

AVIS DE SOUTENANCE
THESE DE DOCTORAT

Présentée par

Mr : ISSAM SAHMOUDI

Spécialité : Informatique

Sujet de la thèse : Contribution à l'accès à l'information en langue arabe : Regroupement thématique des résultats de recherche et indexation à base des phares-clés.

Formation Doctorale : Sciences de l'ingénieur Sciences Physiques, Mathématiques et Informatique.

Thèse présentée et soutenue le samedi 07 juillet 2018 à 15h au centre des conférences devant le jury composé de :

Nom Prénom	Titre	Etablissement	
Nour Eddine RAISS	PES	Faculté des Sciences Dhar El Mehraz de Fès	Président
Karim BOUZOUBAA	PES	Ecole Mohammedia des Ingénieurs de rabat	Rapporteur
Azzedine MAZROUI	PES	Faculté des Sciences d'Oujda	Rapporteur
Abderrahim BENABBOU	PH	Faculté des Sciences et techniques de Fès	Rapporteur
Abderrahim ELQADI	PES	Ecole Supérieure de Technologie de Salé	Examineur
Abdelhak LEKHOAJA	PES	Faculté des Sciences d'Oujda	Examineur
Violetta CAVALLI-SFORZA	PH	Al Akhawayn University Ifran	Examineur
Abdelmonaime LACHKAR	PES	Ecole Nationale des Sciences Appliquées de Tanger	Directeur de thèse

Laboratoire d'accueil : Laboratoire ISA.

Etablissement : Ecole Nationale des Sciences Appliquées de Fès.

Titre de la thèse : Contribution à l'accès à l'information en langue arabe : Regroupement thématique des résultats de recherche et indexation à base des phares-clés.

Nom du candidat : Issam SAHMOUDI

Spécialité : Informatique

Résumé de la thèse

Cette thèse s'inscrit dans le cadre d'un projet, qui vise à mettre en place et améliorer les différents composants d'un Système d'Indexation et Recherche d'Information, dans le but de remédier aux différents problèmes d'accès à l'information en langue Arabe. Il s'agit du problème de Navigation et celui de l'Indexation. En effet dans la phase de la Navigation, les systèmes existants tel que Google, Yahoo, Bing proposent de retourner une liste ordonnée d'une dizaine de milliers de snippets (métas-données), les utilisateurs ne consultent que les premières pages, et par conséquent les documents situés à la fin de la liste ne seront jamais consulté bien qu'ils puissent être pertinents. Par ailleurs le problème d'indexation pose aussi un autre défi pour l'accès à l'information en langue Arabe, en effet, les méthodes d'indexation proposées dans la littérature sont basées sur les mots-clés qui peuvent avoir des sens différents et surtout pour le cas de la langue arabe qui présente un degré d'ambiguïté très élevé, et par suite l'indexation à base des mots-clés influence négativement les résultats des systèmes de recherche d'information pour les différentes langues en générale et la langue arabe en particulier.

Pour remédier à ces deux problèmes, dans le cadre de ce travail de thèse, nous avons proposé différentes contributions visant d'une part à faciliter la navigation et d'autre part à améliorer la phase d'indexation. En effet, pour faire face au problème de navigation, nous avons d'abord proposé de mettre en place un système de navigation interactif basé sur l'algorithme Suffix Tree Clustering STC, ce dernier permet le regroupement thématique des résultats de recherches pour assister les utilisateurs de la langue Arabe à accéder facilement à l'information, ensuite nous avons apporté une deuxième contribution au sujet de la navigation basée sur l'utilisation de l'Analyse Formelle de Concept (Formal Concept Analysis) FCA, cette dernière produit un regroupement conceptuel et présente une interface de navigation hiérarchique sur deux niveaux permettant ainsi une navigation rapide, simple et efficace dans une structure arborescente.

Pour remédier au problème d'indexation, nous avons proposé d'adopter une nouvelle approche d'indexation basée sur les phrases-clés, cette dernière permet la diminution de degré d'ambiguïté rencontré lors de l'utilisation de MotsClés simples, ainsi elles reflètent et représentent le mieux possible les principaux thèmes d'un tel document à indexer. Cependant, il faut tout d'abord mettre en place une phase de traitement pour l'extraction des Phrases-Clés, et que cette dernière phase doit être non supervisée par ce qu'elle va être intégrée dans la phase l'indexation. Nous avons d'abord proposé un nouveau système nommé (KeyPhrases Based Suffix Tree) KpST basé sur l'algorithme d'arbre de suffixes pour l'extraction des phrases-clés, ensuite nous avons proposé une autre version améliorée du même système nommé improved-KpST en intégrant une phase de filtrage linguistique. Pour mettre en évidence l'intérêt de la proposition, nous avons bien mené et une étude comparative entre les deux versions KpST et ImprovedKpST et la méthode nommée KP-Miner considérée dans la littérature comme la plus performante. Les résultats obtenus montrent que la méthode Improved KpST améliore largement la performance et dépasse très significativement la méthode de KP-Miner avec un écart de 16%. Finalement, nous avons mis en œuvre l'ensemble des interfaces utilisateurs nécessaires en offrant à l'utilisateur la possibilité de réaliser la phase d'indexation pour ses propres collections de données textuelles. Notons que jusqu'à aujourd'hui cette dernière fonctionnalité n'est pas encore offerte par aucun des systèmes existants. Cette dernière contribution d'indexation peut être exploitée dans différents domaines et applications pour faciliter l'accès à l'information en langue Arabe.

Mots clés :

Langue Arabe, l'accès à l'Information, Regroupement Thématique des Résultats de Recherche, Navigation intelligente, Indexation à base des Phrases-Clés, Suffix Tree Clustering, l'Analyse Formelle de Concept.