Système d'aide à la décision : outil d'analyse multidimensionnelle utilisant la technologie OLAP

Bessai F.Z., Boutouhami Sara, Zeddigha Ismahane
CEntre de Recherche sur l'Information Scientifique et Technique (CERIST)
Division Base de Données et Système multimédia
Rue des Frères Aissiou, Ben Aknoun BP 143 Alger
E-Mail : zbessai@wissal.dz

1- Introduction

L'information est la nouvelle ressource des entreprises du XXI[e] siècle. Jusqu' à une pério de récente l'information était utilisée à des fins de contrôles ou comptables. Mais avec le temps, les entreprises produisent et manipulent de très importants volumes de données. Ces données sont stockées dans les systèmes opérationnels de l'entreprise au sein de bases de données et elles sont gérées selon des processus transactionnels en ligne (OLTP : On Line Transactional Processing). L'exploitation de ces données dans un but d'analyse et de support à la prise de décision s'avère difficile et fastidieuse. Ces systèmes opérationnels paraissent peu adaptés pour servir de support à la prise de décision. Face à cette inadéquation, les entreprises ont recours à des systèmes d'aide à la décision.

L'informatique décisionnelle est apparue au début des années 90 pour permettre à un utilisateur d'accéder de manière simple et ergonomique à un serveur de données et de valoriser l'information récupérée. Les systèmes d'aide à la décision complètent les systèmes opérationnels et offrent aux utilisateurs la possibilité de répondre aux besoins d'aide à la prise de décision.

L'efficacité de la prise de décision repose sur la mise à disposition d'informations pertinentes et d'outils adaptés. Donc, Disposer de l'information utile, en avoir plus que ses concurrents, l'avoir préparée à l'avance, la rendre disponible au moment où l'utilisateur en a besoin dans un format compréhensible, sont des objectifs très importants.

L'objectif du présent travail est la conception et la réalisation d'un Système d'Analyse Multidimensionnelle (SAM) utilisant la technologie OLAP (On Line Analytical Processing) qui permet une analyse multidimensionnelle très efficace pour répondre à des requêtes de type «What if» [Pen 01]. Ce type d'analyse organise les données sous forme de cubes multidimensionnels, dont les cellules contiennent des mesures pré- calculées, et utilise des opérateurs spécifiques aux cubes pour répondre de manière pertinente aux requêtes des utilisateurs.

2- Architecture générale d'un système d'aide à la décision

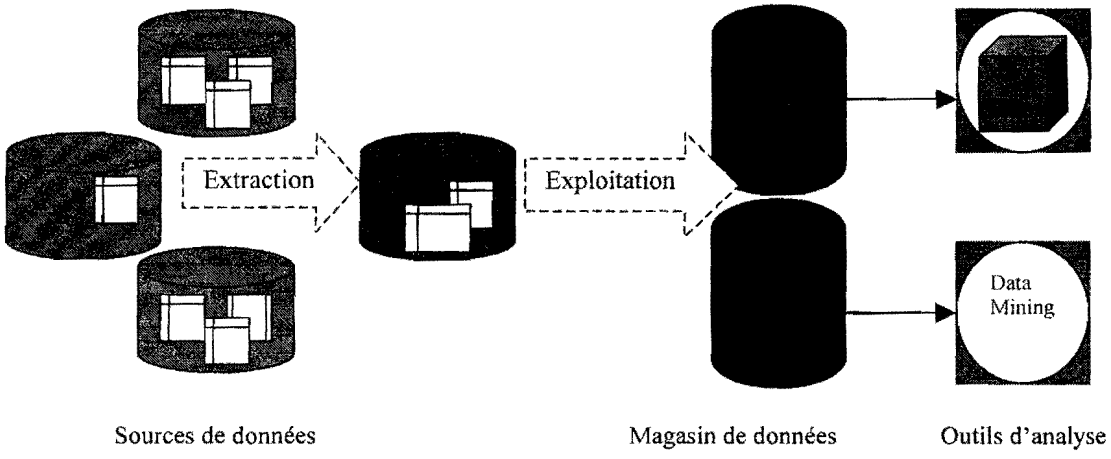

Figure 1 *: Architecture des systèmes décisionnels*

La plupart des systèmes d'aide à la décision reposent sur un espace de stockage centralisé, appelé entrepôt de données (Data Warehouse) dont le rôle est d'intégrer et de stocker l'information utile aux décideurs et de conserver l'historique des données pour supporter les analyses effectuées lors des prises de décision [Tes 00].

Un Data Warehouse ou entrepôt de données est une collection de données consolidant des informations en provenance des différents systèmes de productions ou opérationnels, ces données sont orientées sujet, intégrées, non volatiles, historisées et organisées pour le support d'un processus d'aide à la décision. L'entrepôt de données est le lieu centralisé de toute information pertinente [Bou 98].

La difficulté liée à la gestion d'un Data Warehouse principal et unique a fait naître en complément et souvent des bases de données d'un coût raisonnable ciblées sur quelques sujets limités pour valider rapidement le concept d'information décisionnelle [Gou 97]. Ces bases de données sont appelées Data Marts ou magasins de données [Gar 00]. C'est une petite structure très ciblée et pilotée par les besoins utilisateurs. Le magasin de données a la même vocation que l'entrepôt de données, mais vise une problématique précise avec un nombre d'utilisateurs plus restreint. En effet, un magasin de données est spécialisé pour un type d'activité, un type d'analyse et un groupe d'utilisateurs. En général, c'est une petite base de données organisée en relationnel ou en multidimensionnel. Dans le cas pratique, un magasin de données consiste à extraire une partie de l'information décisionnelle contenue dans l'entrepôt de données; les données extraites doivent correspondre à une structuration adaptée à l'outil d'analyse utilisé.

Les Data Marts sont organisés en modèle relationnel ou multidimensionnel.

3- Le modèle multidimensionnel

Le modèle multidimensionnel comporte les éléments suivants: les tables dimensionnelles et une ou plusieurs tables de faits [Don 99].

La table de faits est la table principale de tout modèle multidimensionnel, destinée à héberger des données permettant de mesurer l'activité. Un fait modélise un sujet de l'analyse. Il est formé de mesures correspondant aux informations de l'activité analysée [Tes 00]. Ces mesures sont de nature numérique et sont généralement valorisées de manière continue. La table de faits contient une clé multiple composée d'un ensemble de clés étrangères. Chaque clé étrangère permet de relier la table de faits à une table nommée table dimensionnelle ou dimension.

Il existe trois types de faits :

Faits additifs: Les faits additifs sont des faits additionnables suivant toutes les dimensions.

Exemple: Chiffre d'affaire et le coût, quantité.

Faits semi-additifs : Les faits semi-additifs se sont des faits additionnables seulement suivant certaines dimensions.

Exemple : Soient deux faits (même magasin et même fait), nous avons vendu des serviettes pour 20 clients et des mouchoirs pour 40 clients. La somme du nombre de clients sur la dimension produit n'a pas de signification, car un client peut avoir acheté des serviettes et des mouchoirs.

Faits non additifs : Les faits non additifs sont non additionnables quelque soit la dimension.

Exemple: Température.

Une table dimensionnelle (ou dimension) modélise une perspective de l'analyse. Une dimension se compose de paramètres correspondant aux informations faisant varier les mesures de l'activité. Les dimensions servent à enregistrer les valeurs pour lesquelles sont analysées les mesures de l'activité. Une dimension est généralement formée de paramètres (ou attributs) textuels et discrets sur lesquels portent les clauses de conditions et de groupement au sein de requêtes.

Chaque dimension est définie par sa clé primaire qui assure l'intégrité référentielle avec la(es) table(s) de faits à laquelle(s) elle est reliée.

Le concepteur intervient dans le choix des dimensions de la base de données, il doit collecter les attributs fortement corrélés entre eux dans une dimension. Ces tables dimensionnelles sont équivalentes, en effet, toutes peuvent être vues comme des points d'entrée, symétriquement identiques, dans la table de faits.

Chaque modèle multidimensionnel doit être constitué de la dimension temps, pour garder trace de l'historique des données.

3.1- Présentation des données multidimensionnelles

La modélisation multidimensionnelle permet la représentation explicite des hiérarchies et même la possibilité de manipuler à la fois le contenu et la structure des données. La modélisation multidimensionnelle offre des opérateurs spécifiques associés à ce modèle de données (Drill-down, slice, dice...).

Les étapes nécessaires pour la modélisation d'une base de données multidimensionnelle sont les suivantes [Gys 96] [Tes 00] :

Détermination des faits représentant les sujets analysés,

Les étapes nécessaires pour la modélisation d'une base de données multidimensionnelle sont les suivantes [Gys 96] [Tes 00] :
Détermination des faits représentant les sujets analysés,
détermination des dimensions représentant les perspectives de l'analyse,
définition des granularités des données de l'analyse,
organisation des paramètres des dimensions selon des dépendances de hiérarchie pour supporter les analyses à différents niveaux de détail.

Le modèle multidimensionnel peut être parfaitement mis en œuvre sur une plate forme relationnelle. Dans ce cas, les données peuvent être présentées sous trois schémas possibles:

a- Le schéma en étoile : Le schéma en étoile comporte une table de fait principale contenant une clé multiple composée d'un ensemble de clés étrangères. Chaque clé étrangère permet de relier la table de fait à la table dimensionnelle.
La dimension dans ce cas, peut être dénormalisée c'est à dire, les attributs qui ont un lien fonctionnel entre eux (hiérarchie) peuvent cohabiter la même table.

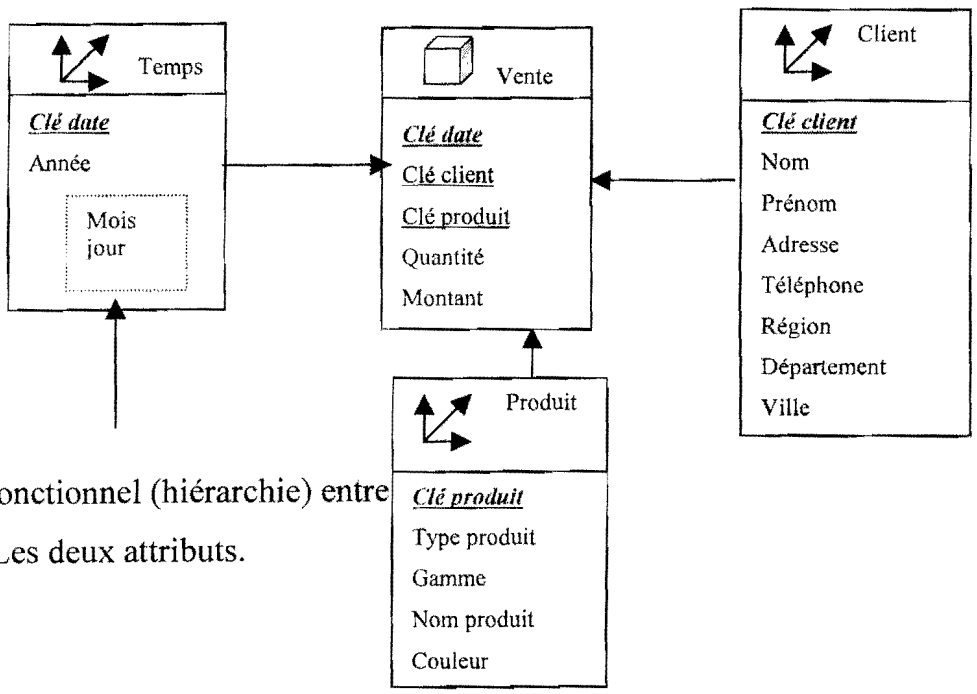

Figure 2: *Modélisation en étoile*

dimensions du modèle en étoile en sous hiérarchie. Une hiérarchie organise les attributs d'une dimension selon une relation « est_plus_fin » conformément à leur niveau de détail.

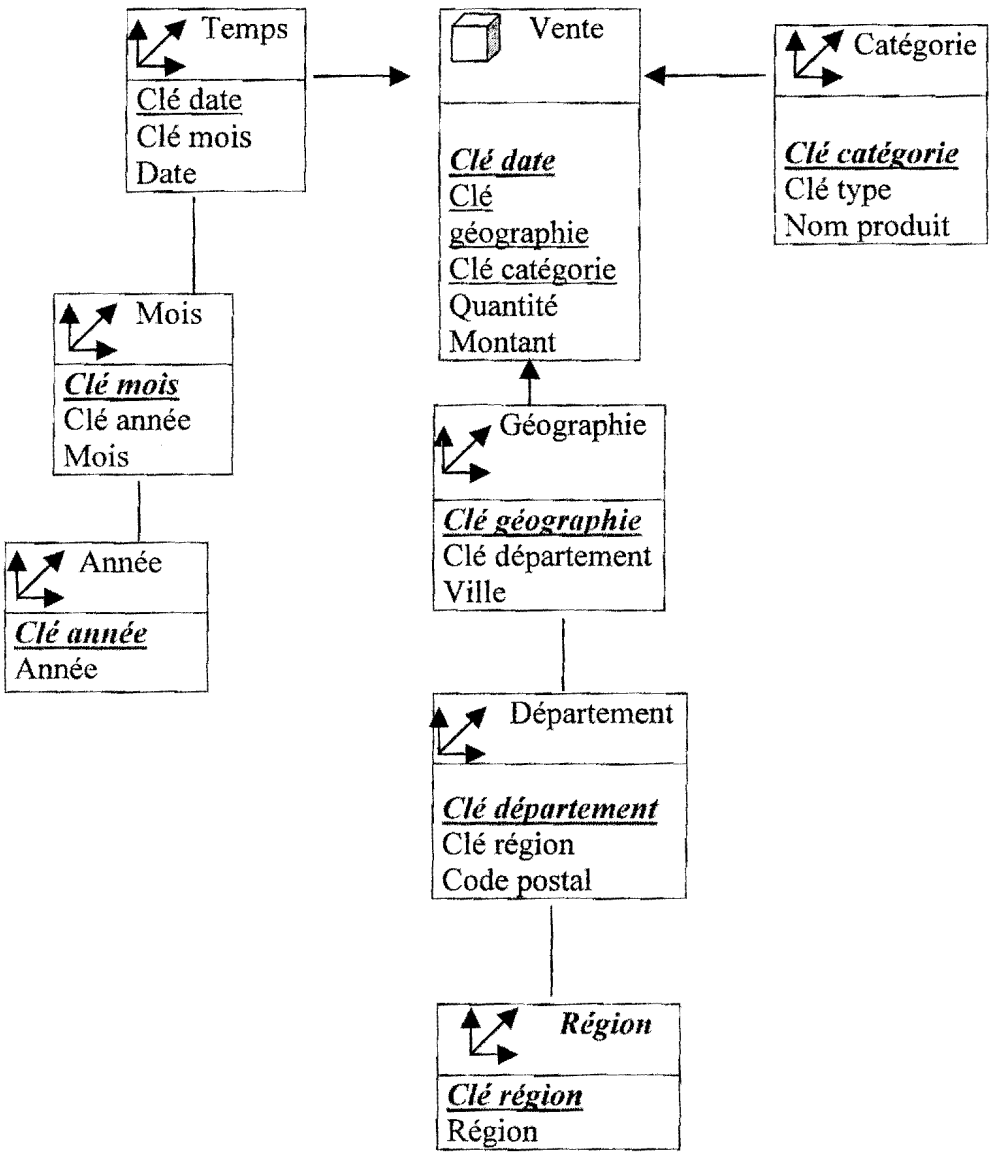

Figure 3: *Exemple d'une modélisation en flocon*

c- Schéma en constellation : Le schéma en constellation est le fusionnement de plusieurs schémas en étoile, nous aurons plusieurs tables de faits et des dimensions communes entre plusieurs schémas en étoile.

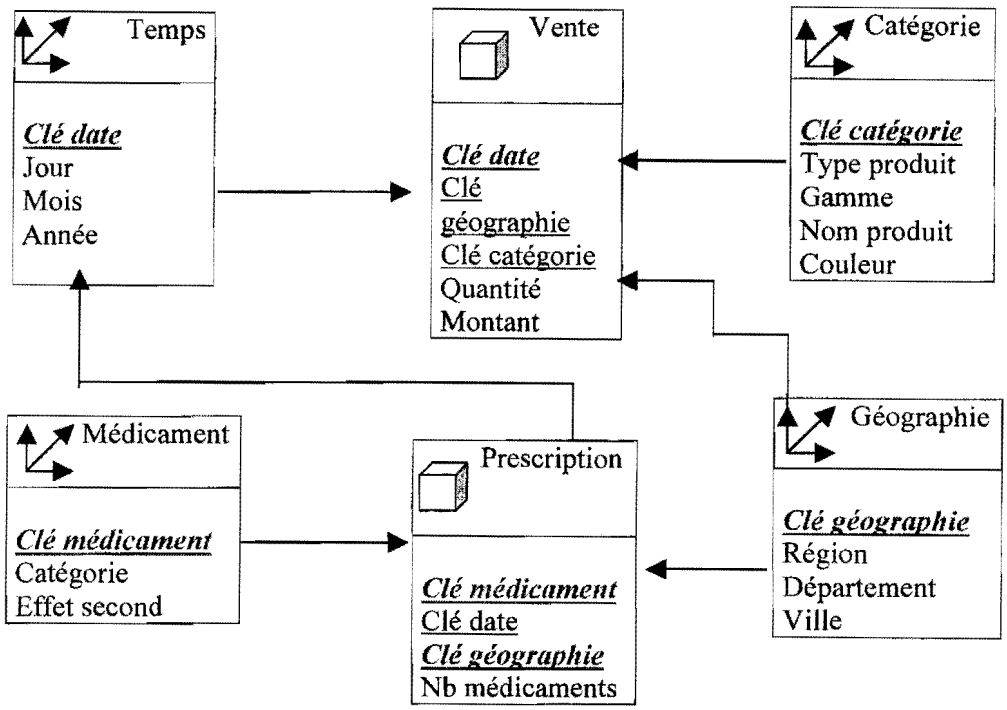

Figure 4: *Exemple d'une modélisation en constellation*

Les applications d'aide à la décision utilisent la technologie OLAP (On Line Analytical Processing) qui aide efficacement les décideurs à prendre les meilleures décisions, en leurs fournissant les informations nécessaires.

Les Data Marts sont organisés en modèle multidimensionnel pour supporter efficacement des processus de type OLAP.

4- Technologie OLAP

OLAP est une technologie qui permet aux analystes et aux administrateurs d'accéder de façon rapide, consistante et interactive à un grand volume de données.

La technologie OLAP correspond aux traitements visant à interroger, à visualiser et à synthétiser les données, ces traitements concernent un nombre d'enregistrements importants. La technologie OLAP organise les données sous

une forme cubique ou d'hypercube (cube de plus de trois dimensions) [Pen 01] [Har 96].

La figure suivante représente les données dans un espace à trois dimensions: la dimension catégorie, la dimension temps et la dimension région. Chaque intersection de ces dimensions représente une cellule comportant le montant des ventes (fait).

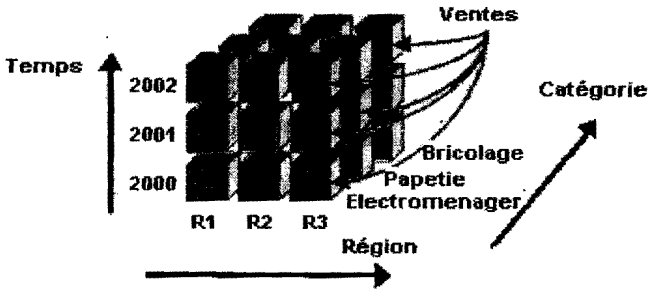

Figure 5: *Représentation multidimensionnelle*

Les points à l'intérieur d'un cube servent à enregistrer les mesures d'activités pour chaque combinaison donnée entre les dimensions. Ces points sont appelés cellules, une cellule peut être vide.

La technologie OLAP présente trois avantages majeurs [Pen 01] [Sap 99]:
L'accès aux données contenues dans les cellules d'un hyper-cube est plus direct que dans toute autre structure, la notion de jointure n'existe pas, puisque toutes les données sont dans un seul tableau. Les temps de réponses sont donc très courts.

L'administration d'un schéma matriciel est beaucoup plus simple que celle d'une base de données relationnelle. Les traitements des agrégats pré-calculés sont automatiques et transparents.

Permet aux analystes et aux administrateurs d'accéder de façon rapide, consistante et interactive à un grand volume de données.

5- Conception et réalisation du Système d'Analyse Multidimensionnelle (SAM)

Les applications OLAP fournissent des informations nécessaires pour la prise de décision. L'indicateur de succès d'une application OLAP est sa capacité de

fournir les informations nécessaires au moment souhaité. Pour aboutir à ce niveau de succès, il faut se baser sur plusieurs niveaux de détails de données, ce qui est assuré par la modélisation multidimensionnelle.

Pour la conception du prototype SAM, nous avons opté pour une architecture qui repose sur un Système de Gestion de Bases de Données Relationnel (SGBDR) et les cubes ne seront créés qu'au moment souhaité, ce qui signifie que ces cubes ne prennent pas d'espace mémoire qu'au moment de l'exécution. Cette architecture s'implémente sur une base de données multidimensionnelle organisée en étoile, en flocon ou en constellation.

Le schéma étoile se caractérise par [Sap 99] :

Toutes les dimensions ont une liaison directe avec la table de faits.

Nombre de jointures limité.

Facilité de navigation.

Vu les caractéristiques du schéma étoile, c'est ce dernier qui a été retenu pour la conception et l'implémentation du prototype SAM.

L'architecture générale du prototype SAM est comme suit :

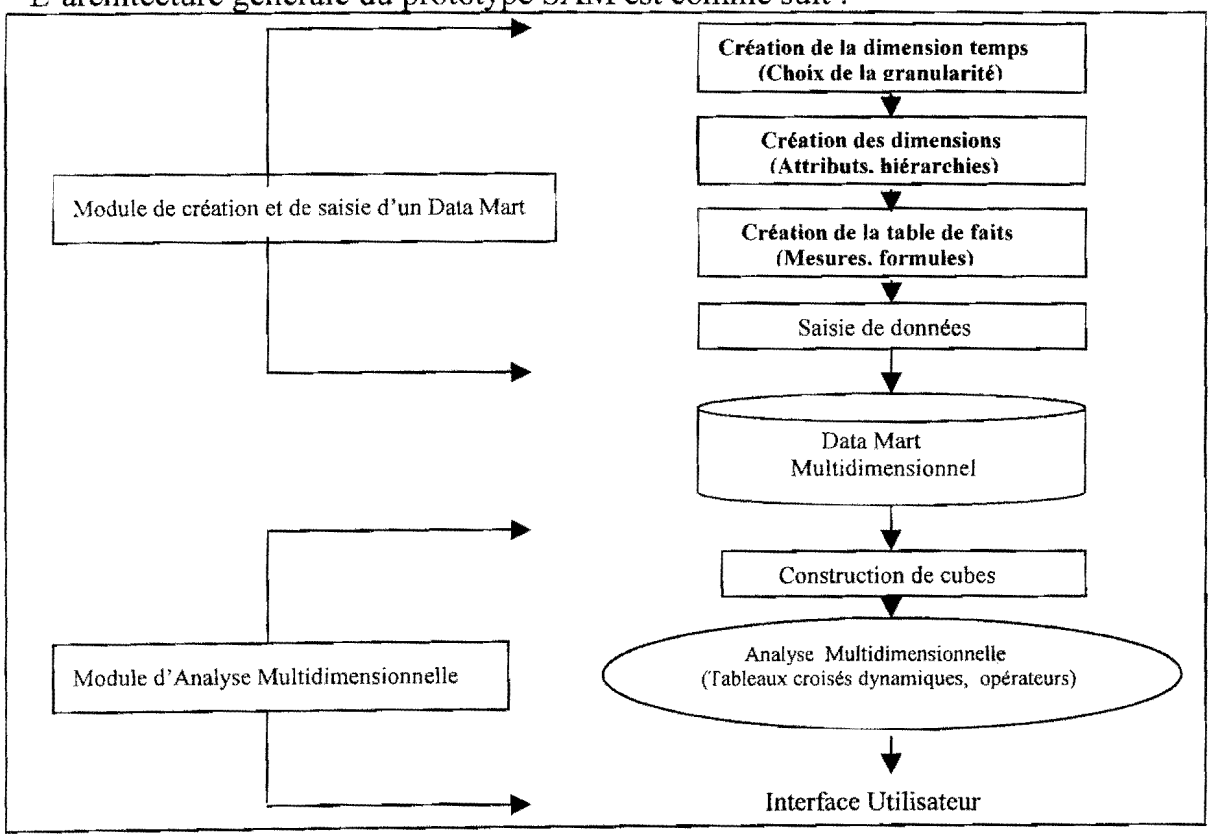

Figure 6: *Architecture générale du prototype SAM*

Le prototype SAM est composé principalement de deux modules : module de création et de saisie d'un Data Mart et le module d'analyse multidimensionnelle.

a- Module de création et de saisie d'un Data Mart

Ce module permet la création d'un Data Mart selon le modèle étoile tout en commençant par la création et la saisie de la dimension temps qui sera définie par sa granularité (année, trimestre, mois, jour) et par la durée sur laquelle portera l'analyse multidimensionnelle. Ensuite, la structure de chaque dimension doit être créée en définissant ses attributs et en donnant éventuellement l'hiérarchie choisie, pour la dimension. La dernière étape dans la création d'un data Mart est la création de la table des faits avec les clés de substitution des différentes dimensions déjà créées.

La saisie permettra l'approvisionnement du Data Mart ainsi créé.

b- L'analyse multidimensionnelle

Une fois le Data Mart multidimensionnel créé, l'analyse multidimensionnelle correspondra à la création d'un ou de plusieurs cubes en utilisant les attributs des dimensions, du Data Mart concerné, comme dimensions du cube et les cellules seront remplies par un fait donné.

Comme la représentation sous forme de tableau est la vision la plus simple et la plus intuitive [Tes 00], le cube ainsi créé sera représenté dans un tableau croisé dynamique tels que les éléments de ce tableau représentent le fait à analyser par rapport aux attributs -des dimensions- qui seront disposés en lignes et colonnes.

L'analyse multidimensionnelle sera effectuée sur le tableau croisé dynamique en utilisant des opérateurs de navigations, spécifiques à la manipulation des cubes, offerts par le prototype SAM.

c- Opérateurs spécifiques aux cubes

A fin d'effectuer une analyse multidimensionnelle à plusieurs niveaux de détail, le prototype SAM offre deux types d'opérateurs: des opérateurs sur les dimensions et des opérateurs sur les faits.

- Opérateurs sur les dimensions :

Drill-down: Cette opération permet le forage vers le bas dans une hiérarchie.

Drill-up: Cette opération permet le forage vers le haut dans une hiérarchie.

Slice: Cette opération permet d'analyser un fait par rapport à une valeur spécifique d'une dimension.

Dice: Cette opération réduit le nombre de dimensions en conservant l'information par un Drill-up.

Activer: Permet de visualiser une dimension dans le tableau croisé dynamique.

Permutation : Permet de permuter deux dimensions dans le tableau croisé dynamique.

En ligne: Permet de mettre une dimension en ligne, selon la position choisie par l'analyseur.

En colonne: Permet de mettre une dimension en colonne, selon la position choisie par l'analyseur.

Nest: permet à une hiérarchie donnée d'emboîter la hiérarchie supérieure.

- Opérateurs sur les faits :

Synthèse : permet d'afficher le fait lui-même.

SUM : permet d'afficher le fait en plus des sommes.

Count : permet d'afficher le nombre de valeurs du fait sélectionné.

AVG : permet d'afficher la moyenne du fait sélectionné.

Pour le développement du prototype SAM, nous avons opté pour le langage C++ BUILDER (Version 5, Anglaise, Entreprise). Ce langage dispose d'une bibliothèque de classes qui facilitent grandement la programmation Windows (Objet Windows Library), et utilise une programmation orientée objet. Aussi, et parmi les points essentiels du choix de ce langage, est qu'il dispose de nouveaux composants appropriés à l'analyse multidimensionnelle tel que le « cube décision ».

Le prototype SAM gère deux types d'utilisateurs:

- Administrateur : il a la possibilité de:
  - Créer et gérer les Data Marts,
  - Créer et gérer les cubes,
  - Créer et gérer les utilisateurs du système.

- Utilisateur : les seules fonctions possibles pour un utilisateur sont:
  - Consulter les données du Data Mart dont il a le droit d'accéder.
  - Effectuer l'analyse multidimensionnelle sur les cubes du Data Mart dont il a le droit d'accéder.

L'interaction entre l'utilisateur et le prototype SAM s'effectue via des interfaces. La fenêtre principale de l'administrateur est organisée sous forme de quatre menus, les principaux d'entre eux sont: **Data Mart**, **Gestion des utilisateurs** et **Cube** qui permettent la création, la gestion et la manipulation des Data Marts.

L'interface d'accès administrateur au prototype SAM est comme suit:

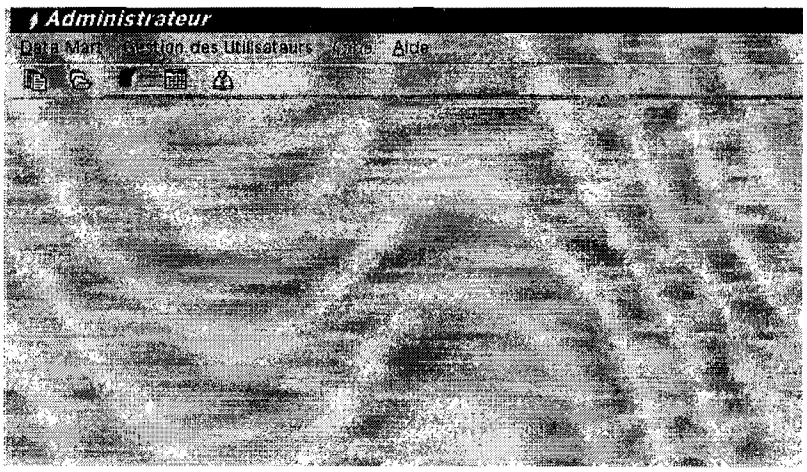

***Figure 7**: Fenêtre principale de l'administrateur*

La fenêtre contenant les fonctions possibles d'un utilisateur est la suivante:

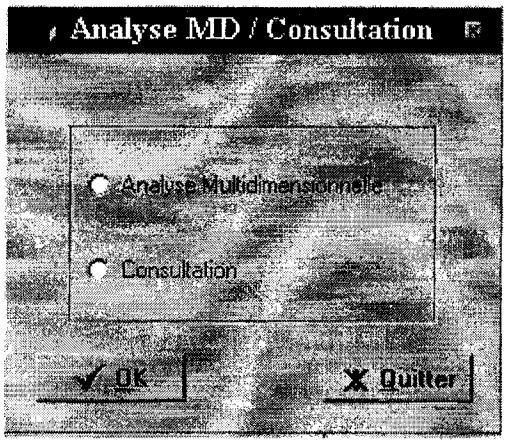

***Figure 8:** Fenêtre du choix d'utilisateur*

6- Conclusion

La prise de décision est devenue une tâche primordiale dans n'importe quelle entreprise, surtout dans le domaine du business. Les systèmes d'aide à la décision sont devenus la clé gagnante des entreprises: Une bonne décision implique une bonne continuité.

Les bases de données multidimensionnelles servent de base à la technologie OLAP.

Le Système d'Analyse Multidimensionnelle (SAM), ainsi conçu et réalisé, assure la création et la gestion de plusieurs bases de données multidimensionnelles, basées sur la technologie OLAP, permettant une analyse performante pour l'aide à la prise de décision.

Le prototype SAM gère deux types d'utilisateurs: l'administrateur ayant comme fonctionnalités la création et la gestion des Data Marts, des cubes et des utilisateurs, et l'utilisateur ayant comme fonctionnalités: l'analyse multidimensionnelle et la consultation des données du Data Mart choisi.

Le prototype SAM est ouvert sur plusieurs SGBD. Il utilise les représentations multidimensionnelles ce qui facilite l'analyse. Le prototype SAM permet la hiérarchie dans les dimensions ce qui rend possible l'analyse sur plusieurs niveaux de détail en exécutant les opérateurs Drill-UP ou Drill-Down.
En gérant l'accès des utilisateurs au système, ce dernier assure la sécurité et la confidentialité des données. Cependant, l'utilisation d'un mécanisme gérant les matrices creuses est nécessaire afin d'augmenter les performances du système.

Bibliographie

[Bou 98]: M. Bouzeghoub
« Data Warehouse architecture, fonctionnalités, outils, conception ».
Séminaire sur les Bases de Données Avancées, BDA'98, Tunis, 1998.

[Don 99]: D. Donsez
« Conception de bases décisionnelles ».
Université de Valenciennes, 1999.
Email :donsez@univ-valenciennes.fr

[Gar 00]: G. Gardarin
« Internet / Intranet et bases de données (Data Web, Data Media, Data Warehouse, Data Mining) ». Edition Eyrolles, 2000.

[Gou 97]: J. M. Gouarné
«Le projet décisionnel «Enjeux, modèles, architecture du Data Warehouse »
Edition Eyrolles, Novembre 1997.

[Gys 96]: M. Gyssens & V.S. Lakshmanan
« A foundation for multi-dimensional databases », 1996.
University of Limburg (Belgium) & University of Montreal (Canada).
Email: gyssens@charlie.luc.ac.be et laks@cs.concordia.ca

[Har 96]: V. Harinarayan, A. Rajaraman, J.D. Ullman
« Implimenting data cubes efficiently », 1996.
Department of computer science, Stanford University.
Emails: {venky, anand, ullman}@db.stanford.edu

[Pen 01]: N.Pendense
« OLAP Architectures ». 2001.
Site: www.OLAPReport.com

[Tes 00]: O. Teste
« Modélisation et manipulation d'entrepôts de données complexes et historiseés ». Thèse de doctorat. Laboratoire IRIT- Pôle SIG 2000.

[Sap 99]: C. Sapia, M. Blaschka, G. Höfling
« An Overview of Multidimensional data warehouse for OLAP »,
Février 1999. Email: {sapia, blaschka, dinter} @forwiss.tu-muenchen.de

Un Système de reformulation de requêtes pour la recherche d'Information

H. Aliane[*], *Z. Alimazighi*[**], *R. O. Boughacha*[*], *T. Djelliout*[*]
* *Centre de Recherche sur l'Information Scientifique et Technique, Alger, Algérie*
E-mail : haliane@mail.cerist.dz
** *Université des Sciences et de la Technologie Houari Boumedienne, Alger, Algérie.*
E-mail : alimazighi@wissal.dz

1. Introduction

Un système de recherche d'information SRI est un système qui gère une collection d'informations organisées sous forme d'une représentation intermédiaire reflétant aussi fidèlement que possible le contenu des documents grâce à un processus préalable d'indexation, manuelle ou automatique. La recherche d'information désigne alors le processus qui permet, à partir d'une expression des besoins d'information d'un utilisateur, de retrouver l'ensemble des documents contenant l'information recherchée *(Abbadeni et al., 1998)* et ce par la mise en œuvre d'un mécanisme d'appariement entre la requête de l'utilisateur et les documents ou plus exactement entre la représentation de la requête et la représentation des documents. La notion de document est prise ici au sens large et peut représenter une combinaison multimédia.

1.1. Notions de base dans un SRI

On distingue quatre notions de base dans un SRI (Abadenni et *al. ,1998*) :

- ***La notion de document :*** L'ensemble des documents sur lesquels portera la recherche est stocké dans une banque de données (sur le Web). Un document est le type d'objet de base géré par le système.
- ***La notion de besoin d'information d'un utilisateur*** : Ce besoin est exprimé par une requête spécifiée dans un formalisme propre au système. Le formalisme de spécification de la requête peut être en langage naturel.

- ***Notion de correspondance entre la requête et les documents*** : Une fois la requête spécifiée, le système tente de retrouver les documents qui correspondent à la requête en se basant sur une mesure de similarité.
- ***La notion de contexte de l'application :*** Le contexte de l'application représente l'univers dans lequel le système fonctionne. L'univers est nécessaire aux systèmes de recherche d'information pour une bonne compréhension des besoins des utilisateurs.

Un SRI doit être capable de retrouver les documents pertinents à partir d'une banque de données (Web) satisfaisant la requête posée par un utilisateur et traduisant un besoin d'information donné.

1.2. Approches de recherche d'information

Les approches de recherche d'information peuvent être classées en trois catégories génériques *(Aliane, 2001), (Ihadjaden, 1994)*:

- ***les approches statistiques :*** consistent à analyser un document, en évaluant les éléments d'un document par leur fréquence d'occurrence dans ce document. Ces statistiques peuvent être utilisées pour créer des index ou extraire les concepts d'un domaine en vue de sa modélisation.
- ***les approches linguistiques :*** visent l'indexation par la compréhension du sens des textes mais pour le moment elles ne permettent pas d'atteindre les objectifs visés et restent très coûteuses à réaliser.
- ***les approches intelligentes basées sur un modèle du domaine :*** D'après Sparck Jones *(Smail, 1998)*, un SRI intelligent est un système manipulant une base de connaissances portant sur les stratégies de RI et capable d'inférer des relations sémantiques entre la requête et les documents. En particulier, nous nous intéressons à l'application des techniques d'Intelligence Artificielle :

à la représentation du contenu des documents;
au traitement de la requête de l'utilisateur.

Dans le cas du web, ces deux points sont d'autant plus importants qu'en plus des bruits et de silences classiques dans un SRI, l'utilisateur se trouve livré à lui-même devant la grande masse d'information disponible.

2. La re-formulation de requête

Les utilisateurs d'un catalogue comme ceux qui utilisent les moteurs de recherche, ne sont pas des professionnels de la documentation. L'utilisateur ne sait pas choisir les bons termes qui expriment le mieux ses besoins d'information *(Aliane, 2001), (Ihadjaden, 1994), (Smail, 1998)*. En introduisant la reformulation de requête, la RI est alors envisagée comme une suite de formulations et de re-formulations de requêtes jusqu'à la satisfaction du besoin d'information de l'utilisateur, la requête initiale permettant rarement d'aboutir à un résultat qui satisfait ce dernier. Il s'agit en particulier d'ajouter des termes à la requête initiale de l'utilisateur et on parle alors d'expansion de la requête de l'utilisateur *(Smail, 1998), (Gauch, 1992)*. On distingue trois niveaux permettant de différencier entre les techniques d'expansion de requêtes *(Ihadjaden, 1994), (Gauch, 1992)* :

- *La source des termes utilisés dans la reformulation* et qui peuvent provenir des résultats de recherches précédentes ou d'une base de connaissance (réseau sémantique, thesaurus).
- *Le choix de la méthode* ou de l'algorithme qui permet de sélectionner les termes à ajouter à la requête initiale.
- *Le rôle de l'usager* dans le processus de sélection des termes et qui peut être actif ou passif.

2.1. La re-formulation manuelle

Cette approche est associée aux systèmes de recherche booléens. On peut procéder à la re-formulation de requête en utilisant un vocabulaire contrôlé (thésaurus ou classification) pour permettre à l'utilisateur de trouver les bons termes pour compléter sa requête.

2.2. La re- formulation automatique

Lorsque le feedback de pertinence s'accompagne d'une adjonction (et/ou) suppression de termes, on parle de re-formulation automatique. La requête de l'utilisateur est remaniée automatiquement, pour intégrer les descripteurs des documents jugés pertinents ou rejetés.

On trouve différentes variantes de cette technique : celles qui sont utilisées automatiquement pour reformuler la requête en augmentant le poids des termes présents dans les documents jugés pertinents et inversement pour diminuer les poids des termes jugés non pertinents.
Le problème avec la re-formulation automatique est l'estimation des « bons » termes qui peuvent conduire effectivement à une amélioration du processus de recherche car l'introduction des termes inappropriés peut entraîner un silence ou au contraire augmenter un bruit.

2.3. La re- formulation interactive

Dans une reformulation interactive, l'utilisateur joue un rôle actif. A l'inverse de la reformulation automatique, ici, ce sont le système et l'utilisateur qui sont, ensemble, responsables de la détermination et du choix des termes candidats à la reformulation. Le système joue un grand rôle dans la suggestion des termes, le calcul des poids des termes et l'affichage à l'écran de la liste ordonnée des termes. L'utilisateur examine cette liste et décide du choix des termes à ajouter dans la requête. C'est donc l'utilisateur qui prend la décision ultime dans la sélection des termes.

3. Approche proposée

Le SRI proposé est composé principalement des trois sous systèmes suivants :
- un système de reformulation de requêtes,
- un système de recherche,
- un système d'indexation.

Par ailleurs le système est doté de deux sortes d'interfaces, la première pour l'utilisateur final qui exprime son besoin d'information à travers une requête, la seconde pour l'expert administrateur de la base de connaissances. L'architecture générale du système est décrite par la figure suivante :

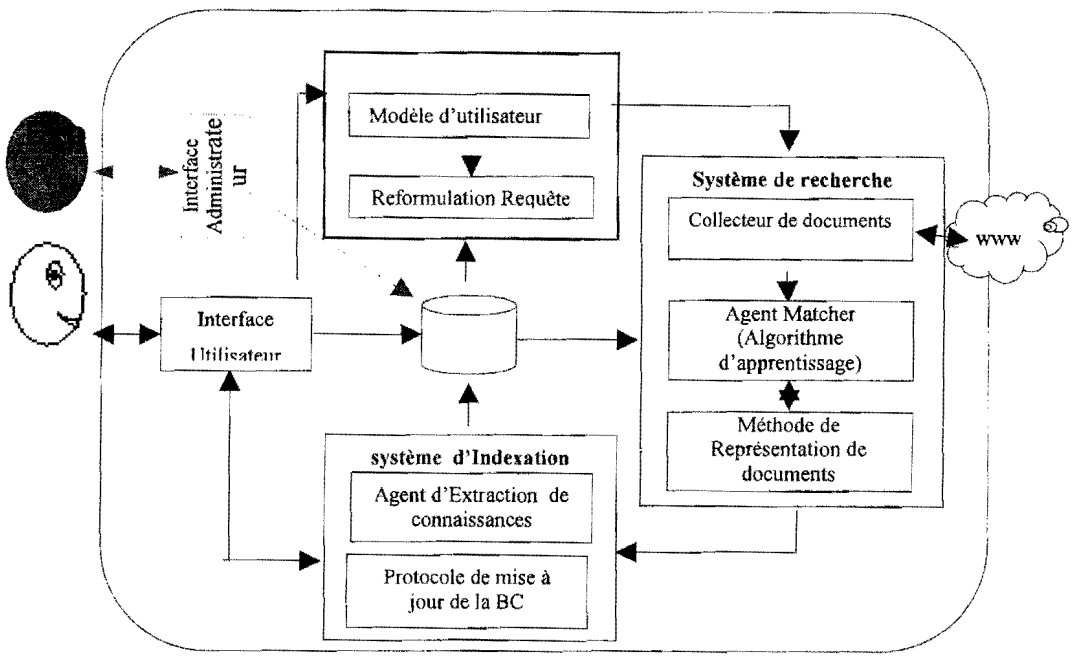

Figure 1 : *Architecture générale du système*

3.1. Fonctionnement du système

Le système est construit autour d'une base de connaissances sous forme d'un réseau sémantique modélisant l'univers du domaine d'application et repose sur une architecture à base d'agents (Abadeni et al, 1998), (Aliane, 2001) pour la prise en charge des différentes tâches (re-formulation, recherche, extraction, ...). Rappelons qu'un réseau sémantique est un ensemble de nœuds et d'arcs. Les nœuds représentant les concepts du domaine et les arcs les relations entre ces concepts (Bonnet, 1984).

Le réseau sémantique est initialisé manuellement par un expert humain qui dispose de toutes les fonctionnalités nécessaires à la gestion d'une base de connaissances (ajout, suppression, mise à jour).

Ultérieurement, le réseau sémantique peut aussi être alimenté par le système d'indexation qui indexe les documents restitués par le système de recherche d'information pour en extraire les concepts pertinents. Le processus de re-formulation ainsi que le processus de recherche utilisent un profil de l'utilisateur :

- L'utilisateur exprime sa requête au système de recherche d'information à travers une interface utilisateur.
- L'agent chargé de la reformulation de la requête récupère les informations du profil utilisateur et de la base de connaissances pour reformuler la requête et la transmettre au système de recherche.
- L'agent collecteur de documents du système de recherche collecte les documents à travers la banque de données (le web).
- L'agent matcher du système de recherche évalue les documents pertinents en utilisant la base de connaissances, le profil de l'utilisateur et un algorithme d'apprentissage.
- L'agent chargé de l'extraction extrait les connaissances à partir des documents pertinents à l'aide d'un algorithme d'extraction et met à jour la BC en prenant en compte les évaluations de l'utilisateur.

Les différents agents du système communiquent par envoi de messages. Un message peut être une requête reformulée, un document html, une représentation de documents... . L'architecture du système de re-formulation de requête est illustrée dans la figure 2 ci dessous.

3.2. Le processus de re-formulation

- Chaque terme de recherche dans la requête initiale représente un mot clé sur lequel l'utilisateur veut ou non de l'information. Le désir de l'utilisateur est représenté par les notions de « mot-clé positif » et « mot-clés négatifs » selon qu'il veut ou non de l'information sur un terme donné.
-Les termes initiaux de la recherche de l'utilisateur sont la meilleure indication de ses centres d'intérêt.
- Quelques termes de la base de connaissances peuvent être utiles.
- Le système ne doit jamais éliminer les mots-clés pour lesquels l'utilisateur a indiqué un intérêt.

3.2.1. Le profil utilisateur

Le profil utilisateur est obtenu une fois que l'utilisateur remplit le formulaire par le biais de l'interface utilisateur. Chaque utilisateur a un profil qui dépend de ses besoins d'information. Le système distingue entre les différents utilisateurs par leur profil en utilisant une approche statistique. Chaque utilisateur est identifié par un descripteur qui est utilisé dans la re-formulation de la requête.

3.2.2. La base de données

La base de données du système est une collection de descripteurs décrivant des documents html (url, titre, description, mots-clés). Le critère utilisé par l'algorithme de recherche est le critère mots-clés, selon un algorithme statistique basé sur le calcul des fréquences d'apparition des termes dans les documents.

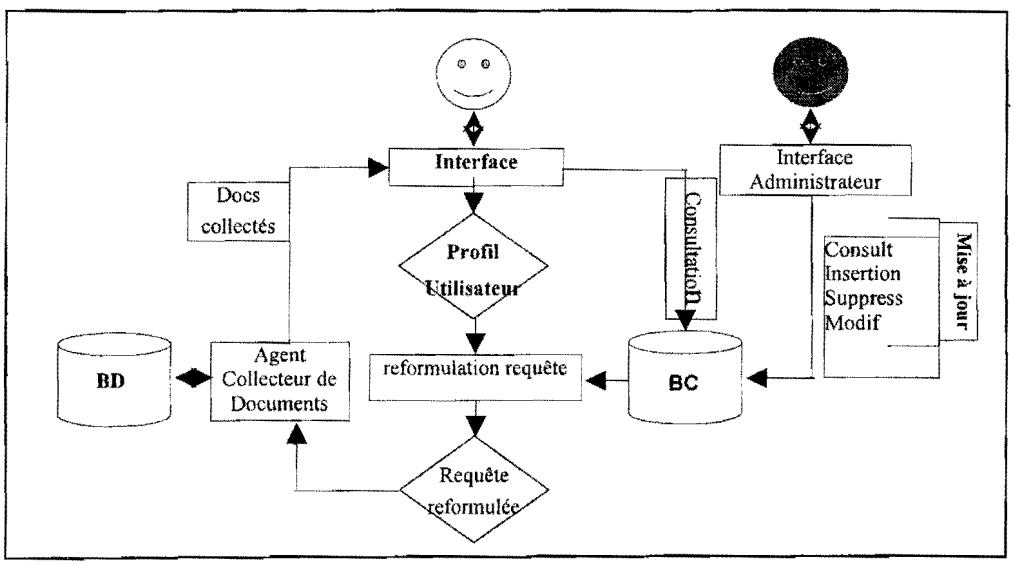

Figure 2 : *Architecture du système de re-formulation de requête*

3.2.3. La re-formulation automatique

L'approche que nous avons choisi pour la reformulation est l'approche interactive. La requête initiale est exprimée sous forme de deux listes choisies par l'utilisateur : la liste des mots-clés positifs et la liste des mots-clés négatifs.

- la liste des mots clés positifs : ce sont les termes proposés par l'utilisateur ou proposés par le système. Ils sont ordonnés selon leurs fréquences à partir de résultats de recherches précédentes.

- La liste des mots-clés négatifs : ce sont les termes pour lesquels l'utilisateur n'a pas un besoin d'information. Ils sont proposés par l'utilisateur ou proposés par le système.

Un mot-clé ne doit pas apparaître dans les deux listes en même temps.

3.2.4. L'expansion de la requête

Les termes de la requête proviennent des deux listes décrites ci-dessus. Pour élargir la requête initiale, des termes issus de la recherche sont ajoutés aux mots-clés. Les termes ajoutés à partir de la base de connaissance à un mot-clé positif dépendent du mot-clé lui-même, s'il appartient ou non à la base de connaissances.

- si le mot-clé positif n'appartient pas à l'ensemble des concepts du réseau, on l'élargit avec le concept racine.
- si le mot-clé positif appartient à l'ensemble des concepts du réseau, il est élargi selon son emplacement dans le réseau : si le concept n'a pas de concept fils, il est élargi par le concept père, sinon il est élargi par les concepts appartenant au sous réseau constitué par les fils. Les concepts ainsi ajoutés sont ceux qui ont une fréquence d'apparition élevée dans les requêtes précédentes de l'utilisateur.

4. Conclusion

Nous avons présenté dans cet article, un système de reformulation de requêtes utilisant une approche interactive pour l'expansion de la requête initiale d'un utilisateur exprimée sous forme de mots-clés. Le prototype réalisé reste à valider sur un corpus réel.

Par ailleurs, nous envisageons dans une étape ultérieure de traiter des requêtes exprimées en langage naturel et d'améliorer les algorithmes de recherche et d'indexation en combinant des outils linguistiques aux algorithmes statistiques actuels.

5. Bibliographie

N. Abbadeni, D. Ziou, S. Wang " Recherche d'images basée sur leur contenu" , Rapport de recherche, université de Sherbrooke, 1998.

H. Aliane " Towards a knowledge based plat-form for automatic indexing and information retrieval" Séminaire sur l'automatisation du trésor de la langue arabe, alger, 2001.

Bonnet A., Intelligence Artificielle : promesses et réalités, InterEditions ,1984.

Ferber J., Les systèmes multi-agents : vers une intelligence collective, Inter Editions, 1995.

Ferber J., «Les systèmes multi-agents : un aperçu général », Technique et Science Informatiques, vol. 16, n°8, 1997.

Gauch S., Smith J.B., An expert system for automatic query reformulation, Technical report, University of north california, 1992.

Ihadjaden M., Conception, réalisation et évaluation d'un système de recherche et de catégorisation automatique d'information textuelle sur Internet, Thèse de l'université ParisIV, 1994.

Smail M., «Vers des systèmes évolutifs de recherche d'information : un état de l'art » Technique et Science Informatiques, vol. 17, n°10, 1998.