Compte rendu : comité de suivi de thèse 1ere année Fabien Amarger

26/09/2013

Introduction

Le jeudi 26/09/2013 a eu lieu le comité de suivi de fin de première année de thèse de Fabien Amarger. Ce comité s'est tenu à l'université Toulouse le Mirail, avec comme participants :

- Jean-Pierre Chanet, Irstea Clermont-Ferrand, encadrant
- Jean-Pierre Chevallet, LIG Grenoble, évaluateur
- Ollivier Haemmerlé, IRIT Toulouse, encadrant
- Nathalie Hernandez, IRIT Toulouse, encadrant
- Chantal Reynaud, LRI Paris XI, évaluateur
- Catherine Roussey, Irstea Clermont-Ferrand, encadrant

Ce comité a permis à Fabien Amarger de présenter les travaux effectués lors de sa première année de thèse. Les évaluateurs ont fait des retours positifs sur ces premiers travaux et ils ont proposé plusieurs points d'amélioration. Les orientations du projet de thèse ont été discutées pour les deux prochaines années.

Présentation du sujet

Le sujet de la thèse est : "Vers un système intelligent de capitalisation de connaissances pour l'agriculture durable : Construction d'une base de connaissances agricoles par transformation de sources existantes et interrogation intelligente de données".
Ce sujet pointe deux objectifs : le premier est la construction d'une base de connaissances agricoles à partir de sources existantes et le deuxième est l'interrogation de cette base à l'aide du système SWIP développé à l'IRIT. La première année de thèse de Fabien a été effectuée à l'Irstea de Clermont-Ferrand. Son travail s'est centré sur l'élaboration d'une méthode pour construire une base de connaissances agricoles à partir de deux types de sources : thésaurus et bases de données. Cette méthode distingue une première étape de filtrage des informations extraites et ensuite une validation de la conceptualisation issue de l'extraction. Pour cette validation, une approche utilisant des scores de confiance est envisagée afin d'obtenir un consensus entre les différentes sources en entrée.

Retours sur la première année

Les retours sur cette première année furent globalement positifs. Les idées présentées par Fabien ont été jugées intéressantes. La première année fut riche : plusieurs articles ont été publiés, dont un dans une conférence internationale. Le développement d'un prototype est en cours et des premiers résultats sont déjà visibles. Les évaluateurs ont noté qu'une bonne dynamique est enclenchée. Néanmoins plusieurs points restent à améliorer.

Manque de conceptualisation et de formalisation

Les idées présentées lors de ce comité furent abordées sous un aspect trop “ingénierie”. Une approche orientée ingénierie n’est pas un problème en elle-même. Néanmoins, le travail demandé lors d’une thèse se doit d’avoir un niveau d’abstraction supérieur. Fabien doit présenter ses travaux sous un aspect plus conceptuel pour formaliser ses propositions. Ceci permettra de donner plus de valeur à ses contributions en s’appuyant sur des théories déjà existantes, notamment des travaux concernant la théorie des graphes. C’est aussi le cas pour la validation conceptuelle des informations qui doit s’appuyer sur des théories déjà existantes en bases de données ou en logique floue.

Pointer les objectifs

L’objectif des travaux présentés étant encore assez large, il est difficile d’en identifier les points forts. La présentation fait état d’un certain nombre de problématiques sans pour autant cibler les problématiques auxquelles la thèse apportera une réponse. Il faudra donc que Fabien, en lien avec ses encadrants, détermine précisément les points forts qui seront les contributions défendues dans sa thèse. Fabien devra aussi clairement présenter ses travaux sous l’angle applicatif agricole. Il pourra ainsi plus facilement justifier les points forts de la thèse mais aussi les raisons de la construction d’une base de connaissances dans le domaine de l’agriculture. En effet, cette construction se fait pour un objectif fixé au préalable : la participation au LOD agricole concernant les observations d’attaques de bio-agresseurs (champignons, maladies, …) sur les grandes cultures en France.

Limitation du sujet

Le planning initial de la thèse prévoyait une participation au projet SWIP pour améliorer l’interrogation des données dès la deuxième année. Cependant, les travaux sur le premier objectif ne sont pas aboutis ; il n’est donc pas possible d’aborder l’interrogation des données. La partie construction de la base de connaissances étant plus complexe que prévue, apporter une contribution sur toutes les problématiques soulevées semble très difficile. Les membres du comité étaient tous d’accord pour affirmer qu’une limitation du sujet devait être opérée. Les travaux sur l’ajout des agrégats au projet SWIP sont donc mis en suspens pour l’instant. L’objectif est de finaliser la construction de la base convenablement avant de l’interroger avec le système SWIP. L’interrogation avec le système SWIP devient un des usages de la base agricole, permettant de justifier la méthode de construction.

Clarifier le vocabulaire utilisé

Le vocabulaire utilisé par Fabien était encore trop ambigu. Il est nécessaire que Fabien travaille sur cet aspect en posant très clairement ses définitions des différentes notions utilisées dans sa thèse. Ce manque de clarté du vocabulaire entraîne une difficulté de compréhension de ses propositions. C’est notamment le cas pour la définitions de “donnée”, “information”, “connaissance” ou “relation”. Le fait de travailler sur plusieurs sources de données différentes (thésaurus, BD, ontologie etc…) pour construire la base agricole implique que plusieurs communautés scientifiques seront approchées dans la thèse de Fabien : terminologie, intégration de données, recherche d’information, etc. Il sera donc nécessaire de bien clarifier les

notions partagées par chacune de ces communautés scientifiques. Par exemple JP Chevallet a demandé “qu’est ce qu’un terme?”. Un terme dans un thésaurus est un terme d’indexation. Un terme dans une terminologie est un terme utilisé dans une langue dit “terme d’usage”. Les termes d’indexation ne sont pas tous des termes d’usage. Il est aussi important d’utiliser un certain nombre de mots clefs pour bien marquer que cette thèse appartient au domaine du web sémantique. Par exemple, durant sa présentation, Fabien n’a pas parlé de “raisonnement”, alors que c’est une notion très importante dans ce domaine.

L’évaluation

Chantal Reynaud a soulevé le problème de l’évaluation qui n’a pas été évoqué durant la présentation. L’analyse des résultats obtenus grâce au prototype développé n’est pas une question simple. Fabien devra réfléchir à une méthode d’évaluation pour pouvoir confirmer son approche. Un premier aspect à vérifier est que la méthode d’enrichissement produit bien une base de connaissances répondant à des critères de qualité qu’il faudra définir. Plusieurs métriques devront alors être proposées pour évaluer ces critères. Une autre idée qui a émergé des discussions est d’utiliser le système SWIP et de proposer cette interface d’interrogation à des experts du domaine. Un cadre d’évaluation devra être défini de façon à quantifier l’apport pour les agriculteurs (utilisateurs réels) de l’interrogation de la base de connaissances constituée. Les résultats de l’interrogation de la base pourront par exemple être comparés aux résultats d’un système de recherche d’information documentaire.

Orientation du projet de thèse

Ces différentes remarques ont abouti à des objectifs à court et moyen termes pour la suite de la thèse de Fabien. Tout d’abord, comme évoqué précédemment, l’objectif principal pour l’instant est la finalisation de la construction de la base de connaissances agricoles. L’amélioration du système SWIP pour l’interrogation à l’aide d'agrégats ne sera traitée que si Fabien en a le temps. Sa priorité sera pour l’instant d’orienter ses recherches sur la construction de la base de connaissances. Fabien devra ensuite se concentrer sur la définition des différentes étapes nécessaires dans l’élaboration de sa méthodologie d’enrichissement de base de connaissances, c’est à dire qu’il devra bien spécifier tous les processus nécessaires à la construction de la base de connaissances. Une fois ces processus définis, il faudra qu’il choisisse quelles étapes des processus constitueront les contributions de sa thèse. Afin de se rendre compte de la complexité des étapes, Fabien devra effectuer une transformation manuellement, sur un exemple simple. En parallèle, il devra travailler sur l’élaboration d’une liste de définitions précises des concepts abordés dans ses travaux. Il continuera également le développement du prototype de filtrage pour obtenir des résultats probants. Enfin il est nécessaire que Fabien se fasse une culture scientifique plus générale concernant certains sujets abordés, tels que la théorie des graphes. Pour cela, une lecture d’articles scientifiques de synthèse peut être un atout pour améliorer la formalisation de ses travaux. L’état de l’art ne remonte pas assez loin ; il est pourtant intéressant d’étudier d’anciens systèmes experts ou de traitement automatique du langage naturel. Deux notions sont abordées dans ses propositions mais restent encore à préciser : la confiance et l’alignement. Il est donc nécessaire de lire des

papiers les concernant pour déterminer quelles théories peuvent être utilisées dans sa thèse. Néanmoins, il est important de préciser que la thèse de Fabien n'est pas dirigée vers l'un ou l'autre de ces deux domaines. Pour ces deux aspects, les évaluateurs ont proposé un certain nombre de références (sur la qualité dans les bases de données, qui peut s'apparenter à la confiance ici, et une thèse sur l'alignement) qui peuvent intéresser Fabien dans ses recherches.

Conclusion

Le comité de thèse de fin de première année de Fabien Amarger aura donc permis de pointer du doigt les améliorations à apporter à son projet de thèse. Fabien doit finaliser et valider sa méthode d'enrichissement de base de connaissances avant d'aborder la problématique de l'interrogation. Pour faciliter la compréhension, Fabien doit améliorer la formalisation de ses propositions et clarifier le vocabulaire employé. Pour conclure, le comité est satisfait des travaux présentés et a remarqué qu'une bonne dynamique était enclenchée.