Construction et usage d'une ontologie de compétences pour l'identification de réseaux collaboratifs d'entreprises

Kafil Hajlaoui, Xavier Boucher, Michel Beigbeder

Construction et usage d'une ontologie de compétences pour l'identification de réseaux collaboratifs d'entreprises

Kafil Hajlaoui — Xavier Boucher — Michel Beigbeder

Ecole Nationale Supérieure des Mines de Saint Etienne
158 cours Fauriel
42023 Saint Etienne
Centre G2I
{hajlaoui, boucher, mbeig}@emse.fr

RÉSUMÉ. *La finalité de ce travail de recherche est de contribuer à la mise en œuvre d'aides à la décision pour la constitution de réseaux collaboratifs d'entreprises. Notre contribution se focalise sur les mécanismes d'extraction d'information utilisés préalablement aux outils décisionnels. Comme hypothèse de base, nous travaillons sur un univers ouvert de partenaires, constitué par le réseau internet. Les travaux développés dans ce papier visent à extraire des informations clés sur les compétences des partenaires potentiels, en vue d'analyser ultérieurement la similarité entre les compétences d'entreprises. Cette similarité de compétences est l'un des paramètres à prendre en compte dans la recherche de réseaux collaboratifs. L'objectif de l'article est de mettre en évidence la méthode de construction d'une ontologie des compétences adoptée pour ce travail, puis de proposer un formalisme basé sur les "patterns" destiné à utiliser de cette ontologie en vue d'extraire des traces de compétences d'entreprises.*

ABSTRACT. *The research goal is to contribute to the implementation of decision-making aids for the constitution of networks collaborative companies. This contribution is focused on the mechanisms for extraction information adapted to decision-making tools. As a basic assumption, we work on an open universe of partners, composed by Internet network. The work developed in this paper aims at extracting key information on competences of the potential partners, in order to analyze later on the similarity between company's competences. This similarity of competences is one of the principal parameters to establishing collaborative networks. The objective of the article is to highlight the construction method of competence ontology adopted for this work, then to propose a formalism based on the "patterns" intended to use this ontology in order to extract competence traces of companies.*

MOTS-CLÉS : *Extraction d'information, Ontologie, Compétence, Réseaux d'entreprises*

KEYWORDS: *Information extraction, Ontology, Competence, Enterprise networks*

1. Introduction

Dans cet article nous présentons des travaux de recherche visant à mettre au point des techniques de recherche d'information, susceptibles d'apporter un support efficace au développement de coopération inter-entreprises. Ces recherches s'inscrivent dans la perspective de fournir des méthodes de traitement de l'information et des outils d'aide à la décision destinés à améliorer les processus décisionnels de création et de mise en place de réseaux collaboratifs d'entreprises. Nous nous situons ainsi dans un contexte de système d'information décisionnel : la contribution développée dans cet article se positionne dans le champ des mécanismes d'extraction d'information, destinés à fournir des informations clés sur des partenaires potentiels de réseaux d'entreprises, informations nécessaires afin de déployer certaines aides décisionnelles.

Différentes approches d'extraction d'information ont été mobilisées au sein de travaux antérieurs, en vue de mettre en évidence des informations caractérisant des réseaux collaboratifs (Camarinha-Matos, 2003) (Plisson, 2007) (Ermilova, 2005). Ces approches de recherche et d'extraction d'information ont la vocation de devenir la pierre angulaire de systèmes décisionnels support de la gestion dynamique des cycles de vie de ces organisations collaboratifs. D'un point de vue recherche d'information, ont peut distinguer 2 grandes catégories d'univers informationnels :

- Une recherche d'information en environnement fermé, c'est-à-dire dans un contexte où un nombre délimité des partenaires potentiels développe un niveau de confiance mutuel suffisant pour partager des informations internes sus un format et une structure de données homogènes. Dans ce cas l'information à traiter s'avère semi-structuré et les frontières de l'ensemble des partenaires potentiel sont délimitées.
- Une recherche d'information en environnement ouvert, où les entreprises ne se connaissent pas a priori et utilise une information hétérogène, faiblement structurée et éventuellement publique (par exemple via le web). De plus ce type de contexte correspond souvent à des situations où le processus collaboratif en lui-même, et le type de partenaires potentiels reste très ouvert. Dans ce cas, la recherche d'information 'avère plus délicate, en raison de documents et d'informations non standardisés et mal structurés.

Le travail présenté se situe dans ce second contexte, puisque nous allons nous intéresser à la recherche d'information au sein de sources publiques fournies par le web. Si l'on se positionne par rapport au cycle de vie des organisations collaboratives, notre contribution ne se situe pas sur l'étape de création d'une entreprise virtuelle pour répondre à une opportunité spécifique du marché, mais au contraire dans une phase préalable du cycle de vie consistant à créer un réseau de partenariat à long terme (communément appelé Virtual Breeding Environment, VBE) préalable au lacement d'entreprises virtuelles plus ciblées sur le court terme.

L'aide à la décision recherchée au final consiste donc à créer des réseaux de partenariats à long terme, de type VBE. Dans étape de création d'un réseau collaboratif, nous considérons par hypothèse que les partenaires ne sont pas connus par avance, et qu'il n'existe pas de relation de confiance préalable, qui pourrait favoriser le partage d'information. Nous nous situons dans un environnement ouvert où le nombre de partenaires potentiels à analyser est très large, ce qui renforce deux orientations de nos travaux : l'utilisation de l'information publique disponible sur le web et la volonté de développer des mécanismes d'extraction d'information automatisés.

Les travaux scientifiques en recherche et extraction d'information ont conduit dans le passé à de nombreux résultats opérationnels. Cependant, dans un contexte d'information disponible sur le web, directement produite par des acteurs métiers selon leur sensibilité propre, sans standardisation ni pré-structuration lié au besoin décisionnel final, les mécanismes de recherche et extraction d'information restent encore bien souvent peu performants. L'enjeu de ces

K. Hajalaoui, X.Boucher, M. Beigbeder, « Construction et usage d'une ontologie de compétences pour l'identification de réseaux collaboratifs d'entreprises », Ingénierie des Système d'Information, Numéro spécial « Ingénierie d'Entreprise et de Systèmes d'Information de la revue ISI », sous la coordination Selmin Nurcan, Khalid Benali, Hervé Pingaud, vol. 15, numéro 4, juillet-août 2010.

recherches consiste également à tirer partie des spécificités des métiers concernés, afin d'améliorer la performance finale des systèmes de recherche ou d'extraction d'information. Dans le cadre de cet article, nous nous situerons dans le contexte métier correspondant au domaine d'activité de l'industrie mécanique. La spécificité de ce contexte sera utilisée (sous forme de ressources sémantiques propres au métier comme nous le verrons ci-dessous) en vue d'améliorer les performances des approches proposées ci-après.

Cet article est constitué de 5 sections centrales, complétées par l'introduction et la conclusion. La section 2 fournit des éléments d'état de l'art sur le domaine de l'extraction d'information, dans un contexte informationnel correspondant à nos hypothèses. Nous mettrons l'accent sur l'usage des ontologies et des patrons, en lien aux développements proposés ultérieurement. La section 3 présente une vue globale du dispositif d'extraction d'information proposé dans ces recherches. Cette synthèse permet de positionner une contribution spécifique concernant l'extraction d'information sur les compétences d'entreprises, qui constitue le cœur de l'article. Ainsi, la section 4 explicite les besoins d'extraction d'information sur les compétences, et l'architecture générale du système d'extraction proposé. Ce système s'appuie sur une ontologie, explicitée en section 5. L'utilisation de cette ontologie pour extraire des informations sur les compétences d'entreprise est développée dans la section 6. La section 7 vient conclure l'ensemble de ce travail.

2. Besoin d'extraction d'information

2.1 Recherche d'information et Extraction d'information

L'extraction d'information consiste à remplir automatiquement une banque de données à partir de textes écrits en langue naturelle (Pazienza, 1997). Il ne s'agit pas de donner un texte brut à l'utilisateur, mais de l'analyser et d'apporter des éléments précis aux questions de l'utilisateur par le remplissage d'un formulaire ou d'une base de données. Ainsi, ce n'est pas une question de retourner le texte approximatif à l'utilisateur, mais de relier les éléments, pour construire l'information complète et structurée. Cependant, l'extraction d'information s'oppose à la recherche documentaire (recherche d'information), qui vise à retourner à partir d'une base de document, un ensemble pertinent au regard d'un besoin d'information formulé sous une requête (Salton, 1983) (Voorhees, 1999). Pour une première phase de recherche, nous avons contribué à une approche de recherche basée sur des techniques de recherche d'information (Hajlaoui, 2008) (Hajlaoui, 2009a). Cette approche se base sur l'idée qu'il existe un rapport entre le contenu véhiculé par un texte et les mots utilisés dans un texte. Que ce rapport est en fonction de la fréquence d'usage des mots et qu'il existe une relation entre la capacité d'un mot à être choisi comme terme d'indexation et sa fréquence d'emploi. Cependant, les techniques de recherche d'information ne répondent pas à l'ensemble des besoins d'extraction en vue de fournir des aides à la décision pertinente pour les décideurs. Typiquement, cet article se focalise sur l'extraction de compétences d'entreprises, pour lequel nous allons employée des techniques linguistiques abordée en 2.2.

2.2 L'utilisation des ontologies et des patrons pour l'extraction d'information

La notion de compétence est une notion pluridisciplinaire abordée selon différents points de vue. Selon le but de l'étude, et selon une analyse issue de la sociologie, de l'économie industrielle, ou du management des organisations, la définition et la caractérisation de la notion de compétence pourront être distinctes. Cette complexité de la notion de compétence, rend difficile la mise au point des mécanismes d'extraction consistant à détecter une information spécifique à partir de fragments de texte (mot, expression, phrase). Dans le cadre de nos recherches, nous n'avons pas pu identifier l'existence des ressources sémantiques répondant à notre besoin de caractérisation des compétences globales et susceptibles de servir de support pour l'extraction d'information. Ce contexte, renforcé par le caractère non structuré de l'information disponible sur le web, nous a conduit à nous travailler sur des techniques émergentes permettant un traitement linguistique des textes : les Ontologie et les patrons lexico-syntaxiques.

Selon B. Bachimont, « *Une ontologie est une représentation linguistique et formelle des concepts d'un domaine pour un contexte applicatif. L'aspect linguistique renvoie au fait que les concepts sont tirés de la langue du domaine et doivent rester intelligibles pour les spécialistes. L'aspect formel renvoie au fait que les concepts doivent être manipulables par la machine et produire un comportement prédictible.* ». Plusieurs chercheurs (Gomez, 1999) (Guarino,

2002) (Fürst, 2004) ont pu démontrer que le concept d'ontologie permet d'analyser et de traiter le savoir dans un domaine en modélisant les concepts pertinents. Les ontologies, comme ressource sémantique, sont utilisées pour aider à l'exploration de corpus. Souvent l'information pertinente se présente dans le voisinage d'un concept particulier du domaine traité, ce qui nécessite une exploration conceptuelle du texte pour la localiser. L'ontologie a notamment pour rôle de valider les entités informationnelles identifiées dans le texte. Dans notre travail, compte tenu de l'absence d'ontologie répondant réellement au besoin, il a été nécessaire de construire une ontologie concernant les compétences d'entreprises.

La construction d'ontologie est un processus complexe qui nécessite la mise en place de nombreux principes et critères. Plusieurs méthodes et pratiques de développement d'ontologies ont été développées dans la littérature, sans toutefois de consensus sur la meilleure méthode à adopter (Mizoguchi, 1998) (Staab, 2001) (Psyché, 2003) Si l'on considère ici qu'une méthodologie est un ensemble de principes d'ingénierie, appliqué avec succès par un auteur au sein d'un processus rationnel de construction d'une ontologie, (Mendes, 2003) a pu dénombrer un total de trente trois méthodologies existantes. Ces méthodologies sont analysées selon le type du processus de construction : à partir du début, par intégration ou fusion avec d'autres ontologies, par re-ingénierie, par construction collaborative ou par évaluation des ontologies construites. Dans bien des cas, les corpus utilisés pour appliquer la construction d'ontologies sont des corpus relativement structurés, composés de textes bien formés du point de vue linguistique (par exemple corpus d'articles de presse). Dans notre cas le choix de la méthode d'ingénierie d'ontologie doit répondre à d'autres exigences : les textes composant le corpus ne suivent aucune structure standard ; la sémantique du vocabulaire utilisé est très liée au domaine métier (vocabulaire contextualité) ; la structure linguistique des textes est parfois absente ; l'ensemble de ces facteurs induisent de forts risques d'ambiguïté. De plus, le choix de la méthode doit prendre en compte le fait que nous ne nous appuyons sur aucune ontologie initiale. Pour répondre à ces critères, notre choix s'est fixé sur la méthode ARCHONTE (Bachimont, 2002). ARCHONTE est la méthodologie qui propose l'approche la plus structurée et la plus complète en vue de maîtriser la spécification de la sémantique des termes, ce qui est indispensable pour traiter la problématique d'ambigüité lors du processus ultérieur d'extraction.

En extraction d'information, la détection au sein d'un texte de la présence d'un concept issu d'une ontologie n'est pas une condition suffisante pour délimiter et confirmer l'information pertinente. Des phénomènes linguistiques peuvent biaiser le sens des mots et un même mot peut prendre deux sens différents selon son contexte d'utilisation. Pour lever cette ambigüité contextuelle, en complément à l'ontologie, nous aurons recours à l'utilisation de patrons linguistiques.

Les patrons linguistiques sont le résultat de la construction d'une signature contextuelle. L'utilisation de cette technique est basée sur les principes de la sémantique distributive qui admet que la signification d'un mot est fortement corrélée aux contextes dans lesquels il apparait. D'une façon plus élaborée un patron lexico-syntaxique identifie la relation recherchée plus précisément en définissant également des contraintes syntaxiques ou typographiques sur le contexte des termes (Grabar, 2004). En linguistique, les approches par patrons sont utilisées pour associer des régularités structurelles à des informations sémantiques. Hearst (Hearst, 1992) est la première à utiliser les patrons dans le contexte de l'extraction d'information. Elle a proposé des ensembles des patrons lexico-syntaxiques qui sont facilement repérables dans le texte et qui apparaissent fréquemment dont le but de reconnaître certaines relation lexicales sans ambigüités. Les patrons linguistiques sont utilisés aussi pour l'enrichissement des ontologies (Chagnoux, 2008)(Ben Mustapha, 2008). L'objectif est d'exploiter la projection des patrons sur un corpus textuel pour enrichir une ontologie existante.

Les patrons ont pour fonctions d'extraire les relations entre les concepts présents dans cette ontologie ou d'extraire des nouveaux concepts qui seront ajoutés dans cette ontologie. Pour repérer les relations sémantiques dans un texte, les patrons lexico-synyaxiques qui décrivent une expression régulière, formée de mots, de catégories grammaticales ou

sémantiques, et de symboles visant à identifier des fragments de texte sont utilisés (Aussenac, 2008) (Auger, 2008) (Buitelaar, 2005). Ces fragments de texte identifiés caractérisent des formes linguistiques dont l'interprétation est relativement stable et correspond à une relation sémantique entre termes. L'identification est basée sur l'analyse des étiquettes morpho-syntaxiques ou sémantiques attribués aux mots lors d'une phase préalable de prétraitement du texte (Lemmatisation, analyse syntaxique…).

3. Présentation générale de l'approche d'extraction d'information adoptée

Par hypothèse, la recherche d'information que nous effectuons intervient dans un environnement ouvert où les organisations ne se connaissent pas et ont une information hétérogène publique et non restreinte (Hajlaoui, 2008) : en l'occurrence nous utilisons les sites web d'entreprises comme seule source d'information. Des travaux antérieurs au sein de notre équipe ont proposé une typologie des modes de coordination entre les différentes entreprises du réseau (Benali, 2005). Cette typologie est basée sur deux paramètres : la complémentarité des activités et la similarité des compétences. En référence à la théorie économique de la coordination inter-entreprise développée par (Richardson, 1972), ainsi qu'à d'autres travaux récents sur la coordination, ces deux paramètres ont été identifiés comme étant discriminants pour justifier le choix d'un type de coopération industrielle (Burlat et Benali, 2007). C'est pourquoi, pour l'ensemble de ce projet de recherche, nous focalisons notre objectif d'extraction d'information sur ces deux paramètres. Nos travaux visent à mettre au point deux mécanismes d'extraction d'information distincts, l'un orienté sur l'identification des secteurs d'activités d'entreprises et l'autre orienté sur le repérage des compétences d'entreprises (Figure 1). Cette extraction d'information doit permettre l'application ultérieure de méthodes d'aide à la décision pour la construction de réseaux collaboratifs.

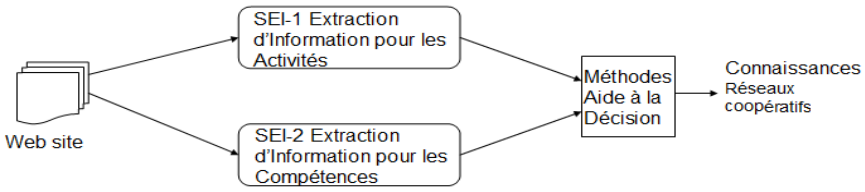

Figure 1. *Deux systèmes d'extraction d'information pour les entreprises*

Le système d'extraction des « secteurs d'activités d'entreprises » (SEI-1) a déjà fait l'objet de publications (Hajlaoui, 2008), (Hajlaoui et al., 2009). Nous nous limiterons donc à une brève synthèse de ces résultats dans le cadre de la section 3. Le reste du papier sera consacrée à l'extraction d'information sur les compétences d'entreprises (SEI-2) qui utilise des techniques plus avancées de recherche et d'extraction d'information.

3.1. *Identification des activités d' entreprises*

3.1.1. Procédure de recherche d'information, basée sur l'indexation contrôlée

L'objectif de la procédure de recherche d'information est d'extraire une information pertinente caractérisant les secteurs d'activités d'une entreprise à partir de son site web. Nous nous situons dans un contexte de recherche d'information caractérisé par une information non standardisée, non pré-formatée et peu structurée. En revanche, le contexte métier ciblé présente l'avantage de pouvoir fournir des ressources sémantiques externes (précisées ci-après) susceptibles d'améliorer les performances de l'extraction. Dans cette perspective, compte tenu de l'existence de ces ressources sémantiques, nous avons pu tester l'adaptation des méthodes classiques d'extraction d'information. Notre méthode se compose de quatre étapes habituelles: extraction, lemmatisation, indexation et mesure de similarité sémantique. Pour tester cette approche (Figure 2), nous avons construit un corpus de données formé d'une centaine de sites web d'entreprises, exerçant dans le même domaine industriel : la mécanique.

Le fait de cibler un domaine métier bien spécifique comme la mécanique nous permet de chercher des ressources sémantiques susceptibles de le caractériser. Tel que nous l'avons d'ores et déjà justifié dans (Hajlaoui et al., 2008), nous avons sélectionné comme ressource sémantique le standard national Code NAF (Nomenclature des Activités Françaises), en limitant son utilisation au domaine industriel de la mécanique. Le code NAF nous fournit une représentation conceptuelle hiérarchisée de tous les secteurs d'activités de ce domaine industriel : c'est une structure hiérarchique de classes et sous-classes de secteurs d'activités. Ce code NAF va être utilisé comme ressource sémantique externe, afin d'améliorer l'expressivité du besoin d'information avant de le soumettre au système de recherche d'information. Après avoir été soumis à une lemmatisation et à une élimination des mots vides, le NAF nous a permis de constitué un Vocabulaire Contrôlé Hiérarchique (VCH), c'est à dire un ensemble de termes utilisés ultérieurement pour réaliser une indexation contrôlée.

La figure 2 synthétise brièvement l'ensemble des étapes de cette procédure d'extraction d'information. L'objectif final consiste à identifier un ou plusieurs codes NAF caractéristiques d'une entreprise, à partir des informations disponibles sur son site-web. Nous procédons par un appariement (mesure de similarité) entre des vecteurs d'informations distincts. D'une part, nous définissons un ensemble de « Vecteurs documents », chacun caractérisant une classe ou bien sous-classe du code NAF. D'autre part nous caractérisons chaque site-web d'entreprise par un vecteur d'information nommé « Vecteur requête ». Des mesures de similarité, entre le vecteur requête et les vecteurs documents vont permettre d'identifier les codes NAF probables des entreprises.

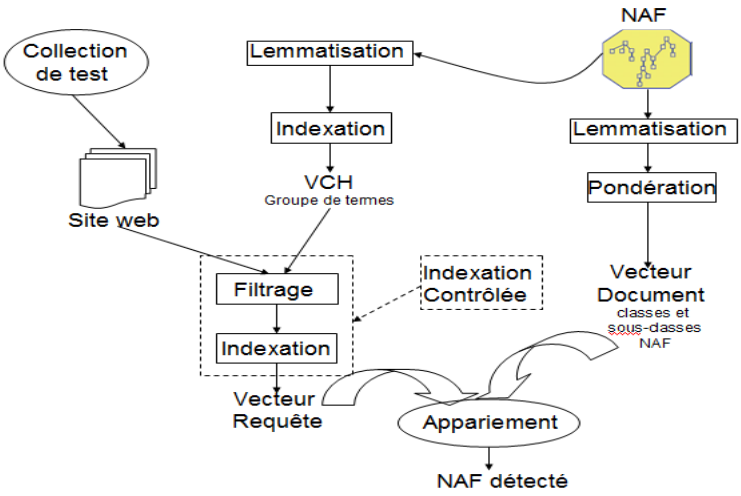

Figure 2. Approche de détection des activités d'entreprises SEI-1

Les étapes les plus importantes pour la performance finale de ce système d'extraction d'information sont les étapes d'indexation et d'appariement.

L'étape d'indexation consiste à analyser un document afin de produire un ensemble de mots clés, appelées aussi « descripteurs », susceptibles d'être utilisés dans le processus de recherche d'informations. Afin d'améliorer les performances du système, en tirant partie des informations sémantiques propres au domaine métier considéré, nous avons déployé une indexation contrôlée, en utilisant le VCH constitué à partir du code NAF. L'utilisation du VCH pour réaliser

une indexation contrôlée a pour objectif de pénaliser les termes porteurs d'ambiguïté et qui ont un impact direct sur la performance du système. La deuxième raison de l'utilisation du VCH est l'exploitation de la force informationnelle et représentative du code NAF, pour améliorer l'exploration du contenu des sites web des entreprises. Pour l'implémentation, nous avons utilisé le système d'indexation SMART[1] (System for the Mechanical Analysis and Retrieval of Text) aussi appelé (Salton's Magic Automatic Retrieval Technique) qui est un système d'indexation pour la recherche d'information.

L'étape d'appariement, quant à elle, constitue une phase de calcul de similarité entre « vecteur requête » et « vecteur document ». Elle se déroule en deux étapes : dans un premier temps on cherche à se positionner sur la classe du NAF la plus pertinente pour l'entreprise ; dans un deuxième temps on explore les sous-classes de cette classe pour se positionner de nouveau sur une sous-classe. En vue d'optimiser la performance finale, nos travaux nous ont conduits à développer d'une part des mesures de similarité utilisant trois fonctions traditionnelles de la recherche d'information (Produit scalaire, Cosinus et Jaccard) et d'autre part une mesure de similarité basée sur un modèle connexionniste (mise en place d'un réseau de neurones pour le calcul de l'appariement (Hajlaoui, 2009)). Nous nous limitons ci-dessous à synthétiser les performances obtenues.

3.1.2. Performances obtenues sur l'identification des secteurs d'activités

L'évaluation de la performance du système est basée sur le calcul des deux indicateurs de performance « précision » et « rappel ». La précision est la capacité du système à rejeter les documents non pertinents, le rappel est la capacité du système à retrouver les documents pertinents. Notre objectif est d'augmenter la précision du système ainsi que son rappel, mais aussi éviter le plus possible d'avoir des valeurs nulles qui signifient que le système ne retrouve pas de documents pertinents. La performance du système dépend non seulement de la mesure de la similarité mais de la façon de définir l'ensemble final de documents appropriés.

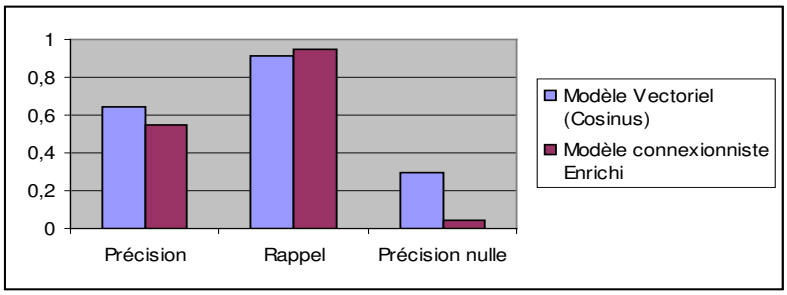

Figure 3. *Evaluation de l'identification des secteurs d'activités*

La figure 3 ci-dessus compare les performances entre la mesure de similarité basée sur la fonction cosinus (plus performante que les fonctions « Produit Scalaire » ou « Jaccard » suite à nos expérimentations), et un appariement mise en œuvre par un réseau de neurones (tel que spécifié dans (Hajlaoui, 2009)). Ces résultats mettent en évidence que la précision est légèrement meilleure pour la fonction cosinus, mais que le modèle connexionniste a permis d'améliorer le rappel et l'indicateur de précision nulle. Concernant la comparaison entre ces 2 appariements, d'autres expérimentations seraient nécessaires dans le futur pour obtenir des conclusions plus définitives. En revanche, nous pouvons d'ore et déjà confirmer que la capacité à identifier correctement un code NAF est d'ores et déjà élevée (tableau 1). Ainsi, ces techniques de recherche d'information s'avèrent efficaces, lorsqu'elles sont enrichies par l'usage d'une ressource sémantique externe spécifique au métier, du type du code NAF.

	Modèle vectoriel (cosinus)	Modèle connexionniste (réseau de neurones)
Pourcentage d'identification correcte de classes NAF	**92 %**	**88 %**

1. SMART : ftp://ftp.cs.cornell.edu/pub/smart/

Pourcentage d'identification correcte de sous-classes NAF	**76%**	**88 %**

Tableau 1. *Performances obtenues pour des deux modèles vectoriel et connexionniste*

3.1.3. Application à la construction de réseaux d'entreprises

Cette application constitue un test de faisabilité sur l'usage de l'information extraite des sites web sur les codes NAF d'entreprise. L'usage que nous testons concerne la recherche de réseaux collaboratifs inter-entreprises. Comme nous l'avons introduit plus haut, il s'agit ici de tester la possibilité d'appliquer une méthode déjà existante et développé par (Benali, 2005). En ce qui concerne l'analyse des activités d'entreprises, cette méthode préconise d'identifier un graphe de complémentarité des activités, afin d'appliquer à ce graphe un algorithme de clustering. Nous synthétisons brièvement les résultats de ces tests.

Afin de déterminer un graphe de complémentarité entre les secteurs d'activités d'entreprises considérées, nous avons utilisé les caractéristiques du code NAF. LE code NAF présente l'intérêt d'être générique, et reconnu par les acteurs du monde industriel (notamment en mécanique). Nous avons ainsi recueilli des informations complémentaires auprès d'experts du domaine de la mécanique pour évaluer un degré de complémentarité générique entre les classes ou sous-classes du code NAF. (Hajlaoui et al., 2008b). L'information recueillie auprès de ces experts a été formalisée sous forme d'une matrice de degrés de complémentarité entre les différents secteurs d'activités associés aux codes NAF (domaine mécanique). Cette Matrice peut être également représentée sous forme d'un graphe (figure 5).

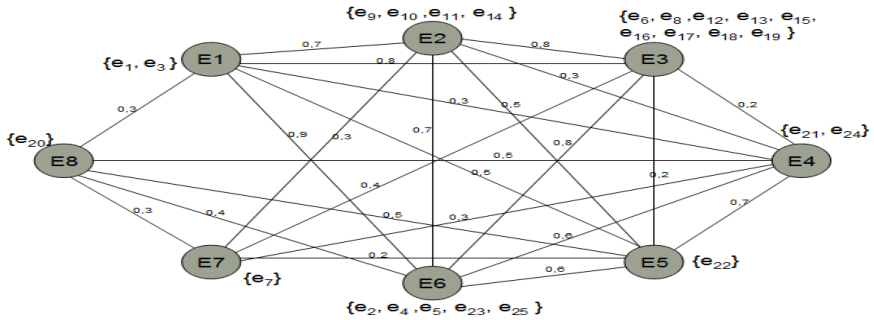

Figure 4. *Résultat du positionnement automatique des 25 entreprises sur le GCA*

Les 25 compagnies étudiées lors de l'étape d'extraction d'information (3.1.2) sont distribuées sur 8 secteurs d'activités. Une entreprise représentative est choisie pour chaque secteur. Nous réutilisons l'algorithme proposé par (Burlat et Benali, 2007). L'objectif de cet algorithme est d'isoler des sous-graphes fortement interconnectés en minimisant la perte d'information (perte d'arcs, perte de complémentarité potentielle). Ces sous-graphes représenteront les entreprises très complémentaires qui permettront de justifier d'une relation de type « réseau proactif » ou de type « firme ». L'algorithme est basé sur du partitionnement et il prend en compte plusieurs aspects spécifiques du graphe de complémentarité des activités établi par l'expert. Il prend en compte non seulement la quantité d'information perdue, mais aussi la qualité d'information : la quantité d'information c'est le nombre d'arcs éliminés et la qualité d'information

est donnée par le degré de complémentarité. L'algorithme regroupe les entreprises en petits réseaux (dénommés clusters) en éliminant le moins des arcs possibles et les moins significatifs (de poids faible). Un indicateur caractéristique des regroupements obtenus est l'indicateur I qui traduit l'intensité de coopération au sein d'un cluster.

L'algorithme de division procède par itérations successives pour déterminer les sous-groupes d'entreprises qui ont des activités complémentaires. Le nombre pertinent d'itérations peut être choisis lors de l'expérimentation en référence à deux indicateurs de qualité de la solution fournie (Hajlaoui et al., 2008b) Par exemple après six itérations les sous-groupes suivants sont obtenus. :

$$G_1 = \{E7\}; G2 = \{E8\}; G3 = \{E5, E4\}; G4 = \{E1, E2, E3, E6\}$$

L'entreprise E7 se retrouve isolée dès le début des itérations, c'est-à-dire avec une très faible intensité de coopération. Nous avons pu vérifier a posteriori la cohérence de ce résultat, compte tenu du code NAF de E7. Deux clusters G3 et G4 apparaissent. Ces 6 itérations correspondent à une intensité de coopération moyenne (I=0,54), au sein des clusters, avec une logique de coopération intra-cluster de type Réseau Proactif (Burlat et Benali, 2007).

Cette première application, met en évidence la faisabilité l'ensemble de la démarche proposée, consistant à extraire des sites web d'entreprises une information synthétique sur les secteurs d'activités d'entreprises, afin d'appliquer dans une seconde étape des outils d'aide à la décision facilitant l'émergence de réseaux collaboratifs inter-entreprises. Cependant les premiers résultats ainsi synthétisés dans la section 3 montrent les limites de n'utiliser qu'une information concernant les « secteurs d'activités » : en effet nous avons souligné que les 25 entreprises de l'expérimentation se répartissaient au final sur 8 secteurs d'activités. Cette donnée seule ne suffit pas à traiter la problématique de création des réseaux collaboratifs. Dans cette optique, la suite de l'article se focalise sur l'extraction des compétences d'entreprises.

4. Problème d'extraction des compétences d'entreprises

4.1. Identification et caractérisation des compétences d'entreprises

Un certains nombre de travaux utilisent une description « théoriquement » disponible des compétences d'entreprises, sans aborder la problématique de mise à disposition de ces informations. Cette approche a également été adoptée dans les travaux de (Burlat et Benali. 2007) auxquels nous nous sommes référés (cf. section 3). Les auteurs ont en effet proposé des méthodes et des outils d'aide à la décision pour la construction des réseaux d'entreprises basés sur la collecte et le traitement des données concernant les entreprises. Dans le cadre de l'expérimentation réalisée, ces données étaient collectées manuellement à partir d'un questionnaire rempli par les dirigeants d'entreprises. Les possibilités de systématisation de cette collecte d'information n'avaient pas été traitées jusqu'à présent.

D'autres travaux sur les techniques d'acquisition des compétences se sont intéressés à l'analyse des textes présentant des données homogènes et structurées qui décrivent le concept de la compétence dans l'entreprise ou l'organisation concernée (Camarinha-Matos et Afsarmanesh 2007) (Ermilova et Afsarmanesh, 2007) (Vanderhaegen et Loos, 2007), à partir d'une information privée directement fournie par les entreprises. Ces textes sont soit des documents qui caractérisent les compétences de l'entreprise, soit des interviews (par courriel en direct) faites avec les experts des entreprises pour décrire ces compétence. Cependant ce type de travaux est inadapté à notre contexte, où l'information structurée n'est pas rendue disponible et où, par hypothèse, nous nous limitons à utiliser une information publique non structurée.

Dans cette perspective, certains auteurs cherchent à identifier des compétences d'entreprise en inférant des règles d'expertises. (Blanchard, 2004) et (Laukkanen, 2005) emploient quelques « règles expertes » basées sur les similitudes entre la définition des compétences, comme par exemple « si les compétences C1 et C2 sont semblables, alors si un individu a acquis C1 alors il a acquis le C2 ». Un exemple réel de l'utilisation de cette règle concerne les compétences Java et C++ qui peuvent être considérées comme semblables. Et si quelqu'un a la compétence Java on peut lui associer aussi la compétence C++. (Sure. 2000) propose d'autres genres « de règles expertes » basées sur l'expérience professionnelle individuelle : un exemple de « règle » est (« si un individu a participé à plusieurs projets traitant Java, alors cet individu peut être considéré compétent en Java »). Un autre exemple est fourni par (Corby, 2004) où l'annotation sémantique est également employée sur les documents produits par l'employé. Cette technique d'identification des

compétences est essentiellement basée sur des règles faites manuellement par l'expert du domaine. Ces règles sont à couverture limitée et pour identifier la compétence dans les données disponibles (documents, dossiers et données), on doit à chaque fois reconstruire une ontologie spécifique au domaine traité.

L'ensemble de ces travaux conservent certaines faiblesses par rapport à nos objectifs. Aucune des méthodes ne présente des résultats de performance issue d'une application réelle liée à un contexte de description des compétences. Elles se basent sur l'analyse des données homogènes collectées soit manuellement, soit automatiquement. Ces méthodes exigent aux entreprises de fournir toutes les données dans une forme structurée. Dans le cadre des hypothèses de travail choisies (information publique non structurée pour la construction de réseaux longs termes), il reste nécessaire de se doter d'une méthode automatique de collecte et de traitement d'information afin d'extraire des traces synthétiques des compétences d'entreprises.

Compte tenue de la complexité de la notion de compétences, et notamment des problèmes d'ambiguïté dans l'identification des compétences, nous proposons de mobiliser des techniques d'extraction d'information basées sur l'usage d'une ontologie et des patrons. L'ontologie utilisée, ainsi que les patrons sont des résultats de cette recherche, comme souligné ci-dessous.

4.2. Système UNICOMP

Nous avons appelé UNICOMP notre système dédié à l'extraction des traces de compétences d'entreprises à partir de leur site web par contraction entre le nom de l'outil d'analyse linguistique UNItex[2] et le terme « COMPétences ». Il prend en entrée le site web de l'entreprise et utilise une ontologie générale décrivant toutes les compétences d'entreprises (La description de l'ontologie et sa méthode de création font l'objet de la section suivante). Cette ontologie est structurée sous la forme de classes conceptuelles abstraites, de classes moins abstraites et des instances de chaque classe. Dans la section suivante, nous détaillons cette ontologie ainsi que la méthode que nous avons suivie pour la créer. En sortie UNICOMP fournit la liste des classes activées dans l'ontologie qui valident chacune l'existence d'un certain type de compétence au sein de l'entreprise.

Figure 5. *Le système UNICOMP*

Pour concevoir le système UNICOMP, nous avons cherché à comprendre le comportement d'experts humains, lorsqu'ils réalisent une tâche similaire. Nous avons ainsi établi et appliqué un protocole scientifique où on fournissait à des experts des fragments de textes issus de sites web (préalablement choisis comme susceptibles de porter une information pertinente), en les mettant en situation d'identifier les classes de l'ontologie métier jugées comme caractérisant les compétences de l'entreprise concernée. Dans ce but l'ensemble des classes de l'ontologie métier était soumis aux experts. Ce protocole nous a permis de comprendre les stratégies cognitives utilisées par les experts, que nous avons cherché à reproduire ensuite dans les mécanismes de fonctionnement de UNICOMP : les experts s'appuient sur des

2. Unitex est une application qui utilise un ensemble de logiciels permettant de traiter des textes en langage naturel en utilisant des outils linguistiques. Ce logiciel est téléchargeable depuis le site: http://www-igm.univ-mlv.fr/~unitex/

termes de référence (marqueurs) décrivant la notion de compétence (par exemple : spécialiser, expérience, savoir faire…). A partir de ces termes, ils identifient des passages autours de ces marqueurs qui contiennent de l'information pertinente. Ensuite, ils interprètent ces passages pour identifier quelles sont « les classes de compétences » effectivement trouvées (activées) dans le texte.

Nous essayons dans cette approche de reproduire le comportement humain. Si la recherche des marqueurs est une opération facilement automatisable, par contre l'identification des passages et leur interprétation est plus délicate. Plusieurs phénomènes d'ambigüité ont été relevés. Cette ambiguïté est surtout liée au contexte d'utilisation du terme ou du concept. La désambiguïsation humaine faite par l'expert est reproduite dans UNICOMP par le recours à la construction de patrons linguistiques pour chaque marqueur de concept d'une compétence. Dans les sections suivantes, nous allons présenter l'ontologie que nous avons construite et les expressions (marqueurs) associées à ces classes. Puis nous expliquons la construction et l'utilisation des patrons linguistiques qui servent de déclencheur de compétence.

5. Construction d'une ontologie des traces de compétences

Pour la construction de l'ontologie des traces de compétences d'entreprises pour le domaine de la mécanique (Hajlaoui, 2009b), nous avons commencé par former un corpus textuel. Ce corpus est l'ensemble des sites web des entreprises sur lesquels nous avons effectué une étape d'extraction pour générer des textes purs. Ce corpus a été soumis à une première étape d'acquisition automatique de termes. Ces termes sont destinés à une étape de normalisation proposée par la méthodologie d'ingénierie des ontologies ARCHONTE (Bachimont, 2002). En suivant les étapes d'ARCHONTE, notre objectif est de construire une ontologie concernant "les traces de compétences d'entreprises". Une trace de compétence est une signature des ressources internes de l'entreprise composant sa compétence. Concrètement, c'est l'ensemble des expressions types qui présentent des informations sur la compétence de l'entreprise. Pour l'acquisition des termes de l'ontologie, nous n'utilisons qu'un sous-corpus textuel formé de 25 sites web à partir de la collection initiale formée de 100 sites web (Hajlaoui, 2008).

Dans notre travail, l'ontologie de traces de compétences est composée de deux parties que nous appelons dans la suite « ontologie générique» et « ontologie métier». L'ontologie générique permet de représenter et modéliser le concept de traces de compétences sous forme abstraite et générique qui reste indépendante de tout domaine métier concerné. L'ontologie métier fournit une extension de cette ontologie spécifique à un domaine métier (mécanique dans notre cas). Les classes de concept de l'ontologie générique sont détaillées en classes de concepts propres au métier ; ces classes de l'ontologie métier sont notamment destinées à regrouper des termes clés susceptibles d'être identifiés dans les sites web et serviront de support à l'extraction de traces de compétences.

5.1. L'ontologie générique

L'ontologie générique est une ontologie de modélisation de la compétence pour la modélisation de l'entreprise. Cette ontologie générique est construite suivant une approche de construction descendante qui consiste à établir un modèle général pour définir la compétence d'entreprise, ensuite ce modèle sera raffiné en sous classes conceptuelles génériques.

Notre modèle de compétences d'entreprises (Hajlaoui, 2009b) se base sur deux notions principales : une compétence émerge comme une combinaison de capacités internes. Ces capacités elles-mêmes sont le résultat de la mobilisation de différents types de ressources que possèdent l'entreprise. Pour raffiner ces deux notions, nous partitionnons les ressources en quatre types de base de ressources : ressources humaines, technologiques, informationnelles et organisationnelles ; En outre, nous distinguons deux types de base de capacités: capacités technologiques se rapportant à la création de la valeur ajoutée basée sur l'utilisation de ressources et de processus techniques, et capacités méthodologiques lié à la valeur ajoutée fournie par les méthodes de travail employées par une entreprise pour fournir ses produits ou services.

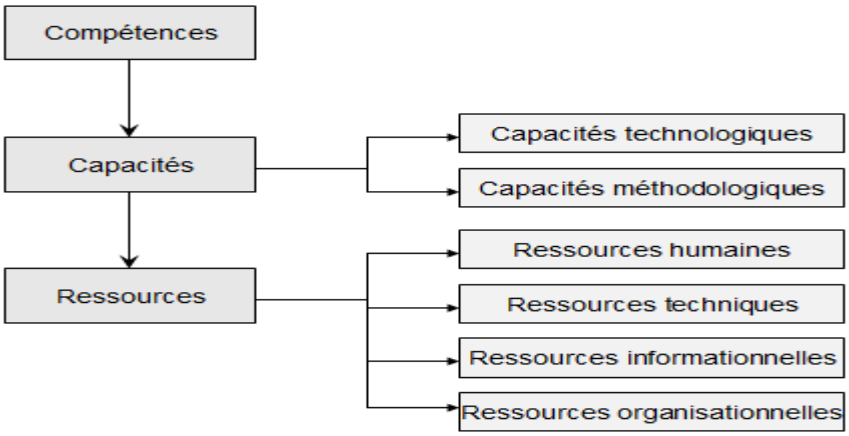

Figure 6. *Modèle de base pour la compétence d'entreprise*

Ce modèle couvre les éléments importants qui constituent le concept de compétence à la fois au niveau individuel et au niveau collectif et qui se résume en savoir, savoir faire et savoir être présents dans les deux types de capacités (technologique et méthodologique). Au même temps le concept des compétences d'entreprises est représenté en vue interne comme en vue externe à savoir tous les facteurs de capacité de collaboration avec des organisations externes qui se génère dans notre modèle par les capacités méthodologiques.

Ce modèle sera raffiné par l'ontologie générique puis, dans chaque domaine spécifique, par l'ontologie métier. Afin de décrire l'ensemble des éléments contribuant à la compétence d'entreprise. On décrit ainsi les compétences reliées au savoir, savoir-faire de la technologie : équipements, procédés de production, ressources techniques… Les capacités méthodologiques regroupent les compétences reliées à l'acteur (individu ou une groupe d'individus) et qui reflètent leurs connaissances, leur savoir et savoir faire, leur expertise et qualification. Les capacités méthodologiques recouvrent une double vision ; en interne de l'entreprise pour modéliser son savoir organisationnel du travail et en vue externe pour modéliser sa capacité de réactivité, d'écoute du client et de d'adaptation à ses besoins.

Ce modèle de compétence à un niveau très abstrait est notre point de départ pour construire notre ontologie.

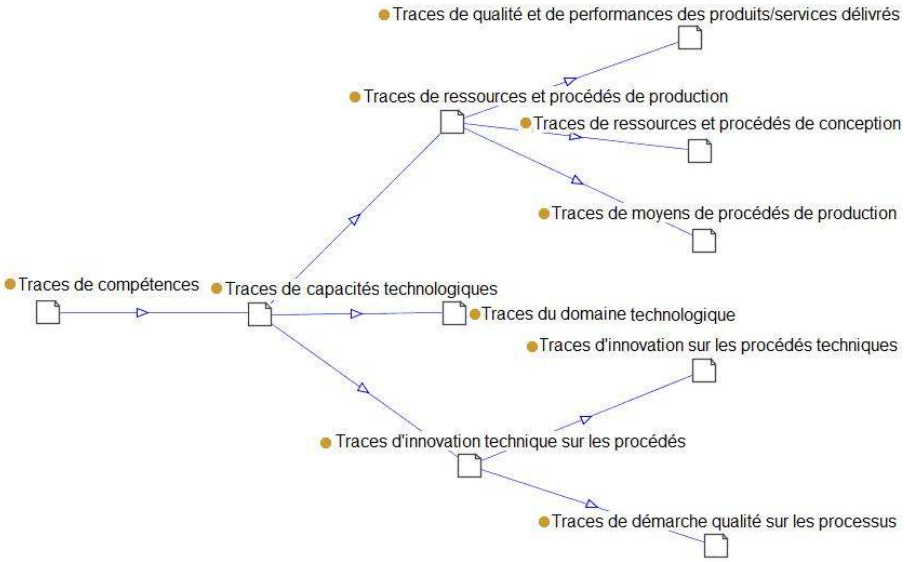

Figure 7. *Extrait de l'ontologie générique*

L'ontologie générique (figure 8) est composée de deux niveaux : un premier niveau qui manipule les concepts abstraits (traces de capacités techniques) et un deuxième niveau composée des concepts structurants (traces de ressources et processus techniques, traces du domaine technologique etc.) décrivant plus en détail les potentiels du modèle de compétence (figure 7) sous forme de classes conceptuelles génériques. Elle est construite selon une approche descendante en partant du modèle de la compétence de l'entreprise tout en spécifiant à chaque fois les classes conceptuelles générées. Une classe conceptuelle peut être définie comme une entité qui regroupe tous les caractéristiques sémantiques liées à une idée dans un domaine précis (dans notre cas compétence des entreprises). Cette idée est exprimée en fonction d'un terme ou d'une expression. Durant la construction de cette ontologie, nous nous sommes confrontés à plusieurs choix conceptuelles possibles qui peuvent donner les mêmes instances mais selon plusieurs points de vue. Cette problématique de diversification et de choix est sollicitée à chaque fois par une nécessité pragmatique qui se résume dans la nature, la qualité et la quantité de l'information manipulée et présentée dans notre corpus (site web de l'entreprise).

5.2. L'ontologie métier

L'ontologie métier permet la description et la classification des concepts moins abstraits. Ces concepts sont dépendants du domaine métier de l'entreprise. L'ontologie métier est un ensemble de classes de concepts permettant de regrouper des marqueurs concrets uilisés pour exprimer une compétence (technique. individuel…) de l'entreprise. Un marqueur est un terme ou une expression qui permet d'introduire (déclencher) une idée liée au domaine de connaissance (compétence d'entreprise) et signaler la présence d'une compétence dans le corpus étudié. Par exemple les termes marqueurs *outils, outillage …* déclarent la présence d'une compétence technique. Cette ontologie est considérée comme inter- changeable d'un domaine à un autre, afin d'assurer la qualité de la trace de compétence extraite qui dépend fortement du métier. Par exemple les classes de marqueurs (types de concepts) du domaine mécanique (outils. outillage …) ne sont pas les mêmes que celles du domaine de l'informatique (système d'exploitation. programmation. base de données…). Dans la suite, nous présentons un extrait de notre ontologie métier construite pour le domaine de la mécanique :

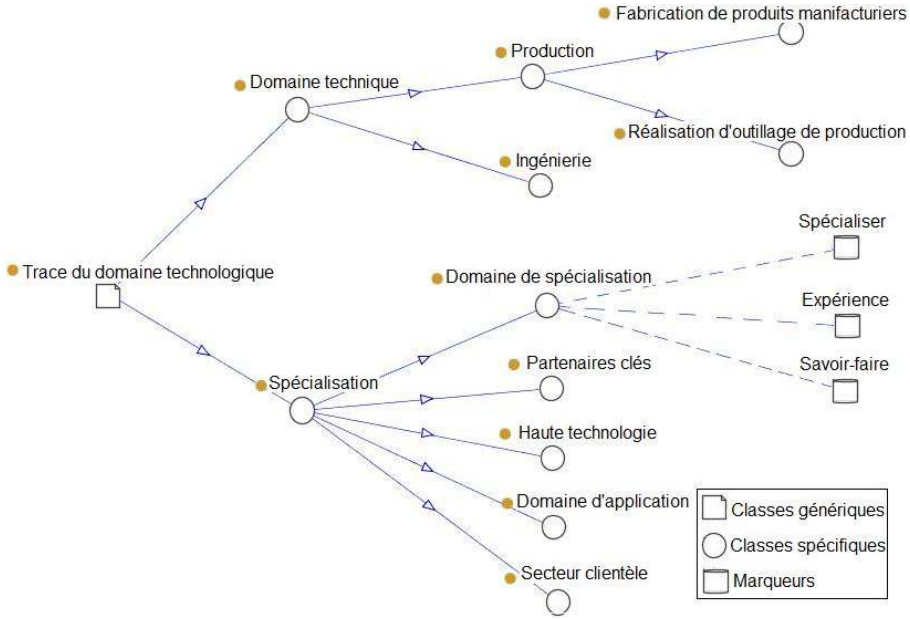

Figure 8. *Extrait de l'ontologie métier*

Les classes de l'ontologie métier dérivent toutes d'une classe de niveau plus abstrait qui appartient à l'ontologie générique. On trouve par exemple la classe technologie dont dérivent trois autres classes (usinage, traitement de surface, assemblage). Nous remarquons qu'il ne s'agit pas d'une classification des compétences d'une entreprise dans le domaine de la mécanique mais plutôt une classification des termes et marqueurs qui génèrent un concept de compétence. Sur les sites web des entreprises nous pouvons croiser par exemple les phrases suivantes:

Phrase 1 : *« Nous sommes spécialisés dans l'usinage haute vitesse »*

Phrase 2 : *« Nous utilisons la technologie laser »*

Dans la première phrase le marqueur *« usinage »* introduit une compétence dans le domaine technologique qui est la *« haute vitesse »*. Dans la deuxième phrase c'est le marqueur *« technologie »* qui déclenche la compétence *« laser »*. Ces deux marqueurs sont les briques d'information qu'on cherche à retrouver dans le texte du site web de l'entreprise. Rappelons que le but final n'est pas l'identification détaillée d'une carte de compétences de l'entreprise, mais plutôt l'identification d'une trace de compétences pour extraire une information synthétique qui est la similarité entre deux entreprises en termes de compétence. Cette trace de compétences est identifiée à partir de la double communication Ontologie générique-Ontologie métier et Ontologie métier-Corpus.

6. Extraction des traces de compétences via l'usage des patrons

6.1. Les patrons linguistiques

Les patrons linguistiques consistent à observer les possibilités de schématiser le contexte lexical et syntaxique d'un fragment de texte. Cette schématisation constitue un patron lexico-syntaxique et permet d'extraire des coupoles de mots à partir d'un corpus textuel (Hearst, 1992). En linguistique, les approches par patrons sont utilisées pour associer des régularités structurelles à des informations sémantiques. C'est Hearst qui est la première à utiliser cette méthode dans le contexte de l'extraction d'information. Elle a proposé des ensembles des patrons lexico-syntaxiques qui sont facilement repérables dans le texte et qui apparaissent fréquemment dont le but de reconnaitre certaines relations lexicales sans ambigüités. Hearst montre à partir de l'exemple de la phrase : « *The bow lute, such as the Bambara ndang, is plucked […]* » sans savoir ce que sont un *Bambara ndang* et un *bow lute*, le lecteur est capable d'indiquer qu'un *Bambara ndang* est une sorte de *bow lute*. Dans cette phrase la relation d'hyponymie peut être reconnue par le patron suivant : « un terme suivi par such as et un autre terme ». Elle est abstraite au sein du patron suivant :

X such as Y, ou X et Y sont des syntagmes nominaux

6.2. Construction des patrons

La stratégie mise en œuvre cherche d'abord à filtrer les séquences pertinentes du corpus autour de l'ensemble des marqueurs de l'ontologie. Trouver des mots sémantiquement proches puis s'assurer que ceux-ci se trouvent au sein d'une structure syntaxique spécifique. Cette phase d'acquisition se compose de trois étapes :

6.2.1. Normalisation du corpus et filtrage de phrases pertinentes

Cette étape consiste à remplacer les mots qui ont le même sens et qui sont pertinents pour le domaine par un seul terme ou expression indiquant le nom de la classe sémantique générale. Ainsi le terme « *entreprise* » peut être exprimé par différentes expressions : « notre entreprise », « notre société », « nous »., « le nom de l'entreprise » etc. Ces expressions sont remplacées par le nom de la classe sémantique qui est « Représentant entreprise ». Des exemples issus du corpus sont :

ATTAX **conçoit**, industrialise et commercialise des dispositifs de fixations destinés à toutes les industries.

qui devient

Représentant entreprise* **conçoit, industrialise et commercialise des dispositifs de fixations destinés à toutes les industries.

La société MECADEX est **spécialisée** dans le décolletage de précision.

qui devient

Représentant entreprise est **spécialisée** dans le décolletage de précision.

Au cours de cette étape de normalisation, les marqueurs « conçoit » «spécialisée » sont aussi identifiés dans le corpus par une recherche automatique de leur lemme grâce à la fonction *Locate* du système Unitex avec une expression régulière. Par exemple : `<spécialiser>`
Reconnaît toutes les entrées qui ont `spécialiser` comme forme canonique ;

Suite à cette normalisation, l'expert filtre ensuite les phrases pertinentes. Il s'agit de ne retenir que les phrases potentiellement pertinentes pour éviter d'analyser de larges pans de texte.

6.2.2. Identification d'exemples représentatifs

C'est la détermination, parmi les phrases filtrées, des syntagmes représentatifs. Il s'agit de l'ensemble des termes qui peuvent et/ou doivent être corrélés au marqueur pour définir un sens pertinent pour notre recherche. A cette étape, seule l'expertise humaine est capable de déterminer et d'évaluer la pertinence d'un syntagme. L'identification des syntagmes

est basée sur une analyse par ambigüité. Cette analyse consiste à chercher toute les ambigüités que peut avoir le marqueur dans le contexte identifié du corpus. Quelques ambigüités qui ont été levés du corpus :

Exemple : **Représentant entreprise** industrie, basée à Genas (Lyon. France), est spécialisée dans la **conception** et la réalisation de machines d'assemblage...

Ambigüité : Il faut faire la différence entre conception des produits et conception des outils de production.

Le tableau suivant résume les cas d'ambigüité qui s'oppose dans des conceptuelles de l'ontologie :

Classe conceptuelle	**Ambigüité**
Traces de ressources et procédés de conception	Comment identifier que ce sont les moyens et non des produits ?
Traces de ressources et procédés de production	Comment identifier que ce sont des moyens de production utilisés et non des équipements vendus à d'autres entreprises ?
Traces de qualité et de performance des produits/services	Comment identifier que ces notions de qualité/performance concerne bien les produits et services ?
Traces d'innovation sur les procédés techniques	Comment identifier que cela concerne les procédés et non l'innovation produit ?
Traces de démarche qualité sur les processus	Certaines sous-classes requièrent une analyse linguistique pour l'extraction d'autres informations que le marqueur

Tableau 2. *Récapitulatif d'ambigüité lors de la communication Ontologie-Corpus*

Une fois cette ambiguïté est identifiée, l'expert propose les possibilités de l'utilisation de ces concepts sans ambiguïté : ce raisonnement permet de construire autour de chaque marqueur un ensemble de schéma linguistique structurelle. Pour rester toujours lier à notre domaine d'application, nous avons limité la généralisation de ces schémas structurelle pour ne pas encombrer le nombre des patrons dans le système avec des cas qui ne sont pas utile pour notre tâche.

6.2.3. Génération des patrons

La génération de variantes de patrons a pour rôle d'étendre la couverture du système en proposant des structures sémantiquement équivalentes. Cette étape se base sur l'expertise humaine ainsi que sur la recherche et l'analyse d'autres exemples d'entreprises sur le web. Les patrons peuvent être utilisés pour trois cas d'usages :

La Détection de la Présence d'un Concept (DPC) : le plus souvent, ce sont des patrons constitués par des termes simples (patrons simples) utilisés pour signaler la présence d'un concept. Ce type de patrons est appliqué sur les marqueurs dépourvus d'ambigüité.

➲ Haute Technologie : est un patron simple qui permet la détection d'une de la présence d'une compétence de technologie

La Désambigüisation entre deux concepts (DEC) : ce type de patrons est utilisé pour détecter un type d'entreprise qui sera utilisé ultérieurement pour éviter d'autres ambigüités. Ils permettent la classification de l'entreprise parmi l'un des deux principales classes conceptuelles (Réalisation d'outillage de production ou Fabrication des produits manufacturiers). Evidemment certaines entreprises peuvent être classées dans les deux types puisqu'elles peuvent exercées les deux types de productions. Ce type de patrons repose surtout sur les marqueurs production et ingénierie (verbe d'action) qui s'insèrent dans des patrons comme suit :

➲ Représentant Entreprise – verbe d'action – COD

Ce patron nécessite l'extraction et l'analyse du COD dans la phrase puisque c'est lui qui va déterminer le type du produit délivré par l'entreprise. Le verbe d'action dans ce patron peut être le verbe produire, fabriquer, industrialiser, concevoir … Ce patron permet de résoudre l'ambigüité autour de la conception des produits et la conception des outils de production en analysant le COD. Si le(s) nom(s) qui se greffent au COD font référence à la liste des marqueurs d'outils de production (exemple : machine, bancs de test, chaine de production…) alors c'est de la conception d'outils de production si non c'est de la conception des produits puisque notre hypothèse est que par défaut toutes les entreprises dans le domaine de la mécanique font de la production.

L'extraction d'Information Complémentaire Rattachée au Concept (EICRC) : Certains patrons sont utilisés pour extraire de l'information (patrons enrichis). Par exemple avec le marqueur « spécialiser » on cherche à savoir et extraire la spécialité de l'entreprise et non pas une simple détection de la présence d'une spécialité. C'est pourquoi le patron proposé est :

➲ Représentant Entreprise – forme verbale passive incluant spécialisé – PREP- GN

Nous avons besoin d'extraire le GN pour savoir quelle est la spécialité de l'entreprise.

Le résultat de l'acquisition des patrons à partir du corpus constitue une bibliothèque avec 35 patrons enrichis complétés par 100 patrons simples non ambigus. L'ensemble des patrons est transcodé dans un langage formel compréhensible par la machine. Ainsi les patrons sont décrits sous la forme des grammaires locales qui représentent un moyen puissant pour représenter la plupart des phénomènes linguistiques. Unitex permet de représenter un ensemble d'expressions linguistiques sous forme d'automates.

6.3. Projection des patrons sur le corpus

La projection des patrons sur le corpus est la recherche des schémas linguistiques, traduites sous la forme d'automates, dans le texte de l'entreprise. On se base sur le programme *« locate »* d'Unitex qui permet d'effectuer cette projection et de trouver les occurrences dans le texte.

Dans l'annexe, nous illustrons graphiquement un exemple de projection d'un patron enrichi qui permet de typer l'entreprise selon qu'elle réalise ou non des activités de production. L'exemple de projection du patron (illustré graphiquement en annexe) montre la possibilité d'extraire une information pertinente qui décrit la compétence de l'entreprise. La projection du patron "production" permet de typer l'entreprise comme une entreprise qui réalise des machines de production. Cette décision est validée à partir de l'occurrence trouvée : MECASONIC conçoit et fabrique des machines ultra-sons. L'ensemble des *occurrences trouvées de tous les patrons projetés sur le texte d'une telle entreprise constitue une trace des compétences.* Dans cet exemple la projection du patron sur le texte de l'entreprise donne lieu à deux occurrences.

6.4 Résultats et performances du système UNICOMP

Le résultat final de la projection de l'ensemble des patrons sur le texte de l'entreprise est un ensemble d'occurrences retrouvées. Une occurrence retrouvée est un schéma syntaxique valide qui relève la présence d'un marqueur de compétence élaboré des éléments d'information pertinents. A l'aide d'un processus d'activation basé sur des règles déterministes, chaque occurrence retrouvée active un concept de l'ontologie initiale de traces de compétences. Le résultat de l'activation par toutes les occurrences constitue la trace de compétence de l'entreprise qui est un sous-arbre ontologique. Pour donner une idée claire sur la définition et la structure d'une trace de compétence, nous modélisons dans

l'arbre ontologique (figure ci-dessous) la structure globale de l'ontologie des traces des compétences pour les capacités techniques. Elle contient 4 niveaux dont chacun contient différentes classes conceptuelles. Chaque classe est représentée en un petit cercle. En vert la trace activée par le système UNICOMP.

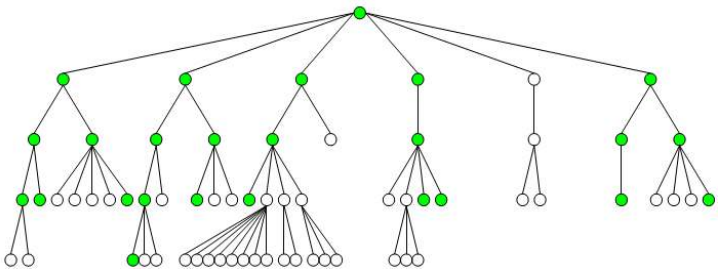

Figure 9. *Trace de compétence : un sous-arbre ontologique*

L'évaluation de la performance du système UNICOMP est basée sur les deux indicateurs Précision et Rappel (Le tableau ci-dessous). Nous avons doté d'une sous-collection de 10 entreprises dans le domaine de la mécanique dont nous avons construit par expertise, à partir de leurs textes extrait du site web et l'ontologie initiale de trace de compétences, leurs traces de compétence c'est-à-dire leurs sous arbre ontologique.

Entreprises	**Précision**	**Rappel**
E1	0,81	0,56
E2	0,92	0,7
E3	0,87	0,5
E4	1	0,66
E5	0,75	0,54
E6	1	0,7
E7	0,76	0,83
E8	0,8	0,66
E9	0,87	0,77
E10	0,88	0,57
Moyenne	**0,87**	**0,64**

Tableau 3. *Indicateurs de précision et rappel obtenus pour UNICOMP*

Les deux indicateurs précision et rappel traduisent la capacité de couverture du système. La construction automatique de la trace de compétence est basée sur la recherche des occurrences des patrons construits autours des instances (marqueurs) de l'ontologie métier. Plus que le nombre des instances est élevé (plus de patrons de recherche), plus qu'on a une couverture plus large du domaine et par conséquent un rappel élevé. Les performances du système dépendent de la

granularité de l'ontologie. En effet, nous estimons qu'un enrichissement automatique ou semi-automatique des instances de l'ontologie est capable d'augmenter le rappel et la précision du système. Ainsi pour augmenter la précision, il faut construire, sans ambiguïté, toutes les réalisations linguistiques dans le domaine traité (la mécanique dans notre cas) qu'on peut construire avec chaque instance de l'ontologie initiale (tâche presque impossible). Un autre facteur intéressant qui peut influencer directement sur ces indicateurs est la qualité (ambigüité sémantique, nombreuses abréviations, phrases asyntaxiques, ponctuation inexistante) du texte fournit par le site web et les phénomènes linguistiques (temps, anaphore, connecteurs, métaphore) qui peut être compréhensible par humaine mais ils ne sont pas encore tous résolus par le traitement automatique de la langue.

7. Conclusion et Perspectives

Dans cet article, un ensemble de contributions ont visé à confronter des techniques du traitement de l'information à la problématique de construction des réseaux d'entreprises collaboratifs. Ces contributions concernent la recherche et l'extraction d'information sur le web. L'identification des compétences d'entreprises est avérée comme un facteur clé pour une aide à la décision en vue de construire des réseaux d'entreprises. L'approche d'extraction des compétences que nous adoptons, est basée sur une chaîne de traitement des textes pour l'extraction d'information à l'aide d'ontologies et de patrons lexico-syntaxique. Cette contribution d'extraction des informations sur les compétences donne lieu au système UNICOMP dont une évaluation des performances a été étudiée.

A la suite de ce travail, un nombre considérable de perspectives peuvent être dégagées :

- Enrichir automatiquement l'ontologie des traces de compétences : proposer une approche basée sur le traitement automatique du corpus et d'une liste de concepts décrivant une première version de l'ontologie initiale à enrichir. Commencer à générer des règles d'association pour la détection de la corrélation entre les concepts de l'ontologie et les mots du corpus. Puis enrichir automatiquement l'ontologie initiale par les concepts appris selon des paramètres validés expérimentalement.
- Evaluer la robustesse de la communication entre l'ontologie métier et l'ontologie générique en testant avec d'autres domaines d'activité (l'informatique par exemple).
- Dans notre travail, les patrons syntaxiques sont construits manuellement à partir des marqueurs qui présentent les instances de l'ontologie métier. Il serait très intéressant de pouvoir construire une méthode automatique permettant d'extraire le patron autour du concept. Une des idées qui peut être étudiée dans ce cadre est d'effectuer une analyse linguistique très fine sur la phrase ou l'ensemble des mots corrélés au concept pour détecter les éléments focus d'information.

Pour appliquer les méthodes d'aide à la décision pour la construction des réseaux d'entreprises en collaboration, il faut savoir quantifier la proximité entre deux traces différentes (c'est-à-dire entre deux sous-arbres ontologiques) pour répondre à la question: Quel est le degré de similarité entre deux entreprises en termes de compétence ? Nos travaux en cours s'intéressent à cette mesure pour tester dans un deuxième temps l'application de la méthode d'aide à la décision.

8. Bibliographie

N. Aussenac-Gilles, S. Despres, S. Szulman. The TERMINAE Method and Platform for Ontology Engineering from texts. Bridging the Gap between Text and Knowledge -Selected Contributions to Ontology Learning and Population from Text. P. Buitelaar, P.Cimiano (Eds.), IOS Press, p. 199-223, 2008.

A. AUGER, BARRIERE C. (2008), Pattern based approaches to semantic relation extraction: a state-of-the-art. Terminology, John Benjamins, 14-1,1-19.

B. Bachimant, A. Issac, and R. Troncy. Semantic commitment for designing ontologies. *13th International Conference on Knowledge Engineering and Knowledge Management* (EKAW), LNAI 2473 :114–121, 2002.

P. Buitelaar, P. Cimiano, B. Magnini, Ontology Learning From Text: Methods, Evaluation and Applications. IOS Press (2005).

M. Benali and P. Burlat. Framework to draw cartographies of coordination modes within smes network. *In 18th International Conference on Production Research*, Salerno, Italy, 2005.

P. Burlat and M. Benali. A metology to characterise co-operation links for networks of firms. *Production Planning and Control*, Vol. 18 No. 2 :156–168, March 2007.

E. Blanchard, M. Harzallah, Reasoning on competence management, *Workshop on Knowledge Management and Organizational Memories of the 16th European conference on Artificial Intelligence (ECAI'04)*, Valence 22-27 août 2004.

N. Ben Mustapha, R. Soussi, H.B. Zgal, and M. Aufre. A metaontology for domain ontology enriching in an information retrieval system. In JFO (Journées Francophones sur les Ontologies) 2008 Lyon-France, 2008.

LM. Camarinha-Matos, H. Afsarmanesh (2007). A comprehensive modeling framework for collaborative networked organizations, *Journal of Intelligent Manufacturing*, 18:529-542.

LM. Camarinha-Matos, H. Afsarmanesh. Elements of a base VE infrastructure, (2003). *Computers in Industry,* 51, pp. 139-163.

O. Corby, R. Dieng-Kuntz and C. Faron-Zucker, Querying the Semantic Web with the CORESE search engine. In: R. Lopez de Mantaras and L. Saitta, Editors, *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI'2004), subconference PAIS'2004*, Valencia, Spain, IOS Press (2004), pp. 705–709

M. Chagnoux, N. Hernandez, and N. Aussenac. From text to ontologies : Non taxonomical relation extraction. In JFO , Lyon-France, 2008.

E. Ermilova et H. Afsarmanesh, *Modeling and management of profiles and competences in VBEs*. Journal of Intelligent Manufacturing 18, 561-586, 2007.

E. Ermilova, N. Galeano, H. Afsarmanesh (2005). *ECOLEAD deliverable D21.2a*. Specification of the VBE competency/profile management, 2005.

F. Fürst. PhD Thesis, Contribution à l'ingénierie des ontologies: une méthode et un outil d'opérationalisation. Université de Nantes, France, Novembre 2004.

N. Guarino and Luc Schneider. Ontology-driven conceptual modelling. In ER, page 10, 2002.

P.A. Gomez, D. Rojas-Amaya, "Ontological Reengineering for Reuse", Fensel D., Studer R., Eds., 11th European Workshop on Knowledge Acquisition, Modeling and Management (EKAW-99), vol. 1621 de LNAI, Berlin, 26-29 1999, Springer, p. 139-156.

N. Grabar and T. Hamon. Les relations dans les terminologies structurées : de la théorie à la pratique. Revue d'intelligence artificielle, 18(1) :57–85, 2004.

K. Hajlaoui. Information extraction procedure to support the constitution of virtual organisations. *In IEEE International Conference on Research Challenges in Information Science (RCIS 2008)* Marrakech, Morocco, 2008.

K. Hajlaoui and X. Boucher. Neural network based text mining to discover enterprise networks. *In 13th IFAC Symposium on Information Control Problems in Manufacturing (INCOM'2009)*. Moscow, Russia, 2009a.

K. Hajlaoui, X. Boucher, and J.J Girardot. Competency ontology for network building. *In 10th IFIP Working Conference on Virtual Entreprises (PRO-VE'09).* Thessaloniki, GREECE, 2009b.

Hajlaoui K., Boucher X., Mathieu M. (2008). Data Mining To Discover Enterprise Networks, *9 th IFIP Working Conference on Virtual Entreprises (PRO-VE'08)* Poznan, POLAND, 8-10 September 2008b.

M.A. Hearst. Automatic acquisition of hyponyms from large text corpora. In In A. Zampolli, editor, Computational Linguistics (CoLing'1992), pages 539-545, Nantes, France,, 1992.

M. Laukkanen and H. Helin, Competence management within and between organizations, *Proceeding of the CAISE'05 Workshops, Enterprise Modelling and Ontologies for Interoperability Workshop, vol. 2* Porto, Portugal, June 13–17 (2005), pp. 359–362.

O. Mendes. État de l'art sur les méthodologies d'ingénierie ontologique. PhD thesis, Montréal, Québec, Canada : Centre de recherche LICEF, 2003.

R. Mizoguchi. A step towards ontological engineering. In the 12th National Conference on AI of JSAI, 1998.

G. Salton, M. Mcgill, Introduction to modern information retrieval. McGaw Hill Publications, New York, 1983

Y. Sure, A. Maedche and S. Staab, Leveraging corporate skill knowledge: from ProPer to OntoProPer, *Proceedings of the 3rd International Conference on Practical Aspects of Knowledge Management* Basel, Switzerland (2000).

J. Plisson, P Ljubic, I Mozetic, N. Lavrac (2007). An ontology for Virtual Organisation Breeding Environments. *To appear in IEEE Trans. On Systems, Man, and Cybernetics.*

V. Psyché, R. Mizoguchi, and B. Bourdeau. Ontology development at the conceptual level for theory-aware its authoring systems. In Conference on Artificial Intelligence in Education (AIED03), 2003.

M.T Pazienza, Information extraction (a multidisciplinary approach to an emerging information technology), Springer-Verlag, Berlin, Heidelberg, 1997

G.B. Richardson. The organization of industry. Economic Journal, vol.82 :883, 1972.

Vanderhaegen, and Loos. (2007). Distributed model management platform for cross-enterprise business process management in virtual enterprise networks. *Journal of Intelligent Manufacturing* 18:553-559.

E.M Voorhess, Naturel language processing and information retrieval, dans M.T. Pazienza, Information extraction, toword scable, adaptable systems, Springer-Verlag, Heidelberg, p. 32-48, 1999.

Annexe. *Exemple de projection d'un patron enrichi sur un texte d'entreprise pour l'extraction d'une compétence.*