Compréhension de recettes de cuisine utilisateurs par extraction de connaissances intrinsèques

Damien Leprovost, Thierry Despeyroux, Yves Lechevallier

*Inria – Rocquencourt, Équipe-projet AxIS, BP 105
78153 Le Chesnay Cedex, France
{prénom.nom}@inria.fr

Résumé. Sur les sites Web communautaires, les utilisateurs échangent des connaissances, en étant à la fois auteurs et lecteurs. Nous présentons une méthode pour construire notre propre compréhension de la sémantique de la communauté, sans recours à une base de connaissances externe. Nous effectuons une extraction de la connaissance présente dans les contributions analysées. Nous proposons une évaluation de la confiance imputable à cette compréhension déduite, afin d'évaluer la qualité du contenu, avec application à un site Web de partage de recettes de cuisine.

1 Introduction

Le Web 2.0 favorise le développement des sites collaboratifs, où les utilisateurs échangent des connaissances, se structurent en communautés, développent des codes, des usages et une sémantique qui leurs sont propres. Dans le cadre de la recherche traditionnelle de fouille de la connaissance, cette évolution en relative autonomie peut se révéler problématique : il n'existe aucune garantie que cette dernière se structure autour d'une sémantique qui soit en adéquation avec les bases de connaissance de référence traditionnelles. La pertinence des conclusions peut alors ne pas ou peu refléter l'évolution réelle du comportement des utilisateurs et de la sémantique de leurs échanges.

Nous proposons une méthode pour construire notre propre compréhension des contributions des utilisateurs, basée uniquement sur les données de celles-ci, afin d'extraire la sémantique des utilisateurs. Nous évaluons cette approche par la mesure d'une valeur de confiance. Nous effectuons notre analyse dans le contexte des recettes de cuisine, dont les sites de partage communautaires sont nombreux et très populaires sur le Web français comme mondial.

2 État de l'art

La recette de cuisine est un type de données particulier, composé d'un ensemble d'ingrédients et de procédures d'exécution. Ce type de données est exploité par de nombreux systèmes de recommandation. Le *Cooking Assistant* (Sobecki et al., 2006) définit un système de recommandation démographique basé sur une inférence à logique floue, efficace pour fournir une réponse globale à un besoin général. Mais la généralisation des caractéristiques conduit à une

recommandation également généralisée. Pour prendre en compte la spécificité des ingrédients, Freyne et Berkovosky utilisent la relation de composition qui existe entre ingrédients et recettes pour propager des évaluations et déterminer un comportement utilisateur (Freyne et al., 2011). Cela nécessite néanmoins une phase constante de normalisation, un travail d'expert consistant à vérifier ou annoter les ingrédients afin qu'ils correspondent à une liste de référence.

Les recettes de cuisine ont également été traitées par des approches de raisonnement à partir de cas. Le système *CHEF* (Hammond, 1986) est un système d'adaptation par la critique, qui permet de prendre en compte la spécificité du type de données qu'est l'ingrédient, en relevant les problèmes découlant d'une substitution. En revanche, une importante phase d'apprentissage est requise. Le système *MIKAS* (Khan et Hoffmann, 2003) propose de contourner ce besoin par un recours à l'expert. Cette aspect de la transmission de connaissance de l'expert au système par l'expérience plutôt que par le déclaratif est vu comme plus efficace et plus adapté aux conditions réelles. Il ne permet toutefois pas une évaluation indépendante des contenus, car dépendant des connaissances propres de l'expert.

3 Extraction de l'information et structure a priori

La première étape consiste à extraire l'information depuis des lignes d'ingrédients librement saisies. Pour exploiter ces données, nous définissons la structure a priori comme étant : *quantité – unité – ingrédient*. En fonction de l'existence d'une valeur dans les champs *quantité* et *unité* nous obtenons les classes 1, 2, 3 et 5 du tableau 1, la classe 5 regroupant les lignes où il a été impossible d'extraire de l'information pour les champs *quantité* et *unité*. Cette classe est ensuite ventilée par une phase d'apprentissage en deux étapes :

- Recherche des incohérences dans les éléments identifiés : la présence d'un ingrédient complexe en classe 1 permet de mettre en évidence une erreur de détection dans les autres classes. Par exemple, la ligne « 500g de corned beef » permet d'identifier la ligne « corned beef » comme *ingredient seul*, corrigeant la première identification de « corned » comme *quantificateur* [1].
- Pour toutes les lignes de la classe 5 restantes, nous cherchons à les faire correspondre aux cas précédemment rencontrés.

Dans le cas du jeu de données Marmiton, la phase d'apprentissage réduit la classe 5 de 20 à 1,9 % de la population (figure 1).

4 Évaluation de la confiance des ingrédients et des recettes

Ordonnant nos ingrédients en fonction de leur fréquence, nous observons sur la fonction cumul des ingrédients un effet longue traîne, phénomène commun à bon nombre de sites sociaux à usages libres. Eu égard à la distribution en loi de puissance de nos données sociales, nous appliquons à notre modèle le principe de Pareto, où 80 % des effets sont le produit de 20 % des causes. Dans le cas où les a ingrédients contenus dans 80 % des lignes représentent moins de 20 % des ingrédients les plus fréquents, nous définissons $b = 5a$ ingrédients comme ensemble représentatif des ingrédients. La valeur de confiance C_i est alors attribuée à chaque

1. Une *unité* sans présence de *quantité* est appelée *quantificateur*.

classe	ligne
1	quantité–unité–ingrédient
2	quantité–ingrédient
3	quantificateur–ingrédient
4	ingrédient seul déduites
5	ingrédient seul supposées

TAB. 1 – *Classes de lignes analysées*

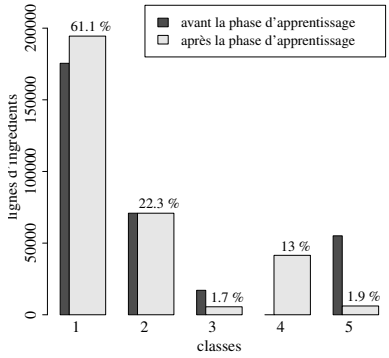

FIG. 1 – *Distribution des classes*

ingrédient, en fonction de sa fréquence d'utilisation par la fonction :

$$C_i = \begin{cases} 1 & \text{si} \quad i < a \\ \frac{i-b}{a-b} & \text{si} \quad a < i < b \\ 0 & \text{si} \quad i > b \end{cases} \tag{1}$$

La figure 2 illustre la confiance ainsi calculée des ingrédients de Marmiton. La valeur de confiance par recette C_x est égale à :

$$C_x = \frac{\sum_{i \in I_x} (\mu_{u(i,x),i}) * C_i}{||I_x||}$$

où I_x est l'ensemble des ingrédients de la recette x et $u(i,x)$ est l'unité associée à l'ingrédient i dans la recette x. $\mu_{u(i,x),i}$ est alors la fréquence de l'unité $u(i,x)$ dans les lignes contenant l'ingrédient i et C_i la confiance de l'ingrédient i.

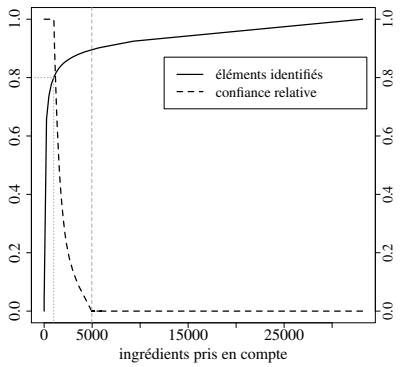

FIG. 2 – *Confiance ingrédient*

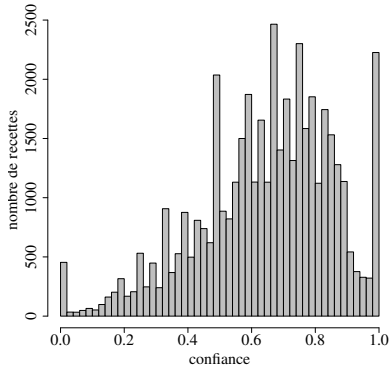

FIG. 3 – *Confiance par recette*

5 Expérimentation

Pour valider notre approche, nous avons développé un robot d'indexation et un analyseur syntaxique dédié pour parcourir le site Marmiton.org, collectant 44 169 recettes distinctes. Ces recettes présentent 354 856 lignes d'ingrédients, qui elle-mêmes comportent 33 177 ingrédients distincts. La figure 3 donne la distribution des valeurs de confiance par recettes.

6 Conclusion et travaux futurs

Nous avons présenté une méthode pour évaluer la confiance d'une publication utilisateur, comme étant la probabilité qu'un autre utilisateur en saisisse la sémantique. Notre approche est indépendante de toute base de connaissance externe, afin de raisonner directement sur les termes manipulés. Cette méthode présente l'avantage de ne pas être dépendant de la langue, ni de souffrir des problèmes de pertinence ou de couverture relatifs aux bases de connaissance. Nous projetons d'exporter la connaissance extraite des contributions utilisateurs, ce qui permettra de définir sans apport extérieur l'ontologie du système analysé, ou d'enrichir une base extérieure. Enfin, l'application de méthode de partitionnement de fouille de données, guidées par nos mesures de confiance, permettra d'évaluer une structure interne de la sémantique du système et des relations déductibles qui existent entre les différents ingrédients (proximité) ou recettes (variantes, alternatives).

Références

Freyne, J., S. Berkovsky, et G. Smith (2011). Recipe recommendation : Accuracy and reasoning. In *User Modeling, Adaptation, and Personalization, 19th International Conference*, Girona, Spain, pp. 99–110. Springer.

Hammond, K. J. (1986). Chef : A model of case-based planning. In *AAAI*, pp. 267–271.

Khan, A. S. et A. Hoffmann (2003). Building a case-based diet recommendation system without a knowledge engineer. *Artificial Intelligence in Medicine 27*(2), 155–179.

Sobecki, J., E. Babiak, et M. Slanina (2006). Application of hybrid recommendation in web-based cooking assistant. In *Knowledge-Based Intelligent Information and Engineering Systems, 10th International Conference, Proceedings, Part III*, pp. 797–804. Springer.

Summary

On community websites, users share knowledge, being both authors and readers. We present a method to build our own understanding of the semantics of the community, without the use of any external knowledge base. We perform a by knowledge extraction from analyzed contributions. We propose an evaluation of the trust attributable to that deduced understanding to assess the quality of user content, on cooking recipes provided by users on sharing websites.