Catégorisation et Désambiguïsation des Intérêts des Individus dans le Web Social

Coriane Nana Jipmo*, Gianluca Quercini*
Nacéra Bennacer*

*Laboratoire de Recherche en Informatique (LRI)
CentraleSupélec, Univ. Paris-Saclay
Gif-sur-Yvette, France
{coriane.nanajipmo, gianluca.quercini, nacera.bennacer}@lri.fr

Résumé. Cet article présente une approche pour la catégorisation et la désambiguïsation des intérêts que les individus renseignent sur les réseaux sociaux en utilisant Wikipédia.

1 Introduction

Dans cet article, nous présentons une étude préliminaire sur le problème de caractérisation et de catégorisation des intérêts que les individus renseignent sur les réseaux sociaux, comme *reading*, *jogging*, *java*, etc. Ces intérêts étant exprimés en langage naturel, nous sommes confrontés au problème de la désambiguïsation dans un contexte limité, compte tenu du peu des informations accessibles dans les profils des individus. Les approches visant la désambiguïsation des *tags* dans les *folksonomies* sont confrontées aux même problème, bien qu'elles peuvent s'appuyer sur les ressources faisant l'objet des tags pour avoir un contexte plus riche [Garcia-Silva et al. (2012)]. Nous explorons une approche permettant de désambiguïser un intérêt d'un individu par la détermination d'un article Wikipédia qui contient la description de celui-ci. La désambiguïsation d'un intérêt se fera en utilisant les autres intérêts renseignés par l'individu comme contexte. Les résultats que nous avons obtenus sur 392 intérêts issus de 50 profils utilisateurs du réseau social *LiveJournal* sont encourageants.

2 Désambiguïsation des intérêts

Soit $\mathcal{I}_u = \{I_1, I_2, ..., I_n\}$ l'ensemble des intérêts qu'un individu α renseigne sur son profil, exprimés sous la forme d'une chaîne de caractères en Anglais (*Computer*, *Music*, ...). Notre objectif est d'associer à chaque intérêt I_j, $j = 1, ..., n$, un article Wikipédia décrivant I_j ; à cet effet, pour chaque intérêt I_j notre approche sélectionne la page Wikipédia P ayant par titre I_j, si elle existe. Deux cas peuvent se présenter : (i) P est un article décrivant I_j, ou (ii) P est une page de désambiguïsation. Dans le premier cas, le mot décrivant l'intérêt I_j n'est pas ambigu (par ex., *Music*), au point que la majorité des utilisateurs Wikipédia ont trouvé un accord quant à son *interprétation* (ou, signification) par défaut (*Music* désigne une forme d'art) et lui ont associé un article. La page P sera alors choisie comme la seule interprétation de l'intérêt I_j.

Dans le deuxième cas, le mot décrivant l'intérêt I_j est ambigu et la page de désambiguïsation P permet d'avoir la liste des articles Wikipédia représentant les interprétations possibles de I_j. Afin de choisir une interprétation pour les intérêts ambigus, notre approche construit un *graphe des interprétations* $\mathcal{G}$ comme suit. Pour chaque intérêt I_j, on ajoute un nœud dans $\mathcal{G}$ pour chaque interprétation de I_j ; un arc est établi entre deux nœuds correspondant à des interprétations de deux intérêts différents dont la similarité dépasse un seul fixé τ. La similarité de deux nœuds est calculée dans le graphe de Wikipédia en utilisant la mesure de similarité WLM [Milne et Witten (2008)]. Ensuite, l'algorithme *PageRank* est utilisé pour affecter un score d'importance à chaque nœud de $\mathcal{G}$; à chaque intérêt on affecte son interprétation ayant le score le plus élevé. L'intuition derrière l'utilisation de *PageRank* sur le graphe des interprétations est que l'interprétation de chaque intérêt vote pour les interprétations similaires des autres intérêts ; la co-occurrence de deux intérêts similaires (par ex., *c++* et *Java*) est donc prise en compte pour choisir leurs interprétations correctes.

3 Evaluation

Nous avons collecté 50 profils du réseau social *LiveJournal* avec un total de 392 intérêts ; 257 intérêts distincts ont un article Wikipédia (page par défaut et/ou page de désambiguïsation), 36 n'ont aucun article. Les interprétations correctes de chaque intérêt ont été déterminées manuellement par deux évaluateurs. Les premiers résultats montrent que la page par défaut correspondant à un certain intérêt est très souvent l'interprétation correcte de l'intérêt et l'interprétation correcte d'un intérêt ambigu figure normalement parmi les trois meilleurs résultats proposés par notre approche.

Références

Garcia-Silva, A., O. Corcho, H. Alani, et A. Gomez-Perez (2012). Review of the State of the Art : Discovering and Associating Semantics to Tags in Folksonomies. *The Knowledge Engineering Review 27*(01), 57–85.

Milne, D. et I. H. Witten (2008). An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *In Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence : an Evolving Synergy*, pp. 25–30. AAAI Press.

Summary

In this paper, we present a preliminary study of the problem of characterizing and categorizing the interests (e.g., *reading*, *jogging*, *java*) that the individuals disclose on social networks. As these interests are expressed in natural language, we are confronted with a disambiguation problem in a limited context, as social network profiles have usually limited textual content. The approach we present here disambiguates an interest of an individual by determining a Wikipedia article that describes it; the other interests disclosed by the individual form the context. The results that we obtained on 392 interests of 50 user profiles of the social network *LiveJournal* are encouraging.