele) »
615	« Voisinage »

suite sur la page suivante

N°	**Articles visités**
616	« Historique du territoire »
617	« Modeles statistiques », « Variables qualitatives »
618	« Fixisme Mobilisme », « Causalite »
619	« Christaller (modele) », « Constructivisme »
620	« Teledetection », « Paysage »
621	« Metropolisation »
622	« Organisation de l'espace », « Distance », « Localisation selon F.Durand Dastes », « Region polarisee », « Voisinage », « Accessibilite »
623	« Carte choroplethe », « Discretisation », « CARTOGRAPHIES », « Mesure », « Figuration cartographique »
624	« Geosimulation », « Geopolitique », « Mondialisation », « Centre Peripherie »
625	« Analyse spatiale », « Analyse de donnees », « Statistique spatiale », « Modeles statistiques », « Systeme d'information geographique (S.I.G) », « Mesure », « Voisinage »
626	« Representation », « carte », « Site », « Figuration cartographique »
627	« Representation », « Longitude », « Meridien », « carte », « Echelle »
628	« Representation », « Le territoire selon Maryvonne Le Berre »
629	« Geographie », « CARTOGRAPHIE THEMATIQUE »
630	« Segregation »
631	« Etagement », « Coupe (Transect) »
632	« Complexite », « La dynamique des systemes selon J.W. Forrester », « Constructivisme »
633	« Frontiere »
634	« Historique du paysage »
635	« Polarisation »
636	« Mondialisation », « Communautes fermees (gated communities) »
637	« Aire d'influence », « Polarisation »
638	« Latitude », « TOPOGRAPHIE »
639	« Production du paysage », « Perception des paysages », « Paysage »
640	« Communaute transnationale », « Localisation selon F.Durand Dastes », « Diaspora »
641	« Connectivite », « Metropolisation »
642	« Nord », « Mondialisation »
643	« Diffusion spatiale »
644	« Orient »
645	« Mondialisation »
646	« Metropolisation »
647	« Complexite », « Theories de l'Analyse Spatiale »
648	« Figuration cartographique »
649	« Mondialisation », « Processus »
650	« Aire d'influence », « carte »
651	« Biocenose »

suite sur la page suivante

N°	**Articles visités**
652	« Figuration cartographique »
653	« TOPOGRAPHIE »
654	« Insularite », « Ile »
655	« Territoire »
656	« CARTOGRAPHIES »
657	« Historique du territoire », « Le territoire selon Claude Raffestin »
658	« Analyse de donnees », « Modeles statistiques »
659	« Anthropisation », « Hydrosysteme »
660	« Systeme d'information geographique (S.I.G) »
661	« Discretisation », « Theories de l'Analyse Spatiale », « Analyse spatiale », « CARTOGRAPHIE THEMATIQUE », « Systeme d'information geographique (S.I.G) »
662	« Representation »
663	« Paysage selon le laboratoire THEMA », « Oekoumene », « La trajection paysagere »
664	« Frontiere », « Le territoire selon Claude Raffestin »
665	« Seuil », « Equite territoriale »
666	« carte », « CARTOGRAPHIE THEMATIQUE »
667	« Organisation de l'espace »
668	« Orient », « Mondialisation »
669	« Biotope », « Biocenose »
670	« Bassin versant », « Estuaire »
671	« Mondialisation »
672	« Carte choroplethe », « Discretisation », « TOPOGRAPHIE », « CARTOGRAPHIE THEMATIQUE »
673	« Mondialisation », « Lieu », « Centre Peripherie »
674	« Analyse de donnees », « Modeles statistiques », « Variables qualitatives »
675	« Paysage selon le laboratoire THEMA », « Paysage visible », « Constructivisme »
676	« Equite territoriale », « Echelle »
677	« Discontinuite », « Equite territoriale », « Oekoumene »
678	« Contingence », « Mondialisation »
679	« Modele »
680	« Littoral »
681	« Mondialisation », « Centre Peripherie », « Constructivisme »
682	« Communautes fermees (gated communities) »
683	« Utilisations des paysages »
684	« Front pionnier », « Oekoumene »
685	« Mondialisation »
686	« CARTOGRAPHIE THEMATIQUE »
687	« carte »
688	« Aire d'influence »

suite sur la page suivante

N°	**Articles visités**
689	« Montagne », « Etagement »
690	« Christaller (modele) », « Lieux centraux »
691	« Metropolisation »
692	« Seuil », « Discontinuite »
693	« Agregation », « Nomothetisme », « Equite territoriale »
694	« CARTOGRAPHIE THEMATIQUE »
695	« Diffusion spatiale »
696	« Topographie (historique de la notion) », « Constructivisme »
697	« Topographie (historique de la notion) », « Localisation selon F.Durand Dastes », « La region espace vecu »
698	« Segregation », « Analyse spatiale »
699	« Systeme morphogenetique »
700	« Mondialisation »
701	« Le territoire selon Maryvonne Le Berre »
702	« Historique du paysage »
703	« Systeme spatial », « Auto organisation »
704	« Biome », « CARTOGRAPHIE THEMATIQUE »
705	« Geomatique », « Geosimulation »
706	« Reseau », « Metropolisation »
707	« Representation », « La region espace vecu »
708	« Insularite »
709	« Le territoire selon Claude Raffestin »
710	« Geographicite »
711	« Systeme », « Modele »
712	« Equite territoriale »
713	« Le territoire selon Claude Raffestin », « Territoire »
714	« Aire de chalandise », « Equilibre », « Metropolisation »
715	« Resilience », « Hydrosysteme »
716	« CARTOGRAPHIES », « Systeme d'information geographique (S.I.G) »
717	« Biome », « Biocenose »
718	« Organisation de l'espace », « Representation », « Imaginaire spatial »
719	« Spatialite », « Localisation selon F.Durand Dastes », « Lieu »
720	« CARTOGRAPHIES », « CARTOGRAPHIE THEMATIQUE »
721	« Carte choroplethe »
722	« Topographie (historique de la notion) », « Site »
723	« Biocenose »
724	« Postmodernisme »
725	« carte »
726	« Bilan hydrique »

suite sur la page suivante

$N°$	**Articles visités**
727	« Complexite », « La dynamique des systemes selon J.W. Forrester »
728	« Biocenose », « Geosysteme »
729	« Evolution des entites geographiques »
730	« Statistique spatiale »
731	« CARTOGRAPHIE THEMATIQUE »
732	« Anthropisation », « Systeme morphogenetique »
733	« Centralite », « Lieux centraux »
734	« Spatialite », « Voisinage »
735	« Carte choroplethe », « Theories de l'Analyse Spatiale », « Analyse spatiale », « Distance », « Statistique spatiale », « Lieux centraux »
736	« Christaller (modele) »
737	« Historique du territoire », « Geopolitique », « Statistique spatiale », « Systeme d'information geographique (S.I.G) »
738	« Geomatique », « Analyse spatiale »
739	« Region polarisee »
740	« Theories de l'Analyse Spatiale », « carte »
741	« Theories de l'Analyse Spatiale », « Diffusion spatiale »
742	« Connexite », « Carte choroplethe », « La region espace vecu », « Systeme d'information geographique (S.I.G) », « Variables quantitatives »
743	« Realisme », « Perception des paysages »
744	« Representation », « Constructivisme »
745	« Theories de l'Analyse Spatiale », « Barriere (effets de) », « Equite territoriale », « Centre Peripherie », « Diffusion spatiale »
746	« Terrain », « Teledetection »
747	« Ecotone »
748	« Terrain », « Latitude », « CARTOGRAPHIES »
749	« TOPOGRAPHIE », « Topographie (historique de la notion) »
750	« Contiguite », « Interaction spatiale », « Geosimulation », « Statistique spatiale », « Voisinage »
751	« Systeme spatial », « Carte choroplethe », « Statistique spatiale », « Polarisation », « Autocorrelation », « Localisation », « Lieux centraux »
752	« Terrain »
753	« Latitude », « TOPOGRAPHIE », « Longitude », « carte », « Systeme d'information geographique (S.I.G) »
754	« Configuration », « Paysage selon le laboratoire THEMA », « Representation », « Organisation de l'espace », « Paysage visible », « Diaspora », « Geosysteme », « Paysage »
755	« Biotope », « Antipode », « Terre », « Meridien », « Oekoumene »
756	« Historique du territoire »
757	« Paysage selon le laboratoire THEMA », « Production du paysage », « Utilisations des paysages », « La trajection paysagere »

suite sur la page suivante

$N°$	**Articles visités**
758	« Espace transfrontalier »
759	« Biocenose »
760	« Structure spatiale »
761	« Bassin versant », « Hydrosysteme »
762	« Situation », « Interaction »
763	« Perception des paysages »
764	« Littoral », « Milieux et couts de mise en valeur », « Littoral »
765	« Contrainte »
766	« Etagement », « Discontinuite »
767	« Communautes fermees (gated communities) »
768	« Representation », « carte »
769	« carte », « Systeme d'information geographique (S.I.G) »
770	« Ecotone »
771	« Interaction spatiale », « Autocorrelation », « Voisinage »
772	« Analyse de donnees »
773	« Littoral »
774	« Representation », « CARTOGRAPHIES », « Topographie (historique de la notion) »
775	« Fondements epistemologiques », « Territoire »
776	« Christaller (modele) », « Lieux centraux »
777	« Paysage visible », « Paysage »
778	« carte »
779	« Situation », « Nord »
780	« Littoral »
781	« Aire d'influence », « Aire de chalandise »
782	« Bassin versant », « Hydrosysteme »
783	« Ville frontaliere »
784	« Analyse spatiale », « Lieu »
785	« Teledetection »
786	« carte »
787	« Equite territoriale »
788	« carte », « CARTOGRAPHIE THEMATIQUE »
789	« Ile »
790	« Figuration cartographique »
791	« Autocorrelation »
792	« Paysage selon le laboratoire THEMA », « Production du paysage »
793	« La trajection paysagere »
794	« Representation », « CARTOGRAPHIES », « La region espace vecu »
795	« Discontinuite », « Discontinuite », « Discontinuite »

suite sur la page suivante

$N°$	**Articles visités**
796	« Montagne », « Bilan hydrique »
797	« carte », « carte », « Systeme d'information geographique (S.I.G) », « CARTOGRAPHIE THEMATIQUE »
798	« Auto organisation », « Distance »
799	« Nomothetisme »
800	« Postmodernisme »
801	« Orient », « EST »
802	« Biotope »
803	« Bifurcation », « Causalite »
804	« Longitude », « Meridien »
805	« Geographicite »
806	« Vallee »
807	« Nature et culture »
808	« Analyse spatiale », « Fondements epistemologiques »
809	« Carte choroplethe », « Figuration cartographique »
810	« Geomatique », « Carte choroplethe »
811	« Nature et culture »
812	« La region espace vecu »
813	« Metropolisation »
814	« CARTOGRAPHIES », « Meridien », « Figuration cartographique »
815	« Biotope », « Biocenose »
816	« Cindynique »
817	« Vallee »
818	« Orient », « EST »
819	« Distance », « Longitude », « Lieu »
820	« Discretisation », « Analyse de donnees », « Variables qualitatives »
821	« Evolution des entites geographiques »
822	« Carte choroplethe », « CARTOGRAPHIE THEMATIQUE », « Figuration cartographique »
823	« Analyse de donnees », « Modeles statistiques », « Variables qualitatives »
824	« Region »
825	« Coupe (Transect) »
826	« Teledetection »
827	« Equite territoriale »
828	« Seuil »
829	« CARTOGRAPHIES »
830	« Seuil », « Densite »
831	« Contingence », « Systeme morphogenetique », « Anthroposysteme »

Bibliographie

Bibliographie

[Abo 10] F. Abou Latif, N. Delestre, N. Malandain, and J.-P. Pécuchet. *"Similarity measure to identify users' profiles in web usage mining"*. In : *Actes du XXVIIIř congrès INFORSIDè*, pp. 77–92, 5 2010. 95

[Agr 93] R. Agrawal, T. Imielinski, and A. N. Swami. *"Mining Association Rules between Sets of Items in Large Databases"*. In : *SIGMOD Conference*, pp. 207–216, 1993. 65, 67

[Agr 94] R. Agrawal and R. Srikant. *"Fast algorithms for mining association rules"*. In : *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pp. 487–499, 1994. 65, 67

[Ake 10] R. Akerkar and P. Sajja. *Knowledge-Based Systems*. Jones and Bartlett Publisher, 2010. 26

[Bes 73] M. Besson. *"A propos des distances entre ensembles de parties"*. Mathématiques et Sciences Humaines, **42** (1973) 17-35. 109

[Bor 98] J. Borges and M. Levene. *"Mining association rules in hypertext databases"*. In : *Proceedings Conference on Knowledge Discovery and Data Mining (KDD'98)*, pp. 149–153, 1998. 65, 67

[BOU 11] N. BOUSBIA. *Analyse des traces de navigation des apprenants dans un environnement de formation dans une perspective de détection automatique des styles d'apprentissage*. PhD thesis, L'UNIVERSITE PIERRE ET MARIE CURIE (France) ET L'ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE (Algérie), 2011. 151

[Bra 97] A. P. Bradley. *"The use of the area under the ROC curve in the evaluation of machine learning algorithms"*. Pattern Recognition, **30** (1997) 1145–1159. 58

[Bru 98] P. Brusilovsky, A. Kobsa, and J. Vassileva, Eds. *Adaptive HyperText and Hypermedia*. Kluwer Academic Publishers, Norwell, MA, USA, 1998. 23

[Buc 84] B. G. Buchanan and E. H. Shortliffe, Eds. *Rule Based Expert Systems : The Mycin Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, 1984. 26

[Can 05] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy. *"SVM and Kernel Methods Matlab Toolbox"*. Perception Systèmes et Information, INSA de Rouen, Rouen, France, 2005. 51

[Cau 01] G. Cauwenberghs and T. Poggio. *"Incremental and decremental support vector machine learning"*. In : *Advances in neural information processing systems 13 : proceedings of the 2000 conference*, pp. 409–415, The MIT Press, 2001. 51

[Che 01] J. Cheng and R. Greiner. *"Learning Bayesian Belief Network Classifiers : Algorithms and System"*. In : *Advances in Artificial Intelligence*, pp. 141–151, Springer, 2001. 33

[Che 98] M. syan Chen, J. S. Park, and P. S. Yu. *"Efficient data mining for path traversal patterns"*. IEEE Transactions on Knowledge and Data Engineering, **10** (1998) 209-221. 66, 67

[Che 99] J. Cheng and R. Greiner. *"Comparing Bayesian network classifiers"*. In : *Proceedings UAI*, Citeseer, 1999. 33

[Cla 83] W. J. Clancey. *"GUIDON."*. Journal of Computer-Based Instruction, **10(1-2)** (1983) 8-15. 27

[Cla 86] W. J. Clancey. *"From GUIDON to NEOMYCIN and HERACLES in twenty short lessons"*. AI Magazine, **7** (1986) 40-60. 27

[Col 73] A. Colmerauer, H. Kanoui, P. Roussel, and R. Pasero. *"Un système de communication homme-machine en Français"*. Tech. Rep., rapport de recherche, Groupe de recherche en Intelligence Artificielle, Marseille, 1973. 27

[Coo 00] R. Cooley. *Web Usage Mining : Discovery and Application of Interesting Patterns from Web Data*. PhD thesis, University of Minnesota, 2000. 63

[Del 00] N. Delestre. *METADYNE, un Hypermédia Adaptatif Dynamique pour l'Enseignement*. PhD thesis, Université de Rouen, 2000. 23

[Del 07] N. Delestre and N. Malandain. *"Analyse et représentation en deux dimensions de traces pour le suivi de l'apprenant"*. Sticef, **14** (2007) 1764-7223. 109

[Eli 07] B. Elissalde and C. Kosmopoulos. *"De la navigation au savoir. Réflexions sur les nouveaux comportements de diffusion et d'acquisition des connaissances à travers l'expérience d'Hypergeo"*. Ametist, (2007) 14. accessible sur : http ://ametist.inist.fr/sommaire.php ?id=234. 81

[FAO 07] A. FAOUR. *Une architecture semi-supervisée et adaptative pour le fitrage d'alarmes dans les systyemes de détection d'intrusions sur les réseaux*. PhD thesis, Institut National des Sciences Appliquées de Rouen, 2007. 45

[Faw 06] T. Fawcett. *"An introduction to ROC analysis"*. Pattern Recognition Letters, **27** (2006) 861 - 874. 57, 58

[Fis 88] R. Fisher and M. Marshall. *"Iris Plants Database"*. 1988. `http://archive.ics.uci.edu/ml/datasets/Iris`. 40, 43, 52

Bibliographie

[Gao 08] W. Gao and O. R. L. Sheng. *"Mining Characteristic Patterns to Identify Web Users"*. 2008. 68, 80

[Gui 04] E. Guichard. *Mesure de l'internet*. Les Canadiens en Europe, 2004. 62

[Haa 04] B. Haasdonk and C. Bahlmann. *"Learning with Distance Substitution Kernels"*. In : *Pattern Recognition*, pp. 220–227, Springer Berlin / Heidelberg, 2004. 98

[Han 00] J. Han, J. Pei, and Y. Yin. *"Mining frequent patterns without candidate generation"*. SIGMOD Rec., **29** (2000) 1-12. 70, 72

[Han 04] J. Han, J. Pei, Y. Yin, and R. Mao. *"Mining frequent patterns without candidate generation : A frequent-pattern tree approach"*. Data mining and knowledge discovery, **8** (2004) 53-87. 70, 74

[Han 06] J. Han and M. Kamber. *Data mining : Concepts and techniques*. Morgan Kaufmann, 2006. 64, 65

[Han 09] D. J. Hand. *"Measuring classifier performance : a coherent alternative to the area under the ROC curve"*. Machine Learning, **77** (2009) 103-123. 60

[Has 98] T. Hastie and R. Tibshirani. *"Classification by pairwise coupling"*. The annals of statistics, **26** (1998) 451-471. 47

[Hin 03] G. Hinton and S. Roweis. *"Stochastic neighbor embedding"*. Advances in neural information processing systems, (2003) 857-864. 54

[JAC 06] C. JACQUIOT. *Modélisation logique et générique des systèmes d'hypermédias adaptatifs*. PhD thesis, Université Paris XI, 2006. 22

[Jai 88] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, N.J., March 1988. 64

[Jen 96] F. Jensen. *An introduction to Bayesian networks*. Vol. 210, UCL press London, 1996. 34

[Koh 01] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag New York, Inc., 3rd Ed., 2001. 42

[Koh 89] T. Kohonen. *Self-organization and associative memory : 3rd edition*. Springer-Verlag New York, Inc., New York, NY, USA, 1989. 65

[Kos 00] R. Kosala and H. Blockeel. *"Web Mining Research : A Survey"*. SIGKDD Explorations, **2** (2000) 1-15. 62

[Kou 05] M. Koutri, N. Avouris, and S. Daskalaki. *"A survey on web usage mining techniques for web-based adaptive hypermedia systems"*. Adaptable and adaptive hypermedia systems, (2005) 125–149. 64, 67

[Lan 98] T. K. Landauer, P. W. Foltz, and D. Laham. *"An introduction to latent semantic analysis"*. Discourse processes, **25** (1998) 259-284. 152

[Ler 06] P. Leray. *"Réseaux bayésiens : Apprentissage et diagnostic de systemes complexes"*. Habilitation à diriger les recherches, 11 2006. 32, 33

[Maa 08] L. van der Maaten and G. Hinton. *"Visualizing data using t-SNE"*. Journal of Machine Learning Research, **9** (2008) 2579-2605. 56

[Maa 09] L. van der Maaten, E. Postma, and J. van den Herik. *"Dimensionality Reduction : A Comparative Review"*. Tech. Rep., Tilburg centre for Cognition and Communication, Tilburg University, 2009. 54

[Mac 67] J. MacQueen. *"Some methods for classification and analysis of multivariate observations"*. In : *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, p. 14, California, USA, 1967. 39

[McC 96] R. McCammon. *"PROSPECTOR II-an expert system for mineral deposit models"*. International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts, **33** (1996) 267A-267A(1). 26

[Mob 07a] B. Mobasher. *"Data mining for web personalization"*. Lecture Notes in Computer Science, **4321** (2007) 90-135. 23, 66

[Mob 07b] B. Mobasher. *"Web Usage Mining"*. In : *Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data*, pp. 449–483, Springer Berlin Heidelberg, 2007. 62

[Mob 96] B. Mobasher, N. Jain, E.-H. S. Han, and J. Srivastava. *"Web mining : Pattern discovery from world wide web transactions"*. Tech. Rep., Technical Report TR-96050, Department of Computer Science, University of Minnesota, 1996. 65, 67

[Nai 04] P. Naïm, P.-H. Wuillemin, P. Leray, O. Pourret, and A. Becker. *Réseaux bayésiens*. Eyrolles, Paris, 2004. 32

[Nas 00] O. Nasraoui, H. Frigui, R. Krishnapuram, and A. Joshi. *"Extracting web user profiles using relational competitive fuzzy clustering"*. International Journal on Artificial Intelligence Tools, **9** (2000) 509–526. 64, 67

[OLe 88] D. E. O'Leary. *"Methods of validating expert systems"*. Interfaces, **18** (1988) 72-79. 27

Bibliographie

[Oli 06] F. Olivier. *De l'identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplètes*. PhD thesis, Institut National des Sciences Appliquées de Rouen, 2006. 33

[Pea 88] J. Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, 1988. 32, 34

[Pei 00] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu. *"Mining access patterns efficiently from web logs"*. Lecture notes in computer science, **Volume 1805/2000** (2000) 396-407. 66, 67

[Pek 02] E. Pekalska, P. Paclik, and R. P. W. Duin. *"A generalized kernel approach to dissimilarity-based classification"*. Journal of Machine Learning Research, **2** (2002) 175-211. 98

[Per 00] M. Perkowitz and O. Etzioni. *"Towards Adaptive Web Sites : Conceptual Framework and Case Study"*. Artificial Intelligence, **118** (2000) 245-275. 64, 67

[Per 98] M. Perkowitz and O. Etzioni. *"Adaptive Web Sites : Automatically Synthesizing Web Pages"*. In : *In Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pp. 727–732, 1998. 64, 67

[Pre 09] P. Preux. *"Fouille de données Notes de cours"*. Note de Course, 8 2009. 40, 43, 46, 50

[PRO 10] E. PROCHASSON. *Alignement multilingue en corpus comparables spécialisés. Caractérisation terminologique multilingue*. PhD thesis, Université de Nantes, 2010. 152

[Que 67] M. Queen. *"Some methods for classification and analysis of multivariate observations"*. In : *Proc. 5th Berkeley Symp. Math. Stat. Proba*, 1967. 64

[Roo 02] D. Roobaert. *"DirectSVM : A simple support vector machine perceptron"*. The Journal of VLSI Signal Processing, **32** (2002) 147-156. 51

[Row 00] S. T. Roweis and L. K. Sau. *"Nonlinear dimensionality reduction by locally linear embedding"*. Science, **290** (2000) 2323. 54

[Sch] H. Schmid. *"TreeTagger"*. Institute for Computational Linguistics of the University of Stuttgart. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`. 130

[Sch 95] B. Schölkopf, C. Burges, and V. Vapnik. *"Extracting support data for a given task"*. In : *Proceedings, First International Conference on Knowledge Discovery & Data Mining. AAAI Press, Menlo Park, CA*, pp. 252–257, 1995. 47

[Sch 99] B. Schölkopf, C. Burges, and A. Smola. *Advances in kernel methods : support vector learning*. The MIT press, 1999. 49

[Sha 04] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. 48, 50

[Sha 97] C. Shahabi, A. M. Zarkesh, J. Adibi, and V. Shah. *"Knowledge discovery from users Web-page navigation"*. In : *Research Issues in Data Engineering, 1997. Proceedings. Seventh International Workshop on*, pp. 20–29, 1997. 64, 67

[Sor 08] A. Soriano-Asensi, J. D. Martín-Guerrero, E. Soria-Olivas, A. Palomares, R. Magdalena-Benedito, and A. J. Serrano-López. *"Web mining based on Growing Hierarchical Self-Organizing Maps : Analysis of a real citizen web portal"*. Expert Systems with Applications, **34** (2008) 2988 - 2994. 65, 67

[Sri 00] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. *"Web usage mining : Discovery and applications of usage patterns from web data"*. ACM SIGKDD Explorations Newsletter, **1** (2000) 12–23. 23, 62, 63

[Sua 07] F. Suard and A. Rakotomamonjy. *"Mesure de similarité de graphes par noyau de sacs de chemins"*. In : *21 colloque GRETSI sur le traitement du signal et des images*, septembre 2007. 102

[Tac 06] B. Taconet, A. Zahour, S. Ramdane, and W. Boussellaa. *"Classification des k-ppv par sous-voisinages emboîtés"*. In : *Colloque International Francophone sur l'Ecrit et le Document (CIFED'06)*, HAL - CCSD, 2006. 46

[Tan 05] D. Tanasa. *Web Usage Mining : Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support*. PhD thesis, Nice Sophia Antipolis University, 2005. 64

[Ves 00] J. Vesanto and E. Alhoniemi. *"Clustering of the self-organizing map"*. IEEE Transactions on Neural Networks, **11** (2000) 586-600. 42

[Vin 03] R. Vinot, N. Grabar, and M. Valette. *"Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet"*. Actes de TALN 2003, (2003) 257–284. 46

[Vis 02] S. Vishwanathan and M. Narasimha Murty. *"SSVM : a simple SVM algorithm"*. IJCNN '02. Proceedings of the 2002 International Joint Conference on Neural Networks, **3** (2002) 2393 - 2398. 51

[W3C 10] W3C. *"HTTP - Hypertext Transfer Protocol"*. 2010. `http://www.w3.org/Protocols/`. 63, 69

Bibliographie

[W3C 99] W3C. *"Web Characterization Terminology & Definitions Sheet"*. 1999. http://www.w3.org/1999/05/WCA-terms/. 63

[Wan 06] J. Wang. *Encyclopedia of data warehousing and mining*. Idea Group, page 1206-1210, 2006. 62

[Wit 05] I. H. Witten and E. Frank. *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann Pub, 2005. 38

[Won 01] C. Wong, S. Shiu, and S. Pal. *"Mining fuzzy association rules for web access case adaptation"*. In : *Workshop on Soft Computing in Case-Based Reasoning, International Conference on Case-Based Reasoning (ICCBR'01)*, Citeseer, 2001. 66

Résumé

L'objectif de cette thèse est d'identifier le profil d'utilisateur d'un hypermédia afin de l'adapter. Ce profil est déterminé en utilisant des algorithmes d'apprentissage supervisé comme le SVM.

Le modèle d'utilisateur est l'un des composants essentiels des hypermédias adaptatifs. Une des façons de caractériser ce modèle est d'associer l'utilisateur à un profil. Le Web Usage Mining (WUM) identifie ce profil à l'aide des traces de navigation. Toutefois, ces techniques ne fonctionnent généralement que sur de gros volumes de données. Dans le cadre de volumes de données réduits, nous proposons d'utiliser la structure et le contenu de l'hypermédia. Pour cela, nous avons utilisé des algorithmes d'apprentissage à noyau pour lesquels nous avons défini l'élément clé qu'est la mesure de similarité entre traces basée sur une « distance » entre documents du site. Notre approche a été validée à l'aide de données synthétiques puis à l'aide de données issues des traces des utilisateurs du site Hypergéo (site web encyclopédique spécialisé dans la géographie). Nos résultats ont été comparés à ceux obtenus à l'aide d'une des techniques du WUM (l'algorithme des motifs caractéristiques). Finalement, nos propositions pour identifier les profils *a posteriori* ont permis de mettre en évidence cinq profils. En appliquant une « distance sémantique » entre documents, les utilisateurs d'Hypergéo ont été classés correctement selon leurs centres d'intérêt.

Mots-clés : Fouille de données d'usage du Web, apprentissage supervisé et non supervisé, algorithmes de projection, distance et dissimilarité, hypermédia adaptatif.

Abstract

This thesis is devoted to identify the profile of hypermedia user, then to adapt it according to user's profile. This profile is found by using supervised learning algorithm like SVM.

The user model is one of the essential components of adaptive hypermedia. One way to characterize this model is to associate a user to a profile. Web Usage Mining (WUM) identifies this profile from traces. However, these techniques usually operate on large mass of data. In the case when not enough data are available, we propose to use the structure and the content of the hypermedia. Hence, we used supervised kernel learning algorithms for which we have defined the measure of similarity between traces based on a "distance" between documents of the site. Our approach was validated using synthetic data and then using real data from the traces of Hypergéo users, Hypergéo is an encyclopedic website specialized in geography. Our results were compared with those obtained using a techniques of WUM (the algorithm of characteristic patterns). Finally, our proposals to identify the profiles a posteriori led us to highlight five profiles. Hypergéo users are classified according to their interests when the "semantic distance" between documents is applied.

Keywords : Web usage mining, supervised and unsupervised learning, visualization, dimensionality reduction, distance and dissimilarity, adaptive hypermedia.