Apport du Web sémantique dans la réalisation d'un moteur de recherche géo-localisé à usage des entreprises

Frédéric Triou*, Fabien Picarougne*
Henri Briand*

*LINA CNRS FRE 2729 - Équipe COD
École Polytechnique de l'Université de Nantes
rue Christian Pauc, 44306 NANTES Cedex 3, France
{frederic.triou, fabien.picarougne, henri.briand}@univ-nantes.fr
http://www.sciences.univ-nantes.fr/lina/fr/research/teams/ECD/index.html

Résumé. La recherche d'une entreprise sur le Web, relative à un savoir-faire particulier, n'est pas une tâche toujours facile à mener. Les outils mis à la disposition de l'internaute ne donnent pas entièrement satisfaction. D'un côté les moteurs de recherche éprouvent des difficultés à faire ressortir clairement le résultat escompté. De l'autre côté, les annuaires spécialisés (type Pages Jaunes) sont tributaires d'une organisation figée, nuisant à leur efficacité.

Face à ce constat, nous nous proposons de créer un nouveau moteur spécialisé dans la recherche d'entreprise, associant Web sémantique et géo-localisation. Cette approche novatrice nécessite l'implémentation d'une ontologie ayant pour objectif la formalisation des connaissances du domaine.

Cette tâche a mis en évidence l'intérêt des structures économiques, maintenues par l'INSEE, et leur utilisation au sein de l'ontologie. Les nomenclatures économiques ont été retenues pour gérer la classification des activités et produits pouvant être dispensés par les entreprises. La structure des unités administratives, telle que gérée au sein du fichier SIRENE, s'est avérée judicieuse pour répondre à la problématique de géo-localisation des entreprises. Une opération de désambiguïsation est réalisée en associant à chaque nœud d'activité les mots clés et synonymes lui correspondant.

Enfin, nous comparons les résultats obtenus par notre moteur à ceux obtenu par le principal moteur de recherche d'activités géo-localisées en France : les Pages jaunes. Que ce soit au niveau de la précision et du rappel, notre moteur obtient des résultats significativement meilleurs.

1 Introduction

Les moteurs de recherche classiques sur le web ont des caractéristiques étonnantes : ils possèdent des milliards de documents dans leur index, ils peuvent traiter des millions de requêtes quotidiennement, ils donnent des réponses très volumineuses quasiment en temps réel et ils nécessitent des ressources informatiques et humaines considérables. On peut dire aujourd'hui

qu'il est pratiquement impossible de concevoir une approche alternative pour un moteur de recherche sans passer par l'un de ces « géants ». Même si les points forts de ces moteurs sont nombreux, ils ont aussi des faiblesses, comme des requêtes très simples, une présentation des résultats souvent pauvre en information, ou encore la nécessité pour l'utilisateur d'explorer un à un les nombreux liens qu'ils donnent en sortie.

En effet, lorsque l'on recherche une information précise ou spécialisée, comme par exemple trouver une entreprise répondant parfaitement à notre préoccupation du moment, les moteurs classiques se révèlent difficile d'usage et peu pertinents à la fois. Pour arriver à ses fins, l'internaute a alors le choix entre plusieurs types d'outils destinés à l'aider dans sa tâche. D'un côté, les méta-moteurs de recherche spécialisés, bien que de plus en plus sophistiqués, se heurtent à deux écueils principaux : le Web invisible et la masse gigantesque d'informations gérées. D'autre part, les annuaires professionnels représentent une alternative intéressante par une meilleure exhaustivité des données managées. Cet avantage est cependant tempéré par le cloisonnement des entités à l'intérieur de rubriques préétablies, manquant de discernement. Ces différentes solutions montrent leurs limites dans le manque d'efficacité en terme de pertinence.

Nous proposons dans cet article un outil de recherche spécialisé dans la recherche d'entreprises mêlant efficacité de formulation de la demande initiale et pertinence des résultats affichés. Afin d'assister le mieux possible l'internaute, ce nouveau moteur de recherche dispose d'une fonctionnalité de géo-localisation autorisant la restriction géographique de la requête et le repérage graphique des entreprises atteintes.

L'élaboration d'un moteur de recherche, à usage des entreprises, bâti sur une infrastructure Web sémantique constitue une nouvelle voie qu'il convenait d'exploiter. L'idée maîtresse est de donner une signification au contenu des documents présents sur le Web. De cette façon, les machines sont en mesure de comprendre le sens des documents et d'effectuer des raisonnements automatisés. La réalisation de ce moteur de recherche « intelligent » passe par la modélisation d'une ontologie. Celle-ci est nécessaire à la formalisation des connaissances du domaine, permettant leur partage et leur interprétation opérationnelle.

La suite de cet article est organisée comme suit ; la section 2 présente un état de l'art de différents systèmes de recherche d'information se basant sur le principe du web sémantique et mettant en œuvre des ontologies. La section 3 détaille l'architecture retenue pour l'élaboration de notre moteur de recherche ainsi que la description des différentes ontologies permettant d'organiser l'ensemble des données accessible. La section 4 donne des résultats expérimentaux obtenus par comparaison au principal outil du genre : *les Pages Jaunes*. La section 5 conclut sur les nombreuses perspectives qui découlent de ce travail.

2 Systèmes de recherche d'information et ontologie

Le Web sémantique, proposé par le W3C (World Wide Web Consortium), est une nouvelle approche qui vise, à partir de la structure actuelle du Web, à donner un sens au contenu des pages. Selon Tim Berners-Lee, inventeur du Web et directeur du W3C, "*The semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation*" Burners-Lee et al. (2001). C'est une manière de donner un sens bien défini aux informations permettant une interprétation aussi bien par les machines, que par les humains.

Par ce biais là, l'objectif du Web sémantique est de décharger les utilisateurs d'une grande partie de leurs tâches de recherche et d'exploitation des résultats. Mais pour que le Web sémantique fonctionne il est nécessaire que les machines aient accès à des collections structurées d'informations et de règles d'inférence qu'elles peuvent utiliser pour parvenir à un raisonnement automatisé. Cette modélisation des données, appelée ontologie, est destinée à jouer un rôle clé car dispensant une connaissance commune, et partagée, du domaine. Le monde des sciences de l'information s'est approprié ce terme pour désigner "*une spécification formelle et explicite des termes d'un domaine ainsi que des relations que ces termes entretiennent entre eux*" Gruber (1993). Il est important de compléter cette définition en précisant qu'une "*ontologie est indépendante des considérations d'exécution, son objectif principal étant de spécifier la conceptualisation du domaine sous-jacent à l'application*" Welty et Guarino (2001).

Dans le cadre du Web sémantique, l'ontologie a ainsi pour enjeu de "*proposer une compréhension partagée et commune, pour un domaine donné, qui peut être transmise aussi bien aux personnes qu'aux applications*" Davies et al. (2002). Elle doit traduire un certain consensus, explicite, de manière à être partagée par la communauté l'ayant construite et acceptée. Ceci est vital pour permettre l'exploitation des ressources présentes sur le Web par différentes applications ou autres agents logiciels. Le Web sémantique doit ensuite ajouter de la logique, c'est-à-dire lui donner la possibilité d'utiliser les règles pour faire des inférences. Par l'exploitation de ces règles, les moteurs d'inférence peuvent raisonner intelligemment et offrir des réponses automatiques à des questions posées par une personne. Ils doivent être en mesure de déduire de nouvelles informations à partir d'informations et de ressources déjà existantes dans l'environnement.

Nous voyons apparaître depuis quelques années des outils exploitant ces données dans des cadres divers et notamment concernant les systèmes de recherche d'information. Les informations contenues dans les ontologies permettent de déterminer un sens non ambigu aux différents éléments rencontrés. Les résultats produits sont par conséquent plus pertinent que ceux issus des moteurs de recherche traditionnels se basant sur des considérations purement statistiques pour effectuer leurs recherches Lawrence et Giles (1999); Brin et Page (1998); Baeza-Yates et Ribeiro-Neto (1999).

Un des premiers domaine à avoir bénéficié des avancées des ontologies sur le Web est certainement celui des annuaires, tel *Yahoo !*. Pour ces applications, la localisation d'une information est assurée par un système de catalogues thématiques hiérarchiques (classification) consultables à l'aide de mots clés ou par navigation de thèmes en sous thèmes. Par exemple, Labrou et Finin (1999) utilise cette classification afin de décrire les documents à la manière d'une ontologie.

L'annotation de chaque document permet de rendre le sens compréhensible par des outils automatique. En utilisant un mécanisme d'inférence, il devient alors possible d'améliorer la qualité des résultats produits par les moteurs de recherche sémantiques Mayfield et Finin (2003). En particulier Shah et al. (2002) utilise le texte sémantiquement enrichi afin de lever certaines ambiguïtés du texte *libre* et procède ensuite par inférence pour améliorer la qualité de l'indexation.

Guha et al. (2003) a, quand à lui, développé un système de recherche d'information utilisant une ontologie afin d'améliorer les résultats des moteurs de recherche classiques en ajoutant des sources issus des concepts de l'ontologie associés aux résultats originaux.

Cette information sémantique peut aussi être utilisée afin d'affiner des mesures utilisés dans les algorithmes de scoring des moteurs de recherche. L'indexeur *Swoogle* Ding et al. (2004), Ding et al. (2005), découvre, indexe et analyse les ontologies des documents du web. Il utilise les informations sémantiques contenus dans les méta-données des documents afin de produire une mesure de similarité la plus pertinente possible. OntoSearch, Gao et al. (2005), analyse les méta-données afin de déterminer des poids dans un vecteur de concept pour chaque document. La pertinence des documents à la requête de l'utilisateur est alors mesurée en calculant une similarité entre vecteur de document et vecteur de requête.

3 Un moteur de recherche d'activités géo-localisées

Le domaine de la recherche d'informations est, en partie, lié aux langues que ce soit lors de l'interprétation d'une requête ou de l'analyse des documents traités. Il apparaît selon plusieurs auteurs (de Loupy (2000)), qu'un système de recherche d'informations devra, pour être efficace, conjuguer l'approche statistique avec un traitement linguistique. De nombreux problèmes de polysémie et de synonymie limitent en effet l'efficacité d'une recherche purement statistique par mots clés. Les relations sémantiques précédentes sont génératrices de non-conformité des résultats produits par une recherche. Une des solutions destinée à lever ces ambiguïtés sémantiques est d'utiliser des liens thématiques de Loupy et Crestan (2004). La démarche consiste à regrouper les termes par affinités : par exemple, le domaine *services informatiques* pourrait regrouper les termes *infogérance*, *développement* et *maintenance logiciels*.

Afin de cataloguer et de gérer les ambiguïtés pouvant intervenir dans la formulation des activités, nous avons décidé de créer une ontologie de description de ce domaine de connaissance. A chaque activité recensée, nous associerons plusieurs termes synonymes nous permettant de lever une partie des problèmes de polysémie et de synonymie. Et de manière analogue nous organiserons la connaissance des entreprises dans une ontologie particulière.

La recherche d'une entreprise par son activité consistera donc à rechercher, dans un premier temps, l'activité recensée la plus proche de ce que recherche l'utilisateur en utilisant les termes synonymes introduits précédemment, puis de lister les références des entreprises dont au moins une activité y est associée.

3.1 Ontologie de description d'activités géolocalisée

Pour produire une ontologie relativement complète des différentes activités d'entreprises pouvant exister, nous nous sommes appuyés sur les données recueillies auprès des organismes officiels chargés d'enregistrer les déclarations de toute entreprise française : l'INSEE. Une initiative similaire à déjà été utilisée dans Marquet et al. (2003) afin d'unifier l'accès à des ressources médicales en se référant aux terminologies standard du domaine.

3.1.1 Répertoire des activités

L'INSEE (Institut National de la Statistique et des Etudes Economiques) est chargé de maintenir le répertoire SIRENE (décret du 14 mars 1973) dont l'objectif est l'identification des établissements ainsi que des personnes physiques et morales, sur le plan national. En marge du répertoire SIRENE, l'INSEE assure, entre autres, les missions d'attribution d'un identifiant

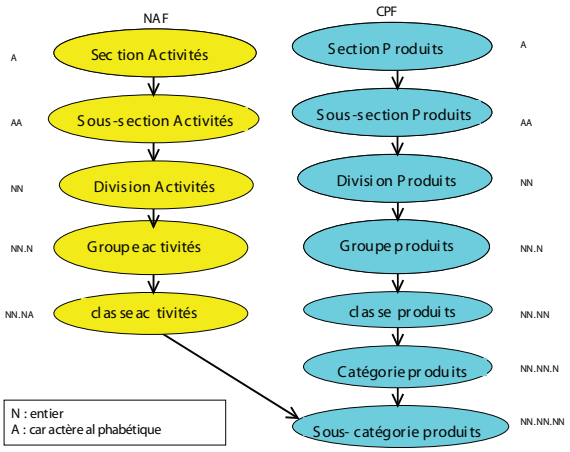

FIG. 1 – *Emboîtement NAF / CPF INSEE.*

unique, le numéro SIREN, aux personnes morales et physiques ainsi que le numéro SIRET à chacun de leurs établissements. Un code APE représentant l'Activité Principale Exercée est également attribué à chaque établissement.

Afin de faciliter l'organisation de l'information économique, nécessaire à ces différentes tâches, l'INSEE maintient un document de référence intitulé "Nomenclature française d'activités et de produits". Celui-ci est constitué par deux sources 1) la Nomenclature Française d'Activités (NAF) et 2) la Classification Française des Produits (CPF).

Ces différentes structures, constituent une source capitale, dans le cadre du présent projet. Les nomenclatures économiques présentent l'avantage d'être maintenues par des experts reconnus du domaine ce qui certifie exhaustivité et qualité du vocabulaire employé. Elles comprennent aussi bien la liste complète des activités pouvant être exercées par les entreprises que les produits développés. En outre, toute unité économique exerçant en France est rattachée à la NAF, via le code APE.

La NAF et la CPF ont été élaborées dans un cadre européen, harmonisé, afin de clarifier l'information sur le marché unique européen. Elles sont organisées sur plusieurs niveaux hiérarchiques : sections et sous-sections comme le montre la figure 1.

Les nomenclatures permettent le classement de toutes les activités économiques et de tous les produits (biens et services). Elles constituent un outil pour ordonner l'information économique mais proposent, aussi, un langage commun présentant un intérêt dans de nombreux domaines.

Le code APE représente l'activité principale exercée par l'entreprise ce qui correspond au code de la classe issue de la nomenclature française des activités. Dans l'hypothèse d'une entreprise exerçant plusieurs types d'activités, l'INSEE, par le biais d'une estimation statistique, détermine celle qui demeure prédominante. Il est important de souligner que, dans le cas d'une entreprise disposant de plusieurs établissements, chacun d'eux dispose d'un code APE.

Cette notion d'activité principale de l'entreprise est importante, notamment en droit, pour déterminer, par exemple, les champs d'application des conventions collectives.

3.1.2 Géo-localisation et description des entreprises

Les attributs caractéristiques d'une entreprise, sur le Web, se matérialisent par les pages constituant son site. En pratique, une entreprise peut être spécialisée dans plusieurs activités et plusieurs produits localisés sur plusieurs lieux. Il en résulte qu'une page donnée peut être liée à un ou plusieurs attributs du répertoire des activités. En outre, la construction d'un moteur de recherche à usage des entreprises implique que les sociétés retenues dans l'index ne possèdent pas nécessairement de site internet. Dans ce cas précis, une adresse Email de contact sera prise en compte pour référencer l'activité ou le produit concerné. On peut imaginer une entité commerciale exploitant plusieurs activités avec, pour chacune d'elles, un responsable possédant une adresse Email.

L'entité de l'entreprise prise en compte dans notre organisation suit alors l'organisation du répertoire SIRENE de l'INSEE. Ce répertoire se base sur la notion d'unité administrative dans laquelle un *établissement* est localisé géographiquement et rattaché à une *unité légale* (entité juridique déclarée aux administrations compétentes) qui est elle-même rattachée à un *groupe financier*. Le numéro de SIREN identifie alors de manière unique une *unité légale* tandis que le numéro SIRET l'*établissement* en tant qu'unité géographiquement localisée.

La structure du répertoire SIRENE par son organisation spécifique va répondre de façon satisfaisante à notre problématique de géo-localisation. En effet, la notion d'établissement, vue par l'INSEE comme une unité géographiquement localisée, via le code SIRET, nous garantit le référencement géographique, sans ambiguïté, de ce type d'unité administrative.

3.1.3 Organisation de l'ontologie

Les différentes classes constituant l'ontologie (décrite dans la figure 2) répondent, chacune d'elles, à un ou des services spécifiques mis en évidence durant la modélisation. Ceux-ci peuvent être classés à l'intérieur de quatre catégories principales : 1) la classification des produits, 2) la nomenclature des activités, 3) l'organisation interne de l'entreprise et 4) l'organisation géographique. Si les deux premières catégories s'imposent de façon triviale, car issues directement du modèle INSEE, les deux suivantes ont été constituées pour mieux structurer l'ontologie définitive. Ces quatre ensembles peuvent être considérés comme des sous-ontologies que nous utiliserons respectivement dans le but de

1. rechercher les concepts liés à un mot clé ; généraliser, spécialiser un concept lié à une activité ou un produit ; récupérer les produits associés à une activité particulière ; sélectionner les pages Web, Email d'une activité, d'un produit,
2. chercher les établissements rattachés à une page Web, un Email,
3. accéder à la société d'un établissement donné ; retrouver les établissements d'une société ; localiser géographiquement un établissement particulier,
4. déterminer les établissements situés dans une zone géographique.

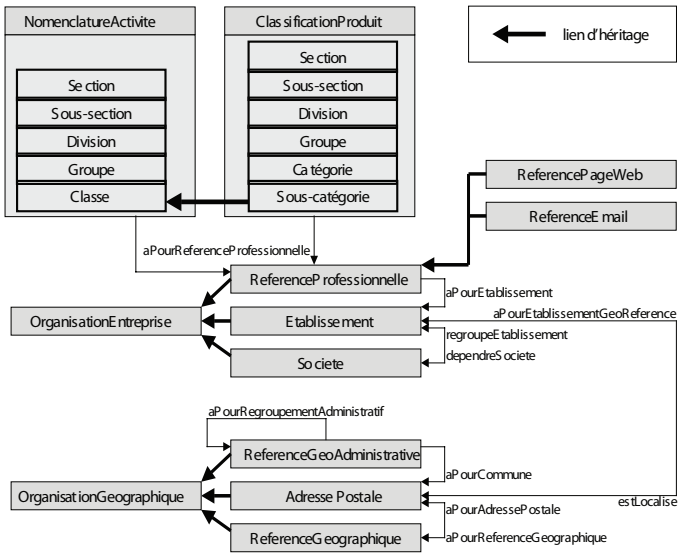

FIG. 2 – *Organisation générale de l'ontologie.*

3.2 Architecture du moteur de recherche

L'architecture retenue en vue de la réalisation du démonstrateur envisagé s'appuiera sur la plate-forme Sesame Broekstra (2005) associée à une base de données MySql. Cette solution présente l'avantage d'être robuste en terme de volumétrie ce qui convient pour la partie moteur de recherche. Le langage de requêtage exploité sera RDQL qui permet de traiter la représentation de l'ontologie comme une base de données. Il autorise l'exécution de requêtes complexes, utilisant des modèles et des contraintes sur les triplets RDF, tout en permettant des jointures.

Pour permettre à l'utilisateur de retrouver facilement l'information qu'il recherche, notre système doit être capable 1) de récupérer rapidement les concepts liés aux mots clés de l'utilisateur, 2) d'offrir un module de désambiguïsation sémantique, 3) de permettre la navigation dans l'ontologie et 4) d'afficher des résultats sous la forme définitive (raison sociale, descriptif entreprise, référence professionnelle).

Le système propose à l'utilisateur de saisir sa requête à l'aide de 2 champs : le *Quoi*, obligatoire permet de préciser les mots clés d'activité à rechercher, et le *Où*, facultatif, définissant la limitation géographique concernée. Les mots-clés saisis par l'utilisateur feront l'objet d'une désambiguïsation sémantique. La composante géographique sera limitée à la saisie d'un ou plusieurs codes postaux.

La récupération des concepts, à partir des mots clés saisis par l'utilisateur, demeure une phase fondamentale du processus. Celle-ci, outre le fait de rechercher les activités et produits rattachés à une liste de mots clés, se charge de supprimer, de factoriser et de définir un ordre de pertinence sur les concepts.

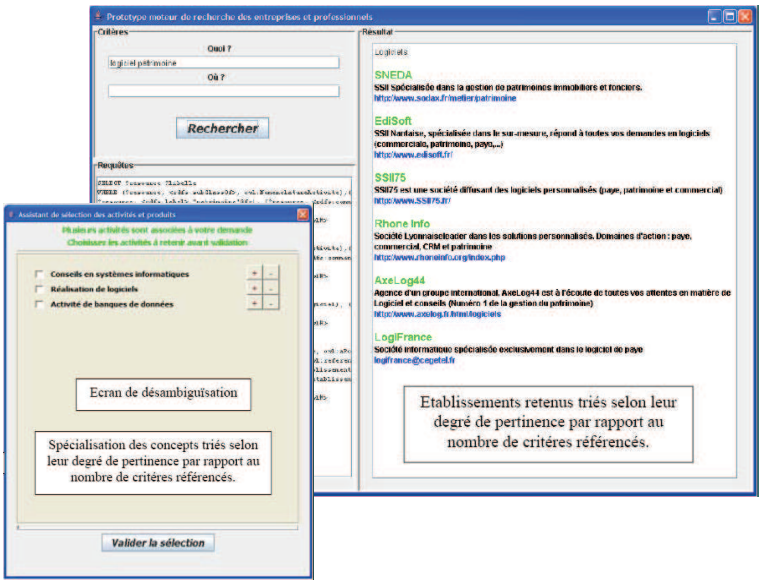

FIG. 3 – *Interface de l'application présentant l'écran de sélection des concepts pour la désambiguïsation.*

Suppression des concepts généraux Ce traitement consiste à ne considérer que le concept le plus spécialisé en cas de concurrence sur une même branche de l'arbre de concepts. Cela permet préciser au mieux la requête de l'utilisateur en éliminant les termes généraux.

Factorisation des concepts La factorisation a pour objectif de ne sélectionner que les concepts les plus généraux en cas de concurrence sur deux branches d'un même sous-arbre. Lorsque la distance hiérarchique entre le concept général et les deux concepts fils est très proche (1 à 2 branches traversées), il peut être judicieux de résumer la pensée de l'utilisateur par le concept général.

Ordre de pertinence des concepts Pour donner un ordre de pertinence aux concepts sélectionnés, nous déterminons un poids représentatif du nombre de mots clés issus de la requête utilisateur associés à chaque concept. Ceux dont le poids est le plus important apparaissent alors en tête de la liste proposée par l'écran de désambiguïsation sémantique.

Lorsqu'un mot clé saisi par l'utilisateur correspond à plusieurs concepts dans l'ontologie, nous proposons de manière interactive un module de désambiguïsation sémantique. Il appartient alors à l'utilisateur de déterminer dans quel champ d'application il souhaite réaliser sa recherche. La liste des concepts associée sera affichée en tenant compte de leur pertinence respective (taux de mots clés associés). Une sélection multiple sera admise.

De plus, nous permettons dans ce cas à l'utilisateur de naviguer par spécialisation / généralisation dans la hiérarchie des concepts afin de préciser au mieux sa pensée. La figure 3 illustre

l'interface de notre application comprenant les champs *Quoi* et *Où*, la zone de résultat et la fenêtre de désambiguïsation.

4 Premiers résultats

4.1 Méthodologie de test

Pour valider notre méthode, nous avons entrepris de construire un jeu d'essaie réduit à une région particulière. Nous avons ensuite étudié les réponses obtenues aux différentes questions proposées dans la table 1 et les avons comparés à celles obtenus par un des moteurs de recherche d'activité géo-localisé les plus utilisés en France : les *Pages Jaunes*.

	Quoi	Activité	Vol. P	Vol. PJ
R1	Logiciel santé	Réalisation logiciels	78	112
R2	Dépannage ordinateur dimanche à domicile	Entretien réparation matériel informatique	31	37
R3	Restaurant menu ouvrier	Restauration de type traditionnel	14	19
R4	Agent Peugeot	Commerce de véhicules	2	4

TAB. 1 – *Requêtes prises en compte lors de nos tests. Ces requêtes intègrent une question (Quoi) qui porte sur une Activité localisée sur Nantes (44000). (Vol. P) représente le nombre d'instances de l'activité retournée par notre prototype. (Vol. PJ) représente le nombre d'instances de l'activité retournée par le moteur des Pages Jaunes.*

Lors de cette phase de test, nous n'avons pas mis en œuvre le module de désambigüation afin de ne pas avantager notre modèle en guidant plus précisément l'utilisateur dans sa formulation de mots clés.

4.2 Analyse des résultats

L'ensemble des résultats obtenus par notre prototype et les *Pages Jaunes* sont donnés en terme de précision/rappel dans le tableau 2. Afin de valider la pertinence du classement des résultats retournés par chaque moteur, nous avons évalué la précision et le rappel pour chaque requête en ne considérant dans un premier temps que les 5, puis les 10, puis 20, puis 30 premiers résultats.

Il est évident que dans les résultats obtenus, notre approche apporte systématiquement une meilleure précision et un meilleur rappel que les *Pages Jaunes*. Ceci est du en particulier aux opérations de suppression et de factorisation de concepts permettant de cibler d'avantage le souhait d'un utilisateur. On peut toutefois noter que l'annuaire testé fait apparaître une qualité hétérogène des réponses obtenues. Dans certains cas, les résultats pertinents du corpus sont pris en compte mais noyés dans l'ensemble des réponses. Dans d'autres exemples, certains éléments corrects ne sont pas considérés dans les résultats.

Requête	#res	Précision		Rappel	
		Prototype	PJ	Prototype	PJ
R1	5	0.80	0	1	0
	10	0.40	0	1	0
	20	0.20	0	1	0.25
	30	0.13	0.03	1	0.25
R2	5	1	0.20	0.83	0.17
	10	0.60	0.20	1	0.33
	20	0.30	0.20	1	0.33
	30	0.20	0.20	1	1
R3	5	0.20	0	1	0
	10	0.10	0	1	0
	20	0.05	0.05	1	1
R4	5	0.20	0	1	0

TAB. 2 – *Résultats comparatifs obtenus en terme de précision/rappel entre notre prototype et les Pages Jaunes (PJ) sur 4 requêtes de test en fonction du nombre de résultats pris en compte (#res).*

À travers les exemples traités, deux grandes familles de mots-clés se dégagent : les mots-clés directement rattachés à une activité ou un produit, ce qui induit une forte adhérence avec les nomenclatures de type NAF (ex : *restaurant*, *logiciel*) ; les autres mots-clés qualifiant un terme commun généraliste (ex : *agent*, *dimanche*). Il apparaît, clairement, que l'annuaire est performant tant qu'il s'agit d'interpréter des mots-clés liés à une activité ou un produit. Son architecture repose, manifestement, sur les nomenclatures de type NAF et les mots-clés associés. En revanche, l'annuaire est dans l'incapacité de traduire des mots-clés de la deuxième famille. Cette lacune se traduit par un manque de précision pouvant conduire à des excès en terme de bruit et de silence. Par opposition, le prototype, construit sur une ontologie, est capable d'interpréter tout type de mots-clés. Ceci suppose que les concepts, toutefois en partie basés sur les nomenclatures activités et produits, soit sémantiquement affectés aux mots-clés indépendamment de la famille d'appartenance.

Il est également important de souligner que le prototype permet une sélection géographique régionale ce que ne permet pas l'annuaire testé. Pour parvenir à ce résultat, il a été nécessaire d'implémenter plusieurs stratégies. Tout d'abord, il était important que la modélisation de l'ontologie tienne compte des trois caractéristiques suivantes : dualité des mots-clés traités, séparation concepts concernés / classement des résultats et relation entreprise / lieu géographique. Et il a été également essentiel de mettre à jour l'ontologie en associant les mots-clés aussi bien du côté des nomenclatures activités et produits que du côté des entreprises indexées.

5 Conclusion

Nous avons décrit dans cet article un nouveau moteur de recherche géo-localisé à usage des entreprises. La méthodologie que nous avons adoptés, issue du web sémantique, nous a permis

d'améliorer significativement l'efficacité de solutions du domaine largement répandu comme les *Pages Jaunes*, en intégrant une ontologie. La phase de modélisation de l'ontologie a mise en évidence l'intérêt des structures économiques maintenues par l'INSEE. Les nomenclatures économiques ont été retenues pour gérer la classification des activités et produits pouvant être dispensés par les entreprises tandis que la structure des unités administratives, telle que gérée au sein du fichier SIRENE, s'est avérée judicieuse pour réponse à la problématique de géo-localisation des entreprises.

Nous avons élaboré un démonstrateur mettant en œuvre différentes stratégies se servant de l'ontologie afin de guider l'utilisateur dans la formulation de sa requête. Nous envisageons de poursuivre dans ce sens en expérimentant notre prototype sur des bases de données plus importantes et en menant des études afin de déterminer plus précisément l'amélioration qu'apporte ce système de *feed-back*. Actuellement, la base de travail est alimentée manuellement et il serait également intéressant de profiter de la connaissance présente dans l'ontologie afin d'introduire une indexation automatique plus pertinente que celle réalisée par les moteurs de recherche classiques. Le prototype réalisé dans ce papier sert de base à l'élaboration d'un outil de recherche géo-localisé plus complet (*Géoternet*) développé par la société IP&moteur.

Références

Baeza-Yates, R. A. et B. A. Ribeiro-Neto (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Brin, S. et L. Page (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems 30*(1–7), 107–117.

Broekstra, J. (2005). *Storage, Queyring and Inferencing for Semantic Web Languages*. Ph. D. thesis, Université d'Amsterdam.

Burners-Lee, T., J. Hendler, et O. Lassila (2001). The semantic web. *Scientific American 284*(5).

Davies, J., F. van Harmelen, et D. Fensel (Eds.) (2002). *Towards the Semantic Web : Ontology-driven Knowledge Management*. New York, NY, USA : John Wiley & Sons, Inc.

de Loupy, C. (2000). *Évaluation de l'apport de connaissances linguistiques en désambiguïsation sémantique et recherche documentaire*. Ph. D. thesis, Université d'Avignon et des Pays de Vaucluse.

de Loupy, C. et E. Crestan (2004). *Systèmes de recherche d'information*, Chapter Traitement automatique des langues et systèmes de recherche d'information, pp. 139–158. Éditions Hermès.

Ding, L., T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi, et J. Sachs (2004). Swoogle : A Search and Metadata Engine for the Semantic Web. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*. ACM Press.

Ding, L., T. Finin, A. Joshi, Y. Peng, R. Pan, et P. Reddivari (2005). Search on the Semantic Web. *IEEE Computer 10*(38), 62–69. A longer version of this paper is available.

Gao, M., C. Liu, et F. Chen (2005). An ontology search engine based on semantic analysis. In *ICITA '05 : Proceedings of the Third International Conference on Information Techno-*

logy and Applications (ICITA'05) Volume 2, Washington, DC, USA, pp. 256–259. IEEE Computer Society.

Gruber, T. R. (1993). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino et R. Poli (Eds.), *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands. Kluwer Academic Publishers.

Guha, R., R. McCool, et E. Miller (2003). Semantic search. In *WWW '03 : Proceedings of the 12th international conference on World Wide Web*, New York, NY, USA, pp. 700–709. ACM Press.

Labrou, Y. K. et T. Finin (1999). Yahoo as an Ontology - using Yahoo categories to describe documents,. In *Proceedings of the 1999 ACM Conference on Information and Knowledge Management (CIKM'99)*.

Lawrence, S. et C. L. Giles (1999). Accessibility of information on the web. *Nature 400*(6740), 107–109.

Marquet, G., C. Golbreich, et A. Burgun (2003). From an ontology-based search engine towards a more flexible integration for medical and biological information. In *Proceedings of the Semantic Integration Workshop*, Volume 82.

Mayfield, J. et T. Finin (2003). Information retrieval on the Semantic Web : Integrating inference and retrieval. In *Proceedings of the SIGIR Workshop on the Semantic Web*.

Shah, U., T. Finin, et A. Joshi (2002). Information retrieval on the semantic web. In *CIKM '02 : Proceedings of the eleventh international conference on Information and knowledge management*, New York, NY, USA, pp. 461–468. ACM Press.

Welty, C. A. et N. Guarino (2001). Supporting ontological analysis of taxonomic relationships. *Data Knowledge Engineering 39*(1), 51–74.

Summary

The search for a company on the Web, relating to a particular ability, is not an easy task. The tools, placed at the disposal of the Net surfer, do not give complete satisfaction. On a side the search engines have some difficulties of emphasizing clearly the result. Other side, the specialized directories (like Yellow Pages) are tributary of a fixed organization, harming their effectiveness.

In this paper, we propose to create a new search engine, associating semantic Web and geolocalization. This innovative approach requires the implementation of an ontology having for objective the formalization of knowledge of the domain.

This task highlighted the interest of the economic structures, maintained by INSEE, and their use within ontology. The economic nomenclatures were retained to manage the classification of the activities and products produced by the companies. The structure of the administrative units, as managed within the file 'SIREN', proved to be judicious to answer the problems of geolocalization of the companies. We associate the keywords and synonymous corresponding to the activity concerned with each node of the ontology.

Lastly, we compare the results obtained by our engine with those obtained by the main search engine of geolocalised activities in France: the Yellow Pages. Our engine obtains significantly better results on both precision and recall.