Annotation sémantique de clusters

Nicolas Fiorini[1], Sebastien Harispe[1], Sylvie Ranwez[1], Jacky Montmain[1], Vincent Ranwez[2]

[1] Centre de recherche LGI2P de l'École des mines d'Alès, site de Nîmes, Parc G. Besse, F-30035 Nîmes cedex 1
{prenom.nom}@mines-ales.fr

[2] UMR AGAP, Montpellier SupAgro/CIRAD/INRA, 2 place Pierre Viala, F-34060 Montpellier
ranwez@supagro.inra.fr

Mots-clés : *Annotation sémantique, clustering.*

1 Introduction

L'annotation de clusters est un traitement important pour l'interprétation des résultats de clustering [1]. Elle peut se faire en tenant compte uniquement des informations propres à chaque cluster (c'est l'annotation interne) ou bien en tenant compte des autres clusters (c'est l'annotation différentielle) [2]. Ces deux points de vue peuvent cohabiter dans la phase de clustering même, où certaines approches vont se focaliser sur les similarités entre les documents alors que d'autres utiliseront aussi leurs différences. La littérature souligne que l'approche différentielle fournit dans la plupart des cas des résultat plus pertinents [2]. Ainsi, bien que d'une complexité algorithmique plus importante, cette approche est généralement préférée. Nous présentons ici les premiers résultats relatifs à la définition d'une approche (hybride) pour l'annotation de clusters composés de documents caractérisés par une représentation de connaissance (e.g., ontologie).

La principale contribution visée dans cette étude porte sur la modélisation des deux critères à maximiser : (i) la *pertinence* des labels associés aux clusters sous une contrainte de (ii) *différentiation* des clusters.

2 Méthode

Puisque nous considérons des documents annotés sémantiquement, il est possible d'utiliser la notion de similarité sémantique [3] afin d'évaluer à quel point une annotation représente un cluster et le différencie des autres.

2.1 Annotation interne

Nous définissons $\mathcal{C}$ un ensemble de concepts partiellement ordonnés dans une ontologie, $\mathcal{D}$ un ensemble de documents indexés par une fonction $index : \mathcal{D} \rightarrow \mathcal{P}(\mathcal{C})$ où $\mathcal{P}$ désigne une partition, et $\mathcal{G}$ un ensemble de clusters (groupes) G composés d'un ensemble de documents, i.e. $G \subseteq \mathcal{P}(\mathcal{D})$. L'objectif de l'annotation (interne) de cluster est de définir une fonction $annot_{int} : \mathcal{C} \times \mathcal{G} \rightarrow \mathbb{R}$. Nous considérerons par la suite le postulat précisant que l'annotation A_G d'un cluster $G \in \mathcal{G}$ est optimale lorsqu'elle maximise cette fonction, qui est la similarité moyenne avec les concepts qui indexent les documents du cluster, soit :

$$A_G = \arg\max_{A \subseteq \mathcal{C}} \left(annot_{int}(A, G) \right), \quad annot_{int}(A, G) = \frac{1}{|G|} \times \sum_{d \in G} sim(A, index(d)) \qquad (1)$$

La fonction $sim : \mathcal{P}(\mathcal{C}) \times \mathcal{P}(\mathcal{C}) \rightarrow [0, 1]$ mesure la similarité sémantique de deux groupes de concepts ; de nombreuses formulations ont été proposées dans la littérature [3].

2.2 Annotation différentielle

Nous considérons que les clusters forment une partition, i.e. les n clusters sont disjoints et leur union correspond à $\mathcal{C}$. On s'attend tout naturellement à ce que la stratégie d'annotation des clusters soit respectueuse de cette partition. Il est donc essentiel que les annotations des clusters soient toutes distinctes et plus généralement qu'elles soient aussi différentes les unes des autres que possible. Cette contrainte n'est pas définie dans l'Equation 1. On suppose que l'on dispose d'une mesure de dissimilarité de deux annotations conceptuelles de cluster A_G et A'_G. A priori cette mesure est symétrique, et on la notera donc $d(A_G, A'_G)$. L'annotation optimale d'un cluster n'est plus définie de manière autonome comme pour $annot_{int}$, le problème est maintenant de trouver un ensemble d'annotation qui soient globalement optimale. Etant donner un ensemble d'annotation $\mathcal{A} = A_1, ..., A_i, ..A_n$ avec $n = |\mathcal{G}|$, on peut estimer la pertinence globale de cette annotation par rapport au partitionement $\mathcal{G}$ des documents par :

$$pertinence(A_1, ..., A_i, ..A_n|\mathcal{G}) = (n-1) \sum_{1 \leq i \leq n} sim(A_i, G_i) + 2\lambda \sum_{1 \leq i < j \leq n} d(A_i, A_j) \qquad (2)$$

La recherche des annotations optimales des clusters revient alors à chercher les n annotations A_i qui maximisent la fonction pertinence ci-dessus. C'est donc un compromis, géré via le paramètre λ, entre la pertinence de A_i pour annoter le seul cluster G_i et la distinction entre l'annotation de ce cluster et les annotations des autres clusters.

Cette équation n'est pas sans rappeler la proposition de Gollapudi et al. [4], *max-sum diversification*, dans le domaine de la diversification des résultats en recherche d'information :

$$S^* = \arg\max_{S \subseteq U} \left((k-1) \sum_{u \in S} w(u) + 2\lambda \sum_{u,v \in S} d(u,v) \right) \qquad (3)$$

où S est un sous-ensemble de l'univers U de documents du corpus, $k = |S|$, $w(u)$ représente la pertinence du document u pour la requête donnée, $\lambda > 0$ est un paramètre ajustant l'importance de la diversité face à la pertinence et $d(u, v)$ est une distance entre deux documents u et v.

3 Conclusions et perspectives

L'Équation 2 présente une façon de trouver, pour chaque cluster, une annotation pertinente tout en contrôlant sa différence avec les annotations des autres clusters. L'aspect combinatoire de cette fonction peut être résolu par l'utilisation d'une heuristique s'inspirant de celle présentée par Gollapudi et al [4] pour implémenter leur approche *max-sum diversification*. La particularité de notre méthode réside dans le fait que nous utilisons des données annotées sémantiquement, nous permettant d'apprécier la similarité ou la distance entre deux documents, deux clusters, etc. Cette approche devrait donc pouvoir être étendue à l'annotation de clusters hiérarchisés, pour lesquels le niveau d'abstraction de l'annotation devra prendre en compte la spécificité du cluster dans la hiérarchie.

Références

[1] F. Role and M. Nadif. Beyond cluster labeling : Semantic interpretation of clusters' contents using a graph representation. Knowledge-Based Systems, 56, 141–155, 2014

[2] C. D. Manning, P. Raghavan and H. Schütze. Introduction to information retrieval (Vol. 1, p. 363). Cambridge university press, 2008

[3] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi and J. Montmain. A framework for unifying ontology-based semantic similarity measures : A study in the biomedical domain. Journal of Biomedical Informatics, Volume 48, Elsevier, pp. 38–53, April 2014

[4] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. Proceedings of the 18th International Conference on World Wide Web - WWW '09, 2009

View publication stats