



UNIVERSITE MOHAMED V-AGDAL
FACULTE DES SCIENCES
Rabat



HABILITATION UNIVERSITAIRE

Spécialité : Informatique

Recherche d'Information, et Optimisation des requêtes dans les entrepôts de données

Présentée par :

Abderrahim EL QADI

Doctorat National de la Faculté des Sciences de Rabat

Professeur Assistant à l'EST de Meknes

Soutenue le 21 Juin 2010 devant la commission d'examen :

Président	: Driss Aboutajdine	Professeur à la Faculté des Sciences de Rabat
Examineurs	: Nouredine Zahid	Professeur à la Faculté des Sciences de Rabat
	Rachid Oulad Haj Thami	Professeur à l'ENSIAS de Rabat
	Omar El Beqqali	Professeur à la Faculté des Sciences de Fès

Résumé

Les travaux présentés dans cette Habilitation Universitaire contribuent à plusieurs aspects. Nous avons présenté les différents enseignements assurés et les responsabilités scientifiques, pédagogiques et administratives assumées depuis janvier 2003. Egalement, nous avons décrit l'ensemble des activités de recherche scientifiques, d'encadrement détaillant mes contributions dans des travaux de thèses, de DESA et de Master, et des recherches en cours développées sur la base des travaux menés depuis l'obtention du doctorat national.

Les travaux de recherche développés dans cette HU se situent dans deux axes. Le premier s'inscrit dans le cadre de l'amélioration des performances des Systèmes de Recherche d'Information (SRI). Le deuxième s'attache à l'optimisation des requêtes dans les entrepôts de données décisionnels, en se basant sur les concepts de la génétique.

L'objectif principal des SRI est de répondre au besoin en information des utilisateurs. Les utilisateurs interrogent, au moyen d'une requête, une base de documents numériques et le SRI leur renvoie une liste de documents susceptibles de répondre à leur besoin. De nombreux modèles et stratégies de recherche d'information ont été proposés dans la littérature. Chacun d'eux utilise les éléments d'une théorie formelle afin de résoudre les problèmes inhérents à la recherche d'information : représentation du sens des documents et requêtes, traduction des liens sémantiques entre concepts, estimation de la pertinence ...

Afin valoriser au mieux l'ensemble des informations disponibles, les méthodes existantes de RI doivent être adaptées ou de nouvelles méthodes doivent être proposées. Le modèle d'analyse sémantique latente (LSA) a été proposé comme un mécanisme permettant de se greffer à un modèle de base, afin d'adapter la description de la requête à l'environnement linguistique du système. Dans les évaluations effectuées sur des collections d'essai, nous avons montré la performance du modèle LSA, et nous avons jugé l'intérêt d'appliquer conjointement le traitement automatique de la langue et la pondération des unités lexicales, pour pouvoir améliorer la performance de la méthode LSA. Nous avons confirmé que l'utilisation d'un modèle postérieur (LSA) au modèle VSM, la pondération okapi BM 25, et la pseudo-racination est également pertinente lorsque le corpus est en langue arabe. Sur la collection d'évaluation établie dans le domaine de l'environnement, nous avons abouti à une amélioration moyenne de 17% lorsque la racination et un dictionnaire sont utilisés.

Les entrepôts de données sont dédiés aux applications d'analyse et de prise de décision. Le processus d'analyse est réalisé à l'aide de requêtes complexes comportant de multiples jointures et des opérations d'agrégation sur des tables volumineuses. Les performances de ces requêtes dépendent directement de l'usage qui est fait de la mémoire secondaire. En effet, chaque entrée-sortie sur disque nécessitant jusqu'à une dizaine de millisecondes, l'accès à la mémoire secondaire constitue de ce fait un véritable goulot d'étranglement. L'administrateur, dans le but de minimiser le coût d'exécution de ces requêtes, sélectionne un ensemble de vues matérialisées et un ensemble d'index. Cette sélection diminue le coût des requêtes, mais entraîne un autre problème : les tables, les vues matérialisées et les index occupent une place très importante, et en conséquence ils ne peuvent pas être stockés en totalité dans la mémoire centrale. Dans un tel environnement, le nombre des entrées-sorties peut être grand si de bonnes techniques d'optimisation ne sont pas mises en oeuvre.

Par conséquent, il y a un besoin de développer des techniques qui peuvent faciliter l'exécution efficace de ces requêtes OLAP pour un grand entrepôt de données. À cet égard, nous avons présenté une nouvelle technique basée sur les AGs. Elle consiste à fragmenter un schéma relationnel d'un entrepôt de données horizontalement, ensuite verticalement afin de réduire le coût d'exécution de requêtes. Nous avons formalisé le problème de sélection de schéma de fragmentation verticale comme un problème d'optimisation avec contrainte. Cette dernière représente le nombre de fragments verticaux que l'administrateur peut maintenir. Les résultats générés par notre modèle de fragmentation basé sur les AGs ainsi que leur implémentation sous Oracle 10g sont intéressants et montrent que la fragmentation mixte peut être utilisée dans les entrepôts de données ayant des tables de dimension importantes.