Vers une Architecture Multi-agent à Base des Réseaux Petits Mondes Hiérarchiques et des Réseaux Possibilistes pour les Environnements Riches en Informations

Bilel Elayeb[*], Fabrice Evrard[**], Montaceur Zaghdoud[*] et Mohamed Ben Ahmed[*]

[*] Laboratoire RIADI-GDL, Ecole Nationale des Sciences de l'Informatique, Manouba 2010 Tunisie.
{Bilel.Elayeb, Mohamed.Benahmed}@riadi.rnu.tn, Montaceur.Zaghdoud@ensi.rnu.tn
[**] IRIT-ENSEEIHT, 02 rue de Charles Camichel, B.P. 7122, 31071 Toulouse Cedex 7 France.
Fabrice.Evrard@enseeiht.fr

Résumé :

La problématique majeure de la Recherche d'Information (RI) consiste à extraire à partir d'une collection de documents, ceux qui répondent à un besoin utilisateur en se basant souvent sur des informations pauvres. Les différents modèles connus de la RI (booléen, vectoriel, probabiliste, bayésien) représentent les documents et les requêtes sous forme de listes de termes pondérés puis mesurent une valeur de pertinence (similarité vectorielle, probabilité de pertinence) en se basant sur ces termes et leurs poids. La pondération des termes est à notre sens l'élément fondamental de tous les modèles de RI actuels (Ribeiro-Neto et al., 1996), (Sparck, 1998). Lorsqu'elle est calculée automatiquement, cette pondération est obtenue à partir de combinaisons des fréquences d'apparition des termes dans les documents (tf), des fréquences d'apparition des termes dans la collection (idf) et de la longueur des documents (dl) (Salton et al., 1994) (Singhal et al., 1996). Quel que soit le modèle, la réponse à une requête est une liste de documents ordonnés selon cette valeur de pertinence. Certaines approches considèrent les poids des termes comme des probabilités de pertinence. Dans ces modèles, l'incomplétude de l'information, intrinsèque à la représentation vectorielle d'un document, n'est pas considérée lors de son évaluation pour une requête donnée. En réalité, on ne distingue pas entre les notions de possibilité ou de certitude lors des calculs de la pertinence. Les méthodes actuelles, relativement pauvres, utilisées pour représenter les documents (ensemble de termes et de leurs poids) ainsi que pour représenter le besoin utilisateur ne sont pas totalement compatibles avec une définition précise de la pertinence.

L'objectif de ce travail est donc de proposer une approche possibiliste mixte (quantitative et qualitative) basée sur les mesures de nécessité et de possibilité dans un modèle de Recherche d'Information. Un tel modèle devrait être capable de répondre à des propositions du type :

- Est-il plausible à un certain degré que le document d_i constitue une bonne réponse à la requête R_j?
- Est-il nécessaire, certain (dans le sens possibiliste), que le document d_i réponde à la requête R_j?
- Le document d_1 est préférable au document d_2 ou l'ensemble $\{d_1, d_2\}$ est-il préférable à l'ensemble $\{d_3, d_4\}$?

Le premier type de proposition vise à éliminer les documents faiblement plausibles de la réponse. La seconde réponse se focalise sur les documents qui seraient réellement pertinents. Le dernier type de proposition suggère que la liste ordonnée des documents en réponse à un besoin utilisateur peut être traitée d'une manière qualitative, et que des approches ordinales pourraient être utilisées dans la représentation des documents et des requêtes. La définition de la pertinence d'un document vis à vis d'une requête, en fonction des données dont nous disposons, est difficilement exprimable (ou traduisible) par une unique mesure de probabilité. En effet, celle-ci ne tient pas compte des notions d'imprécision et de vague intrinsèques à la pertinence (Brini et Boughanem, 2003). En réalité, une mesure de probabilité portant sur un événement et son contraire est quelque peu restrictive. Dans le modèle proposé, un document contenant tous les termes de la requête constitue une réponse possiblement pertinente à la requête. Cette plausibilité doit être renforcée par une certitude provenant de la mesure de nécessité. La mesure de possibilité est utile pour filtrer les documents et la mesure de nécessité pour renforcer la pertinence des documents restants. L'usage de la théorie des possibilités en RI avait déjà été suggéré par (Prade et Testemale, 1987) qui proposaient un nouveau modèle d'indexation sous forme de groupes de mots-clés, pondérés par des degrés de possibilité et de nécessité.

La problématique générale de ce travail de recherche s'intéresse, au début, à l'Internet, aux systèmes multi-agents (SMA) et aux systèmes d'information qui font appel à ce contexte. Nous distinguons, en fait, trois points principaux :

Le premier point est que lorsqu'on lance une requête sur Internet utilisant des mots clés (grâce au moteur de recherche Google, par exemple) on obtient comme résultat une liste en général considérable (plusieurs milliers) de pages web répondant à cette requête. On peut par une utilisation de mots-clés complémentaires rechercher quelque chose de plus pertinent dans l'ensemble des résultats précédents. Ainsi de suite jusqu'à obtention (du moins on l'espère) du résultat recherché. Or on pourrait procéder différemment en structurant l'ensemble des

pages web obtenues au premier résultat. La structure consisterait en fait à classer tous ces résultats par domaines et sous domaines. Une technique très prometteuse émerge aujourd'hui et fait appel au réseaux petits mondes hiérarchiques (RPMH) (Watts et Strogatz, 1998) (Nguyen et Martel, 2003) (Newman, 2003) (Matthew et Brian, 2004).

Ainsi chaque page web serait un noeud d'un gigantesque graphe dont les arcs seraient les liens hypertextuels d'une page vers une autre. Certains calculs sur ce graphe sont à même de faire apparaître des regroupements thématiques (pages web qui "parlent" plus ou moins d'un même sujet). Ainsi chercher une information sur le web ne se ferait plus au hasard. Mieux encore : une requête sous forme d'une description même approximative de ce que l'on cherche ferait aboutir dans un tel "cluster" thématique et même sur la meilleure page web de ce cluster.

Le deuxième point est que chaque page web est à ce jour un document inerte qui ne "réagit" pas à l'utilisateur qui l'a demandé. Or on pourrait conférer à toute page web des caractéristiques, des comportements, des moyens d'interaction avec l'utilisateur pour peu que ces pages web soient considérées comme des agents communicants au sens des SMA. La technique précédente aurait organisé toutes ces pages web comme une société d'agents hiérarchisés et organisés en petits mondes. Un utilisateur devrait pouvoir dialoguer au travers d'un langage de communication avec ces agents mais surtout il devrait pouvoir spécifier un agent fictif chargé de communiquer avec eux afin de constituer un nouvel agent (une nouvelle page web) contenant toutes les informations que l'utilisateur recherche : informations obtenues par échanges multiples entre cet agent fictif et cette communauté thématique d'agents (Ferber, 1995).

Le croisement de ces deux techniques : une approche SMA appliquée à un réseau de pages web organisées en petits mondes, permet d'envisager de façon nouvelle les problèmes de recherche d'informations, de fouille de données, de construction raisonnée de connaissances, et de les résoudre par des méthodes orientées par les buts de l'utilisateur. Ces buts sont exprimés via un langage de communication et résultent d'un discours utilisateur (constitué d'une suite d'énonciations dans un ACL) faisant office de programme de construction d'un nouvel agent.

Dans cette perspective, nous proposons une nouvelle architecture multi-agent de recherche d'information sur Internet, baptisée SARIPOD, à base des deux RPMH et des Réseaux Possibilistes (RP). Le premier RPMH consiste à structurer les réponse "à la Google" en zones denses de pages web très fortement liées les unes aux autres. Nous faisons ainsi apparaître des nuages denses de pages qui "parlent" plus ou moins du même sujet et qui répondent toutes fortement à une requête. Le second RPMH est celui qui consiste à ne pas prendre les mots-clés tels quels mais à considérer une requête comme multiple en ce sens qu'on ne cherche pas seulement le mot-clé dans les pages web mais aussi ses synonymes. Les Réseaux Possibilistes engendrent le mixage de ces deux RPMH afin d'organiser les documents recherchés selon les préférences de l'utilisateur. En effet, ce système présente une nouvelle approche possibiliste mixte pour un système de Recherche d'Information. Ce système, qui voit la RI comme un problème de diagnostic, traduit à l'aide de réseaux possibilistes naïfs (Dubois et Prade, 1987, 1998) (BenFerhat et al., 2002) des relations de dépendance entre les documents et les termes de la requête. Ces relations sont quantifiables par deux mesures : la possibilité et la nécessité de pertinence. Le processus de recherche restitue les documents plausiblement ou nécessairement pertinents pour un utilisateur. De plus, si l'approche de base de (Brini et al., 2004) tient compte ici de l'aspect quantitatif et ne tient pas compte de la dépendance entre les termes de la requête, notre système permet de l'étendre au cadre qualitatif possibiliste, en introduisant la dépendance entre les termes de la requête.

L'évolution de l'Internet et l'apparition des entrepôts de données couplés à la nature dynamique et hétérogène de l'information font en sorte qu'il est de plus en plus difficile de trouver l'information pertinente et à jour malgré son abondance. Une approche prometteuse à la résolution de ce problème consiste à utiliser des agents logiciels qui coopèrent pour trouver la réponse adéquate à une requête d'information. Le système SARIPOD est une architecture basé multi-agent pour la recherche d'informations sur Internet.

Le fait qu'on a affaire à des sources d'informations collectées à partir du réseau Internet, nous a fait opter pour le développement d'un agent crawler capable d'explorer le web. Il nous a paru également intuitif d'interfacer l'utilisateur au moyen d'agents d'interface. Finalement, et comme nous l'avons souligné plus haut, le fait qu'on a affaire à des environnements ouverts et dynamiques nous a fait opter pour le développement d'une couche d'agents intermédiaires. Nous avons donc fait apparaître trois niveaux d'abstraction au niveau de l'architecture multi-agent abstraite du système SARIPOD : la couche de communication avec l'utilisateur, la couche de traitement d'informations et la couche d'extraction d'informations.

Nous avons fait les tests de SARIPOD sur des sites web locales contenant 10000 pages et pour des mots-clés ayant des nombres variables de mots proches. Nos résultats préliminaires se limitent uniquement aux occurrences verbales dans les définitions des verbes dans le RPMH de dictionnaire.

Suite à ces expérimentations, nous avons remarqué que la détermination des mots proches de mots-clés est très dépendante du degré de nettoyage du dictionnaire français utilisé (Le Grand Robert) comme source de données pour le RPMH de dictionnaire. Nous avons remarqué aussi que toute augmentation du nombre de synonymes donne plus de chance de collecter plus d'URLs, ce qui augmente en conséquence la durée de la recherche. Suite à des tests successifs, nous avons opté pour un nombre de mots proches qui nous semble optimal égal à trois en tenant compte de la qualité des documents collectés ainsi que de la durée d'une requête de recherche.

Nous avons signalé aussi que l'écart entre le degré de pertinence possibiliste de la page la plus pertinente ($DPM(d_1)$) et celui de la page la moins pertinente ($DPM(d_N)$) de la collection diminue lorsque le nombre des URL sélectionnées augmente, ce qui prouve que le but premier qui nous a motivé dans l'usage des RPMH est bien vérifié ici : faire en sorte que les réponses renvoyées suite à une requête ne soient plus le vrac "à la Google", mais quelque chose de structuré en RPMH de sorte que si une page parmi les réponses renvoyées semble pertinente alors toutes celles qui lui sont "proches" dans ce RPMH le seront aussi. Ainsi, nous augmentons très considérablement le nombre de pages récupérées pour les mots-clés recherchés et nous changeons le PageRank (au sens de Google) des pages résultats de la requête.

Nous présentons, comme évaluation du système SARIPOD, d'une part une justification de l'utilisation des RPMH dans le domaine de la recherche d'information, et d'autre part son évaluation par rapport aux moteurs classiques de recherche d'information sur Internet tels que Google (Vise et Malseed, 2006). En fait, nous distinguons deux usages très importants de ces deux RPMH ainsi que leur mixage dans le système SARIPOD :

Le premier est celui qui consiste à structurer les réponses "à la Google" en zones denses de pages web très fortement liées les unes aux autres. On fait ainsi apparaître des nuages denses de pages qui "parlent" plus ou moins de la même chose et qui répondent toutes fortement à une requête. Pour un autre nuage de pages web fortement liées les unes aux autres il en va de même, elles répondent toutes à cette même requête. La différence essentielle est que chaque nuage de pages web répond fortement d'une manière particulière à la requête.

Par exemple, la requête "jouer", dans le RPMH des synonymes des mots du français, donne quatre nuages de verbes proches de jouer : le premier nuage concerne A = {parier, risquer, miser, hasarder,...}, le deuxième B = {tromper, mystifier, abuser, berner,...} etc. pour les deux autres. Pour le web il en va de même une requête (exprimée avec quelques mots-clés) renvoie un ensemble de pages web (réponses à la Google) qu'il faut organiser en RPMH de sorte à faire apparaître quelques grands nuages de pages web parmi toutes ces réponses. Chaque nuage regroupe ainsi un lot de pages qui répondent toutes de façon pertinente et d'une certaine façon à la requête (comme A répond pertinemment à la requête "jouer" d'une certaine façon : celle qui s'intéresse au "pari", ou comme B qui répond tout aussi pertinemment à la même requête "jouer" mais cette fois d'une façon différente : celle qui s'intéresse à l'"abus", etc.). Pour le web chaque nuage de pages web sera pertinent et, grâce à des mots-clés supplémentaires, il sera possible de sélectionner un nuage particulier ou une partie de ce nuage.

La qualité réside dans le fait que quand on regarde les pages web d'un même nuage, toutes les pages sont pertinentes, mais si ce degré n'est pas encore suffisant, on peut faire des requêtes dans ce seul nuage (contrairement à Google qui n'organise jamais ses 300.000 réponses en nuages) pour obtenir un sous-ensemble de pages web que l'on peut de nouveau (donc récursivement) organiser en sous-RPMHs et ainsi de suite. Au plus profond de cette entreprise de structuration on trouve des pages web seules. L'ensemble des réponses a donc été organisé en RPMH et sous-RPMH de sorte à constituer une sorte d'arbre de décision (ou structure de classification) des pages web en fonction des mots-clés utilisés. Ce que ne fait pas Google qui sait seulement faire des recherches dans l'ensemble des réponses précédentes. En fait, Google est "capable" de renvoyer dans une sous requête des pages que notre système aurait mis dans des nuages différents lors de la première requête.

Le deuxième usage très important des RPMH est celui qui consiste à ne pas prendre les mots-clés tels quels mais à considérer une requête comme multiple en ce sens qu'on ne recherche pas seulement le mot-clé dans les pages web mais aussi ses synonymes. En fait, au delà de la stricte synonymie, on recherchera ce mot-clé mais aussi des mots qui lui sont "proches". Proche au sens du calcul de la proxémie définie par notre approche basée sur l'étude des circuits dans un RPMH de dictionnaire. Les mots considérés comme proches incluent donc les synonymes de ce mot mais ne s'y restreignent pas. On aura potentiellement (en pratique ça sera limité par une borne) tous les mots plus ou moins proches du mot de la requête. Ce nombre de mots est paramétrable (1, 5, 100, ...). Une requête est donc maintenant très flexible puisqu'elle tolère qu'une page web soit une bonne réponse même si elle ne contient pas (à strictement parler) le mot-clé en question.

Or pour pouvoir disposer de cette flexibilité nous avons évidemment besoin d'un dictionnaire et surtout d'avoir structuré ce dictionnaire (l'ensemble des entrées de celui-ci) en RPMH justement pour savoir quel mot est proche de quel autre. Or il y a de nombreuses façons de faire émerger une structure de RPMH à partir d'un dictionnaire, celle de (Gaume et al., 2004) par exemple consiste à se servir des définitions : M_1 est relié à M_2 si et seulement si M_2 appartient à la définition de M_1, à l'aide de cette définition de la relation entre deux mots il en déduit par proxémie la "proximité sémantique" de tout mot à tout autre. Notre système SARIPOD reprend cette définition

et s'appuie sur cette proxémie entre les mots pour rendre les requêtes plus flexibles. On peut à partir de là quantifier les pages web obtenues suite à une requête utilisant certains mots-clés. Chaque page réponse sera caractérisée par un degré d'adéquation ou de pertinence qui résultera de la combinaison des degrés de proxémie aux mots-clés de la requête des mots effectivement présents dans cette page.

Mots-clés :

Agents, Système Multi-Agent, Recherche d'Informations, Réseaux Petits Mondes Hiérarchiques, Réseaux Possibilistes, Préférences Utilisateur, Document Pertinent.

Références :

Benferhat, S., Dubois, D., Garcia, L., Prade, H., 2002. On the transformation between possibilistic logic bases and possibilistic causal networks. In *International Journal of Approximate Reasoning*, 29, n°2, 135-173.

Brini, A., H., Boughanem, M., Dubois, D., 2004. Towards a possibilistic approach for information retrieval. *Data and Knowledge Engineering Proceedings EUROFUSE 2004*, Varsovie, Pologne, p. 92-102.

Brini, A., H., Boughanem, M., 2003. Relevance feedback : introduction of partial assessments for query expansion. *Proc. 2d. EUSFLAT Conf*, Zittau, Allemagne, 67-72.

Dubois, D., Prade, H., 1998. Possibility Theory: Qualitative and Quantitative Aspects. In *Handbook on Defeasible Reasoning and Uncertainty Management Systems*, vol.1, Kluwer, 21-42.

Dubois, D., Prade, H., 1987. *Théorie des Possibilités : Application à la Représentation des Connaissances en Informatique*. Edition MASSON. Paris.

Ferber, J., 1995. *Les systèmes multi-agents, vers une intelligence collective*. InterEditions, Paris.

Gaume, B., Hathout, N., Muller, P., 2004. Désambiguïsation par proximité structurelle. *TALN 2004*, Fès, Maroc.

Newman, M., E., J., 2003. The structure and function of complex networks. *SIAM Review*, volume 45, 167–256.

Ribeiro-Neto, B., Silva, I., Muntz, R., 1996. A Belief Network Model for IR. *Proc. Of the 19th ACM SIGIR Conf. on Research and Development in Information Retrieval*, 253-260.

Prade, H., et Testemale, C. 1987. Application of possibility and necessity measures to documentary information retrieval. *In Uncertainty in Knowledge-Based Systems*. B. Bouchon, R. Yager, Eds., LNCS n°286, Springer Verlag, Berlin, 265-274.

Salton, G., Allan, J., Buckley, C., Singhal, A., 1994. Automatic Analysis, Theme Generation and Summarization of Machine Readable Texts. *Science*, 264, 3, 1421-1426.

Singhal, A., Salton, G., Mitra, M., Buckley, C., 1996. Document Length Normalization. *Information Processing and Management*. 32(5): 619-633.

Sparck, J., K., 1998. A Look Back and a Look Forward. *Proceedings of the 11th annual international ACM SIGIR Conf. on Research and Development in Information Retrieval*, 13-29.

Vise, D., Malseed, M., 2006. *Google story*, Edition Dunod, Paris.

Watts, D., Strogatz, S., 1998. Collective dynamics of 'small-world' networks. *Nature*, (393), 440–442.