

Méthodologie d'Évaluation Intelligente des Concepts Ontologiques

Lobna Karoui*, Marie-Aude Aufaure*
*Supelec, Plateau de Moulon, Gif-sur-Yvette
{Lobna.Karoui, Marie-Aude.Aufaure}@supelec.fr
http://www.supelec.fr/ecole/si/pages_perso/karoui.html

Résumé. Un des problèmes majeurs dans la gestion des ontologies est son évaluation. Cet article traite l'évaluation des concepts ontologiques qui sont extraits de pages Web. Pour cela, nous avons proposé une méthodologie d'évaluation des concepts basée trois critères révélateurs : « le degré de crédibilité »; « le degré de cohésion » et « le degré d'éligibilité ». Chaque critère correspond à un apport de connaissance pour la tâche d'évaluation. Notre méthode d'évaluation assure une évaluation qualitative grâce aux associations de mots ainsi qu'une évaluation quantitative par le biais des trois degrés. Nos résultats et discussions avec les experts et les utilisateurs ont montré que notre méthode facilite la tâche d'évaluation.

1 Introduction

L'évaluation d'une ontologie est une tâche difficile. Ceci explique l'absence de méthodes d'évaluation standard ou de mesures d'évaluation servant à valider l'ontologie. Dans cet article, nous focalisons notre intérêt sur l'évaluation des concepts de l'ontologie de domaine (appelés concepts ontologiques) qui sont extraits des pages Web. Nous travaillons sur les documents HTML écrits en français, dans le domaine du tourisme. Dans un travail précédent, nous avons défini un contexte structurel qui tient compte du document HTML et développé un algorithme de clustering afin de bien rassembler les mots sémantiquement proches (Karoui et al, 2006). Le résultat de ce travail était constitué de classes de mots pour lesquelles les experts ont du réaliser une lourde tâche d'évaluation et d'étiquetage. Pour aider ces derniers et faciliter l'interprétation sémantique de ces classes (concepts), nous avons défini une méthode d'évaluation basée sur trois critères révélateurs. Ces derniers sont le degré de crédibilité, le degré de cohésion et le degré d'éligibilité. Le degré de crédibilité exploite deux types de contextes : un contexte linguistique et un contexte documentaire. En se basant sur ces deux types de contextes, nous calculons le degré de crédibilité associé à chaque classe de mots et à chaque contexte. Le degré de cohésion calcule le degré de rapprochement des mots d'une classe en utilisant les documents du web. Le degré d'éligibilité sélectionne ou suggère le mot de la classe qui peut être son concept ou qui peut orienter le raisonnement vers le futur concept approprié. Dans ce qui suit, nous détaillerons ces critères ainsi que leur rôle.

2 Les critères révélateurs de l'évaluation intelligente des concepts

Les critères révélateurs assistent l'expert du domaine durant la tâche d'évaluation. Ces critères sont : le degré de crédibilité: le caractère de ce qu'on croit ; le degré de cohésion: le caractère d'une chose dont toutes ses parties sont réunies avec une relation logique entre ses éléments et sans aucune contradiction ; le degré d'éligibilité : le caractère d'un mot qui com-

bine les conditions nécessaires pour être élu comme concept puisqu'il est le mot le plus représentatif de la classe ou qu'il peut orienter le raisonnement, l'interprétation et la tâche d'étiquetage.

Degré de crédibilité. Notre objectif est de pouvoir modifier les classes en cas de nécessité, de les étiqueter et d'estimer le rapprochement sémantique de leurs mots. Pour cela, nous effectuons une « contextualisation progressive » lors du processus d'évaluation, qui exprime la diversité des contextes et se concentre sur comment différents objets (dans notre cas les mots) opèrent et s'adaptent à leurs contextes à travers différentes structures de documents, d'organisations de mots, de méthodes de conception, d'intensions de concepteur, etc. La contextualisation progressive rejette l'idée de l'utilisation d'un unique contexte pour comprendre un objet. Au contraire, elle affirme qu'un discours ou une rédaction ordinaire révèlent plusieurs contextes pour chaque objet étudié. L'interaction de ces contextes permet une interprétation sémantique correcte et écarte les mauvaises évaluations. Dans notre recherche, les différents contextes déduits à partir des documents sont donnés par deux analyses différentes. La première est une analyse linguistique qui permet de fournir les groupes nominaux, groupes verbaux, les associations de mots par une préposition (de, à, etc.) et celles par une conjonction (et, ou, etc.). La seconde analyse est documentaire. Elle se fait à une échelle plus large afin de donner les différentes sections de phrases, phrases, paragraphes et documents. Ainsi, nous définissons deux types de contextes : linguistique et documentaire. En utilisant le contexte linguistique, nous obtenons les mots proches pour chaque groupe de mots étudiés. Le contexte documentaire est plus général. Par conséquent, l'information produite est complètement ou complètement nouvelle pour les mots d'une classe. Grâce au processus de contextualisation progressive et son information produite, l'évaluation de l'expert est plus rapide et se fait en respectant l'ordre suivant : contexte linguistique (Groupe nominal, Groupe Verbal, Groupe Prépositionnel, Groupe de Conjonction) puis contexte documentaire (Sections de phrases, Phrase, Paragraphe, Documents). Nous donnons un exemple d'un extrait de deux contextes générés pour une classe de mots :

Classe = {académie, club, golf, golfeur}; **Granularité** = groupe prépositionnel

Extrait : « académie de golf » ; Extrait : « académie pour le golf »

=> **Association** : (académie, golf)

Extrait : « club de golf » ; => **Association** : (club, golf)

Classe	Classes de Mots
Classe 1	golf, golfeur, club, académie
Classe 2	archéologie, ethnologie, peuple, civilisation
Classe 3	falaise, national, rocher, parc
Classe 4	chapelle, église, évangélisation, génération, mémoire, religieux, culte, lieu, sanctuaire
Classe 5	excursion, pied, personne
Classe 6	randonneur, gorge

TAB. 1 – Exemples de classes de mots

En fournissant tous ces contextes à l'expert, il lui sera difficile d'exploiter les associations entre les mots pour évaluer et étiqueter les classes (au vu de la quantité d'information importante qu'il devra analyser). Afin de lui faciliter la tâche, nous définissons un indice sémantique qui représente la crédibilité de l'association des mots étudiés en considérant les différents contextes. Cet indice est appelé « degré de crédibilité ». Il est calculé automatiquement pour chaque classe de mot et pour chaque contexte.

Notre algorithme, intitulé « Calcul de Degré de Crédibilité » et noté CDC, est exécuté sur un ensemble de classes de mots afin de calculer leurs degrés de crédibilité (DC). Ces derniers sont de la forme suivante :

$DC = \{\text{nombre de mots}\} / \text{nombre d'associations}$ / Exemple : 3_2 (deux associations de trois mots).

Par exemple, en prenant la classe 1 (TAB.1) et en choisissant la phrase comme contexte, l'algorithme génère toutes les combinaisons possibles dans ce contexte c.-à-d. il essaye de trouver les 4 mots {académie, golf, golfeur, club}, puis les associations possibles de combinaisons de trois mots, etc. Pour chaque association trouvée, il présente les mots associés ainsi que le degré de crédibilité qui représente le nombre de fois que cette association a été trouvée. Par exemple, l'algorithme détecte deux associations de trois mots qui sont {académie, golf, golfeur} et {golf, golfeur, club}. Dans ce cas, le degré de crédibilité est 3_2 c.-à-d. deux associations de trois mots. Cet algorithme possède plusieurs fonctionnalités à savoir :

- Trouver les associations disponibles dans les contextes ainsi que le concept de la classe de mots. Par exemple, pour la Classe 2 (TAB.1), le concept est « civilisation » car ce terme de la classe est le plus fréquent dans les associations.

- Détecter les éléments générant du bruit dans une classe pour les supprimer ou les transférer dans leurs classes appropriées. Par exemple pour la classe 6 (TAB.1), le mot 'randonneur' a été trouvé dans de nombreuses associations contextuelles relatives à la classe 5 et pas dans celles de la classe 6. Dans ce cas, l'algorithme peut proposer de changer la place de ce mot de la classe 6 vers la classe 5. Lorsqu'un mot d'une classe apparaît souvent dans les associations de deux classes différentes (par exemple sa classe et une autre), l'algorithme écrit le mot en rouge pour souligner une situation ambiguë.

- Enrichir une classe par d'autres mots provenant des associations produites par les contextes. Pour la classe 3, nous pouvons enrichir la classe par les mots 'naturel' et 'espace' et lui associer le nom du concept suivant : 'espace naturel'.

Cette méthode d'évaluation permet à un utilisateur (expert, ingénieur de connaissances, étudiant, etc.) de commencer par analyser les associations produites par les contextes linguistiques. A défaut de ne pas retrouver l'information utile, il pourra étendre son analyse aux résultats des contextes documentaires. Grâce au degré de crédibilité calculé pour chaque classe et pour chaque contexte, l'utilisateur peut obtenir des informations utiles et dans certains cas suffisantes pour manipuler les classes de mots (supprimer, ajouter, etc.), les évaluer et les étiqueter. Pour une classe, si l'utilisateur trouve les degrés ($5_1, 4_3, 3_8, 2_{15}$) qui sont des indications quantitatives, il devrait commencer par analyser l'association de 5 mots. Si ceci s'avère insuffisant pour lui, il pourra poursuivre avec les associations suivantes (de 4 mots, etc.). Autrement dit, l'algorithme génère un nombre d'informations complémentaires, mais l'utilisateur ne sera pas amené à toutes les utiliser mais plutôt à suivre l'ordre (des contextes linguistiques vers les contextes documentaires et au sein d'un même contexte des associations les plus larges (association de cinq mots 5_1) vers les plus réduites (les associations de deux mots 2_{15})) pour décider de s'arrêter dès qu'il atteindra son objectif.

Degré de cohésion. Notre idée est basée sur le fait qu'il existe une grande quantité de pages web indexées par Google (8 058 044 651). Le web représente des connaissances fournies par différentes personnes ou organismes qui ont des expériences différentes mais certainement complémentaires dans le domaine. Nous sommes intéressées par le calcul du degré de cohésion de chaque classe de mots. Pour cela, nous définissons une distance sémantique entre les mots d'une classe en se basant sur les documents indexés par Google. Notre mesure intitulée 'degré de cohésion', notée Coh-D, utilise le nombre de documents dans lesquels les mots de la classe apparaissent ensemble. Ce critère est défini après une adaptation de la distance

Méthodologie d’Evaluation Intelligente des Concepts Ontologiques

normalisée de Google (Cilibrasi et Vitanyi, 2006) de manière à pouvoir calculer le rapprochement sémantique pour un nombre de mots supérieur ou égal à deux (la distance de Google peut estimer le rapprochement sémantique entre uniquement deux mots). Notre formule est la suivante:

$$X = \frac{\text{Min } \text{Log} (\text{NBH} (W_i))}{\text{Max} (\text{Log} (\text{NBH} (W_j)))}$$

$$\text{Coh} - D = X \times \text{NBH} \{W_1, W_2, \dots, W_j\}$$

- NBH (Wi) est le nombre de réponses donné par Google quand il cherche le mot Wi, Wi est un mot appartenant à la classe {W1, W2, W..., Wj}.

- NBH {W1, W2, W..., Wj} est le nombre de réponses reliées à tous les mots de la classe.

N°	Classe de mots	label	Degré-cohésion
1	académie, club, golfeur, golf	Golf	37135
2	bonheur , reste, touriste camping, caravane,	Camping	15759,59 / 131266,62
3	allant, canotage, passant, population, spécimen, échelle, réserve, vue	inconnu	368,2
4	chapelle, église, évangélisation , génération, mémoire, religieux, culte, lieu, sanctuaire	religion	883,21 / 21413,76
5	avion, banquise, musée, recherche, semaine offerte,	inconnu	518,69
6	boîte, clientèle, francophone, métropole, ville	service	76842,01
7	administration, attention, éducation, particulier, préoccupation, cervidé , faune, habitat, utilisateur	administration	223,52 /81433,26
8	brochet, feuille , jaune, doré, saumon	poisson	1257,59 / 34997,49

TAB. 2 – Des résultats du degré de cohésion

Lors de l’évaluation, l’expert de domaine constate qu’il existe trois types de classes résultantes : les classes acceptables, incorrectes et inconnues (karoui et al, 2006). Une classe acceptable est une classe que l’expert est capable de labelliser et dans laquelle les termes appartenant au même groupe sont proches sémantiquement. Une classe incorrecte est une classe qui présente l’un des deux cas suivants : soit elle contient certains termes qui n’ont pas de relations avec le concept extrait de cette classe, soit elle contient plusieurs concepts clairement identifiés par l’expert. Une classe inconnue est une classe dont les termes n’ont aucune relation sémantique ; l’expert ne peut pas en donner une interprétation sémantique.

Par exemple, les classes 1 et 6 (TAB.2) sont deux classes acceptables puisque tous les mots sont associés au même concept. Les classes 3 et 5 (TAB.2) contiennent des éléments de bruit (mots qui ne doivent pas appartenir à la classe en question). C’est pour cela, que les experts du domaine suppriment plusieurs mots afin de trouver le concept ou ne peuvent pas définir de concept pour la classe. Dans les autres cas, la classe peut contenir des mots de bruit mais qui sont facilement repérables (‘feuille’ dans la classe 8). La question qui se pose est comment peut-on détecter le mot qui doit être supprimé de la classe. En utilisant le moteur de recherche Google et sa base de données, nous calculons le nombre de réponses de chaque mot de la classe en l’associant au mot représentant le domaine (dans notre cas “tourisme”). Le mot qui apparaît le moins est supprimé et nous maintenons cette modification au niveau du calcul du degré de cohésion pour observer son effet. D’autres exemples sont présents dans la TAB.2. Les mots en gras sont ceux qu’une fois supprimés, améliorent le degré de cohésion.

Par exemple, dans la classe 7 (TAB.2), en supprimant le mot “cervidés”, nous obtenons un degré de cohésion plus représentatif (un impact réel puisque le degré de cohésion est passé de 223.52 à 81433.26) et par conséquent une tâche d'évaluation plus facile. D'après nos expérimentations et l'avis des experts, quand le degré de cohésion est faible (<1000), cela indique que le contenu de la classe devra être modifié.

Le degré de cohésion est un critère quantitatif qui aide l'expert du domaine ou un utilisateur novice à voir si la classe est sémantiquement cohérente ou non. Mais l'évaluateur ne peut pas établir un jugement uniquement en se basant sur ce critère. Ce dernier est plus intéressant quand il est considéré avec les autres critères (crédibilité et éligibilité).

Degré d'éligibilité. Le degré d'éligibilité représente le nombre de voix qui indique à l'utilisateur quel est le mot candidat qui peut probablement représenter la classe ou du moins initier le processus de raisonnement qui mène à la définition du concept approprié de la classe. Il est calculé en appliquant la formule suivante :

$$ED = \text{Min} \left(\forall xi \mid \left(\sum_1^n NT(xi) \right) / n - NT(xi) \mid \right)$$

$NT(x)$ est le nombre d'occurrences de x ; n est le nombre de mots dans la classe.

Le mot ayant le degré d'éligibilité le plus proche de la moyenne de la classe désigne le candidat représentant la classe. Il est noté Mot-E-D. Ce degré est calculé uniquement pour les classes acceptables. Vu que nous opérons dans le domaine de la sémantique, ces critères révélateurs ainsi que les constatations résultantes restent des propositions à approuver ou à rejeter par l'expert. Dans les deux cas, nous en tirons profit car même en cas de proposition incorrecte, l'expert raisonnera plus rapidement pour définir le 'bon' concept approprié à la classe.

N°	Classe de mots	Mot-E-D	ED	Mot-Référence
1	académie, club, golfeur, golf	golf	59144750	golf
2	bonheur, reste, touriste camping, caravane	caravane	42828000	camping
3	falaise, mètre, national, rocher, parc	parc	760824000	parc

TAB. 3 – Des résultats

Pour la classe 1 (TAB.3), notre critère trouve le bon terme “golf” qui représente la classe. Dans le cas où le mot désigné par le degré d'éligibilité et celui mentionné par le degré de crédibilité en tant que mot de référence sont identiques, l'algorithme présente le mot en question en un style gras afin d'exprimer une forte suggestion. La classe 1 est une classe acceptable. La classe 3 (TAB.3) contient un mot de bruit qui est “mètre”. En le supprimant, l'algorithme détecte le bon terme représentatif de la classe qui est “parc”. Pour la classe 2, l'algorithme propose “caravane” en tant que concept. Mais le mot de référence (le mot qui est donné par l'algorithme de calcul de degré de crédibilité et qui est le plus présent dans tous les contextes) de cette classe est “camping”. Puisque les deux mots sont différents, l'algorithme les mentionne avec une couleur rouge pour signaler une situation d'ambiguïté. Dans ce cas, l'évaluateur doit faire le choix entre les deux mots ou définir un autre concept soit sur la base des deux ou complètement nouveau. Mais dans les deux premiers cas (les classes 1 and 3 (TAB.3)), l'évaluateur peut décider facilement.

Discussion. Notre méthode d'évaluation des concepts ontologiques, que nous avons caractérisée comme étant 'intelligente', du fait qu'elle pourra à terme s'appliquer automatiquement, donne deux types d'informations révélatrices à savoir :: une indication qualitative basée sur les associations des mots déduites à partir de la projection des mots d'une classe dans un contexte défini ; une indication quantitative représentée par le degré de crédibilité calculé pour chaque classe et pour chaque contexte, le degré de cohésion et celui d'éligibilité. L'évaluation qualitative permet d'aider l'utilisateur au niveau de l'interprétation sémantique. Notre méthode permet à un utilisateur ordinaire d'aider l'expert en manipulant en amont les classes de mots et en suggérant des étiquettes sémantiques. Par conséquent, l'expert peut décider rapidement si le label est convenable ou si la classe est sémantiquement correcte. Ce type d'évaluation permet, également, la réutilisation et l'évolution de l'ontologie puisqu'elle est faite à partir des pages web et tiendra compte de leurs changements possibles. Ainsi, si les documents du web changent, les contextes extraits changent et les résultats de l'algorithme CDC seront différents. D'où l'effet de mise à jour qui sera appliqué. Ces informations présentées à l'expert serviront à l'évaluation en ligne et seront stockées pour une éventuelle utilisation ultérieure. Notre méthode a été appliquée après (pourra être pendant) la construction de la hiérarchie de concepts (pas de l'ontologie complète). Dans ce travail, le choix des contextes a été défini manuellement, par contre l'application de ce choix sur le corpus, l'extraction des contextes et des associations ainsi que les calculs des degrés d'évaluation sont faits automatiquement.

3 Conclusion

Dans cet article, nous avons proposé une méthode d'évaluation conduite par le web qui guide l'interprétation, évite les cas ambigus et aide l'expert ou l'utilisateur. Afin d'atteindre cet objectif, notre méthode se base sur trois critères révélateurs qui sont le degré de crédibilité, le degré de cohésion et le degré d'éligibilité. Chaque critère est calculé en appliquant un algorithme séparé. L'algorithme de crédibilité essaye d'éliminer ou de déplacer les éléments bruits, propose des labels sémantiques et produit des associations de mots par type de contexte et pour chaque classe. Les degrés de cohésion et d'éligibilité informent l'utilisateur de la relation entre les mots de chaque classe et quel serait le possible candidat représentant le concept. Comme perspective, nous travaillons sur la visualisation de ces critères.

Références

- Cilibrasi, R. and P. Vitanyi (2006). *Automatic Extraction of Meaning from the Web*. IEEE International Symposium on Information Theory, Seattle, Washington.
- Karoui, L., Aufaure, M-A., Bennacer, N. (2006). *Context-based Hierarchical Clustering for the Ontology Learning*. The 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI-06).

Summary

In this framework, we propose a new evaluation method based on a large web document collection, several context definitions deduced from it and three criteria. It provides a support for either a domain expert or a novice user. Our contribution is proposing an evaluation framework that does not depend on a gold standard, could be applied to any domain even if expert intervention is not available and provides qualitative and quantitative criteria.