Une perspective analytique pour la Recherche d'Information: Application : conception et évaluation de Tissue Microarrays

Julie Bourbeillon, Catherine Garbay, Françoise Giroud

Une perspective analytique pour la Recherche d'Information

Application : Conception et évaluation de Tissue Microarrays

Julie Bourbeillon — Catherine Garbay — Françoise Giroud

Laboratoire TIMC-IMAG,
IN3S, Faculté de Médecine,
38706 La Tronche cedex
{Prenom.Nom}@imag.fr

RÉSUMÉ. *Par delà les techniques classiques d'exploration d'une collection de documents comme la Recherche d'Information ou la visualisation graphique, nous présentons dans cet article une perspective analytique intégrant l'expression d'une requête centrée objectif et la construction d'un document multimédia virtuel. La construction de ce document de synthèse est considérée comme un problème d'adaptation complexe. Une première approche pour caractériser ce processus est proposée, et les connaissances nécessaires sont introduites. Nous présentons ensuite l'architecture du moteur d'adaptation envisagé. Une application au domaine médical, plus précisément à la technologie des Tissue Microarrays, permet d'illustrer l'intérêt de cette démarche. Des éléments de validation sont fournis dans ce domaine à l'aide de l'outil TreeMaps de l'Université du Maryland.*

ABSTRACT. *Beyond classical techniques to explore a document collection such as Information Retrieval and graphical visualisation we present in this paper an analytic view which combines the expression of a goal-centered query and the construction of a virtual multimedia document. Building this synthetic document is considered as a complex adaptation problem. We propose a first approach to characterise this process and introduce the required knowledge. An application to the medical field and in particular the Tissue MicroArrays technology allows us to illustrate the interest of this approach. A preliminary validation of the proposed concepts is performed by means of the TreeMaps tool from the University of Maryland.*

MOTS-CLÉS : *système d'information adaptatif, document multimédia, document de synthèse, requête analytique, domaine médical, Tissue MicroArrays.*

KEYWORDS: *adaptive information system, multimedia document, synthetic document, analytic query, medical domain, Tissue MicroArrays.*

1. Introduction

Par delà les techniques classiques d'exploration d'un ensemble de documents comme la Recherche d'Information ou la visualisation graphique de la collection, nous présentons dans cet article une étude préliminaire pour un système d'information permettant à l'usager de porter un regard analytique sur un espace documentaire. Dans un contexte de globalisation de l'information, l'expression de requêtes permettant de sélectionner un ensemble de documents pertinents de taille raisonnable devient de plus en plus complexe. De même, l'accroissement du volume de documents disponible, par exemple sur internet, rend de plus en plus difficile l'appropriation de l'ensemble d'une collection de documents par des techniques de visualisation et de navigation dans cet espace.

De plus, l'objectif de l'utilisateur en recherche d'information se limite rarement à un aperçu global de l'ensemble de la collection ni à la liste des documents pertinents à compulser par rapport à une requête simple. Il poursuit souvent un but d'analyse complexe et cherche en général à répondre à une ou des questions, à valider des hypothèses, en se basant sur les documents présents dans la collection. Cet objectif implicite doit pouvoir servir de point d'entrée au système et être explicité facilement. Il nous paraît donc essentiel d'assister l'utilisateur dans la formulation d'une requête qui lui permettra de répondre à ces besoins. Cela implique la modélisation des tâches qu'il pourrait vouloir réaliser et le recours à des modèles de tâche dans la construction de la requête.

Il s'agit alors de réconcilier les approches introduites précédemment et de tirer parti à la fois :

— d'une entrée dans le système par le biais de requêtes comme dans les outils de Recherche d'Information, tout en donnant une dimension fonctionnelle à ces requêtes par le biais d'un modèle de l'étude à réaliser. Il s'agit d'une adaptation à la tâche de l'utilisateur,

— d'une visualisation graphique interactive des résultats sous forme d'un ou des document(s) de synthèse. Cette visualisation permet d'offrir de nouvelles perspectives et encourage la formulation et la validation de nouvelles hypothèses, fournissant au final une meilleure résolution des problèmes analysés et une meilleure connaissance du domaine d'étude.

Nous nous proposons donc de mettre en place un système d'information adaptatif permettant de construire un document de synthèse à partir d'une collection en fonction d'une requête orientée tâche. Pour l'instant, la réalisation d'un tel système est encore à un stade exploratoire, mais l'intérêt de la démarche est illustré par une application au domaine médical. Il s'agit plus précisément de fournir une assistance à la conception de Tissue Microarrays (TMA), une technologie récente déjà très utilisée en oncologie. Cette technologie implique la fabrication de lames histologiques TMA, comportant chacune plusieurs centaines de microéchantillons de tissus organisés en lignes / colonnes, sur lesquelles divers marqueurs moléculaires

sont révélés. Dans ce contexte, les lames TMA ainsi que les données (données cliniques, quantification de marqueurs...) et documents (images de lames...) associés à chaque échantillon constituent le document de synthèse à construire. L'adaptation à la tâche de l'utilisateur consiste à proposer l'élaboration d'un TMA « optimal » au sens d'une problématique de recherche médicale, anatomopathologique ou biologique.

Dans cet article nous présentons une première formalisation du processus d'adaptation. Après une revue de l'état de l'art dans la section 2, la section 3 donne un aperçu de notre approche du problème. La section 4 présente les besoins de représentation des connaissances pour la mise en place du système. La section 5 introduit un début de réflexion sur l'architecture du futur système. La section 6 présente une application au domaine des TMA ainsi qu'une première validation de l'approche envisagée en utilisant l'outil TreeMaps (Shneiderman, 1992) de l'Université du Maryland.

2. État de l'art

La réduction de l'espace documentaire et sa représentation graphique sont des éléments essentiels de l'exploration d'une collection de documents. Celle-ci fait l'objet de nombreux travaux au sein de la communauté des systèmes d'information.

La réduction de l'espace documentaire est principalement assurée par des méthodes de Recherche d'Information (Singhal, 2001). L'objectif est généralement de retourner une liste de documents ou extraits de documents pertinents en fonction d'une requête utilisateur portant sur des éléments présents dans les documents. Les points de vues ainsi générés sur la collection apportent une vision épurée qui limite le nombre de documents à compulser, et facilitent ainsi le travail de recherche.

La formulation de la requête joue un rôle déterminant pour la sélection des éléments d'intérêt, mais il s'agit pour l'utilisateur d'un problème complexe (Aula, 2003), quelle que soit son expertise sur le domaine exploré. Afin de faciliter cette formulation, de nombreux auteurs proposent de s'appuyer sur une expression à un niveau d'abstraction supérieur, par le biais de requêtes dites « conceptuelles » (Bloesch *et al*, 1997), qui permettent l'expression de notions familières aux utilisateurs, ou « sémantiques » (Stuchenschmidt *et al*, 2004), qui permettent l'exploration de volumineuses collections de documents par thèmes en se basant sur des ontologies (Gruber, 1995)(Guarino, 1998) ou thésaurus.

Malgré ces efforts, le mode d'interrogation demeure une requête de sélection par le contenu : ancrée dans les données, elle ne permet pas l'expression fonctionnelle de l'objectif de la recherche et des besoins d'information de l'usager. Les questions de l'adaptation à l'usager sont pourtant reconnues comme importantes, avec de nombreux travaux dans le domaine de l'hypermédia et du Web adaptatifs (Brusilovsky, 2002) (Wu *et al.*, 2000). Il s'agit dans ces travaux d'exploiter des profils construits à partir des traces d'exécution du système (systèmes adaptatifs) et /

ou des préférences saisies par l'utilisateur (systèmes adaptables). Mais, si l'apparence et le contenu des pages sont effectivement générés en fonction du profil de l'utilisateur, peu d'assistance est apportée à l'expression de son besoin d'information effectif.

Un autre aspect important est la visualisation, que ce soit celle de la collection complète de documents ou des résultats d'une requête. La présentation organisée d'une collection peut être porteuse d'informations complémentaires par rapport à un ensemble de documents ou de données. Cette notion est particulièrement exprimée au sein des systèmes de visualisation d'information. Elle est relevée dans (Friendly *et al*, 2003) qui insiste sur l'importance du choix de présentation des informations en statistiques et fouille de données. Une visualisation adaptée avec une organisation pertinente permet de faire ressortir des faits saillants et suggère des inférences qui resteraient insoupçonnées sans une présentation correcte.

Le volume d'éléments à présenter au sein d'une structure géométrique simple reste donc une problématique majeure, même si elle a fait l'objet de nombreuses études (Card *et al*, 1997). Une communauté de chercheurs s'intéresse depuis toujours à ces problèmes : ce sont les géographes et en particulier les cartographes qui, au sein des systèmes d'information géographique, cherchent à présenter de gros volumes de données pertinentes dans un espace à deux dimensions. De nombreux concepts de cartographie peuvent être appliqués à d'autres domaines (Skupin *et al*, 2003), et la symbolique spatiale est souvent utilisée au sein des systèmes de visualisation d'information ou de fouille visuelle de données (Keim, 2002), afin de permettre l'appréhension de concepts abstraits multidimensionnels. Cependant, en visualisation d'information, le système de coordonnées est souvent arbitraire en pratique, et ne consiste finalement qu'en une métaphore graphique pour un algorithme de classification sous-jacent. Dans un contexte où l'espace d'affichage est limité, l'objectif devient plutôt une représentation compacte de l'information comme au sein des cartes de Kohonen (Kohonen *et al*, 1996) ou dans l'approche TreeMaps (Shneiderman, 1992). Mais là encore l'organisation des éléments au sein de la structure d'affichage reste centrée sur des attributs explicitement présents dans les documents.

Dans ce contexte, les problématiques de visualisation des résultats d'une requête prennent une importance de plus en plus grande au sein de la communauté de Recherche d'Information (Baeza-Yates *et al*, 1999), et des concepts comme ceux des TreeMaps sont utilisés pour la représentation de la distribution de termes de la requête au sein de documents semi-structurés par exemple (Großjohann *et al*, 2002). Comme exposé précédemment, les efforts de sémantisation sont majoritairement centrés sur l'espace à explorer. S'ils facilitent la recherche de documents pertinents par rapport à une notion intéressante pour l'utilisateur, ils n'en facilitent pas pour autant l'expression d'une requête lui permettant d'atteindre son objectif. Il nous paraît donc essentiel d'assister l'usager dans la formulation d'une requête « analytique », qui lui permettra de répondre à ses besoins fonctionnels. Cela implique la modélisation

des tâches de l'utilisateur et le recours à des modèles de « tâches analytiques » dans la construction de la requête.

Les notions de but ou tâche de l'utilisateur ont fait l'objet de nombreuses études par la communauté d'Intelligence Artificielle, menant à diverses formalisations et définitions de tâches et algorithmes associés, tels ceux développés autour de la méthodologie CommonKADS (Wielinga *et al*, 1992) ou des méthodes de résolution de problèmes (Problem-Solving Methods ou PSM) (Fensel *et al*, 1998). Mais ces méthodes s'attachent peu à formaliser l'effort de focalisation nécessaire à la résolution du problème, qui s'avère indispensable dès lors que de vastes espaces informationnels sont considérés. Des solutions à ce problème peuvent pourtant être trouvées en s'inspirant des techniques de Recherche d'Information.

Un couplage des deux approches, au sein de ce qui pourrait être appelé « Recherche d'Information Analytique » est donc proposé dans cet article. Quelques tentatives dans ce sens sont déjà décrites dans la littérature. Par exemple, un moteur de recherche qui permet l'interaction de chercheurs avec une base de données expérimentale est présenté dans (Hunter, 2004). Ce système orienté hypothèse combine exploration, intégration, recherche et inférence sur les données. Mais la spécification des hypothèses reste complexe et proche des données et une assistance plus poussée de l'utilisateur paraît nécessaire. Dans un domaine connexe, une spécification générale des buts d'analyse en fouille visuelle de données est définie dans (Nocke *et al*, 2004) mais celle-ci reste centrée sur les aspects visuels et reste encore au stade préliminaire.

Nous envisageons quant à nous de considérer l'expression de l'objectif d'analyse de l'utilisateur comme partie intégrante et explicite de la requête. En conséquence, le système devra construire à la volée une vue sur la collection de documents adaptée à cet objectif, c'est-à-dire présenter la collection obtenue en résultat de la requête non pas sous la forme d'une simple liste, mais comme un document synthétique dont l'organisation répond au besoin fonctionnel exprimé par l'usager. Cette adaptation implique la réalisation d'un assemblage de données hétérogènes et complexes, regroupant textes, images, vidéos, sons, et données brutes et la prise en compte de connaissances du domaine. La présentation de ce document synthétique à l'utilisateur soutiendra un processus ultérieur de reformulation. Des méthodes de fouille de données pourront par la suite être appliquées au contenu de ce document multimédia afin d'en extraire des informations quantitatives complémentaires.

3. La tâche de synthèse : vue générale

3.1. *La synthèse comme conception d'un document multimédia*

L'objectif du système est de fournir un point de vue synthétique et statistiquement exploitable sur une collection de documents. Ce point de vue doit être adapté aux besoins de l'utilisateur. Par exemple le système pourrait être consacré

à l'étude de la gestion du territoire. Dans ce contexte, il devrait permettre une approche analytique sur une collection de documents comprenant des images satellites, des cartes, des éléments cadastraux, des informations concernant les surfaces consacrées à différents types d'exploitations (forêts, cultures, prairies, zones industrielles, zones urbaines...), des textes de directives et réglementations officielles, des rapports économiques... Il s'agirait alors de proposer une sélection organisée de documents permettant de mener l'étude envisagée par l'utilisateur, par exemple la comparaison entre plusieurs régions ou l'analyse de l'évolution d'une zone géographique au cours du temps.

Étant donnée une requête, la représentation à construire peut être considérée comme un arrangement de documents, références à documents ou extraits de documents, chacun de ces éléments étant caractérisé par un ensemble d'informations issues d'un dépôt de données. En conséquence, cette représentation peut être considérée comme une collection de documents multimédia qui est générée en fonction d'une question utilisateur. Ce document de synthèse inclut :

— la requête utilisateur : celle-ci permet de définir l'étude à réaliser sur la collection de documents. Il s'agit d'une requête analytique complexe qui présente à la fois des critères classiques de Recherche d'Information portant sur des attributs des objets de la collection, des concepts du domaine d'étude présents au sein d'une base de connaissances, mais aussi l'objectif de l'utilisateur. Dans le cas de l'exemple précédent, il peut s'agir de l'« évolution au cours du temps de la déforestation de la forêt amazonienne entre 1950 et 2000 »,

— une grille documentaire : il s'agit de l'assemblage de documents sélectionnés et organisés spatialement selon la requête utilisateur. L'utilisation d'une grille permet une visualisation simple et compacte de faibles dimensions du résultat de la recherche, et un aperçu qualitatif des données pertinentes mettant en relief des phénomènes inattendus (émergence de ruptures, lots, saillances) pour une analyse quantitative future. Dans l'exemple précédent, l'assemblage logique serait de consacrer chaque case de la grille à une année entre 1950 et 2000 et de proposer un classement chronologique. Chaque élément de la grille peut être associé à d'autres éléments :

- un document correspondant complet : pour le cas considéré, il s'agirait d'une fiche présentant la déforestation pour une année donnée, comportant par exemple une image satellite, une carte, la surface boisée, la surface défrichée, des liens vers les réglementations, les rapports économiques ou scientifiques émis cette année là...

- d'autres données associées au document : auteur, date de publication, indice de confiance...

Dans le futur on peut envisager d'inclure des références à des requêtes similaires conduites avec l'outil ou des références bibliographiques pertinentes.

Cette description sommaire des éléments d'un document de synthèse peut conduire à la réalisation de modèles de documents de synthèse spécifiques d'un

domaine d'application. Ces modèles de documents pourraient être paramétrables au sein de la requête ou selon des préférences stockées dans un profil utilisateur.

Les documents générés peuvent finalement servir de support pour des publications dans les revues pertinentes du domaine et permettre d'enrichir la base de connaissances.

3.2. *La synthèse comme réponse à une requête analytique*

Le document de synthèse présenté précédemment doit permettre à l'utilisateur d'atteindre un objectif d'analyse complexe et abstrait, ce en quoi il diffère des système de Recherche d'Information classiques. Considérons l'exemple de la gestion du territoire. Des paradigmes de Recherche d'Information permettent à l'utilisateur de retrouver dans la collection des documents indiquant la surface de forêt amazonienne défrichée par exemple en 1990. La recherche de cette information peut avoir été le seul et unique but de l'utilisateur mais souvent son objectif est bien plus complexe. Nous pouvons imaginer que l'usager cherche par exemple à étudier les tendances dans le défrichement de la forêt durant une période donnée : comment évoluent les surfaces défrichées? Quelles sont les causes de ce défrichement? Ces causes sont-elles toujours les mêmes? Étant donnée l'importance écologique de la forêt amazonienne y-a-t-il des réglementations pour limiter le défrichement et quelle est leur influence? Des techniques de Recherche d'Information classiques apportent difficilement une réponse à ce type de questions. Elles impliquent en effet l'observation d'un ensemble hétérogène de documents, ainsi que d'un ensemble de données numériques, et leur organisation chronologique afin d'appréhender une évolution. C'est cet aperçu analytique que vise à proposer le document de synthèse.

Considérant un ensemble d'études que l'utilisateur pourrait vouloir réaliser, il apparaît que nombre d'entre elles sont proches au niveau conceptuel. Par exemple étudier l'évolution du défrichement entre 1950 et 1990 dans la forêt amazonienne ou l'évolution des surfaces de zones urbaines en France entre 1900 et 1950 sont deux études similaires. Il paraît donc pertinent de regrouper ces études similaires en familles d'études que l'utilisateur pourrait vouloir réaliser. Dans le cadre de la gestion du territoire, des familles d'analyses pourraient être :

— « Évolution » : par exemple évolution au cours du temps de la déforestation en Amazonie entre 1950 et 2000,

— « Comparaison » : par exemple comparaison entre deux régions,

— « Bilan régional » : à une date donnée, les documents concernant les aires géographiques de la région étudiée sont organisés sur la grille en fonction de leurs positions géographiques réelles relatives.

Des objectifs composés peuvent en outre être envisagés, comme « Comparaison d'évolutions » : un des axes de la grille est alors utilisé comme axe temporel et l'autre comme axe géographique, par exemple.

A chaque famille on peut associer un modèle exprimant les traits généraux de ce type d'étude. Ce modèle permet de définir comment construire un document de synthèse adapté à un objectif générique donné. Il s'agit d'un **modèle d'adaptation**. Chaque analyse que l'utilisateur peut vouloir réaliser consiste alors en une spécialisation du modèle d'adaptation de la famille correspondante.

3.3. *La synthèse comme tâche complexe*

À partir des données et des connaissances disponibles, l'outil à concevoir doit générer à la volée un document multimédia de synthèse, soit construire une vue sur la collection de documents selon une requête analytique. Il s'agit d'un problème complexe qui peut être décomposé en trois sous-problèmes :

— problème de **sélection** : construire un document de synthèse est assimilable à la recherche, au sein d'une liste d'éléments (l'ensemble originel de documents), d'une collection (la liste des documents pertinents) qui corresponde à la demande (la requête utilisateur) et qui respecte certaines règles générales,

— problème d'**organisation spatiale** : il s'agit ensuite de placer des objets (documents, références à documents ou portions de documents) sur une grille,

— problème de **présentation** : l'ensemble précédemment généré doit être présenté à l'utilisateur sous la forme la plus lisible et conviviale possible, en respectant d'éventuelles préférences.

Afin de résoudre ce problème, il s'agit tout d'abord de réaliser une analyse plus fine de chacune de ces composantes du processus d'adaptation et les décomposer en sous-problèmes élémentaires, au sein d'une **ontologie de la tâche de synthèse**. Il faudra ensuite définir comment chacun de ces sous-problèmes doit être résolu en adéquation avec la requête analytique d'origine. Or cette requête dépend d'une famille d'études pour laquelle est définie un modèle d'adaptation. La spécialisation de ce modèle d'adaptation avec des contraintes issues de la requête permet de construire un **plan d'adaptation** qui guide la résolution du problème : chaque sous problème composant la tâche d'adaptation est mis en correspondance avec des éléments de résolution décrits dans le plan d'adaptation.

La complexité du problème est liée à plusieurs facteurs:

— données à manipuler : ces données peuvent être hétérogènes dans leur type : textes, images, vidéos, données brutes. Certaines peuvent être pérennes et d'autres à validité limitée dans le temps ou l'espace. Certaines peuvent être quantitatives et d'autres qualitatives. Elles peuvent être appréhendées à diverses échelles. De plus, les points de vue sur ces données sont multiples, dépendants de l'utilisateur ou de l'objectif qu'il veut atteindre. Enfin, la difficulté peut être accrue par la combinatoire importante associée, si la collection originelle présente un volume important de documents, et par la diversité des documents à construire,

— but de la requête : la requête n'est pas limitée à des critères d'inclusion / exclusion et sa finalité est différente du simple retour d'informations pertinentes. Centrée sur les besoins de l'utilisateur, elle doit répondre à un objectif complexe abstrait auquel la visualisation des résultats doit permettre de donner une réponse qualitative,

— plan d'adaptation : il est construit par spécialisation d'un modèle d'adaptation et tient également compte des critères de sélection, d'organisation et de présentation exprimés dans la requête. Cette spécialisation conduit donc à la construction d'un sous-ensemble organisé de critères potentiellement très hétérogènes. En effet, à chaque domaine d'application, à chaque tâche analytique peuvent être associées des collections de critères dont le volume peut être important, ce qui rend difficile la définition du plan d'adaptation et la résolution du problème en fonction de ce plan.

4. Ontologies pour la synthèse

La résolution du problème d'adaptation que nous venons de définir implique la spécification de connaissances concernant tout à la fois la tâche de synthèse et le domaine d'application.

4.1. *Ontologie de la tâche de synthèse*

La conception d'une document de synthèse correspondant au mieux aux besoins de l'utilisateur implique tout d'abord une analyse précise de la tâche de synthèse, comme évoqué précédemment. Chacune des étapes du processus de synthèse (sélection, organisation spatiale et présentation) peut faire l'objet d'une décomposition hiérarchique jusqu'à obtention d'une collection organisée de sous-problèmes élémentaires.

Une vue organisée de l'espace du problème de synthèse est présentée dans le Tableau 1. Cette hiérarchie doit s'affiner et se spécialiser en fonction du domaine d'application spécifique pour constituer une représentation complète du problème.

4.2. *Ontologie de domaine*

Afin de proposer un document de synthèse adapté au mieux à la question posée par l'utilisateur il faut ensuite formaliser et représenter les connaissances concernant le domaine d'application : cette formalisation passe par la représentation des objets et concepts du domaine ainsi que des relations existant entre eux, soit une **ontologie de domaine**. Dans l'exemple utilisé précédemment, il faudrait constituer une ontologie géographique, économique, politique, agronomique... La construction de cette ontologie fait apparaître deux catégories distinctes. L'une consiste en un ensemble d'objets et des attributs associés qui sont les éléments d'intérêt pour l'utilisateur. Dans

Étape de la synthèse	*Catégorie de sous-problèmes*
Sélection	Inclusion : il s'agit de définir la liste d'éléments pertinents dans le cadre de l'étude à réaliser : • *Groupes* : il faut inclure des groupes d'individus correspondant aux besoins de l'utilisateur ; il faut inclure des groupes d'individus correspondant à des ensembles homogènes de patients représentatifs de la diversité globale au sein de la population étudiée • *Individus* : les individus à inclure doivent être représentatifs de la diversité au sein de leur groupe
	Exclusion : il s'agit de définir dans quels cas retirer des éléments de la liste des documents pertinents : • *Données manquantes* : il faut définir une stratégie d'exclusion pour les éléments incomplets • *Nombre d'emplacements* : le nombre d'individus à intégrer est conditionné par la taille de la grille et des critères d'exclusion pour s'adapter au nombre d'emplacements doivent être définis • *Validité des données* : selon les études, l'âge des documents peuvent avoir leur importance et des règles concernant l'exclusion des documents anciens doivent être définies
Organisation spatiale	Géométrie : la grille a une conformation géométrique particulière • *Longueur de la grille* • *Largeur de la grille*
	Ordonnancement : il faut définir comment organiser les éléments au sein de la grille: • *Zones de regroupement* : il faut spécifier quelle taille et quelle forme auront les sections de grille attribuées à chaque groupe d'éléments • *Organisation au sein d'une zone* : il s'agit de définir comment organiser les éléments au sein de chaque zone de regroupement de la grille
Présentation	Eléments accessibles : il s'agit de définir le contenu du document de synthèse • *Niveau grille* : il faut spécifier quels éléments seront directement affichés dans la grille • *Niveau documents associés* : il faut spécifier quels documents complémentaires sont accessibles, et comment
	Apparence : il s'agit de spécifier l'aspect décoratif du document de synthèse (palette de couleurs, fontes, taille des images...)

Tableau 1. Vue hiérarchisée de l'espace de la tâche de synthèse

l'exemple considéré, il peut s'agir de la forêt avec des attributs de type surface, espèces des arbres qui la composent... Cette catégorie constitue le **domaine d'étude**. Parallèlement, des connaissances génériques ou techniques concernant l'approche expérimentale de ce domaine d'étude sont définies. Ces connaissances du **domaine expérimental** ne feront en général pas l'objet d'une interrogation directe. Il s'agit par exemple de critères de représentativité statistique, ou de validité temporelle, ou de contraintes particulières de visualisation. Le contexte technique sera également pris en compte (supports d'affichage disponibles en particulier).

La représentation des connaissances du domaine implique aussi la définition de la configuration des documents de synthèse. Comme évoqué précédemment, des **modèles de documents** peuvent être définis pour chaque domaine d'étude. Ces modèles décrivent les éléments à présenter dans le document dans son ensemble et au sein de chaque case de la grille documentaire.

4.3. *Ontologie de la tâche analytique*

Les diverses études que l'utilisateur peut vouloir réaliser, représentées par des requêtes analytiques, peuvent être regroupées en familles d'études. Ces familles d'études peuvent être organisées au sein d'une **ontologie de la tâche analytique**. L'élément essentiel pour chaque famille d'étude est de définir comment manipuler les objets et concepts inclus dans l'ontologie de domaine afin de permettre l'adaptation du document de synthèse à la question de l'utilisateur. Il s'agit de la définition du modèle d'adaptation correspondant à la famille d'études considérée. Pour chaque famille d'étude, il s'agit donc de définir un modèle générique qui doit permettre :

— de faciliter l'expression de son objectif par l'utilisateur, celui-ci n'ayant qu'à sélectionner le modèle approprié et le paramétrer,

— de guider le processus de synthèse par les critères qui lui sont associés.

La réalisation de cette ontologie implique:

— la définition d'un formalisme de représentation pour les modèles d'adaptation. Ceux-ci constituant des modèles de tâche, un formalisme proche de ceux utilisés en Intelligence Artificielle pour représenter ce type d'éléments peut être adopté. Ce formalisme est en cours de définition,

— l'acquisition de connaissances concernant les tâches analytiques et leur représentation avec le formalisme précédemment évoqué. Cette acquisition de connaissances est en cours mais elle est rendu difficile par :

– la quasi-absence de travaux dans ce domaine,

– le panel très étendu de tâches analytiques possibles qui rend difficile leur recensement même partiel et la définition de relations claires entre catégories,

– la dépendance étroite qui existe entre certains types de tâches analytiques et le domaine d'application. Certaines tâches sont en effet spécifiques à un domaine ou différent légèrement dans leur définition d'un domaine à l'autre.

5. Procédure de synthèse

5.1. *Formalisation de la requête*

La formalisation de la requête est un préalable indispensable à la construction du plan d'adaptation. Elle peut s'avérer un problème complexe. Un exemple va nous permettre d'en illustrer les principaux éléments.

Si l'on considère à nouveau le domaine d'application de la gestion du territoire, une étude à réaliser pourrait être l'« évolution au cours du temps de la déforestation de la forêt amazonienne entre 1950 et 2000 ». Pour cette requête de la famille « Évolution », il s'agit de définir :

— la tâche analytique (« Évolution » dans l'exemple considéré);

— les paramètres de la tâche analytique : ici, l'élément à étudier est la déforestation, qui peut être représentée par la surface défrichée. Il s'agit d'une étude au cours du temps, qui peut être représentée au niveau annuel. Les deux paramètres de la tâche analytique sont donc les suivants :

[Évolution] [Paramètre1] = [surface défrichée]

[Évolution] [Paramètre2] = [Temps] [Intervalle] [Année],

— les critères d'inclusion : il s'agit d'une série de contraintes sur des valeurs d'éléments associées aux documents qui permettront de sélectionner les documents pertinents à intégrer. Dans l'exemple de la déforestation, ces critères définissent la région (l'Amazonie), le type d'exploitation (la forêt), et la période de temps à considérer (entre 1950 et 2000). ils sont représentés par les éléments suivants :

[Région] = [Amazonie],

[Type exploitation] = [Forêt],

[Année] => [1950] et [Année] =< [2000],

— le modèle logique de document, squelette de document de synthèse qui permettra de définir quelles sont les informations à présenter. Pour le même exemple, il s'agira d'un modèle générique pour une étude territoriale spécifié par le couple :

[Modèle de Document] = [Analyse territoriale],

— les préférences : ensemble de valeurs de paramètres qui permettront d'affiner le modèle de document ; il s'agit par exemple du nombre d'éléments par page ou par case de la grille. Aucune préférence n'étant spécifiée dans la requête d'exemple, une taille de grille par défaut sera utilisée.

Cette décomposition devrait aider à la construction d'un formalisme pour la requête, comme présenté dans le Tableau 2.

Élément de la requête	**Formalisation ([Élément père]...) [Élément] (= [Value])**
Tâche analytique	[Évolution]
Paramètres de la tâche analytique	[Évolution] [Paramètre1] = [surface défrichée] [Évolution] [Paramètre2] = [Temps] [Intervalle] [Année]
Critères d'inclusion	[Critère d'inclusion] [Région] = [Amazonie] [Critère d'inclusion] [Type exploitation] = [Forêt] [Critère d'inclusion] [Année] => [1950] [Critère d'inclusion] [Année] =< [2000]
Modèle logique de document	[Modèle de Document] = [Analyse territoriale]
Préférences	[Taille Grille] = [défaut]

Tableau 2. Exemple de formalisation d'une requête analytique

5.2. *Décomposition du processus de synthèse*

Etant donnée une requête, la construction du plan d'adaptation implique trois niveaux de spécialisation successifs :

— Niveau Domaine : ce niveau contient les connaissances concernant le domaine d'étude et le domaine expérimental, soit la collection complète des critères pour le domaine d'application considéré,

— Niveau Tâche Analytique : ce niveau contient un extrait de la collection précédente, sélectionné, organisé et paramétré en fonction de la famille d'études correspondante, appelé modèle d'adaptation,

— Niveau Requête : ce niveau contient le plan d'adaptation, spécialisation du modèle d'adaptation selon les éléments spécifiés dans la requête.

Parallèlement, trois étapes de composition, correspondant aux trois sous-problèmes à résoudre sont distingués:

— Niveau Factuel : il correspond à l'étape de sélection où les données ou faits sont analysés afin de proposer une liste d'éléments pertinents,

— Niveau Logique : il correspond à l'étape d'organisation spatiale où un arrangement thématique de la liste précédente est réalisé,

— Niveau de Présentation : il correspond à l'étape de présentation où le document précédent est préparé pour l'affichage en prenant en compte d'éventuelles préférences utilisateur.

La génération d'un document de synthèse adapté à une requête utilisateur implique la manipulation de l'ensemble de ces niveaux par un moteur d'adaptation.

5.3. *Architecture du moteur d'adaptation*

La génération du document de synthèse est un processus en deux phases, réalisé par le moteur d'adaptation présenté Figure 1. Il faut tout d'abord générer le plan d'adaptation en parcourant les trois niveaux de spécialisation décrits précédemment. Ce parcours passe par :

— Le choix de la Collection de Critères, au Niveau Domaine, pour l'application considérée,

— Le choix d'un Modèle d'Adaptation pour la famille au Niveau Tâche Analytique,

— La spécialisation du Modèle d'Adaptation en fonction de la requête, au Niveau Requête.

Il s'agit ensuite de construire un document de synthèse adapté à la requête : ceci s'effectue selon les trois niveaux de composition précédemment distingués :

— Étape de Sélection : au Niveau Factuel, un processus de composition factuel est utilisé pour sélectionner un ensemble de documents pertinents regroupés au sein d'un Document Virtuel Orienté Collection, selon un Modèle Factuel de Document,

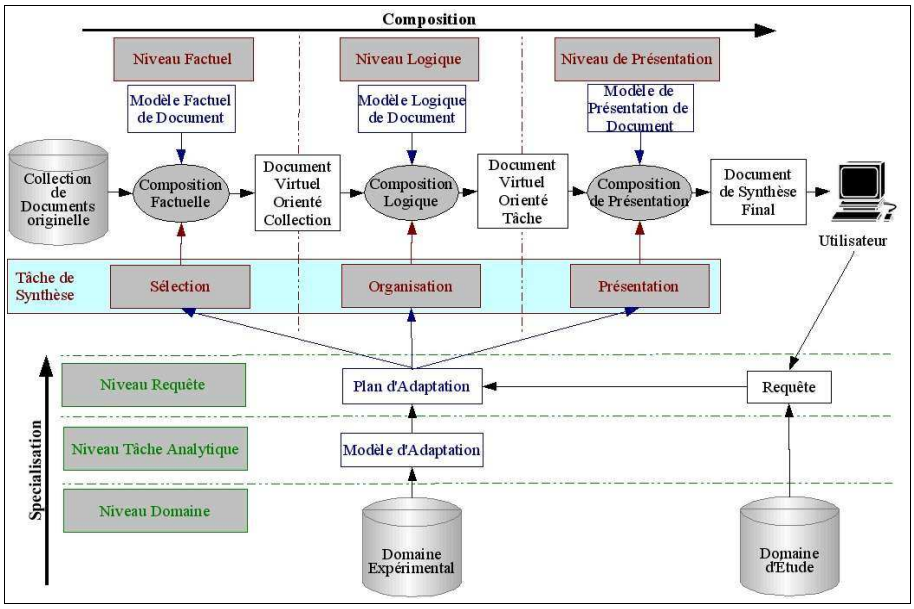

Figure 1. Architecture du moteur d'adaptation

— Étape d'Organisation Spatiale : au Niveau Logique, un processus de composition logique permet d'agencer spatialement la liste précédente en un Document Virtuel Orienté Tâche, en suivant un Modèle Logique de Document,

— Étape de Présentation : au Niveau de Présentation, un processus de composition de présentation sert à préparer l'affichage du Document de Synthèse final, en suivant un Modèle de Présentation de Document.

6. Application au domaine des Tissue MicroArrays

Nous proposons ici une application des éléments décrits précédemment au domaine des Tissue MicroArrays. Après une présentation de la technologie et l'étude de son adéquation aux paradigmes de Recherche d'Information Analytique, nous verrons comment spécialiser les composants génériques présentés pour ce domaine d'application particulier. Nous présenterons enfin une validation préliminaire de notre approche en utilisant ce domaine d'application comme support.

6.1 *Présentation de la technologie*

La technique des « Tissue MicroArrays » (TMA) est une technologie récente, déjà très utilisée en oncologie pour l'aide au pronostic et au suivi thérapeutique (Kallioniemi *et al.*, 2001)(Hoos *et al.*, 2001). Elle permet, en complément d'études moléculaires globales, la détection rapide des cibles moléculaires (séquences ADN, ARN ou protéines) dans des milliers de spécimens de tissu à la fois. Selon cette technique, on sélectionne, en fonction de l'étude à réaliser, des patients dont des biopsies sont disponibles en archive. Un pathologiste analyse une lame histologique de chacune de ces biopsies et détermine les régions d'intérêt de l'échantillon à étudier. Dans le bloc de paraffine de la biopsie (bloc donneur), des carottes de tissu sont prélevées en correspondance avec les zones pertinentes prédéfinies. Ces carottes sont insérées dans un bloc receveur vierge (bloc TMA) à partir duquel des lames sont réalisées et traitées comme le seraient des lames histologiques conventionnelles. Ces lames TMA font alors l'objet d'une acquisition d'image à différents grossissements. Les images sont ensuite partitionnées en images individuelles de spots (correspondant à la coupe de chaque carotte du bloc), qui font l'objet d'une annotation anatomopathologique et d'une analyse d'image pour quantification de marquage.

Par rapport à des études menées avec des techniques classiques, celles utilisant la technologie des TMA permettent des économies de réactifs et de matériel biologique. De plus, le traitement en masse d'une collection d'échantillons de tissus apporte une dimension statistique au travail du pathologiste. Ces deux avantages peuvent être encore accentués par le recours au concept de lame TMA virtuelle : des images de spots existantes peuvent être sélectionnées et réagencées en fonction d'une nouvelle étude sans nécessiter la construction d'un nouveau bloc.

Même si cette technologie semble prometteuse elle souffre d'un manque de connaissances formalisées et d'automatisation de la préparation du plan d'expérience et de la fouille de données. Étant donné la complexité de ces tâches et considérant la grande quantité de données à manipuler il apparaît nécessaire de concevoir un système informatique pour assister à la réalisation de ces opérations.

Or, même si un grand pas a été franchi avec la définition de la « TMA Data exchange specification » (Berman *et al.*, 2003), les outils développés autour de la technique se consacrent surtout à de la gestion de données (Henshall, 2003) (Shergill *et al.*, 2004). Il paraît donc nécessaire de proposer un outil d'assistance à l'utilisation de cette technologie intervenant à deux étapes du cycle présenté ci-dessus :

— aide à la conception de blocs TMA réels, par génération de représentations virtuelles de blocs TMA à fabriquer en fonction de l'étude à réaliser,

— accompagnement de la fouille de données par génération de lames TMA virtuelles associées à des informations pertinentes pour l'étude en cours.

6.2. *L'élaboration de TMA comme tâche de synthèse*

La technologie des TMA constitue un outil de recherche médicale. L'assistance dans la mise en place et l'analyse d'expériences ayant recours à cette technique implique la manipulation d'un gros volume de données et de documents de types variés. L'objectif de cette assistance est de proposer une sélection et organisation des spots ou carottes permettant de répondre à la question biologique ou médicale. En conséquence, les paradigmes de Recherche d'Information Analytique paraissent bien adaptés.

Les éléments construits dans le cadre de cette technique consistent en un arrangement d'échantillons ou images d'échantillons selon une matrice lignes / colonnes. Ils sont donc proches par nature des documents générés par notre système. Cela facilite l'appropriation intellectuelle de la représentation abstraite sous forme de document de synthèse par rapport à l'objet concret qu'est la lame TMA. L'assistance proposée présente la particularité très intéressante de conduire à la construction à la fois d'éléments virtuels mais aussi d'éléments physiques. Cela implique la définition d'une ensemble de critères très particuliers liés à la réalisation d'objets réels, ce qui permet d'expérimenter le système envisagé dans un contexte hautement spécialisé. Au domaine d'étude est donc ici associé un domaine expérimental complexe.

Étant donné une requête analytique représentant l'étude à réaliser, il s'agit donc de proposer une représentation de lame ou bloc TMA, consistant en un arrangement de spots ou carottes sur une grille, chaque spot ou carotte étant caractérisé par un ensemble d'informations extraites d'une base de données. Ces deux types de représentations peuvent être considérés comme des collections de documents multimédia comportant :

— la requête analytique utilisateur, qui permet la définition de l'étude à réaliser en utilisant la technologie (par exemple la comparaison entre deux groupes de patients...),

— une grille TMA, constituée d'un assemblage d'images de spots ou de références de carottes sélectionnés et organisés spatialement en fonction de la requête utilisateur; à chacun de ces éléments peuvent être associées des informations concernant:

– le patient correspondant : dossier clinique...

– l'analyse et l'annotation d'un ensemble complexe d'images : quantification de marquage, description de structures tissulaires...

À l'avenir, on peut envisager d'intégrer des références à des études similaires, de la littérature pertinente (PubMed...), ou des informations concernant les molécules étudiées, tirées de bases de données généralistes telles que SwissProt, GenBank...

Les caractéristiques de la problématique et du document à construire impliquent une bonne adéquation entre le problème de conception de TMA et l'outil de génération de documents de synthèse tel qu'il est envisagé. Les paragraphes suivants présentent donc les ontologies spécifiques à la conception de TMA qui ont été réalisées et les spécificités apparues au sein du processus de synthèse.

6.3.Ontologies pour la conception de TMA

6.3.1. *Ontologie de domaine*

Construire une ontologie de la conception de TMA implique tout d'abord une représentation du domaine d'étude. Celle-ci implique la modélisation du champ de la pathologie étudiée, ici le cancer du côlon. Les ontologies médicales actuellement disponibles sont souvent trop générales (Bodenreider, 2004). Une ontologie du cancer du côlon (Figure 2), répertoriant une centaine de termes du niveau organe au niveau cellule, a été réalisée par un pathologiste (Dr. Simony-Lafontaine, du Centre Régional de Lutte Contre le Cancer de Montpellier).

Il est aussi nécessaire de représenter les autres objets et concepts concernant la technologie, qu'il s'agisse d'éléments purement informatiques, comme des dossiers cliniques, ou d'objets réels, tels les lames histologiques ou blocs TMA. Toutes ces notions ont été intégrées dans une ontologie réalisée avec l'outil Protégé2000 de l'université de Stanford. Cette représentation a guidé la construction d'une base de données relationnelle contenant actuellement les dossiers cliniques et données concernant le matériel biologique pour une centaine de patients.

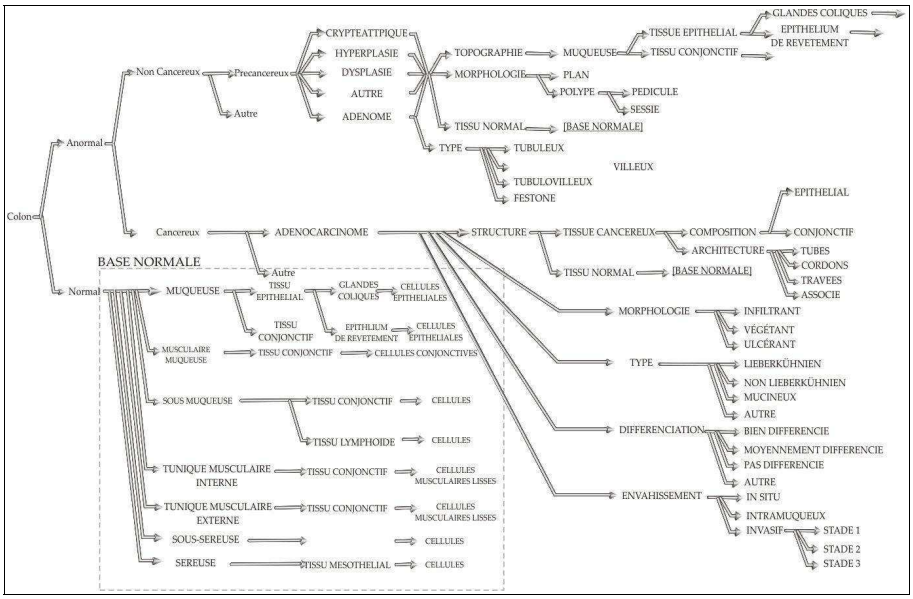

Figure 2. Section de l'ontologie du domaine d'étude correspondant à la description du côlon et ses pathologies

Un modèle logique de document doit aussi être défini, en adéquation avec les besoins exprimés dans le paragraphe précédent. Spécifique à la technologie des TMA, il définit les composants présents dans un document de synthèse associé à ce domaine d'étude particulier.

En complément, il faut aussi représenter le domaine expérimental, qui consiste en une collection de critères techniques et règles de conception spécifiques au domaine considéré. Ces critères à prendre en compte peuvent être organisés en fonction des étapes du processus d'adaptation. Quelques exemples de critères spécifiques au domaine des TMA sont présentés dans le Tableau 3.

6.3.2. *Ontologie de la tâche analytique*

La définition d'une ontologie de la tâche analytique reste encore à construire, et sa spécialisation pour le domaine particulier de la technologie des TMA en est elle aussi à un stade très préliminaire. Néanmoins, une première étude du type d'expériences que les biologistes et médecins pourraient réaliser par le biais de cette technologie a conduit à la réalisation d'une liste partielle de familles d'études.

Type	*Critères*
Sélection	Critères de représentativité : • *Variabilité* : préserver la variabilité des patients en se basant sur les données cliniques et biologiques • *Répétitions* : plusieurs copies de la même carotte au sein du même bloc pour pallier aux pertes de matériel biologique lors des coupes
	Critères de validité : • *Epuisement des blocs* : dans la cas de la construction d'un bloc réel, le matériel biologique correspondant doit être disponible • *Données manquantes* : les patients comportant des données manquantes doivent être intégrés en dernier recours
	Critères d'économie : • *Réutilisation* : en cas de requêtes proches réutiliser si possible des blocs déjà construits
Organisation spatiale	Critères physique : • *Taille des blocs* : dans le cas d'un bloc physique, le format du bloc influence la taille de grille • *Diamètre de l'aiguille* : dans le cas d'un bloc physique, le diamètre de l'aiguille utilisée pour prélever les carottes influence le format de la grille
	Critères d'économie : • *Réutilisation* : en cas de requêtes proches réutiliser si possible des blocs déjà construits
Présentation	Préférences utilisateur : • préférences de taille d'affichage, exprimées dans un profil stocké • préférences de thème de couleurs, exprimées dans un profil stocké

Tableau 3. Exemples de critères spécifiques au domaine des TMA

6.4.Validation préliminaire de l'approche

6.4.1. *Problématique*

Afin de réaliser une validation préliminaire de l'approche présentée précédemment, nous proposons de nous intéresser aux deux points d'ancrage de la méthode : la notion de requête analytique, et la notion de document de synthèse.

— Requête analytique : la problématique majeure à ce niveau est la traduction de la requête sous forme d'un ensemble de critères. À partir d'une requête formulée en langue naturelle, est-on capable de la traduire en une requête analytique, c'est-à-dire un ensemble de critères élémentaires permettant de guider le processus de synthèse ?

— Document de synthèse : cette construction, visualisée sous la forme d'un document multimédia, permet-elle d'inférer des connaissances et est-elle effectivement pertinente pour l'utilisateur ?

Pour ce faire, nous avons réalisé une première expérimentation en utilisant une base de données élaborée dans le cadre d'une collaboration avec le Centre Régional de Lutte Contre le Cancer de Montpellier, et l'outil TreeMaps[1] de l'Université de Stanford.

6.4.2. *Matériel et Méthodes*

Des patients atteints d'un cancer du côlon suivis au Centre Régional de Lutte Contre le Cancer de Montpellier ont été sélectionnés. Pour chacun des 51 individus concernés, nous disposons de leur dossier clinique et de blocs de biopsie prélevés lors de l'ablation de leur tumeur dans des zones tumorales et supposées saines. Un ensemble de marqueurs tumoraux ont été révélés sur des lames histologiques réalisées à partir de ces biopsies. Le pourcentage de cellules marquées, l'intensité et l'hétérogénéité de marquage pour chacune de ces molécules a été évaluée par un anatomopathologiste (Dr. Joëlle Simony-Lafontaine du CRLCC de Montpellier).

Afin de valider l'approche Recherche d'Information Analytique présentée dans cet article, nous nous proposons de nous intéresser à l'un de ces marqueurs tumoraux, la β-caténine. Bien connue comme molécule d'adhésion cellulaire, elle est aussi une protéine capable, après un passage intranucléaire, d'activer la prolifération et d'inhiber la mort cellulaire. Dans le cas de cellules normales, cette molécule est présente largement au niveau membranaire, puis elle est dégradée. Elle se trouve donc en principe peu présente au niveau cytoplasmique, et absente au niveau du noyau cellulaire. Dans les cellules tumorales, des mutations empêchent cette dégradation. La molécule s'accumule dans le cytoplasme puis le noyau où elle accélère la prolifération cellulaire. La localisation de la β-caténine dans la cellule est donc un facteur discriminant entre cellules saines et tumorales.

Pour cette étude, nous avons donc choisi de nous intéresser à une requête analytique qui pourrait s'exprimer en langage naturel sous la forme : « comparaison de la localisation de l'expression de la β-caténine entre cellules normales et tumorales pour l'ensemble des patients ».

Pour construire le document de synthèse correspondant à cette requête analytique, nous proposons d'utiliser l'outil TreeMaps de l'Université du Maryland. Ce logiciel permet la visualisation graphique de hiérarchies de données. A partir de fichiers textes, il conduit à la construction d'un pavage, où chaque case représente une ligne du fichier, soit un individu. La valeur d'une variable pour chaque individu peut être codée par l'intensité de coloration ou la taille de la case. L'organisation hiérarchique des individus selon certaines variables peut être représentée par des regroupements entre cases. Enfin, les individus à observer peuvent être sélectionnés par la spécification de filtres particuliers.

[1] http://www.cs.umd.edu/hcil/treemap/

La validation consistera dans un premier temps à formaliser la requête analytique dans ce cadre puis à représenter un « substitut » de plan d'adaptation sous la forme du paramétrage d'un document construit sous TreeMaps. Il s'agira ensuite d'évaluer la visualisation graphique proposée.

6.4.3. *Résultats*

6.4.3.1. Expression de la requête analytique sous forme d'un ensemble de critères

Afin de valider la première partie du processus, il s'agit tout d'abord de formaliser la requête. Ainsi, pour la requête « comparaison de la localisation de l'expression de la β-caténine entre cellules normales et tumorales pour l'ensemble des patients » :

— Tâche analytique : il s'agit d'une tâche de « Comparaison », qui implique la définition de plusieurs éléments listés ci-dessous,

— Objectif de la comparaison : il s'agit des éléments à comparer. Ici, nous nous intéressons à l'expression de la β-caténine qui peut être représentée par exemple par le pourcentage de cellules marquées.

— Populations à comparer : il s'agit ici d'une comparaison entre cellules normales et tumorales en fonction de la localisation, plus précisément on souhaite construire un document de synthèse organisant les individus (patients) selon ces deux axes d'analyse du problème,

— Critères d'inclusion : on cherche à étudier l'ensemble des patients, donc nous n'inclurons pas de critères particuliers.

— Aucune préférence utilisateur particulière n'est spécifiée ici.

Ayant formalisé la requête comme présenté dans le Tableau 4, l'étape suivante est de construire un plan d'adaptation de type « comparaison » qui tienne compte des paramètres de la requête et du domaine d'étude. Du fait du caractère encore préliminaire de notre travail, et dans le cadre particulier de cette validation, nous nous contenterons d'exprimer un « substitut » de plan d'adaptation sous la forme d'un plan d'exécution de l'outil TreeMaps. En effet, les critères correspondant aux éléments de la requête peuvent être traduits sous forme de paramètres dans cet outil :

— [Comparaison] [Objectif] = [% cellules marquées en β-caténine] : dans TreeMaps, on indique un codage par intensité de coloration du pourcentage de cellules marquées en β-caténine,

— [Comparaison] [Population] = [Biologique] [Nature] : la nature tumorale ou normale des tissus est représentée par un premier niveau de hiérarchie, ce qui implique l'organisation spatiale des individus en deux régions, en fonction de la valeur de cet attribut,

Élément de la requête	**Formalisation ([Élément père]...) [Élément] (= [Value])**
Tâche analytique	[Comparaison]
Paramètres de la tâche analytique	[Comparaison] [Objectif] = [% cellules marquées en β-caténine] [Comparaison] [Population] = [Biologique] [Nature] {tumoral / normal } [Comparaison] [Population] = [Biologique] [Cellule] [Localisation] {membrane / cytoplasme / noyau }
Critères d'inclusion	/
Modèle logique de document	[Modèle de Document] = [TMA]
Préférences	/

Tableau 4. Exemple de formalisation d'une requête analytique dans le contexte de l'outil TreeMaps

— [Comparaison] [Population] = [Biologique] [Cellule] [Localisation] : la localisation du marquage dans la cellule correspond à un second niveau de hiérarchie. Au sein de chacune des régions présentées précédemment, cette définition conduit à un arrangement en trois zones, en fonction de la localisation cellulaire du marquage.

Le modèle d'adaptation « Comparaison », comme tout modèle d'adaption, comprend à la fois de critères paramétrables par la requête, soit des critères du domaine d'étude, et des critères spécifiques issus du domaine expérimental. Dans le cadre de l'analogie avec l'outil TreeMaps, les critères du domaine d'étude décrits ci-dessus sont représentés par des définitions de filtres et hiérarchies. Les critères du domaine expérimental correspondent à l'expression de l'algorithme sous-jacent à l'outil.

Nous avons finalement traduit une requête analytique en un ensemble de critères simples, soit en une ébauche de plan d'adaptation. Cette traduction a été effectuée manuellement, à des fins d'illustration. Cette illustration, quoique très partielle, semble suggérer que la mise en place d'un système de définition d'un plan d'adaptation en fonction d'une requête peut être assurée. Il convient maintenant d'étudier la complexité de ce processus de traduction et d'en définir les formalismes de représentation et les langages de description. Les langages classiquement utilisés dans le cadre du Web sémantique, comme XML, XSLT, RDF et OWL, seront étudiés dans ce sens (Ding *et al.*, 2004).

6.4.3.2. Visualisation graphique des résultats

Il s'agit maintenant d'observer si la représentation graphique des résultats est porteuse d'informations, en construisant un document de synthèse à l'aide de l'outil TreeMaps, selon les spécifications présentées au paragraphe précédent. Le fichier en entrée indique pour 51 patients une estimation du pourcentage de cellules marquées pour la β-caténine en fonction de la localisation cellulaire (membrane, cytoplasme ou noyau) et de la nature du tissu (tumoral ou normal). Ce pourcentage de marquage, qui correspond à l'objectif de l'étude, est codé par une intensité de coloration, du blanc (0% de cellules marquées) au noir (100% de cellules marquées). Afin de représenter les deux niveaux de comparaison requis, une hiérarchisation selon la nature du tissu puis la localisation cellulaire a été définie. Aucun filtre n'a été mis en place, puisqu'on s'intéresse à tous les patients.

Les résultats sont présentés Figure 3. Une observation d'ensemble rapide fait apparaîte que pour les cellules normales (en haut), la localisation de la β-caténine est surtout membranaire (à gauche), légèrement cytoplasmique (au milieu) et quasi inexistante dans le noyau (à droite). Pour les cellules tumorales au contraire (en bas), le marquage est important à la fois dans la membrane et le cytoplasme et elle est présente dans le noyau.

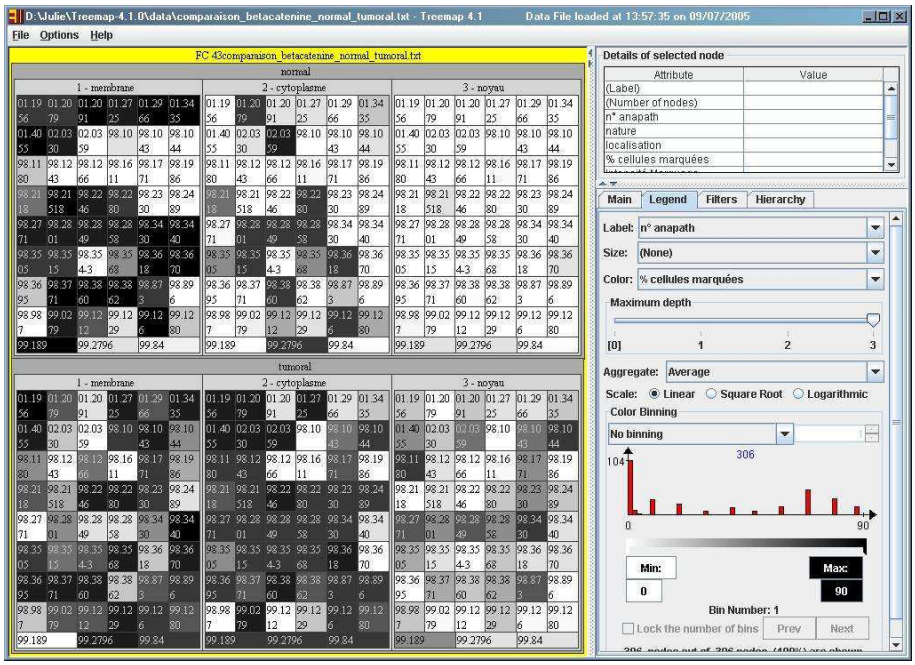

Figure 3. Comparaison de la localisation de l'expression de la - caténine entre cellules normales et tumorales pour l'ensemble des patients en utilisant TreeMaps

Or, biologiquement, nous savons que dans la cellule normale, la β-caténine est exprimée dans la membrane puis est dégradée. L'importante présence membranaire et la faible présence cytoplasmique révélées par la représentation graphique sous TreeMaps sont en accord avec ces connaissances biologiques. De même, dans la cellule anormale, il est connu que des mutations empêchent la dégradation de la β-caténine qui s'accumule dans le cytoplasme puis migre jusque dans le noyau où elle contribue à augmenter la prolifération cellulaire. Là encore la représentation sous TreeMaps fait apparaître visuellement cette accumulation cytoplasmique et nucléaire.

Le document de synthèse construit sous TreeMaps en réponse à la requête analytique initiale offre donc une visualisation graphique d'ensemble à la fois simple et en accord avec des faits biologiques connus. Il s'agit déjà d'un accomplissement important, qui laisse suggérer que de nouvelles hypothèses pourraient être validées par ce biais. Le couplage d'une telle visualisation avec des fonctions de navigation associées à l'espace documentaire ainsi construit offrira de nouvelles perspectives dans le domaine.

6.4.4. *Discussion*

Les résultats obtenus semblent très encourageants. Néanmoins, ils révèlent certaines limites de l'outil TreeMaps, pour chacune des deux composantes étudiées:

— Formulation de la requête :

– la requête n'est pas explicitement exprimée et n'est représentée que de manière dispersée, sous la forme d'un ensemble de filtres et de paramètres, ce qui rend difficile l'appréhension globale de la problématique biologique étudiée et sa relation avec le document qui a été construit,

– inversement, l'expérience réalisée donne un aperçu de la complexité potentielle du processus de traduction qui conduit à l'expression d'un jeu de critères à partir d'une requête analytique formulée en langue naturelle. Elle confirme donc l'apport d'un niveau de conceptualisation permettant l'expression de telles requêtes de manière formelle.

— Construction du document de synthèse :

– l'accent mis sur la notion de hiérarchie comme critère de base pour la visualisation implique la construction d'un fichier particulièrement adapté pour l'accès à certaines tâches. Ainsi dans le cas présenté précédemment, la tâche de « Comparaison » implique de multiples copies des mêmes données, une pour chaque ensemble de niveau le plus fin. Pour étudier une tâche d'« Évolution », il faudrait construire un autre type de fichier, avec une seule copie des données mais des éléments donc la valeur peut être étudiée en fonction d'un autre attribut.

– les possibilités de reformulation de requête sont limitées. Ainsi, par exemple, la visualisation présentée en exemple pourrait conduire le chercheur à

s'interroger sur les causes de l'absence de l'expression de la β-caténine chez certains patients. Il pourrait vouloir visualiser uniquement les patients avec un faible marquage dans toutes les parties de la cellule pour chercher les points communs entre ces patients. Ce type de sélection au sein du jeu de données visualisées n'est pas possible avec TreeMaps.

Le système envisagé devra apporter des solutions à ces problèmes.

7. Conclusion

Nous nous sommes intéressés dans cet article à une classe particulière de documents, que nous appelons « documents de synthèse », encore peu considérée dans la littérature. Nous avons caractérisé le type de requêtes correspondant, dites « requêtes analytiques » : ces requêtes formalisent de manière explicite l'objectif de l'utilisateur, et permettent d'intégrer la notion de tâche analytique dans le processus de recherche d'information. Nous avons proposé une architecture pour un moteur d'adaptation permettant la génération de documents de synthèse en réponse à des requêtes analytiques. L'architecture proposée repose sur une hiérarchie construite selon deux axes:

— Spécialisation : l'adaptation à plusieurs niveaux de spécialisation consiste en un raffinement progressif de la procédure d'adaptation vers un cas particulier. Du point de vue de l'ingénieur, cela facilite la procédure d'acquisition et de représentation des connaissances. Du point de vue de l'utilisateur, cela permet une formulation et reformulation flexible de la requête,

— Composition : l'adaptation à plusieurs niveaux de composition consiste en une construction progressive du document de synthèse final. Du point de vue de l'ingénieur, cela permet une décomposition de la tâche, ce qui facilite sa mise en œuvre. Du point de vue de l'utilisateur, cela apporte une possibilité de visualiser des résultats intermédiaires et facilite la formalisation de l'expertise.

Le recours à des modèles tels que les Modèles d'Adaptation ou Modèles Logiques de Documents apporte enfin une grande souplesse à l'ensemble du processus d'adaptation.

Cette approche, bien qu'encore à un stade exploratoire offre de nombreuses perspectives, que ce soit dans le domaine de la technologie des TMA ou pour la Recherche d'Information. La mise en oeuvre d'un prototype devrait permettre la validation du modèle proposé, et en particulier l'analyse de la faisabilité d'un outil vraiment générique.

Remerciements

Nous remercions la « Ligue Contre le Cancer – Comité de Savoie » et le programme "Bio-Informatique" inter-EPST, pour leur soutien à ces travaux, ainsi que le Dr. Joëlle Simony-Lafontaine, du CRLCC de Montpellier, pour son expertise et les jeux de données qui ont servi de support à cet article.

8. Bibliographie

Aula, A., « Query Formulation in Web Information Search », *Proceedings of IADIS International WWW/Internet 2003 Conference, Volume I*, Algarve, November 2003, IADIS Press, p. 403-410.

Baeza-Yates, R., Ribero-Neto, B., *Modern Information Retrieval*, Addison-Wesley Longman, May 1999.

Berman JJ., Edgerton ME., Friedman BA., « The tissue microarray data exchange specification: a community-based, open source tool for sharing tissue microarray data », *BMC Med Inform Decis Mak*, vol. 23, n° 3, 2003, p. 5.

Bloesch, A.C., Halpin, A., « Conceptual Queries Using ConQuer-II », *Lecture Notes In Computer Science, vol. 1331, Proceedings of the 16th International Conference on Conceptual Modeling*. 1997, p 113-126.

Bodenreider O., « The Unified Medical Language System (UMLS): integrating biomedical terminology », *Nucleic Acids Res., Database issue*, vol. 1, n° 32, 2004, p. 267-270.

Brusilovsky P., « From Adaptive Hypermedia to the Adaptive Web », *Communications of the ACM*, vol. 45, n°2, 2002, p. 31-33.

Card, S.K., Mackinlay, J., « The structure of the information visualization design space », *Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97)*. 1997, p 92-99.

Ding Y., Fensel D., Klein M., Omelayenko B., « The Semantic Web: Yet Another Hip? », *Data & Knowledge Engineering archive*, vol. 41, n°2-3, 2002, p. 205-227.

Fensel, D., Motta, E., « Structured Development of Problem Solving Methods », *Proceedings of KAW'98, Eleventh Workshop on Knowledge Acquisition, Modeling and Management*, April 1998, Banff, Alberta, Canada.

Friendly, M., Kwan, E., « Effect ordering for data displays », *Computational Statistics & Data Analysis*, vol. 43, n°5, 2003, p 509-539.

Großjohann, K, Fuhr, N., Effing, D., Kriewel, S., « A User Interface for XML Document Retrieval », *Informatik bewegt: Informatik 2002 - 32. Jahrestagung der Gesellschaft für Informatik e.v. (GI)*, 2002, p. 166-170.

Gruber T., « Toward principles for the design of ontologies used for knowledge sharing », *International Journal of Human-Computer Studies, Special issue: the role of formal ontology in the information technology*, vol. 43, n° 5-6, 1995, p: 907-928.

Guarino N., « Formal Ontology and Information Systems », *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98*, Trento, Italy, p. 3-15.

Henshall S., « Tissue microarrays », *J Mammary Gland Biol Neoplasia*, vol. 8, n° 3, 2003, p. 347-358.

Hoos A., Cordon-Cardo C., « Tissue microarray profiling of cancer specimens and cell lines: opportunities and limitations », *Lab Invest*, vol. 81, n° 10, 2001, p. 1331-1338.

Hunter, J., Falkovych, K., Little, S., « Next Generation Search Interfaces - Interactive Data Exploration and Hypothesis Testing », *Proceedings 8th European Digital Libraries Conference*, September 2004, p. 86-98.

Kallioniemi OP., Wagner U., Kononen J., Sauter G., « Tissue microarray technology for high-throughput molecular profiling of cancer », *Hum Mol Genet*, vol. 10, n° 7, 2001, p. 657-662.

Keim, D.A., « Information Visualization and Visual Data Mining », *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, n°1, 2002, p. 1-8.

Kohonen, T., Oja, E., Simula, O, Visa, A., Kangas, J., « Engineering applications of the self-organizing map », *Proceedings of the IEEE*, vol. 84, n° 10, 1996, p. 1358-84

Nocke, T., Schumann, H., « Goals of Analysis for Visualization and Visual Data Mining Tasks ». *Prague CODATA Workshop on Information Visualization, Presentation and Design*, 2004

Shergill IS., Shergill NK., Arya M, Patel HR., « Tissue microarrays: a current medical research tool », *Curr Med Res Opin*, vol. 20, n° 5, 2004, p.707-712.

Shneiderman, B., « Tree visualization with Tree-maps: A 2-d space-filling approach », *ACM Transactions on Graphics*. vol. 11, 1992, p. 92-99.

Singhal, A., « Modern information retrieval: A brief overview », *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, vol24, n°4, 2001, p. 35-43

Stuckenschmidt, H., de Waard, A., Bhogal, R., Fluit, C., Kampman, A., van Buel, J., van Mulligen, E., Broekstra, J., Crowlesmith, I., van Harmelen, F., Scerri, T., « A Topic-Based Browser for Large Online Resources », *Lecture Notes in Artificial Intelligence. Proceedings of the Proceedings of the 14th International Conference on Knowledge Engineering and Knowledge Management ({EKAW}'04)*. 2004. p.433-448.

Skupin, A., Fabrikant, S.I., « Spatialization methods: A Cartographic research agenda for non-geographic information visualization », *Cartography and Geographic Information Science.* vol. 30, n° 2, 2003, p. 95-119.

Wielinga, W., Schreiber, G., Breuker, J., « KADS: A modelling approach to knowledge engineering », *Knowledge Acquisition, Special issue "The KADS approach to knowledge engineering"*, vol. 4, n°1, 1992, p. 5-53

Wu H., De Bra P., Aerts A., Houben G-J., « Adaptation Control in Adaptive Hypermedia Systems », *Lecture Notes in Computer Science*. Vol. 1892, 2000, p. 250.