Modèle d'indexation de données peu symboliques dans des documents structurés : L'exemple du graphique dans un corpus de documents techniques

Leïla Kefi[*], Catherine Berrut[*], Eric Gaussier[**]

[*]*Equipe Modélisation et Recherche d'Information Multimédia*
Laboratoire CLIPS-IMAG – BP53- 38041 Grenoble cedex 9, France
Leila.kefi@imag.fr, catherine.berrut@imag.fr

[**]*Xerox Research Centre Europe -6 chemin de Maupertuis - 38240 Meylan, France*
Eric.gaussier@xrce.Xerox.com

RÉSUMÉ. *Cet article s'intéresse à l'indexation des données ayant une sémantique pauvre dans des documents structurés. Le but est d'exploiter le contenu des données symboliques avoisinantes afin d'en extraire les fragments adéquats pour compléter l'indexation de la donnée non symbolique. Cette approche a été abordée dans le cadre concret d'une application dans un contexte professionnel : indexer les graphiques des documents techniques en exploitant le texte qui les accompagne. Cette indexation est articulée autour d'un modèle de représentation des graphiques tenant compte de la finalité de leur utilisation et du professionnalisme de leurs usagers, et d'un modèle d'extraction des termes d'indexation à partir du texte du document technique.*

ABSTRACT. *This paper deals with data with little semantics in structured documents. The aim is to exploit symbolic data in order to extract the adequate fragments to complete the non-symbolic data indexation. This approach has been studied in a concrete application frame that has a professional context: indexing graphics in technical documentation exploiting the textual context in which they appear. This indexing is articulated around a representation model for graphics that takes into account their use and their professional users and an extraction model of the indexing terms from the text of technical documentation.*

MOTS-CLÉS : *indexation, documents structurés, donnée peu symbolique, graphique, contexte professionnel.*

KEYWORDS: *Indexation, structured documents, weakly symbolic data, graphics, professional context.*

1. Introduction

Dans les systèmes de recherche d'information traditionnels, la recherche se fait sur le document en tant qu'entité indivisible. Lorsque le système manipule des documents structurés, ces derniers sont organisés selon une certaine hiérarchie, formant alors une arborescence de blocs, éventuellement un graphe. Un article sera, par exemple, composé d'un titre, d'une introduction, d'un certain nombre de sections, elles-mêmes composées de paragraphes et d'une conclusion. Selon le besoin de l'utilisateur, l'un de ces blocs peut représenter une réponse plus pertinente à sa requête que le document en entier. Il faut donc que les systèmes de recherche d'information puissent donner, en réponse à une requête, une partie d'un document, voire une reconstitution personnalisée de ce document.

Les blocs formant le document structuré, loin d'être indépendants, sont reliés entre eux par des relations qui permettent de donner au document une intégrité sémantique. Certaines recherches ont tenu compte de cette dépendance dans le but d'enrichir l'indexation d'un bloc par l'ajout des index des blocs avoisinants ou supérieurs hiérarchiquement (la notion de portée, par exemple).

Plus particulièrement, dans les documents structurés multimédia, cette utilisation du contenu sémantique des blocs avoisinants pour enrichir l'indexation d'un bloc non textuel est nécessaire. En effet, certains blocs, comme les images, manquent d'un langage permettant d'en exprimer la sémantique. Et même si ces blocs contiennent en eux-mêmes des informations permettant leur description, celle-ci reste pauvre sémantiquement et ne permet pas de représenter les données en question de façon suffisante et par conséquent de les retrouver efficacement.

Dans les approches actuelles, l'index d'un bloc sans sémantique est considéré comme étant fonction de l'ensemble des descripteurs des blocs avoisinants. Nous désirons aller au-delà de cette approche et considérer l'indexation du bloc sans sémantique comme étant fonction de certains fragments extraits des blocs alentours, l'extraction dépendant du bloc sans sémantique. Par conséquent, l'indexation de ce bloc tiendra de ce qui fait sens pour lui et non plus de ce qui fait sens autour de lui dans les blocs alentours. Son indexation dépendra alors de sa nature et de ce que l'on peut éventuellement en savoir, voire en extraire.

Nous avons souhaité aborder cette approche, dans une première étape, dans le cadre concret d'une application. Nous avons choisi l'usage des documents techniques de type manuels d'utilisation (des descriptions d'imprimantes). Ce choix a été, entre autres, motivé par le contexte professionnel dans lequel se situe cette application. Dans une telle application, l'usager et la finalité d'utilisation des documents nécessitent que le système réponde avec précision aux attentes de l'utilisateur. Ces aspects concernant l'usager, l'usage et la précision, nous ont semblé intéressants à considérer dans notre étude.

Dans les documents techniques, nous nous sommes plus particulièrement orientés vers le média graphique comme donnée non symbolique, car bien que des travaux sur l'analyse et l'indexation de ce média existent déjà, peu de recherche s'est concentrée sur l'enrichissement que peut apporter le contenu sémantique d'un autre média à son interprétation. Qui plus est, ce média offre un biais intermédiaire entre le « tout sens » et le « sans sens ». En effet, il a une sémantique pauvre, mais il peut être considéré comme un objet structuré qui offre des informations intrinsèques intéressantes (traits, zooms, flèches, etc.) permettant une réflexion intéressante sur ce qui peut améliorer son indexation. Ainsi, nous désirons aussi décomposer le graphique pour l'indexer de façon précise au lieu de nous contenter de l'indexer globalement.

Nous nous intéressons, dans une première partie, à l'effet du contexte professionnel de l'application sur l'indexation des graphiques (§2). Ensuite, nous considérons le graphique non plus comme un objet fermé et global non analysable, mais comme un objet contenant des informations de natures multiples qui améliorent son indexation (§3). À partir de ces deux points, nous définissons un modèle de représentation des graphiques en adéquation avec leur nature et les besoins des utilisateurs (§4).

Dans une seconde partie, nous nous penchons sur l'enrichissement de la donnée graphique par le contenu des blocs textuels qui l'entourent dans un document technique (§5). Cette partie représente une étude préalable et concrète visant à démontrer la possibilité d'améliorer l'indexation du graphique par des fragments extraits des blocs de textes la commentant.

2. La recherche d'information dans les documents techniques à usages professionnels

Les documents techniques qui nous intéressent sont à usage professionnel. Cela nous amène à considérer trois points :

Tout d'abord, ces documents sont fréquemment consultés par les utilisateurs, ce qui entraîne une certaine mémorisation visuelle des graphiques qu'ils contiennent. (§2.1)

Ensuite, ces documents véhiculent des savoirs et des savoir-faire propres à un champ technique particulier. Ainsi ils représentent aussi bien la description d'une machine, de son fonctionnement et des divers processus la concernant, que la description des procédures de réalisation d'une action technique dans un environnement précis. Cette finalité d'utilisation des documents techniques qui doit être prise en compte dans la description du graphique est détaillée dans §2.2.

Enfin, les utilisateurs de documents techniques sont très exigeants quant à l'information qu'ils recherchent, le système devra donc être orienté précision, d'où

la nécessité d'une indexation spécifique : représenter les éléments intéressants de manière détaillée. (§2.3)

2.1. *Prise en compte des besoins et habitudes des utilisateurs*

Les manuels d'utilisation des composants matériels sont des documents utilisés fréquemment par les professionnels. Ces utilisateurs connaissent le contenu des documents qu'ils manipulent et lorsqu'ils recherchent une information, ils souhaitent accéder à un fragment précis d'information qu'ils ont déjà vu et qu'ils connaissent : ils savent à priori ce qu'ils recherchent. Ils ne désirent pas trouver « **un** graphique décrivant l'écran d'affichage de l'imprimante DocuPrint N17 », mais ils veulent retrouver « **le** graphique qui décrit cet écran » et dont ils se rappellent.

Plus précisément, les professionnels se souviennent de la géométrie du graphique. Par exemple, ils peuvent vouloir retrouver le graphique décrivant le chargement d'incidents papiers dans le magasin d'alimentation manuelle et qui, dans leur souvenir, contient un zoom en haut à droite et une flèche descendante du côté gauche.

En fait, les graphiques des documents techniques contiennent des données géométriques (formes, traits, positions, zooms etc.) et de nombreuses expériences ont montré que lorsqu'une personne est confrontée à un énoncé de "spatialisation" organisant des objets les uns par rapport aux autres, elle construit mentalement une représentation de la scène (Michel D. 1997). Cette représentation mentale du graphique peut s'ancrer partiellement ou totalement dans la mémoire de l'utilisateur. On parlera, dans ce cas, de sa mémoire visuelle. La trace laissée par le graphique dans la mémoire de l'utilisateur peut ainsi représenter une requête probable pour retrouver le graphique en question.

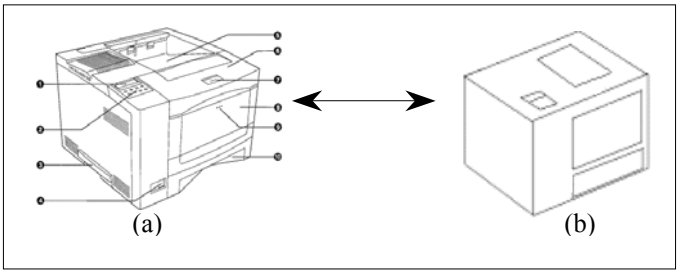


Figure 1 : *Exemple d'empreinte d'un graphique (a) sur la mémoire visuelle (b)*

La prise en compte de la **mémoire visuelle** de l'utilisateur a donc une importance dans la représentation des graphiques des documents techniques et elle doit être prise en compte.

2.2. *Prise en compte de la finalité d'utilisation*

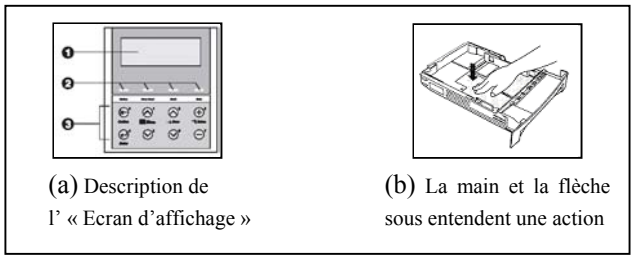
Les documents techniques permettent de répondre à deux grandes classes de requêtes :

- « Qu'est ce que 'x' ?» : Le document fournit alors des informations contribuant à l'énumération des composants de l'objet 'x', la description de leurs propriétés et de leurs fonctions (de l'objet 'x' et de ses composants).

- « Comment faire pour 'x' ?» : le document fournit alors des informations indiquant l'action générale à accomplir, le cas d'application de cette action, les actions intermédiaires à effectuer sur les composants de la machine et les conditions et l'ordre d'exécution de ces actions.

Afin de diminuer l'effort de modélisation de l'utilisateur, les graphiques permettent de schématiser les informations citées ci-dessus et l'utilisateur peut alors facilement les localiser dans le graphique.

On retrouve alors deux types de graphiques. Certains sont à visée descriptive et on y distingue une présentation des objets qu'ils contiennent. D'autres sont à visée opératoire et on y trouve la définition d'une liste d'actions à appliquer sur les objets qu'ils contiennent.



(a) Description de l' « Ecran d'affichage »

(b) La main et la flèche sous entendent une action

Figure 2 : *Exemples de graphiques à visée descriptive (a) et opératoire (b)*

Le graphique est ainsi perçu comme une description de la sémantique de son contenu, avec une distinction entre un aspect descriptif et un **aspect opératoire**. Ce dernier aspect est une particularité des graphiques des documents techniques à usage professionnel et il doit nécessairement être pris en compte lors de la modélisation de tels graphiques.

2.3. *Précision de l'indexation*

Lors de la rédaction d'un texte technique, les auteurs de la documentation suivent des règles bien précises. Ils doivent se montrer particulièrement vigilants sur la pertinence de leurs phrases, l'homogénéité du texte, les styles de phrases et le vocabulaire employé.

Ceci est dû à la finalité d'utilisation des documents techniques et au « professionnalisme » de ses utilisateurs.

De même et pour les mêmes raisons, le langage d'indexation des graphiques doit être spécifique.

D'un autre côté, lorsqu'ils recherchent une information, les utilisateurs des documents techniques désirent accéder au fragment du document qui les intéresse. Le système doit alors donner en réponse précisément ce fragment recherché. Il doit donc être orienté précision. Pour ce faire, l'indexation d'un graphique doit être détaillée et ses index doivent suivre une structure bien précise. Le maximum d'information le concernant devra être représenté (§3) et rangé dans un schéma logique.

3. Le graphique des documents techniques: Des informations multiples à indexer

Le graphique dans les documents techniques est un média qui, une fois considéré comme un objet analysable et non plus comme une donnée élémentaire, contient une multitude d'informations. En effet, outre les informations internes qu'il est possible d'en extraire (traits, flèches, zoom, etc.), il s'agit d'un objet structuré ayant une sémantique qui même si elle n'est pas explicite dans celui-ci est mentionnée dans le texte qui le commente.

Dans ce qui suit, nous résumons les différentes interprétations possibles du graphique contenu dans la documentation technique :

– Le graphique est un objet complexe et multi-facettes, dont les éléments intéressants sont liés entre eux par des relations de composition. Par exemple, L'« Imprimante N17» est composée du « Magasin papier », de la « Poignée », etc. Ces éléments peuvent être soit des composantes matérielles, soit des formes particulières rajoutées par les auteurs pour exprimer une action, une énumération ou un agrandissement, etc.

– Le graphique est accompagné d'une description de la sémantique de son contenu, avec un aspect opératoire et un aspect descriptif.

– Le graphique est perçu comme un ensemble de formes géométriques disposées d'une manière particulière et ayant chacune un degré d'importance particulier.

– Le graphique visualisé s'imprime dans la mémoire visuelle des utilisateurs. Cette mémoire leur fournit alors un moyen d'accès aux graphiques. Certains objets de ce graphique auront plus d'impact sur leur mémoire visuelle que d'autres, selon l'importance de ces objets dans le graphique.

Les différents points énumérés ci-dessus représentent les grandes lignes qui nous ont guidés dans l'élaboration de notre modèle.

4. Proposition du modèle

La multiplicité de la nature des informations à indexer nécessite de définir un modèle pour la représentation des graphiques. Ces derniers étant structurés et multi-facettes, nous avons choisi de nous inspirer d'un modèle supportant ces deux aspects: Le modèle EMIR². (Mechkour M. 1995)

Nous commençons donc par présenter synthétiquement le modèle EMIR² avant de décrire notre modèle, qui est capable de représenter complètement les graphiques des documents techniques.

4.1. *Présentation de EMIR²*

EMIR² est un modèle qui « considère comme représentation du contenu d'une image diverses interprétations de l'ensemble des objets images et des relations qui les lient.» (Mechkour M. 1995)

À chaque interprétation, ou plus exactement vue, est associé un modèle donnant une description des objets contenus dans l'image, des relations qui peuvent les relier et des opérations applicables sur ces descriptions.

EMIR² défini cinq vues :

– la vue physique correspond aux données brutes de l'image (matrice de pixels, dimensions, etc.),

– la vue structurelle représente l'ensemble des éléments pertinents de l'image reliés par des relations de composition,

– la vue symbolique représente l'interprétation sémantique des éléments contenus dans l'image,

– la vue spatiale représente l'ensemble des objets géométriques associés aux contours des éléments contenus dans l'image et leurs inter-relations,

– la vue perceptive représente l'ensemble des attributs visuels objectifs des éléments contenus dans l'image (couleur, texture et brillance).

4.2. *Description de notre modèle*

Contrairement au modèle que nous visons, le modèle EMIR² fait abstraction du domaine des images, de l'application particulière qui les manipule et du type des utilisateurs auquel elle est destinée.

Donc, en tenant compte du média particulier vers lequel nous nous sommes orientés, il est apparu que la vue perceptive est inutile (il n'y a pas de couleurs, ni de textures dans les graphiques) et que des vues structurelles et spatiales adaptées aux graphiques doivent être définis.

Et le fait de dédier le système à des utilisateurs spécialisés entraîne qu'il faut prendre en compte, d'une part, une sémantique particulière, ce que nous ferons dans la vue opératoire, et d'une autre part, la mémoire visuelle de l'utilisateur et nous le ferons dans la vue dite mémoire visuelle.

Nous avons ainsi abouti à un modèle dont les différentes vues et leurs interrelations sont schématisées dans la figure suivante :

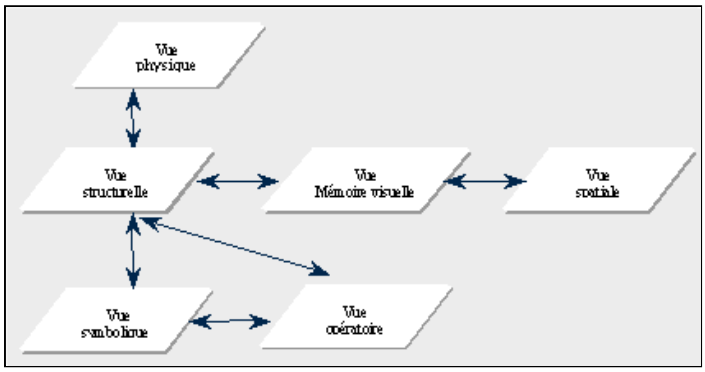


Figure 3 : *Vue synthétique du modèle proposé*

4.2.1. *La vue physique*

C'est la vue la plus élémentaire d'un graphique. Elle rassemble les caractéristiques générales du graphique, telles que ses dimensions, sa résolution, sa matrice de pixels.

4.2.2. *La vue structurelle*

Elle représente la décomposition d'un graphique en objets graphiques. Nous faisons la distinction entre les objets graphiques représentatifs (représentant un objet réel tel que une cartouche) et les objets graphiques illustratifs (représentant un zoom, une énumération, une main, etc.)

À cette vue correspond une relation de composition « CONTIENT » reliant des objets graphiques entre eux. D'autres relations lient les objets illustratifs aux objets représentatifs. Ainsi, la relation « EQUIVAUT » relie un zoom ou une représentation plate et l'objet visé par l'objet illustratif, la relation « NUMEROTE » relie une numération et l'objet pointé, et la relation « ACTION_SUR » relie une flèche ou une main et l'objet sur lequel s'effectue l'action.

La vue graphique est, de ce fait, représentée par un graphe dont les nœuds sont les objets graphiques et les arcs les relations citées ci-dessus.

4.2.3. *La vue spatiale*

Cette vue permet de représenter les formes des objets graphiques et leurs dispositions les uns par rapport aux autres. Elle comporte donc des informations géométriques décrivant les objets spatiaux ainsi que leurs inter-relations (telles que leurs positions relatives). La forme des objets spatiaux est représentée par une combinaison d'éléments géométriques de base qui sont les points, les segments et les polygones.

4.2.4. *La vue symbolique*

Cette vue correspond à la représentation du contenu descriptif d'un graphique. Dans cette vue, sont représentés les propriétés du graphique et des objets graphiques.

À chaque objet structurel est associé un objet symbolique défini dans cette vue. Et à chaque objet symbolique correspondent des attributs décrivant l'objet en question. Des relations symboliques peuvent relier deux objets symboliques.

Il est important de noter que différents objets structurels peuvent avoir un seul objet symbolique qui leur correspond. Il s'agit du cas d'objets structurels équivalents auxquels nous associons un même objet symbolique afin d'éviter la redondance.

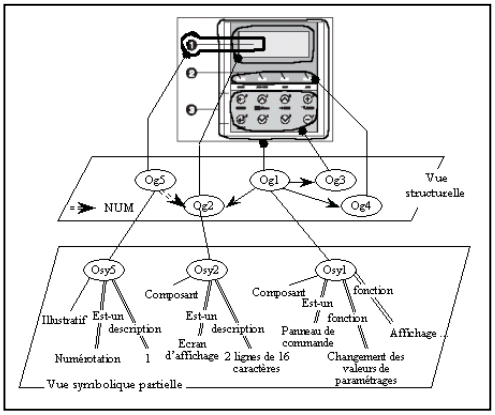


Figure 4 : *Exemple de vue descriptive d'un graphique*

4.2.5. *La vue opératoire*

La perception du contenu des graphiques par un utilisateur spécialisé est fonction non seulement de la description des objets qu'elle contient, mais aussi de la description de son aspect opératoire. Ainsi, un graphique contenant un objet correspondant à un « magasin », est incomplètement décrit si on ignore dans sa représentation l'action qu'on applique sur le « magasin », soit le « chargement de papier ».

La vue opératoire correspond donc aux représentation et description des actions à appliquer sur les objets graphiques. Elle est définie par des objets opératoires associés aux objets structurels, mis à part les objets illustratifs autres que la flèche et la main, ainsi que des relations entre objets opératoires.

À un objet structurel peuvent être associés plusieurs objets opératoires. Ceci est vrai lorsque plusieurs actions doivent être exécutées sur un même composant ou une action parmi plusieurs doit lui être appliquée selon la condition.

Par contre, dans le cas d'un objet structurel et illustratif (représentant une flèche ou une main), cet objet et celui qu'il vise (objet représentant le composant sur lequel s'applique l'action) sont représentés par un seul objet opératoire afin d'éviter la redondance.

Chaque objet opératoire est complété par les valeurs d'une ou plusieurs des propriétés définies ci-dessous :

– l'action générale décrite par le graphique (dans l'exemple : « le chargement de papier »),

– cas spécifique de cette action générale (dans l'exemple : « l'utilisation du magasin d'alimentation manuelle »),

– étape concernant l'action générale (dans l'exemple : « 2 »),

– actions à appliquer sur les objets graphiques que nous appelons les actions intermédiaires (dans l'exemple : « faire glisser »),

– condition d'application des actions intermédiaires (dans l'exemple : « utilisation du papier long format »),

– ordre d'application des actions intermédiaires (dans l'exemple : « 1 », « 3 »).

Les relations liant deux objets opératoires sont des relations d'ordre (Exemple : (avant, O_op2, O_op4)).

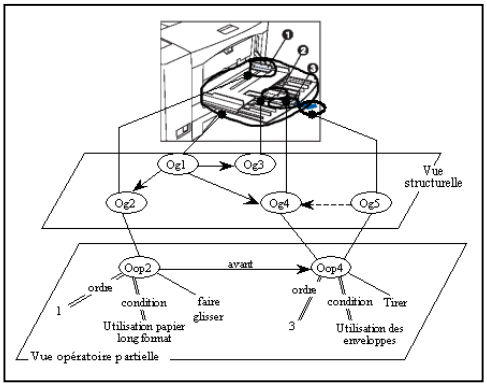


Figure 5 : *Exemple de vue opératoire d'un graphique*

4.2.6. *La vue mémoire visuelle*

Dans cette vue, sont représentés les objets et les relations issus de la vue spatiale ayant un impact sur la mémoire visuelle de l'utilisateur.

Une pondération correspondant à l'importance de l'impact sur la mémoire visuelle de l'utilisateur est associée à chaque objet et relation entre deux objets de cette vue.

On distingue trois types de poids :

– Un poids affecté à l'objet selon l'impact de l'existence de cet objet dans le graphique sur la mémoire visuelle de l'utilisateur. Autrement dit, ce poids représente l'importance pour l'utilisateur de l'existence de l'objet en question dans le graphique.

– Un poids relatif à l'importance de la forme de l'objet graphique pour la mémoire visuelle de l'utilisateur. Ce poids représente donc l'importance pour l'utilisateur de la forme de l'objet.

– Un poids affecté à la relation liant deux objets correspondant à l'importance pour l'utilisateur de la position relative entre les deux objets.

Afin de mieux comprendre ces différents poids, nous prenons l'exemple de la figure 6:

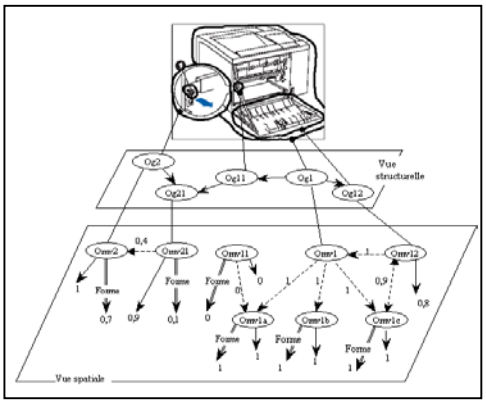


Figure 6 : *Exemple de vue mémoire visuelle d'un graphique*

Dans l'exemple,

– à l'existence de l'objet Omv2 est associé le poids 1 : (Omv2, 1). L'utilisateur se rappellera à 100% du zoom, représenté dans cette vue par l'objet Omv2, présent dans le graphique en question.

– à la forme de l'objet Omv12 est associé le poids 0,8 : (Forme, Omv12, 0.8). L'utilisateur se rappellera à 80% de la forme de l'objet, du graphique en question, représenté dans cette vue par Omv12.

– à la relation entre les objets Omv2 et Omv21 est associé un poids 0,4 : (Omv21, Omv2, 0.4). L'utilisateur se rappellera à 40% que dans le graphique en question, l'objet représenté dans cette vue par Omv21, est contenu dans le contour de l'objet zoom, représenté par Omv2.

5. Enrichissement de l'indexation de la donnée non symbolique par les données symboliques avoisinantes

Dans ce paragraphe, nous nous intéressons à l'extraction des fragments contenus dans les blocs textuels entourant un graphique qui vont servir à son indexation.

En général, afin d'enrichir une donnée non symbolique par les données symboliques avoisinantes, deux problèmes apparaissent :

– Où et comment localiser les données symboliques qui peuvent enrichir la donnée qui nous intéresse, autrement dit celles qui contiennent des informations la concernant.

– Comment extraire les fragments permettant de d'enrichir la description de la donnée non symbolique à partir des blocs de données précédemment localisés.

Dans notre application sur les documents techniques, il s'agit de localiser le texte qui est en relation avec le graphique qui nous intéresse, afin d'en extraire par la suite les termes qui vont servir à « remplir » le modèle que nous avons proposé.

5.1. *Localisation des données symboliques adéquates (les commentaires) :*

La structure d'un document facilite l'identification de ses différentes entités. Ainsi, il est plus aisé de relier les entités symboliques aux entités non symboliques correspondantes.

Dans les documents structurés, il existe une propagation de l'information entre les blocs reliés par une relation de composition. Ainsi, dans les documents techniques, le graphique est relié à la section, au sous-chapitre et au chapitre qui le contiennent. Cette relation est mise en évidence par l'existence d'un lien entre le graphique et les titres respectifs de la section, le sous chapitre et le chapitre.

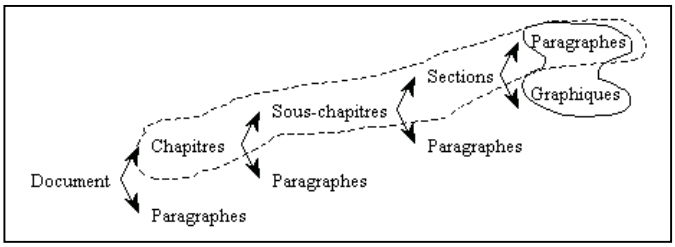


Figure 7 : *Blocs de textes correspondant à un graphique*

En partant de la donnée non symbolique qui nous intéresse et en remontant vers la racine de la hiérarchie du document structuré, les titres des blocs rencontrés semblent être des données permettant d'enrichir notre donnée non symbolique. Il nous reste alors à trouver les blocs de données qui lui correspondent et qui la décrivent avec plus de précision. Autrement dit, dans le cas des documents techniques, il nous reste à trouver le paragraphe commentant le graphique.

Il existe deux types de liens qui peuvent relier un premier bloc et celui qu'il commente:

– les liens explicites: c'est le cas lorsque le texte référence directement le graphique (voir figure X, la figure x illustre…),

– les liens implicites: il s'agit du cas où l'auteur n'utilise pas d'expression linguistique spécifique pour référencer le graphique.

Dans le cas d'une référence explicite, le fait qu'un graphique soit éloigné de son commentaire n'est pas gênant puisque le lecteur peut pointer le graphique à l'aide de la référence. Par contre, dans le cas d'une référence implicite si le graphique est trop éloigné de son commentaire alors l'information se perd au fur et à mesure de la lecture empêchant ainsi le lecteur d'effectuer la relation. Il s'agit, dans ce deuxième cas, d'étudier l'espace géographique dans lequel se situe le graphique afin de retrouver le bloc qui lui correspond.

Nous avons supposé dans un premier temps que un lecteur ne pouvant pas retrouver des liens implicites trop éloignés, l'auteur en a forcément tenu compte lors de la rédaction (ou mise en page). Le texte commentant le graphique est, dans ce cas, celui qui lui est le plus proche. Reste alors à extraire depuis ce bloc, les fragments ayant trait au graphique (selon le modèle que nous avons proposé).

5.2. *Extraction des index dans le commentaire :*

Le critère de qualité majeur d'un document technique est basé sur son efficacité et sa facilité d'utilisation. Pour répondre à cette contrainte, un certain nombre de règles est respecté par les rédacteurs techniques sur le plan de la forme textuelle. Ils doivent se montrer particulièrement vigilants sur la pertinence de leurs phrases, sur l'homogénéité du texte et les styles de phrases utilisés. Les auteurs suivent ainsi des règles bien précises lors de la rédaction d'un texte technique. Cela se traduit par l'existence d'une certaine régularité au niveau du document. Cette régularité peut être exploitée afin de retrouver des règles et des patrons linguistiques exploitables pour le repérage des index des graphiques.

Notre but, à ce stade, étant de démontrer la possibilité d'exploiter la régularité dans la rédaction des documents techniques dans le but d'en extraire les expressions linguistiques correspondantes aux termes d'indexation des graphiques, nous nous sommes limités à l'étude d'un unique manuel d'utilisation. Notre étude à permis d'aboutir à des critères spécifiques à notre corpus. Même si elle ne nous permet pas

de généraliser les règles d'extraction d'index, elle est une preuve de la faisabilité de cette extraction, et dès lors une étude similaire et plus poussée pourra être faite.

Les informations à extraire du texte afin d'enrichir la description du graphique sont celles représentées dans les deux vues symbolique et opératoire, à savoir :

– quelles sont les composantes contenues dans le graphique, leurs propriétés et leurs fonctions ?

– quelles actions doivent leur être appliquées, à quelles conditions, et quel est leur ordre d'exécution ?

– quel est l'action décrite par le graphique, quelle est son étape et dans quel cas est-elle nécessaire ?

Afin d'extraire les termes correspondant à ces informations, une étude du commentaire à été menée pour tenter de mettre en place des règles permettant leur repérage.

Il en a résulté que, d'un point de vue « structure », le commentaire peut être un titre, un ou plusieurs items d'une liste ou un paragraphe classique. Et d'un point de vue « sens », le commentaire peut soit décrire un composant, soit une action à lui appliquer. Ce deuxième point de vue est détaillé dans ce qui suit :

– Lorsqu'il s'agit d'une description d'un composant, on parlera du patron COMDESC (le composant COM et sa description) :

COMDESC :

*(ADJ)** **COM** *(PREP GN)*(ADJ)**[1]

COMPOSANT : COM **DESCRIPTION**: PREP GN/ADJ

Exemple :

« Ecran d'affichage **de 16 caractères** »

COMPOSANT :écran d'affichage **DESCRIPTION**: de 16 caractères

Où **COM** : *(ADJ)*((PREP)* **SUBC** (SUBC)*)* (ADVB)**[2]

Exemples : « carte réseau » : ***SUBC*** *SUBC*, « panneau avant » : ***SUBC*** *ADVB*

– dans le cas de la description d'une action, on parlera du patron ACT (le composant et l'action à y appliquer) :

ACT : Pouvant être ACT1 (action sans condition) ou ACT2 (action avec condition). Nous donnerons seulement un exemple de ACT1 :

---

[1] ADJ : Adjectif, PREP : Préposition, GN : Groupe nominal, * : facultatif

[2] SUBC : Substantif commun, ADVB : Adverbe

ACT1 :

Exemple :

VR *ART* **COMDESC** *(PARE NB PARE)* (**GP**)**[3]

Soulevez le guide d'extrémité (1)

COMPOSANT/ **DESCRIPTION** : DESCCOM **ACTION_INTER**: infinitif de VR

COMPOSANT :Guide d'extrémité **ACTION_INTER**: Soulever

Où **VR** : **VRB** (VRB)* (PREP)*[4]

Exemples : « faites glisser » : ***VRB*** *VRB*, « appuyez sur » : ***VRB*** *PREP*

Selon que nous trouvons l'un de ces deux patrons (COMDESC ou ACT)dans un titre, un item ou un paragraphe, l'extraction des index diffère.

Dans le tableau suivant, nous présentons les règles d'extraction des index à partir des patrons décelés :

DESCRIPTION	PATRON		TERMES EXTRAITS
1er titre en remontant	TITRE1	GN	Graph.CONDITIONGENERALE=GN
2nd titre en remontant	TITRE2	GN	Graph.ACTIONGENERAL=GN
Titre non dans sommaire	SOUSTITRE	GN	Graph.DESCRIPTION=GN
Item dans liste	ITEMD	NB COMDESC	Termes extraits de COMDESC
Item dans liste	ITEMA	NB ACT	ORDRE=NB et termes extraits de ACT
Paragraphe classique	PARA	NB {ACT}	Graph.ETAPE=NB et termes extraits des ACT

5.3. *Quelques chiffres*

Nous avons étiqueté le texte des commentaires manuellement afin de ne pas influencer les résultats par les limites d'un système particulier.

Sur 62 commentaires, en plus des titres, nous avons appliqué les règles définies précédemment et nous avons obtenu les résultats présentés dans le tableau suivant :

Patrons	Fréquence dans les commentaires	Nombre de traductions exactes[5]
ITEMD	42	32 (76%)
ITEMA	96	76 (79%)
TITRE1	53	53 (100%)
TITRE2	53	53 (100%)
SOUSTITRE	7	7 (100%)
PARA	39	32 (82%)

[3] ART : Article, PARE : Parenthèse, NB : Nombre, GP : Groupe nominal ou verbal
[4] VRB : Verbe
[5] Traduction exacte: les termes extraits du patron correspondent aux index adéquats.

Les résultats obtenus sont prometteurs, cependant une étude plus poussée devra être menée.

6. Exemples de requêtes/réponses

Nous avons traduit notre modèle dans le formalisme des graphes conceptuels (Sowa J.F. 1984) et utilisé l'opérateur de projection comme fonction de correspondance.

Nous avons indexé manuellement 20 graphes et proposé trois requêtes :

– Une requête Req1 relative à l'aspect descriptif : « Donnez-moi le graphique décrivant l'écran d'affichage de l'imprimante N17.»

– Une requête Req2 relative à la mémoire visuelle : « Donnez-moi le graphique qui, je m'en souviens, contient un parallélépipède, une flèche à sa gauche, deux mains et deux énumérations.»

– Une requête Req3 relative à l'aspect opératoire et mémoire visuelle « Donnez-moi le graphique qui, je m'en souviens, contient un parallélépipède, une flèche à sa gauche, deux mains et deux énumérations et qui décrit le chargement du papier dans le cas de l'utilisation du magasin1.»

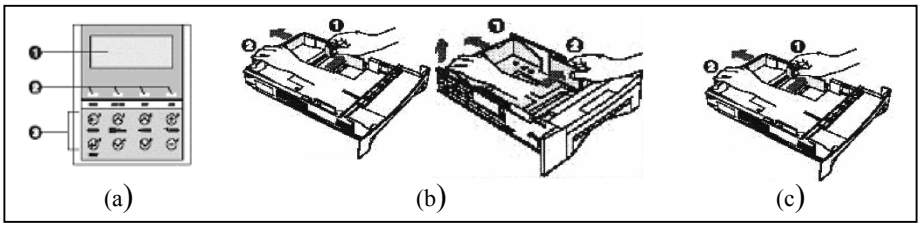


Figure 8 : *Graphiques répondant aux requêtes Req1(a), Req2(b) et Req3(c)*

La requête Req2 permet de retrouver deux graphiques répondant à la requête. En rajoutant à cette requête l'aspect opératoire (Req3), la réponse se précise : nous obtenons ainsi un seul graphique, celui qui parmi les deux nous intéresse réellement.

7. Conclusion et perspectives

La définition d'un bon modèle d'indexation est un problème clé en recherche d'information. Lorsque les données à indexer sont peu symboliques, la tâche est encore plus délicate, car il faut pouvoir retrouver ailleurs les éléments décrivant le mieux ces données. Nous avons proposé un modèle pour décrire les aspects les plus pertinents des graphiques des documents techniques en tenant compte des

particularités propres à un tel média, de leur finalité d'utilisation dans les documents techniques, et du contexte professionnel dans lequel ils apparaissent.

Afin d'extraire les termes d'indexation du graphique à partir du texte qui l'entoure dans le document structuré qu'est le document technique, nous avons mis en place des règles combinant des patrons syntaxiques et les positions structurelles des syntagmes, pour extraire les termes enrichissant la sémantique de ce graphique.

Le modèle d'indexation étant défini, il reste à effectuer des expérimentations plus poussées avec des utilisateurs professionnels et de définir un modèle de correspondance plus adapté. Ce travail sera aussi repris sur un grand corpus de manuels d'utilisation d'imprimantes Xerox.

7. Bibliographie

Badjo-Monne B., Bertier M., Vers un modèle d'indexation des documents techniques, *Document numérique*, Volume4- n°1-2/2000, Hermes.

Joly M., *Introduction à l'analyse de l'image*, NATHAN Université, 1993.

Lorenz O. and Monagan G., « Automatic indexing for storage and retrieval of line drawings», *IS&T/SPIE's Symposium on Electronic Imaging Science & Technology,* San Jose Convention Center, San Jose, California, USA. Feb 1995.

Lorenz O. and Monagan G., « Retrieval of line drawings», *Proceedings of The Third Annual Symposium on Document Analysis and Information Retrieval*, Alexis Park Hotel, Las Vegas, USA, April 1994.

Malandain N., « Automatic geographical hypertext "multi-scaled links" generation», *In Proceedings of Fifth International Workshop on Principles of Digital Document Processing*, Munich, Germany, September 2000.

Mechkour M., « EMIR², un modèle étendu de représentation et de correspondance d'images pour la recherche d'informations. Application à un corpus d'images historiques», Thèse Informatique de l'université Joseph Fourier, Grenoble I, novembre 1995.

Michel D., *Langage et cognition spatiale*, Sciences Cognitives, 1997.

Ouerfelli T., Lallich-Boidin G., « Pratiques d'indexation dans les Bases Textuelles Structurées : Application aux Textes Techniques sous Format HTML», *CAIS 2000: Dimensions of a global information science, Proceedings of the 28th Annual Conference.*

Rowe N. C., Precise and efficient access to captioned picture libraries: The MARIE project. Technical report, Department of Computer Science, Naval Postgraduate School, 1996.

Salton G., McGill M. J., *Introduction to modern information retrieval*, McGraw-Hill, NewYork, 1983

Sowa J. F., *Conceptual Structures*, Addison-Wesley, Reading, MA, 1984.

Wright P., « Presenting technical information: a survey of research findings», *Instructional Science*, 6, 93134, 1977.