Similarité entre listes de termes dépendants et pondérés

Michel Leclère, Michaël Thomazo, Michel Chein

Similarité entre listes de termes dépendants et pondérés

M. Chein, M. Leclère, M. Thomazo

15 janvier 2015

1 Introduction

Il est classique de représenter ce dont parle un document comme une liste pondérée de termes (mots ou concepts). Les termes peuvent être des mots ou des concepts et le poids d'un terme représente l'importance de la notion dans le document. Souvent, une telle description du "contenu" d'un document est considérée comme un vecteur (le coefficient d'un terme étant calculé par tf-idf) et une mesure de corrélation entre vecteurs, comme le cosinus, est utilisée pour mesurer la similarité de deux documents relativement à leur contenu. Une hypothèse sous-jacente à une telle approche est que les termes sont indépendants (cf. Gerard Salton, A. Wong, and C. S. Yang. A vector space model for information retrieval. Communications of the ACM, 18(11) :613–620, November 1975. pour le "Vector Space Model" toujours utilisé depuis 1975). Dans certains domaines, où les objets considérés sont également représentés par des listes de termes pondérés, l'hypothèse d'indépendance entre les termes ne tient pas.

Considérons par exemple plusieurs recettes de cuisine (fictives ...).
R1 : 2 citrons, 200g de farine, 50g de sucre, une gousse de vanille, une cuillère d'huile
R2 : 2 oranges amères, 200g de farine, 50g de miel, 25g de beurre
R3 : 2 citrons, 200g de farine, 50g de sel, un piment, une cuillère de cognac
Sans être un fin gourmet on peut considérer que la recette R2 est plus proche de la recette R1 que la recette R3, ceci parce que les divers ingrédients ont des relations entre eux, certains, comme une cuillère d'huile, peuvent être (plus ou moins bien) remplacés par d'autres, comme 25g de beurre (ou le sucre par du miel) sans dénaturer complètement la recette ...

Dans cet article nous nous intéressons à des cas où les termes se ressemblent plus ou moins et où cette ressemblance est représentée par une fonction numérique σ entre termes que nous supposerons connue. Nous proposons une fonction de similarité sim entre deux listes de termes pondérés, sim étant fonction de la similarité σ entre termes et du poids des termes dans les listes.

Les poids des termes doivent aussi être pris en compte. En effet, considérons par exemple les trois descriptions suivantes : $d_1 = \{(t_1, 0.1), (t_2, 0.9)\}$, $d_2 = \{(t_1, 0.3), (t_2, 0.7)\}$, $d_3 = \{(t_1, 0.7), (t_2, 0.3)\}$, la similarité entre d_1 et d_2 devrait être plus grande que celle entre d_1 et d_3.

Nous avons empiriquement évalué la fonction proposée dans deux cas. Dans le premier les listes représentent des domaines de compétence de personnes, dans le second cas il s'agissait de classifier des sites web.

Le problème fondamental, dont est issue la présente étude, qui nous intéresse concerne l'identification d'objets lorsque ces objets sont décrits par un ensemble d'attributs, certains attributs ayant pour valeurs des listes de termes pondérés. Nous avons donc trois niveaux de similarité. Le premier niveau est celui de la similarité entre termes. Le second niveau est celui de la similarité entre valeurs d'attributs et donc, en particulier, la similarité entre listes de termes pondérés. Le troisième niveau est celui de la similarité entre objets.

Dans notre expérimentation en cours (cf. Projet Qualinca) les objets ne sont pas des recettes de cuisine mais des personnes décrites par un ensemble de méta-données (nom, dates de vie, langues, domaines de compétence, emplois, etc.). Pour chaque dimension d, par exemple une dimension temporelle peut combiner différents attributs temporels, nous définissons une fonction de similarité entre les descriptions, suivant d, de deux objets o et o'. Puis, à partir de ces fonctions de similarité, nous définissons des prédicats logiques représentant qualitativement ces similarités. Enfin, ces prédicats sont utilisées dans des règles logiques pour définir une fonction de similarité qualitative entre les objets eux-mêmes (cf. projet Qualinca).

Dans cet article nous ne nous intéressons qu'à la première étape de notre démarche c'est-à-dire que nous ne considérons qu'une seule dimension. On retrouve une telle situation en classification de documents lorsque l'on ne s'intéresse qu'au contenu des documents et c'est l'exemple que nous utilisons intuitivement dans cet article.

Dans la section **??** nous définissons plus précisément le problème lorsque la similarité entre termes est définie par une fonction numérique connue et nous proposons des propriétés qu'une mesure de similarité entre deux ensembles pondérés de tels termes dépendants devrait satisfaire. Dans la section **??** une mesure de similarité satisfaisant ces propriétés est définie comme l'optimum d'une fonction linéaire sur un ensemble décrit par des contraintes linéaires, cette fonction pouvant aussi se voir comme le coût maximal d'un flot sur un réseau.

Nous n'abordons pas ici le problème initial, celui de la construction de la similarité entre termes. Quand ces termes sont les concepts d'une ontologie de nombreuses mesures de similarité entre termes on été proposées (cf. XXXX).

2 Le problème

2.1 Ensemble de termes muni d'une mesure de similarité

On dispose d'un ensemble fini T de termes permettant de qualifier un objet suivant une certaine dimension. Par exemple, lorsque l'on s'intéresse à la dimension "contenu" (de quoi parle un document) les termes pourront être des concepts d'une ontologie, des mots d'un lexique, des termes d'une terminologie ... Lorsque l'on considère la dimension "co-auteur" d'un auteur les termes seront les noms des co-auteurs. Nous nous intéressons au cas où ces termes ne sont pas indépendants et où l'on dispose sur cet ensemble d'une mesure de similarité, notée σ qui vérifie les propriétés suivantes :

– σ est une fonction de $T \times T$ dans l'intervalle des réels $[0, 1]$;
– σ est symétrique : $\forall t, t' \in V \;\; \sigma(t, t') = \sigma(t', t)$;
– $\forall t, t' \in T \;\; \sigma(t, t') = 1$ ssi $t = t'$;

Deux termes t et t' sont dits *dissimilaires* si et seulement si $\sigma(t, t') = 0$. Intuitivement plus $\sigma(t, t')$ est grand plus t et t' sont similaires, et plus il est petit plus ils sont dissimilaires.

Observation 1 *On ne suppose ni inégalité triangulaire ni son inverse pour σ.*

Observation 2 (A VOIR) *Il faudrait peut-être dire un mot sur la construction de σ ?*

2.2 Description d'un objet par des ensembles de termes pondérés

On dispose d'un ensemble d'objets décrits suivant plusieurs dimensions. La description d'un objet o suivant la dimension d est un ensemble $d(o)$ de couples (terme, poids) vérifiant les propriétés suivantes :

Si $d(o) = \{(t_1, p_1), ..., (t_m, p_m)\}$:
– $\forall (t, p) \in d(o)$ on a $t \in T$ et $p \in]0, 1]$,
– $\forall (t, p) et (t', p') \in d(o), t \neq t'$, (chaque terme apparait au plus une fois),
–

$$\sum_{\forall (t,p) \in d(o)} p = 1$$

Une description est dite *étendue* lorsque l'on autorise des poids nuls et que tous les termes de T sont présents.

Deux objets o et o' sont dits *dissimilaires suivant d* si et seulement si pour tout $(t, p), \in d(o)$ et pour tout $(t', p'), \in d(o')$ on a $\sigma(t, t') = 0$, i.e. tous les termes de o sont dissimilaires de tous les termes de o'.

Observation 3 *on considère d(o) et pas o tout court car on veut pouvoir considérer plusieurs dimensions décrivant le même objet o, i.e. $d_1(o), d_2(o), ...$ On pourrait aussi considérer T_d pour être plus correct car un ensemble de termes dépend de la dimension.*

Observation 4 *On pourrait considérer des termes de poids nul. Le fait d'autoriser des poids nuls simplifierait certaines notations, par exemple $d(o)$ serait une application de T dans $[0, 1]$, mais ça n'entrainerait pas que les $d(o)$ soient des vecteurs à cause de la dépendance, "vecteur" implique indépendance.*

2.3 Propriétés d'une fonction de similarité entre deux descriptions

Soit deux objets o_1 et o_2, on veut définir une similarité entre $d(o_1)$ et $d(o_2)$, notée $sim_{d,\sigma}(o_1, o_2)$, qui ne dépende que de leurs descriptions suivant d et qui satisfasse au moins les propriétés suivantes :

1. c'est une fonction qui "utilise la même échelle" que la fonction σ définie sur T, donc : $sim_{d,\sigma}(o, o') \in [0, 1]$;
2. la symétrie : $sim_{d,\sigma}(o_1, o_2) = sim_{d,\sigma}(o_2, o_1)$;

3. l'identité : $d(o_1) = d(o_2)$ ssi $sim_{d,\sigma}(o_1, o_2) = 1$.
4. $sim_{d,\sigma}$ est une fonction croissante de σ, i.e. pour tout couple d'objets o, o', si $\forall t \in d(o), t' \in d(o')$ on a $\sigma(t, t') \leq \sigma'(t, t')$ alors $sim_{d,\sigma}(o, o') \leq sim_{d,\sigma'}(o, o')$
5. "Convexité". Soient $A = (a_1, ..., a_n)$ et $B = (b_1, ..., b_n)$ deux listes de n nombres réels, une liste $C = (c_1, ..., c_n)$ est dite *entre A et B* lorsque $\forall i = 1, ..., n$ on a $c_i \in [min(a_i, b_i), max(a_i, b_i]$.
   Pour tout triplet d'objets o_1, o_2 et o_3 tels que la liste des poids de la description étendue $d(o_3)$ soit *entre* la liste des poids des descriptions étendues $d(o_1)$ et $d(o_2)$ on a $sim_{d,\sigma}(o_1, o_3) \geq sim_{d,\sigma}(o_1, o_2)$.

Pour simplifier les notations, lorsque d ou σ sont fixés nous noterons sim_σ ou sim_d la fonction $sim_{d,\sigma}$.

3 La similarité comme optimum d'une fonction linéaire

Considérons les descriptions de deux objets suivant une dimension d, $d(o) = \{(t_i, p_i)|i = 1, ..., m\}$, $d(o') = \{(t'_i, p'_i)|i = 1, ..., n\}$.

Definition 1 (Relation de correspondance entre termes) *Un terme t de $d(o)$ est dit en* correspondance *avec un terme t' de $d(o')$ lorsque $\sigma(t, t') > 0$.*

Nous proposons de définir la similarité entre $d(o)$ et $d(o')$ comme une fonction des similarités entre les termes en correspondance, i.e. nous ignorons les termes absents de l'une de ces descriptions ainsi que les couples dissimilaires de termes, mais il faut aussi tenir compte des poids des termes dans les descriptions.
En effet, considérons par exemple les trois descriptions suivantes : $d_1 = \{(t_1, 0.1), (t_2, 0.9)\}$, $d_2 = \{(t_1, 0.3), (t_2, 0.7)\}$, $d_3 = \{(t_1, 0.7), (t_2, 0.3)\}$, la similarité entre d_1 et d_2 devrait être plus grande que celle entre d_1 et d_3.
Nous prenons en compte les poids des termes dans les descriptions de la manière suivante : soit $(t_i, p_i) \in d(o)$ et $(t_{i_1}, p_{i_1}), ..., (t_{i_k}, p_{i_k})$ les éléments de $d(o')$ dont les termes sont en correspondance avec t_i. On note par x_{ii_j} l'importance (le poids) de la similarité entre t_i, de $d(o)$, et t_{i_j} de $d(o')$ dans la mesure de la similarité entre $d(o)$ et $d(o')$. Nous imposons que la somme de ces importances soit inférieure à l'importance de t_i dans $d(o)$, i.e. à son poids p_i. On fait de même pour chaque terme de $d(o')$.

Les précédentes propositions conduisent aux définitions suivantes.

Definition 2 (Similarité entre descriptions) *La fonction de similarité entre $d(o)$ et $d(o')$ est définie par :*
$sim_d(d(o), d(o')) = max(\sum(x_{ij} \times \sigma(t_i, t_j))$*, somme prise pour l'ensemble des couples* (t_i, t_j)*,* $t_i \in d(o)$*,* $t_j \in d(o')$*, et* t_i *et* t_j *en correspondance,*
sous les contraintes linéaires suivantes :
pour chaque (t_i, p_i) *de* $d(o)$ *si* $t_{i_1}, ..., t_{i_k}$ *sont les termes de* $d(o')$ *en correspondance avec* t_i *on a :*
$x_{ii_1} + ...x_{ii_k} \leq p_i$*, de même,*
pour chaque (t_j, p_j) *de* $d(o')$ *si* $t_{j_1}, ..., t_{j_r}$ *sont les termes de* $d(o)$ *en correspondance avec* t_j *on a :*
$x_{ji_1} + ...x_{ji_r} \leq p_j$.

Ce problème d'optimisation linéaire peut être considéré comme un problème de flot de coût maximum sur le réseau défini de la manière suivante.

Definition 3 (Graphe de correspondance de deux descriptions) *Le* graphe de correspondance entre deux descriptions $d(o), d(o')$ *de deux objets* o, o' *suivant une dimension* d *est le graphe biparti* (A, B, E) *associé à la relation de correspondance entre les termes de* $d(o)$ *et ceux de* $d(o')$.

- *L'ensemble de sommets* $A = \{a_1, ..., a_m\}$ *(resp.* $B = \{b_1, ..., b_n\}$*) est en bijection avec l'ensemble des termes* $\{t_1, ..., t_m\}$ *de* $d(o)$ *(resp.l'ensemble des termes* $\{t'_1, ..., t'_n\}$ *de* $d(o')$*) en relation avec au moins un terme de* $d(o')$ *(resp.* $d(o)$*).*
- *L'ensemble des arêtes* E *représente la relation de correspondance entre les termes de* $d(o)$ *et de* $d(o')$.

Definition 4 (Réseau associé à deux descriptions) *Le réseau associé est classiquement construit en orientant les arêtes de* A *vers* B*, en ajoutant deux sommets* S *et* P *ainsi que les arcs de* S *à tout sommet de* A *et ceux de tout sommet de* B *à* P.
Les capacités et les coûts des arcs sont définis de la manière suivante :

- *un arc* (S, a_i) *a pour capacité* p_i *si* (t_i, p_i) *est l'élément de* $d(o)$ *associé à* a_i *et pour coût 0,*
- *un arc* (b_j, P) *a pour capacité* p_j *si* (t_j, p_j) *est l'élément de* $d(o')$ *associé à* b_j *et pour coût 0,*
- *un arc* (a_i, b_j) *a une capacité infinie et pour coût* $\sigma(t_i, t_j)$ *si* (t_i, p_i) *est l'élément de* $d(o)$ *associé à* a_i *et si* (t_j, p_j) *est l'élément de* $d(o')$ *associé à* b_j.

Property 1 $sim_d(d(o), d(o'))$ *est égal au coût maximum d'un flot sur le réseau associé à* $d(o)$ *et* $d(o')$.

Proof 1 *Soit* $d(o), d(o')$ *les descriptions de deux objets et* R *le réseau associé. Considérons une solution* $\{x_{ij}\}$ *correspondant à la définition de* sim_d*, i.e.*
$sim_d(o, o') = \sum_{\forall (a_i, b_j) \in R} (x_{ij} \times \sigma(t_i, t_j))$.
Pour tout $a_i \in A$ *on pose* $x_{Si} = p_i - \sum_{\forall j, (a_i, b_j) \in R} x_{ij}$ *et*
$x_{iP} = p'_i - \sum_{\forall j, (a_j, b_i) \in R} x_{ji}$ *. L'ensemble des* x *définit un flot sur le réseau R. Réciproquement si on considère un flot de coût maximum sur R sa restriction sur les arcs* (a_i, b_j) *est bien une solution du problème correspondant à la définition de* sim_d.

On a les propriétés suivantes.

Property 2 *Pour tout couple d'objets o et* o' *on a :*

1. $0 \leq sim_d(o, o') \leq 1$.
2. $sim_d(o, o') = sim_d(o', o)$.
3. $sim_d(o, o') = 0$ *ssi* $d(o)$ *et* $d(o')$ *sont dissimilaires.*
4. $d(o) = d(o')$ *ssi* $sim_d(o, o') = 1$.
5. *si* $\forall t \in d(o), t' \in d(o')$ *on a* $\sigma(t, t') \leq \sigma'(t, t')$ *alors* $sim_{d,\sigma}(o, o') \leq sim_{d,\sigma'}(o, o')$.

Proof 2 *1. En prenant la valeur 0 pour tous les arcs on a un flot de coût nul donc son coût* $sim_d(o, o')$ *est* ≥ 0*. Comme pour tout* i *et* j *on a* $\sigma(t_i, t'_j) \leq 1$ *et que* $\sum_{(a_i, b_j) \in R} x_{ij} \leq \sum_i p_i \leq 1$ *on a* $sim_d(o, o') \leq 1$.

2. *En considérant le réseau $R^{-}(d(o), d(o'))$ obtenu en inversant le sens de tous les arcs de $R(d(o), d(o'))$ on obtient la symétrie de sim_d.*
3. *Si $d(o)$ et $d(o')$ sont dissimilaires le réseau est vide donc tout flot nul. Réciproquement si un flot de coût maximum est nul c'est que le réseau est vide. En effet, si le réseau R est non vide soit (a_i, b_j) un arc de R on peut considérer le flot ayant pour valeur $min(p_i, \sigma(t_i, t'_i), p'_j)$ sur les arcs $(S, a_i), (a_i, b_j), (b_j, P)$ et nul ailleurs, ce flot a un coût non nul.*
4. *Si $d(o) = d(o')$ on considère le flot de valeur p_i sur (S, a_i), (a_i, b_i), (b_i, P), ceci pour tout i, et nul ailleurs, son coût est égal à 1, il est donc de coût maximum. Réciproquement, supposons que $sim_d(o, o') = 1$. Pour tout i on a $\sum_{j,(a_i,b_j)\in R} x_{ij} \leq p_i$ et $\sum_i p_i \leq 1$ donc tous les arcs (S, a_i) sont saturés dans un flot de coût 1. Le coût $\sigma(t_i, t'_j)$ de tous les arcs tous les arcs (a_i, b_j) de flot non nul doit être égal à 1 donc, comme $\sigma(t, t') = 1$ ssi $t = t'$ et qu'un terme apparait au plus une fois dans une description les deux descriptions sont bien égales.*
5. *Pour la monotonie il suffit de remarquer qu'un flot de coût max pour σ et un flot pour σ'.*

Property 3

La fonction sim définie ci-dessus satisfait la propriété de convexité.

Proof 3

Un ensemble de k dimensions (concernant un ensemble d'objets) munie chacune d'une fonction de similarité est *complet* si lorsque $sim_d(o, o') = 1$ pour toute dimension d alors les objets o et o' sont similaires. $sim(o, o')$ est définie comme une fonction des sim_d. (Suite en cours)