

Le document a connu un véritable essor avec le développement du web. Son contenu est devenu très varié. Il est constitué de texte, d'images et de son et le tout s'articulant autour d'une structure. A leur début, les systèmes de recherche d'information (RI) considéraient les documents comme des unités atomiques indépendantes les unes des autres et constituées d'un ensemble de mots et de phrases. L'avènement de nouveaux standards de représentation des documents, et plus particulièrement XML, a poussé la communauté de RI à exploiter la richesse présente dans ces documents et à développer de nouveaux concepts pour l'indexation et l'interrogation du corpus XML. L'information structurelle des documents peut en effet servir à affiner le concept de granule documentaire. La réponse fournie à l'utilisateur ne se résume plus à un document entier mais à des parties de document apportant une information pertinente à un besoin utilisateur. Notre objectif est d'utiliser toutes les sources d'évidence liées au contenu et à la structure du document pour sélectionner les éléments pertinents répondant à une requête. Nous orientons nos travaux selon deux directions. La première consiste à exploiter la structure hiérarchique contenue dans les documents XML et la seconde consiste à utiliser l'information apportée par les liens de type Xlink et Xpointer reliant les documents XML. Ceci nous a conduit à proposer : (i) une méthode de recherche basée sur la propagation de la pertinence, (ii) une méthode basée sur la propagation des termes et de leur poids. Ces méthodes ont pour but de retrouver les unités d'information les plus exhaustives et les plus spécifiques répondant à une requête utilisateur constituée de mots clés. (iii) Une autre dimension doit s'ajouter aux mesures d'exhaustivité et de spécificité citées ci-dessus pour évaluer la pertinence d'un élément vis-à-vis d'une requête. Il s'agit de l'information apportée par les liens Xlink et Xpointer. Les documents semi-structurés peuvent être représentés sous forme arborescente, le but est alors de trouver les sous arbres de taille minimale répondant à la requête