RECITAL 2004, Fès, 19-22 avril 2004

Nouvelle méthode syntagmatique de vectorisation appliquée au self-organizing map des textes vietnamiens

NGUYEN Tuan-Dang
GREYC, Université de CAEN
Campus Côte de Nacre, F-14032 Caen Cedex
tnguyen@info.unicaen.fr

Résumé :

Par ses caractéristiques éminentes dans la présentation des données, Self-Organizing Map (SOM) est particulièrement convenable à l'organisation des cartes. SOM se comporte d'un ensemble des vecteurs prototypes pour représenter les données d'entrée, et fait une projection, en conservant la topologie, à partir des vecteurs prototypes de n-dimensions sur une carte de 2-dimensions. Cette carte deviendra une vision qui reflète la structure des classes des données. Nous notons un problème crucial pour SOM, c'est la méthode de vectorisation des données. Dans nos études, les données se présentent sous forme des textes. Bien que le modèle général du SOM soit déjà créé, il nous faut de nouvelles recherches pour traiter des langues spécifiques, comme le vietnamien, qui sont de nature assez différente de l'anglais. Donc, nous avons appliqué la conception du syntagme pour établir un algorithme qui est capable de résoudre ce problème.

Mots-clefs :

Self-Organizing Map, text mining, classification, vectorisation du texte, syntagme, évaluation visuelle.

1 Introduction

La sélection des propriétés du texte est l'étape la plus importante dans le processus de vectorisation. La plupart des expérimentations publiées dans ce domaine utilisent la méthode de détermination d'un vocabulaire, qui est des mots, pour représenter le contenu du texte. Néanmoins, les langues sont normalement différentes l'une de l'autre dans leur façon et leur moyen d'exprimer les concepts. Nous trouvons, par exemple, qu'une langue peut utiliser un seul mot pour exprimer une idée quelconque tandis que l'autre ne peut qu'exprimer la même idée par un syntagme ou une proposition ou une phrase ou même plusieurs phrases. Dans une langue monosyllabique comme le vietnamien, d'après la théorie moderne de la grammaire fonctionnelle [CAO Xuan Hao 1991, 1998], à cause de la limite du nombre des mots monosyllabes, le vietnamien est oblige d'utiliser des syntagmes pour exprimer des concepts que les langues indo-européennes expriment par un mot. Par exemple, pour exprimer le concept du « vélo», en français nous utilisons un seul mot mais en vietnamien nous devons utiliser un syntagme « xe đạp ». Dans cet exemple, un syntagme simple qui représente une classe des syntagmes stables, au point de vue de la composition structurelle. En effet, un syntagme peut se composer d'un nombre imprévu des mots. Cependant, le sen du syntagme n'est pas l'intégration du sens des mots composants mais le syntagme peut avoir un sens tout à fait différent. L'idée de la création d'un dictionnaire des syntagmes stables est impossible à cause des

raisons suivantes: a) la combinaison réelle des mots pour créer des syntagmes est illimitée; b) l'utilisation du dictionnaire pour segmenter les syntagmes dans la phrase est extrêmement compliquée parce qu'il n'y a pas des signes formels distinguant les noms, les verbes, les adverbes ...

Pour former une phrase, le vietnamien utilise des syntagmes fonctionnels qui sont structurellement dynamiques. Le sens de la phrase est présenté par les idées composantes; chaque idée est réalisée par un syntagme mais non pas par des mots. Cela signifie aussi que le syntagme est réalisé dans la phrase: il peut être composé par un mot ou plusieurs mots ou même une proposition; tout cela n'est pas important ni obligatoire. Certes, un syntagme de cette conception peut s'analyser en hiérarchie des syntagmes composants. Pour trouver les unités grammaticales représentant des idées composantes de la phrase dans toute la collection des textes, nous créons un algorithme qui détermine des syntagmes et qui les analyse en syntagmes composants, jusqu'au niveau le plus bas que possible. Dans une phrase standard du vietnamien, il y a deux particules formelles, "THÌ" et "LÀ", pour distinguer, au niveau logique le plus haut de la phrase, deux syntagmes importants: thème et commentaire. Le problème difficile est que dans la plupart des phrases vietnamiennes, ces deux particules sont sous-entendues. Pour cette raison, l'algorithme doit utiliser des syntagmes déjà analysés à partir des phrases standard pour supposer des syntagmes existants dans les autres phrases.

2 Self-organizing map

L'algorithme SOM définit une projection non-linéaire à partir d'une espace de R^n à un tableau de 2-dimensions qui contient M neurones. Les vecteurs d'entrée de n-dimensions dans l'espace d'origine sont notés comme r_i et chaque neurone est connecté à un vecteur de référence de n-dimensions w_i. L'algorithme concurrentiel SOM s'appuie sur la recherche des neurones les plus convenables aux vecteurs d'entrée, par le calcul des distances entre un vecteur avec tous les vecteurs référentiels des neurones pour trouver le neurone gagnant. L'adaptation des vecteurs référentiels se passera pour le neurone gagnant et ses voisinages. Alors, les voisinages du neurone gagnant apprennent aussi à chaque vecteur d'entrée. Cet apprentissage local se répète plusieurs fois et aboutit à un ordre global. Ce dernier assure que les vecteurs proches dans l'espace d'origine de n-dimensions apparaissent aux neurones voisins sur une carte de 2-dimensions. Chaque étape de l'apprentissage se compose des étapes suivantes :

1. Choisir au hasard un vecteur d'entrée, calculer la distance entre ce vecteur et tous les vecteurs référentiels des neurones.
2. Choisir le neurone gagnant, qui a la distance la plus petite par rapport au vecteur d'entrée. La distance est prédéterminée.
3. Adapter les vecteurs référentiels au neurone gagnant *j* et à ses neurones voisins. Le voisinage du neurone gagnant est défini par la fonction Gaussienne.

3 Modèle de création de la carte pour les textes vietnamiens

3.1 Pré-traitement

M1. Méthode de seuil de la fréquence absolue: le seuil est de 50, nous choisissons 2757 mots importants.

M2. Méthode de sélection des mots utiles du Rosengren : 2757 mots ayant l'indice KF les plus élevées.

M3. Méthode de détermination des mots clés du Guiraud : 2757 mots clés les plus importantes

M4. Méthode de sélection des groupes de mots : Cette méthode se base sur la notion des courts contextes. Un court contexte du mot *m* est présenté par un mot précédé et un mot succédé du mot *m*. Nous retenons des 5,090 groupes de mots.

M5. Méthode de sélection des syntagmes généraux.

3.2 Algorithme de détermination des syntagmes généraux

ENTRÉE: Ensemble des phrases de tous les textes. Ces phrases sont d'abord décomposées par les virgules séparées entre les propositions.

1. $S = \{\}$.
2. Analyser toutes les phrases possibles en THÈME et COMMENTAIRE, par deux particules "THÌ" et "LÀ". Soit R, ensemble des phrases d'entrée qui ne sont pas encore analysées par ces deux particules. Soit *D*, un ensemble des syntagmes qui sont des THÈMES. Soit *T*, un ensemble des syntagmes qui sont des COMMENTAIREs. On appelle: $C = R + T$.
3. Pour chaque syntagme, on fait:

3a. Étendre le contexte pour syntagme s', avec s' dérivé de s par l'extraction d'un dernier mot dans la structure de s. L'extension du contexte pour s' signifie que nous cherchons toutes les distributions de s' dans tous les contextes possibles.

3b. Si le volume de contextes de s' est supérieur à un seuil, qui est de 10 dans l'expérience, on estime que s' est un syntagme. On s'arrête cette étape pour s actuel et revient à la troisième étape pour un autre s.

3c. Revenir à 3a, jusqu'au moment où s' ne peut plus être composé au moins de deux mots.
Revenir à la troisième étape pour un autre s.

4. Utiliser des syntagmes du S pour décomposer d'autres phrases dans C. $\forall c \in C$, on décompose *c* en syntagme en se basant sur les syntagmes du S:
   - *c* est considéré comme l'ensemble des formes syntagmatiques connues ou inconnues.
   - Les formes syntagmatiques inconnues dans c sont de nouvelles appliquées à la troisième étape.

SORTIE: L'ensemble *S* des formes syntagmatiques connues.

3.3 Vectorisation des textes

La vectorisation est un processus de l'affectation des poids aux éléments du vecteur qui représentent un texte. Chaque élément du vecteur est une propriété du texte. Nous utilisons la fréquence des mots pour déterminer le poids. Nous utilisons la distance Euclidienne pour calculer la similarité des deux vecteurs textuels.

3.4 Réduction de la taille des vecteurs

La méthode Random Mapping [KASKI 1998] est appliquée pour diminuer la taille des vecteurs. La taille réduite est de 100 dimensions. Dans cette méthode, un vecteur n est multiplié par une matrice

R qui dispose des valeurs initiales au hasard. La projection donne le résultat sous forme d'un vecteur des dimensions diminuées : $x = Rn$.

3.5 Construction de la carte SOM

Nous précisons les paramètres suivants:

- La carte comporte des 400 neurones, par la taille de 20 x 20.
- Le nombre des itérations dans l'algorithme : T = 100.000.
- Le voisinage du neurone gagnant est déterminé par la fonction géométrique: $h_j(N_j(t),t) = \eta(t)$.
- La fonction de la proportion d'apprentissage : $\eta(t) = \eta_{\max}(t)(1 - t/T)$, avec $\eta_{\max}$ est initialement de 50% la taille de la carte.

4 Expérimentation

4.1 Données et méthode

Nos données sont des 5,325 textes (full-text) à partir du site du Parti Communiste du Vietnam (http://www.cpv.org.vn).

4.2 Méthode d'évaluation du résultat

1. **Qualité de l'algorithme SOM**: nous utilisons l'évaluation directe par « average quantization error » et « cost function » sur toute la collection des textes.

Average quantization error : $J_{dist} = \frac{1}{L}\sum_{l=1}^{L} \min_{k=1,\dots,M} \left\| x_l - w_k \right\|^2$, L le nombre des vecteurs d'entrée.

Cost function : $E = \sum_k \left\| x_k - n_c \right\|^2 + \sum_i \sum_j h_{ij} N_i \left\| n_i - m_j \right\|^2$, N_i note des vecteurs les plus proches au vecteur référentiel m_i , $n_i = 1/N_i \sum_{x_k \in V_i} x_k$, V_k voisinage Vonoroi correspondant au vecteur référentiel m_i . Quand nous développons cette approximation, nous supposons que $h_{ij} = 1, \forall i, j$.

Qualité de la visualisation de la carte: La création de la carte a pour but de trouver des classes existantes dans les données et de les visualiser sur une carte de 2-dimensions. Pour évaluer la convergence visuelle de l'algorithme et l'avantage de la vectorisation des textes, nous disposons dix groupes des textes qui sont déjà classés par le rédacteur et qui composent de la collection. Bien que la classification humaine soit subjective, nous vérifions la convergence de chaque groupe de données homogène par essai distinct. Notons V l'ensemble des unités (neurones) qui enregistrent au moins un vecteur convergeant, nous définissons les indices suivants:

Indice I_i: l'évaluation de la cohérence moyenne d'une unité sur la carte. $I_c = \frac{\sum_{j \in V} h_j}{|V|}$, h_j est le nombre des unités $i \in V, i \neq j$ voisines de l'unité $j \in V$, h_j est à 8 au maximum et h_j est à 0 au

minimum quand *j* est une unité isolée. La cohérence moyenne n'est pas meilleure quand il y a beaucoup des unités éparses ou isolées.

Indice I_d: l'évaluation de la densité moyenne d'une unité sur la carte. $I_d = \frac{\sum_{j \in V} d_j}{|V|}$, d_j est le nombre des vecteurs convergeant à l'unité.

Indice $I_{cd} = I_c \times I_d$: évaluation de la cohérence et de la densité moyennes des régions sur de la carte.

4.3 Comparaison de cinq méthodes de vectorisation

	M1	M2	M3	M4	M5
Average quantization error	0.02546	0.0247	0.00029072	0.00024752	0.00017996
Cost function	534.65031	422.07605	39.3245502	17.17635	13.06308

Tableau 1 : Résultats de cinq méthodes de vectorisation sur toute la collection

4.4 Comparaison de la qualité de visualisation

4.4.1 Dix groupes de données

No	Nombre des textes	Sujets
1	46	La direction du Parti Communiste du Vietnam
2	152	Collection de Ho Chi Minh, volume 1
3	120	Mémoires de la prison
4	71	Lenin
5	63	Ho Chi Minh et les peuples vietnamiens
6	85	Archives du Parti Communiste du Vietnam, volume 8
7	41	Collection des informations et des textes du Parti Communiste du Vietnam
8	14	Culture révolutionnaire
9	45	Archives des tâches idéologiques
10	399	Etudes du congrès IX

Tableau 2 : Dix groupes de données

4.4.2 Evaluation par l'indice $I_{cd} = I_c \times I_d$

Groupe de données	M1	M2	M3	M4	M5
1	3.38725	3.37191	4.44727	2.41928	4.45161
2	5.34717	7.14337	8.31886	7.41942	9.88057
3	3.91404	4.5660	8.70748	27.5	7.65306
4	3.69093	3.26059	4.9819	6.02170	5.57438
5	3.76460	4.10684	9.6768	2.70703	4.49732
6	3.21898	4.88763	6.20885	3.89583	6.44315

7	1.89815	2.14204	3.96	2.87284	3.96
8	0.42857	1.35714	2.56805	0.85714	1.84722
9	2.3625	2.53472	3.6914	1.92308	4.104
10	10.91355	15.32225	20.52528	19.52008	22.12989

Tableau 3: Evaluation par l'indice $I_{cd} = I_c \times I_d$

4.5 Remarques

L'efficacité est classée: M5, M3, M4, M2, M1. M3 se place devant M4 car M3 se montre mieux que M4 dans l'évaluation indirecte (par l'indice I_{cd}), bien que M4 soit un peu plus mieux que M3 dans l'évaluation directe (par « average quantization error » et « cost function »).

5 Conclusion

En travaillant sur les textes vietnamiens, nous avons évalué et proposé de nouvelles méthodes de la sélection des propriétés du texte, de la vectorisation du texte et de la mesure qualitative de la carte. Nous constatons que la méthode universelle de la sélection des propriétés, en se basant sur un seuil de la fréquence absolue des mots, n'est pas efficace. La notion du syntagme est pratique dans le but de la présentation du contenu des documents écrits en vietnamien. La contribution de notre étude sera utile de créer des cartes utilisées par les moteurs de recherche traitant sur les textes vietnamiens dans la collection.

Références

CAO Xuan Hao (1991), Le vietnamien: essai de la grammaire fonctionnelle, volume 1. Editeur des Sciences Sociales. 254 pages.

CAO Xuan Hao (1998), Le vietnamien: quelques problèmes phonétiques, grammaticales et sémantiques. Editeur de l'Education. 752 pages.

NGUYEN Tuan Dang (2002), Texte mining des documents vietnamiens avec SOM, *mémoire du Master en informatique*, Université des Sciences Naturelles, Hồ Chí Minh ville, Việt nam.

HONKELA T., KASKI S., LAGUS K., KOHONEN T. (1997), WEBSOM self-organizing maps of document collections. *Proceedings of WSOM'97*, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6.

IIVARINEN J., KOHONEN T., KANGAS J., KASKI S. (1994), "Visualizing the clusters on the self-organizing map," in *Proc. Conf. Artificial Intell. Res.Finland*, C. Carlsson, T. Järvi, and T. Reponen, Eds. Helsinki, Finland.

KASKI, S. (1998), Dimensionality reduction by random mapping: Fast similarity computation for clustering. *Proceedings of IJCNN'98*, International Joint Conference on Neural Networks.