PROJET DE RECHERCHE **RESPONdING**

REcherche **S**émantique sur un cor**P**us d**O**cumentaire pour la co**N**ception en **ING**énierie

Entreprise : **TraceParts**- Parc Eco Normandie 76430 Saint Romain
Laboratoire : **LITIS** (équipe MIND) - 76800 St Etienne du Rouvray
Candidat :

Mots-clés :

recherche d'information enrichie par la sémantique, navigation dans un corpus documentaire, ontologies, web sémantique, interaction humain-machine

Contexte

Présentation de la société

TraceParts est spécialiste du contenu numérique 3D pour la conception et l'ingénierie, développant des solutions web et mobiles de configurateurs de produits, de catalogues électroniques et de bibliothèques de composants CAO pour les fournisseurs de produits industriels. Le portail `www.traceparts.com` est utilisé gratuitement en 25 langues par des millions d'utilisateurs dans le monde, offrant des centaines de catalogues et des centaines de millions de modèles CAO et fiches techniques pour l'industrie dans les domaines de la Conception, de la Fabrication, des Achats et de la Maintenance. Les informations sur *TraceParts* et ses solutions sont disponibles sur `www.traceparts.com`.

En utilisant les services *TraceParts* pour créer et diffuser leurs données produits, les fournisseurs de composants industriels bénéficient d'une visibilité marketing mondiale et d'un savoir hautement qualifié.

Le portail TraceParts.com

La première version du portail `www.traceparts.com` a été lancée en en Juillet 2001. Il s'agit d'une plateforme Web unique au monde offrant gratuitement des ressources techniques pour les concepteurs et les ingénieurs, dans 25 langues. Parmi ces données, plus de 110 millions de fichiers 3D issus de plus de 800 catalogues différents, permettent aux utilisateurs de logiciels de Conception Assistée par Ordinateur (CAO), de gagner un temps précieux en n'ayant pas à redessiner les composants industriels dits "du commerce" pour leurs travaux de conception de machines, d'outillages ou d'assemblages.

Le service de téléchargement de fichiers CAO 3D depuis le portail `www.traceparts.com` est financé par les fabricants et distributeurs de composants qui y référencent leurs produits ainsi que par la publicité. Le portail compte aujourd'hui près de 3,7 millions d'inscrits et génère plus de 7 millions de fichiers 3D chaque mois, ce qui le positionne comme l'un des tous premiers sites Web au monde de contenu 3D pour l'industrie.

Au travers de ce portail, l'utilisateur a accès à un catalogue de données CAO 3D, rigoureusement identiques aussi bien pour une utilisation dans l'application desktop `TraceParts DVD` que pour une utilisation sur le portail Web ou dans l'application mobile pour smartphones et tablettes. Les données catalogues

sont articulées autour de formats ouverts et pérennes. Enfin, une indexation centralisée des données permet une recherche et une navigation dans les catalogues, à la fois puissante, rapide et performante.

Le type de données traitées au travers de la plateforme est extrêmement hétérogène (de nombreuses pièces sont configurables à façon et comportent des informations différentes dans leur structure et dans leurs types de valeurs). Cela a pour conséquence d'entraîner une volumétrie importante de données et en constante évolution.

Le moteur de recherche

Le portail `www.traceparts.com` utilise Elacticsearch (ES) comme moteur de recherche. Chaque élément extrait des catalogues fournisseurs est enregistré dans ES et les utilisateurs finaux peuvent rechercher les pièces de fabricants dans tous les catalogues présents sur le site web. Une pièce identifiée par son numéro fabricant peut être contenue dans plusieurs catalogues. Afin de regrouper et classer efficacement les pièces fabricants, *TraceParts* a créé sa propre classification sous la forme d'un catalogue additionnel.

Les utilisateurs finaux interagissent avec l'application via un champ de recherche de texte intégral. Les résultats retournés sont ensuite affichés sur la page de résultats principale selon le classement proposé par défaut par ES.

La page « search » permet d'affiner la liste des résultats grâce à une recherche à facette. Lorsqu'une recherche effectuée ne permet pas d'identifier le catalogue désiré, les facettes sont fondées sur les catégories *TraceParts*. L'arbre de catégories affiché est calculé en temps réel en compilant les résultats de recherche et les hiérarchies de catégories stockées séparément.

Malheureusement, pour certaines recherches de termes courants (i.e : cylindre), l'application retourne des résultats inadaptés en haut de la liste.

La liste de résultats principale contient une entrée distincte pour chaque élément, même s'ils correspondent à des variantes du même élément de base (I.E dimension différente).

Enfin, l'application Web permet des recherches multi-langues. À l'origine, toutes les langues étaient disponibles. Cependant pour des raisons de performances, l'implémentation actuelle utilise la langue courante choisie par l'utilisateur et la langue par défaut (anglais).

Problématique

Dans toute plateforme Web, le moteur de recherche est un des composants les plus importants. Il permet de récupérer des documents en mettant en relation les mots indiqués dans la requête de l'utilisateur et le contenu indexé dans les documents que l'on recherche.

Le nombre d'utilisateurs de la plateforme augmente toujours plus, et doit traiter toujours plus de données, très hétérogènes, toujours plus volumineuses (du fait de la multiplication des conformations de certaines pièces dans l'espace). *TraceParts* doit donc sans cesse trouver de nouvelles méthodes permettant de faire face à cette augmentation constante et massive de données dans le but de fournir un service toujours plus performant et toujours aussi qualitatif. Aujourd'hui, le volume et la taille des données mises en jeu au sein de la plateforme *TraceParts* ne permettent plus dans un avenir proche d'utiliser des systèmes de recherche standard (commerce). La rapidité et la pertinence du moteur de recherche multilingues (23 langues) sont des enjeux stratégiques pour *TraceParts* sur lequel vont être portées une grosse partie des efforts R&D à venir.

Actuellement, la page « search » est de loin la plus utilisée (en moyenne sept itérations, 60% du trafic). La qualité des résultats étant la même entre les pages successives de « search », l'utilisateur passe beaucoup

de temps à chercher son besoin dans les résultats du moteur de recherche. ***La baisse de la moyenne des itérations sur la page « search » est l'un des objectifs quantifiables à atteindre.***

Objectifs de la thèse, verrous scientifiques et résultats attendus

L'objectif principal de cette thèse est de d'améliorer qualitativement le moteur de recherche de *TraceParts* pour qu'un utilisateur trouve, avec le moins d'interactions possibles, les documents pertinents à sa requête et donc, les documents qu'il va télécharger. Cela passe par l'amélioration et l'ajout de fonctionnalités concernant l'aide au requêtage, la pertinence des documents choisis, leur classement et les recommandations que peut faire le système.

Une des premières améliorations concerne la complétion automatique de la requête proposée par le moteur de recherche. L'idée serait ici d'utiliser à la fois les traces des recherches précédentes faites par les utilisateurs du portail et qui ont abouti et une formalisation du corpus de documents en utilisant des algorithmes génératifs de texte [1].

La seconde amélioration concerne la sélection et le classement des documents pertinents pour l'utilisateur. Plutôt que de faire de la recherche plein texte, nous proposons de tirer parti de la représentation formelle des documents proposée précédemment.

La troisième amélioration porte sur la recommandation d'autres documents en fonction de la requête initiale de l'utilisateur. Encore une fois nous proposons d'utiliser les traces pour compléter la formalisation du corpus en ajoutant des relations entre les documents (comme, par exemple, les documents qui sont téléchargés souvent ensemble). Mais nous proposons d'utiliser aussi les connaissances techniques du domaine pour associer les documents des pièces techniquement compatibles.

Enfin la dernière amélioration concerne la reformulation d'une requête utilisateur quand celui-ci n'est pas satisfait des résultats obtenus. Ces méthodes de reformulation seront fondées sur la représentation sémantique du contenu des documents et des liens inter et intra-documents identifiés précédemment dans le corpus de *TraceParts*. Enfin, elles seront également guidées par le contexte de navigation de l'utilisateur (son profil et ses interactions) et aussi par d'autres contextes de navigation sémantiquement proches [2].

TraceParts met à disposition de nombreuses données aussi bien au niveau des traces d'utilisation (environ 2To depuis 2002) que du corpus de documents (environ 1000 documents). Ainsi nous pourrons expérimenter et valider nos hypothèses de recherche dans un contexte réel.

Pour ce faire, nous nous appuierons sur des travaux de notre équipe, réalisés dans le cadre des projets PlaIR 2.0 et PlaIR 2.018 où nous nous sommes intéressés à une problématique de recherche documentaire spécialisée nécessitant une assistance. L'objectif principal était de comprendre, sur la base des interactions avec l'utilisateur, son besoin d'assistance pour décider de l'action la plus pertinente [3]. Nous avons proposé des modèles de personnalisation et d'adaptation pour des utilisateurs de profils différents, experts ou non du domaine, dans l'accès à une base documentaire. Dans cette thèse, nous ajouterons une couche sémantique pour la recherche efficace de documents pertinents dans la base documentaire qui nous intéresse. En effet, les pratiques des utilisateurs du portail de *TraceParts* lors de la recherche d'information pertinente pour leurs besoins seront formalisées grâce à un socle sémantique (connaissances, règles, expériences et métaconnaissances) [4, 5, 6], qui permettra de capitaliser les stratégies de navigation des utilisateurs, tout en donnant un cadre formel pour la représentation du contenu des documents.

De nombreuses approches sémantiques pour la recherche d'information permettent l'exploitation d'un vocabulaire commun décrit dans une ontologie du domaine pour faciliter la mise en correspondance des requêtes utilisateurs et des documents [7, 8] . Certaines de ces approches s'appuient également sur l'ontologie pour définir des mécanismes d'expansion de requêtes [9, 10]. Cependant, aucune de ces approches ne propose un mécanisme de reformulation qui soit à la fois guidée par un ensemble

d'interactions avec l'utilisateur et par des stratégies de reformulation contextuelles découvertes en exploitant le comportement d'autres utilisateurs. Pour aboutir à cet objectif, une méthodologie en trois phases sera suivie : a) définir des approches permettant d'utiliser la représentation sémantique des documents pour découvrir des liens sémantiques inter-documents, b) définir des approches de découverte de liens sémantiques entre contextes de navigation utilisateurs et c) découvrir des règles contextuelles de reformulation de requêtes à partir de contextes de navigation sémantiquement proches.

Au regard des objectifs cités et des données mises à disposition par *TraceParts*, le verrou scientifique à lever dans cette thèse relève du domaine de la représentation formelle de connaissances. D'une part, il s'agira de combiner la formalisation des connaissances à partir de données brutes ou semi formelles sous forme d'ontologie [11]. D'autre part, il faudra compléter l'ontologie obtenue à partir de sources formelles externes et des usages de la plate-forme. L'ontologie finale permettra ainsi d'améliorer les performances de la recherche documentaire et la recommandation d'autres documents aux utilisateurs.

Les principaux résultats attendus à la fin de ce projet de thèse sont les suivants :

1. Un modèle formel pour le domaine de la conception numérique 3D pour l'ingénierie. Ce modèle sera fondé sur la typologie des documents dans le corpus de *TraceParts* et de leur structure. Des standards tels que la classification *ecl@ss* (http://www.eclass.eu) seront utilisés.
2. Une représentation sémantique du contenu des documents du corpus documentaire de *TraceParts*, incluant des liens inter et intra-documents. Cette représentation sera obtenue à partir des résultats de l'objectif 1, complété à l'aide de schémas utilisées dans le monde des données liées, qu'ils soient généralistes, comme par exemple Wikidata, DBPédia, Yago [12], etc., ou spécialisés pour la conception en ingénierie comme *ecl@ss* (http://www.eclass.eu). Cela permettra d'ajouter des propriétés (métier ou pas) à l'ontologie à développer. L'objectif ici est d'avoir une représentation formelle du contenu du document, permettant, grâce à l'ontologie, d'inférer des faits non déclarés dans les documents mais qui seraient susceptibles de répondre à des futures requêtes.
3. Une typologie des utilisateurs finaux du corpus de *TraceParts*, en fonction de leurs activités de navigation et de leur historique de recherche.
4. Des mécanismes de transformation des requêtes informelles des utilisateurs sur le corpus en requêtes formelles, interprétables sémantiquement. Ces mécanismes exploiteront les liens inter et intra-documents indiqués précédemment et les stratégies de reformulation contextuelles découvertes en capitalisant le comportement d'autres utilisateurs.

Organisation des travaux de recherche

Sur la base des objectifs cités plus haut, les travaux incluront les tâches suivantes, déclinées en six catégories et organisées suivant le planning prévisionnel présenté ci-après :

1) Recueil de traces d'usage, extraction d'indicateurs qualitatifs et description semi-formelle du domaine
   a) Définition d'indicateurs qualitatifs du moteur de recherche
   b) Analyse de la structure des documents du corpus
   c) Qualification de la typologie des utilisateurs
2) Modélisation formelle des traces et du corpus
   a) Développement de modèles formels pour le domaine de la conception numérique 3D en ingénierie associé aux corpus étudiés
   b) Formalisation, classification et opérationnalisation des traces
   c) Formalisation du corpus de documents en ontologie
   d) Complétion de l'ontologie à partir de sources formelles externes et des traces
3) Amélioration du moteur de recherche
   a) Complétion dynamique de la requête
   b) Identification et tri des ressources pertinentes
   c) Recommandations de ressources additionnelles

d) Aide à la reformulation de la requête en cas d'échec

4) Développements informatiques
5) Validation
6) Rédaction du manuscrit

Tâche		1ère année				2ème année				3ème année			
1	a												
	b												
	c												
2	a												
	b												
	c												
	d												
3	a												
	b												
	c												
	d												
4													
5													
6													

Références

1. Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic and, Pei-hao Su, David Vandyke, Steve J. Young. *Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue System*s. Journal CoRR 2015. http://arxiv.org/abs/1508.01745
2. Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, Enrico Motta, *Semantically enhanced Information Retrieval: An ontology-based approach*, Web Semantics: Science, Services and Agents on the World Wide Web, Volume 9, Issue 4, 2011, Pages 434-452, ISSN 1570-8268, https://doi.org/10.1016/j.websem.2010.11.003.
3. Louvet, J.-B., Dubuisson Duplessis, G., Chaignaud, N., Vercouter, L. et Kotowicz, J.-P. 2017. *Modeling a collaborative task with social commitments.* In : KES - International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. Marseille, France, p. 377-386.
4. C. Zanni-Merk. *KREM: A Generic Knowledge-Based Framework for Problem Solving - Proposal and Case Studies*, in proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, INSTICC (Eds.), Science and Technology Publications, 2015, p 381-388.
5. Zanni-Merk, C. et Szczerbicki, E. juil. 2019. *Building collective intelligence through experience : a survey on the use of the KREM model.* In : Journal of Intelligent & Fuzzy Systems, p. 1-13. doi : 10.3233/JIFS-179327.
6. Firas Abou Latif, Nicolas Delestre, Nicolas Malandain, Jean-Pierre Pécuchet. *Similarity measure to identify users' profiles in web usage mining*. INFORSID XXVIII°, May 2010, Marseille, France. pp.77-92, 2010.
7. Zahra Vahidi Ferdousi, Dario Colazzo, Elsa Negre. *CBPF: Leveraging Context and Content Information for Better Recommendations*. ADMA 2018: 381-391
8. Li, Z., Raskin, V., Ramani, K.: *Developing Engineering Ontology for Information Retrieval*. J. Comput. Inform. Sci. Eng. (2008)
9. Fu G., Jones C.B., Abdelmoty A.I. *Ontology-Based Spatial Query Expansion in Information Retrieval.* In: Meersman R., Tari Z. (eds) On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE. OTM 2005. Lecture Notes in Computer Science, vol 3761. Springer, Berlin, Heidelberg
10. J. Bhogal, A. Macfarlane, and P. Smith. 2007. *A review of ontology based query expansion.* Inf. Process. Manage. 43, 4 (July 2007), 866-886. DOI=http://dx.doi.org/10.1016/j.ipm.2006.09.003
11. R. Sharman, R. Kishore, R. Ramesh, *Ontologies : A Handbook of Principles, Concepts and Applications in Information Systems*, Springer, 2007.
12. Denny Vrandečić Markus Krötzsch. *Wikidata: a free collaborative knowledge base.* Communications of the ACM VOL. 57 NO. 10 pp78-85. September 2014. https://doi.org/10.1145/2629489

Le candidat

Ce sujet convient aux étudiants qui s'intéressent à l'ingénierie de la connaissance et à la recherche d'information. Il (elle) doit être titulaire d'un diplôme en informatique (Master ou Ingénieur).

Les compétences requises sont notamment les suivantes :

- travail rigoureux,
- autonomie et réactivité,
- excellentes capacités de travail en équipe,
- un très bon anglais écrit et parlé,
- des connaissances en technologies du web sémantique et en apprentissage automatique sont indispensables.

Cette thèse permettra au candidat d'acquérir un important bagage dans le domaine de la modélisation formelle des connaissances et de la recherche d'information, ainsi qu'une expérience efficace de travail dans un environnement dynamique et multidisciplinaire.