Tạp chí Khoa học Giáo dục Kỹ thuật, số 11
Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh

WEIGHTED COMBINATIONS OF ONTOLOGICAL FEATURES AND KEYWORDS FOR DOCUMENT CLUSTERING

Van T.T. Duong[1]

ABSTRACT

Keyword-based information processing has limitations due to simple treatment of words. In this paper, we introduce named entities as objectives into document clustering, which are the key elements defining document semantics and in many cases are of user concerns. First, the traditional keyword-based vector space model is adapted with vectors defined over spaces of entity names, types, name-type pairs, and identifiers, instead of keywords. Then, hierarchical document clustering can be performed using the similarity measure defined as a distance between the vectors representing documents. Experimental results are presented and discussed. Clustering documents by information of named entities could be useful for managing web-based learning materials with respect to related objects.

Keywords: *— named entity; latent semantics; clustering quality.*

I. INTRODUCTION

Clustering, which is to partition and group data points of similar properties together, is not only an important problem in data mining and knowledge discovery, but also a useful technique for information processing in other application areas. Traditionally, for documents or texts, it is only based on keywords (KW) occurring in documents. However, keywords alone are not adequate, because in many domains and cases named entities (NE) constitute the main content of a document. Named entities are those that can be referred to by names, such as people, organizations, and locations, ([12]) which are inherently different from words. Roughly speaking, they represent individuals while words denote general concepts, such as types, properties, and relations. A motivating example for our research work is that, given a set of documents, one may want to classify and extract groups of documents about political people, commercial organizations, or tourist attraction places. For the last group, for instance, both keywords related to tourist attraction and named entities being places mainly determine if a document belongs to that group.

There are different ontological features of named entities that can be of user interest. First, the user may want to have documents about exactly identified named entities, like the *Moscow City* in Viet Nam but not a city of the same name elsewhere[2]. Second is the case when only the name and type of entities are of concern or available, as for documents about people named *McCarthy*, while there are different people having the same name *McCarthy*. Third, one may be interested in documents about entities of a certain type, like company or city. Fourth, it is not uncommon that only entity names are significant. In short, the possible

[1] Faculty of Information Technology & Applied Mathematics, Ton Duc Thang University, Ho Chi Minh City, Viet Nam.

[2] In fact, there are different cities named *Moscow* in the world.

distinct features of named entities in question are names, types, joins of names and types, and identifiers.

In [12], the most significant entity name in a document was used as its label, based on an enhanced version of the *tf.idf* measure. Then the documents with labelling named entities of the same type were grouped together. As such, it was simply classification of documents by the types of their representative entity names, rather than clustering. Consequently, it could not produce a partition each cluster of which was a group of documents having close semantics regarding various named entities occurring in them. Meanwhile, for document searching, in [5] the authors adapted the traditional Vector Space Model (VSM) with vectors over the space of NE identifiers in the knowledge base of discourse and equally linear combination of its NE-identifier-based vector and keyword-based vector. Meanwhile, the latent semantics model proposed in [7] used both keywords and named entities as terms for a single vector space, but only entity names were taken into account. Recently, [4] introduced a multi-vector space model on all the above-mentioned NE features, then explored and evaluated the information retrieval performance of various combinations of keywords and named entities.

Another issue that is not less important to clustering is how to measure the quality of a clustering result. Traditionally, clustering quality was evaluated using two complementary measures: (1) *internal measure* that reflects the average semantic distance between data points within each cluster; the smaller the better; and (2) *external measure* that reflects the average semantic distance between the clusters; the larger the better. In [9], the authors defined the *cluster entropy* and the *class entropy* as the internal and external measures, the Overall Entropy (*OE*) as their linear combination. The smaller the overall entropy is, the better clustering quality is. Ideally, all data points in each cluster have the same label, i.e., the cluster entropy is 0, and all data points of the same label reside in the same cluster, i.e., the class entropy is 0.

Such a measure is thus based on the purity of the resulting partition itself, comprising the purity of each cluster, defined by the cluster entropy, and that of all the clusters, defined by the class entropy. We view it as an *objective measure*, for which a clustering result is not tested against a pre-determined "gold standard" one. In contrast, recently [11] proposed an information theoretic criterion, called Variation of Information (*VI*), to measure how different two partitions were. We view it as a *subjective measure*, which allows one to evaluate the clustering quality of a technique by comparing a partition generated by that technique with a corresponding partition manually constructed by humans for testing. As a significant result proved in this paper, if the cluster and the class entropies have the same weight, i.e., 0.5, these two measures are equivalent in the sense that the *VI* is as twice as the *OE*.

Until now, to our knowledge, there is no clustering method that takes into account all the above-mentioned named entity features in combination with keywords. For the paper being self-contained, Section II summarizes the basic notions and formulation of the previously proposed multi-vector space model combining named entities and keywords. Section III mathematically presents the *OE* and *VI* measures. Section IV presents our experiments and evaluation of clustering quality, with respect to these two measures, when varying the weights of the named entity and keyword components in the model. Finally, Section V draws concluding remarks and suggests further work to be investigated.

II. COMBINED ENTITY-KEYWORD VECTOR SPACE MODELS

Despite having known disadvantages, VSM is still a popular model and a basis to develop other models for information retrieval, because it is simple, fast, and its ranking method is in general either better or almost as good as a large variety of alternatives ([1]). We recall that, in the keyword-based VSM, each document is represented by a vector over the space of keywords of discourse. Conventionally, the weight corresponding to a term dimension of the vector is a function of the occurrence frequency of that term in the document, called *tf*, and the inverse occurrence frequency of the term across all the existing documents, called *idf*. The similarity degree between a document and a query is then defined as the cosine of their representing vectors.

For formally representing documents by named entity features, we define the triple (N, T, I) where N, T, and I are respectively the sets of names, types, and identifiers of named entities in the ontology of discourse. Then:

5. Each document d is modelled as a subset of $(N \cup \{*\}) \times (T \cup \{*\}) \times (I \cup \{*\})$, where '*' denotes an unspecified name, type, or identifier of a named entity in d, and
6. d is represented by the quadruple $(\vec{d}_N, \vec{d}_T, \vec{d}_{NT}, \vec{d}_I)$, where $\vec{d}_N$, $\vec{d}_T$, $\vec{d}_{NT}$, and $\vec{d}_I$ are respectively vectors over N, T, $N \times T$, and I.

A feature of a named entity could be unspecified due to the user intention, the incomplete information about that named entity in a document, or the inability of an employed NE recognition engine to fully recognize it. Each of the four component vectors introduced above for a document can be defined as a vector in the traditional *tf.idf* model on the corresponding space of entity names, types, name-type pairs, or identifiers, instead of keywords. However, there are two following important differences with those ontological features of named entities in calculation of their frequencies:

1. The frequency of a name also counts identical entity aliases. That is, if a document contains an entity having an alias identical to that name, then it is assumed as if the name occurred in the document. For example, if a document refers to *Saigon City*, then each occurrence of that entity in the document is counted as one occurrence of the name *Ho Chi Minh City*, because it is an alias of *Saigon City*.
2. The frequency of a type also counts occurrences of its subtypes. That is, if a document contains an entity whose type is a subtype of that type, then it is assumed as if the type occurred in the document. For example, if a document refers to *Saigon City*, then each occurrence of that entity in the document is counted as one occurrence of the type *Location*, because *City* is a subtype of *Location*.

The similarity degree of a document d and a document (or query) q, with respect to the named entity features, is then defined to be, where $w_N + w_T + w_{NT} + w_I = 1$:

$$w_N.cosine(\vec{d}_N, \vec{q}_N) + w_T.cosine(\vec{d}_T, \vec{q}_T) + w_{NT}.cosine(\vec{d}_{NT}, \vec{q}_{NT}) + w_I.cosine(\vec{d}_I, \vec{q}_I)$$

We deliberately leave the weights in the sum unspecified, to be flexibly adjusted in applications, depending on user-defined relative significances of the four ontological features. We note that the join of $\vec{d}_N$ and $\vec{d}_T$ cannot replace $\vec{d}_{NC}$ because the latter is concerned with entities of certain name-type pairs. Meanwhile, $\vec{d}_{NT}$ cannot replace $\vec{d}_I$ because there may be different entities of the same name and type. Also, since names and types of an entity are derivable from its identifier, products of I with N or C are not included. In brief, here we

adapt the notion of terms being keywords in the traditional VSM to be entity names, types, name-type pairs, or identifiers, and use four vectors on those spaces to represent a document for searching or clustering.

Let $\vec{d}_{KW}$ and $\vec{q}_{KW}$ be respectively the vectors representing the keyword features of d and q, as in the traditional VSM. The similarity degree of d and q is then defined as follows, where $w_N + w_T + w_{NT} + w_I = 1$ and $\alpha \in [0, 1]$:

$$sim(\vec{d}, \vec{q}) = \alpha.[w_N.cosine(\vec{d}_N, \vec{q}_N) + w_T.cosine(\vec{d}_T, \vec{q}_T) + w_{NC}.cosine(\vec{d}_{NT}, \vec{q}_{NT}) + w_I.cosine(\vec{d}_I, \vec{q}_I)] + (1 - \alpha).cosine(\vec{d}_{KW}, \vec{q}_{KW}) \quad \text{(Eq. 1)}$$

Here α represents the weight of the named entity component, and $(1 - \alpha)$ of the keyword component, in defining the similarity of the documents.

The NE-based multi-vector model can be useful for clustering documents into a hierarchy via top-down phases each of which uses one of the four NE-based vectors presented above (cf. [3]). For example, given a set of geographical documents, one can first cluster them into groups of documents about rivers and mountains, i.e., clustering with respect to entity types. Then, the documents in the river group can be clustered further into subgroups each of which is about a particular river, i.e., clustering with respect to entity identifiers. As another example of combination of clustering objectives, one can first make a group of documents about entities named *Saigon*, by clustering them with respect to entity names. Then, the documents within this group can be clustered further into subgroups for *Saigon City*, *Saigon River*, and *Saigon Market*, for instance, by clustering them with respect to entity types. Another advantage of splitting document representation into four component vectors is that, searching and matching need to be performed only for those components that are relevant to a certain query. Meanwhile, the KW-based vector is complementary to the NE-based vectors in representing the salient points in the content of a document.

III. MEASURES OF CLUSTERING QUALITY

Formally, for the objective measure Overall Entropy introduced above, suppose $C = C_1 \cup C_2 \cup \ldots \cup C_k$ is a partition on the set of N data points taking labels in the set $\{l_1, l_2, \ldots, l_{k^*}\}$. Ideally, each cluster C_i contains only data points labeled l_j. Let n_j be the total number of data points of label l_j, and n_{ij} be the number of data points labeled l_j in cluster C_i. Then, the cluster entropy E_c and the class entropy E_l are defined as follows:

$$E_c(C) = -\sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{n_{ij}}{N}\log\frac{n_{ij}}{|C_i|} \qquad E_l(C) = -\sum_{j=1}^{k^*}\sum_{i=1}^{k}\frac{n_{ij}}{N}\log\frac{n_{ij}}{n_j} \quad \text{(Eqs. 2)}$$

It can be observed that, for k-means clustering ([8]), if the pre-specified number of clusters k increases, then the class entropy tends to increase but the cluster entropy tends to decrease. Meanwhile, if the value of k decreases, then the class entropy decreases while the cluster entropy increases. So, the overall entropy is defined as a linear combination of the cluster and class entropies as below, where $\beta \in [0, 1]$ is empirically determined:

$$E(C) = \beta.E_c(C) + (1 - \beta).E_l(C) \quad \text{(Eq. 3)}$$

Meanwhile, for subjective measure Variation of Information, assume $C^* = C^*_1 \cup C^*_2 \cup \ldots \cup C^*_{k^*}$ is the pre-constructed correct partition of the dataset of discourse. The information variation between C and C^* is defined by:

$$VI(C, C^*) = H(C \mid C^*) + H(C^* \mid C) = H(C) + H(C^*) - 2I(C,C^*)$$

$$I(C,C^*) = \sum_{i=1}^{k}\sum_{j=1}^{k^*} \frac{|C_i \cap C_j^*|}{N} \log \frac{|C_i \cap C_j^*| / N}{(|C_i| / N).(|C_j^*| / N)}$$

$$H(C) = -\sum_{i=1}^{k} (\frac{|C_i|}{N} \log \frac{|C_i|}{N})$$

$$H(C^*) = -\sum_{j=1}^{k^*} (\frac{|C_j^*|}{N} \log \frac{|C_j^*|}{N}) \qquad \text{(Eqs. 4)}$$

Here $H(C|C^*)$ is referred to as *clustering conditional entropy* of C given C^*, $I(C, C^*)$ is called *clustering mutual information* between C and C^*, and $H(C)$ and $H(C^*)$ are respectively *clustering entropies* of C and C^*. As proved in [11], VI is a metric, for which $VI(C, C^*) = 0$ if and only if $C = C^*$.

IV. EXPERIMENTAL RESULTS

In the scope of this paper, for experiments we focus on the type feature of named entities, because many named entities in various texts may have the same type. That hidden ontological feature is ignored in the traditional keyword-based information processing, which affects clustering quality. That is, our experiments

are performed on vectors of the form $\alpha.\ cosine(\vec{d}_T, \vec{q}_T) + (1 - \alpha).cosine(\vec{d}_{KW}, \vec{q}_{KW})$.

The value of α is varied in the experiments to find how significant the NE and KW components are to clustering quality; $\alpha = 0$ means purely keyword-based clustering, while $\alpha = 1$ means purely named entity-based clustering.

For testing clustering quality with respect to the VI measure, we use the Reuters 21578 dataset, which contains 21,578 documents. In this dataset, the header of each document, besides its body text, has the topic tag TOPICS containing the main keywords representing the topic of the document, and the named entity tags PEOPLE, ORGS, PLACES, and EXCHANGES respectively containing the main people, organizations, places, and stock exchange agencies that the document is presumably about. Below is an example of the header of a document in this dataset:

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN"

CGISPLIT="TRAINING-SET" OLDID="12925" NEWID="742">

<DATE> 2-MAR-1987 15:46:40.19</DATE>

<TOPICS><D>grain</D><D>wheat</D></TOPICS>

<PLACES><D>usa</D><D>australia</D></PLACES>

<PEOPLE><D>lyng</D><D>yeutter</D></PEOPLE>

<ORGS></ORGS>

<EXCHANGES></EXCHANGES>

<TEXT>

<TITLE>U.S. WHEAT GROUPS CALL FOR GLOBAL ACTION</TITLE>

<DATELINE>WASHINGTON, March 2 - </DATELINE>

<BODY>....</BODY>

</TEXT>

</REUTERS>

It specifies that the document is about the topics *grain* and *wheat*, the places *USA* and *Australia*, and the people *Lyng* and *Yeutter*.

From this dataset, we select a sub-set of 500 typical documents for hard clustering experiments, such that the content of each of them is clearly about named entities of a particular type. Such a size of a testing dataset is common in clustering experiments (cf. [12]). At first, approximately 7,000 documents each of which has only one named entity tag are automatically filtered. Next, we manually select 500 documents each of which is clearly about an entity type. Some tagging errors in the original dataset are also fixed during this document selection process. Further, we employ the ontology and NE recognition engine of KIM ([10]) to automatically annotate named entities in the selected documents. Then we obtain a testing dataset for hard clustering with 4 clusters based on the named entity tags. The distribution of the 500 documents across the four NE tags is as follows:

PLACES : 195 documents
PEOPLE : 105 documents
ORGS : 129 documents
EXCHANGES : 71 documents

Clustering by entity types combined with keywords

In this experiment, on the entity type testing dataset of 500 documents and 4 clusters, *k*-means is applied with *k* varying from 2 to 10 and, for each value of *k*, α varying from 0 to 1 on 0.1 incremental steps; $\alpha = 0$ means purely keyword-based clustering, while $\alpha = 1$ means purely entity type-based clustering. That is because words like "*Saigon*" and "*Bangkok*" are different keywords while, viewed as representing named entities, they are of the same type PLACES. Figure 4.1 illustrates the *VI* and *OE* diagrams with $k = 4$, showing that the best performance is achieved at $\alpha = 0.9$, which is much better than the purely keyword-based case at $\alpha = 0$. It also shows that the *VI* measure is as twice as the *OE* measure as theoretically proved above.

Table 4.1

Clustering Measures	**α=0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
VI	2.28	1.67	1.44	1.24	1.16	1.15	1.14	1.15	1.16	**0.95**	1.31
OE	1.14	0.84	0.72	0.62	0.58	0.58	0.57	0.58	0.58	**0.47**	0.66

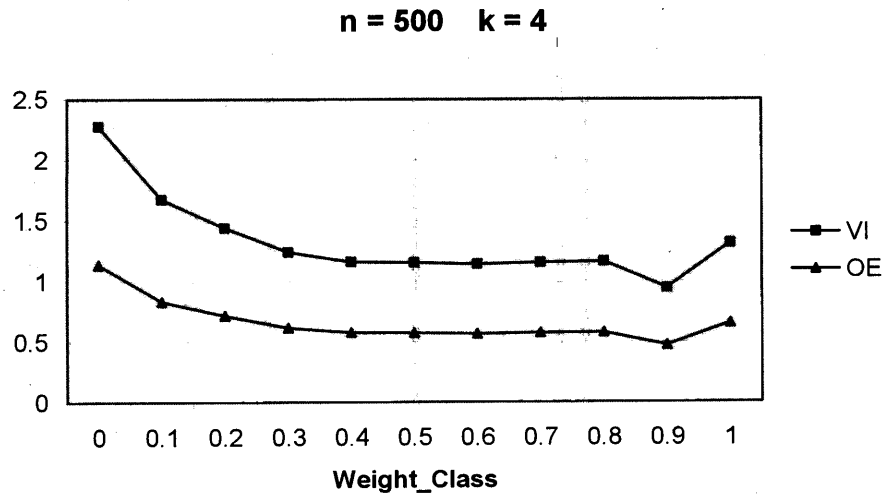

Fig. 4.1

Also, we compare the best cases for each value of k as plotted in Figure 4.2. As expected, $k = 4$ is the optimal value for the testing dataset with 4 clusters.

Table 4.2

Clustering Measures	***k*=2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
VI	1.56	1.15	**0.95**	1.17	1.44	1.57	1.76	1.85	1.88
OE	0.78	0.58	**0.47**	0.59	0.72	0.79	0.88	0.92	0.94

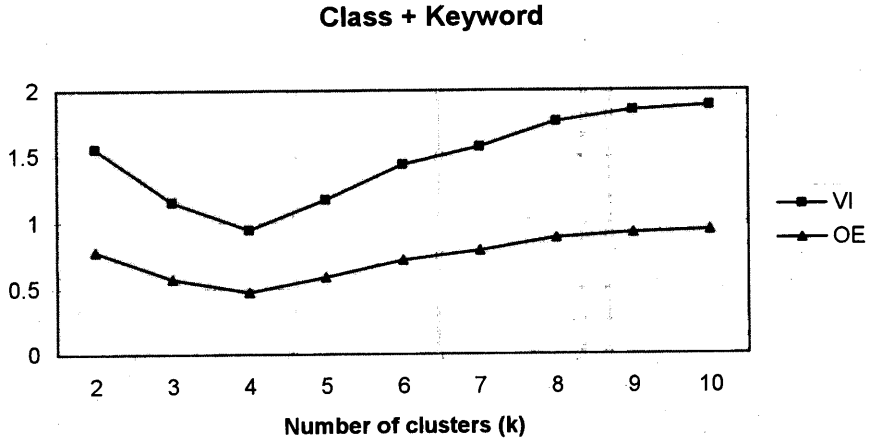

Fig. 4.2

Clustering by entity names combined with keywords

In this experiment, on the entity name testing dataset of 500 documents and 24 clusters, *k*-means is applied with *k* varying from 4 to 40 and, for each value of *k*, α varying from 0 to 1 on 0.1 incremental steps; $\alpha = 0$ means purely keyword-based clustering, while $\alpha = 1$ means purely entity name-based clustering. Figure 4.3 illustrates the *VI* and *OE* diagrams with $k = 24$. One can see that, not as for clustering by entity types, purely entity name-based clustering (i.e., $\alpha = 0$) and purely keyword-based clustering (i.e., $\alpha = 1$) are not much different in performance. That can be explained by the fact that using entity names will have advantage only if many entities in the documents to be clustered have aliases, which is apparently not the case of the selected testing dataset. Otherwise, an entity name is just like a character string as any keyword.

Table 4.3

Clustering Measures	α=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
VI	3.49	3.29	3.19	2.99	2.89	3.17	2.81	2.77	3.24	3.24	**2.66**
OE	1.74	1.65	1.59	1.5	1.44	1.58	1.4	1.39	1.62	1.62	**1.33**

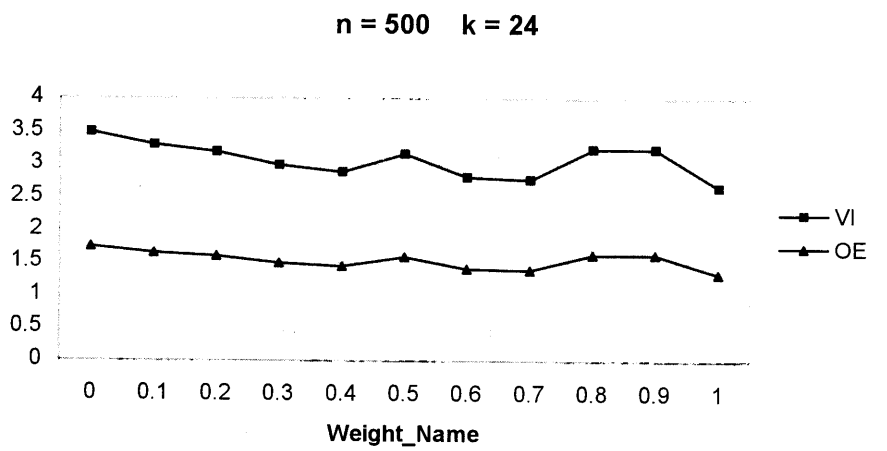

Fig. 4.3

Clustering by entity identifiers combined with keywords

This experiment is also conducted on the entity name testing dataset of 500 documents and 24 clusters, with *k* varying from 4 to 40 and, for each value of *k*, α varying from 0 to 1 on 0.1 incremental steps; $\alpha = 0$ means purely keyword-based clustering, while $\alpha = 1$ means purely entity identifier-based clustering. Figure 4.4 illustrates the *VI* and *OE* diagrams with $k = 24$. As for clustering by entity names, clustering by entity identifiers does not substantially outperforms purely keyword-based clustering when there not many entities having aliases in the testing dataset. Then an identifier is just a unique string rather than one representing different entity names.

Table 4.4

Clustering Measures	α=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
VI	2.28	1.67	1.44	1.24	1.16	1.15	1.14	1.15	1.16	**0.95**	1.31
OE	1.14	0.84	0.72	0.62	0.58	0.58	0.57	0.58	0.58	**0.47**	0.66

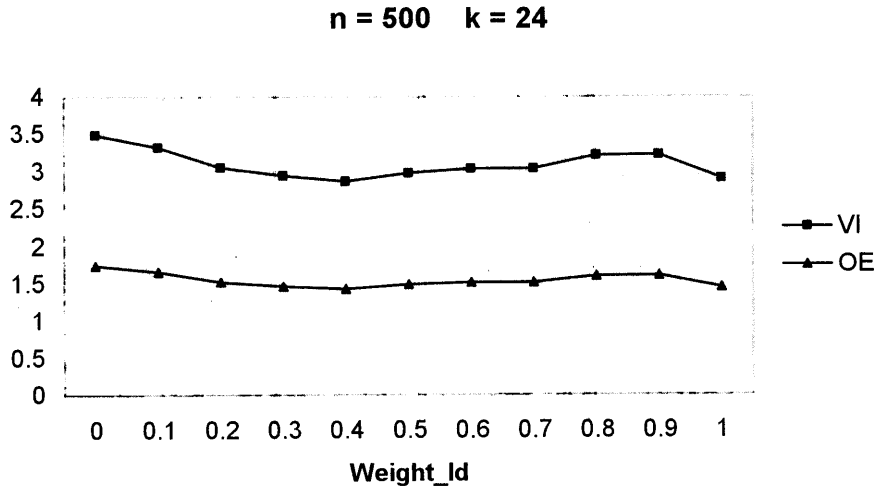

Fig. 4.4

V. CONCLUSION

We have proposed a multi-vector space model for NE-based information processing, as an adaptation of the traditional keyword-based VSM with vectors over NE spaces. Each document (or query) is represented by four component vectors over the four spaces of entity names, types, name-type pairs, and identifiers, allowing searching and clustering documents by various NE features. Vector dimensional weights are computed in accordance to the *tf.idf* scheme and with respect to each of those four features of named entities. Similarity between two documents is then defined as a distance between their representative vectors. As compared to other NE-based models, the essential of our proposed model is that distinct features of named entities, type subsumption, and name aliases are all taken into account.

We have applied the proposed model to document clustering with respect to occurring named entities. Experiments show that NE-based clustering is complementary to keyword-based one, giving a new perspective for user needs and producing meaningful layers and groups of documents, regarding different features of named entities mentioned in the documents. It could be useful for web-based learning where, in many subjects, named entities together with general concepts constitute the main contents of a document.

As presented in the paper, the label of a document is currently defined as the set of the dominant feature values in the document, and two labels are considered to be totally different if their two defining sets are different, though they may share most of feature values. It would be more reasonable, and bring better clustering quality, if the overlapping of those two sets is taken into account. In another aspect, combination of both keyword-based clustering and NE-

based clustering is to be investigated because, for instance, we may want to group documents about pollution, by keywords, and divide the group into subgroups each of which is about pollution in a particular city, by entity identifiers.

REFERENCES

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley (1999).
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American (2001).
3. Cao, T.H., Do, H.T., Hong, D.T., Quan, T.T.: Fuzzy Named Entity-Based Document Clustering. In: Proceedings of the 17th IEEE International Conference on Fuzzy Systems (2008) 2028-2034.
4. Cao, T.H. (2008) PRICAI'08
5. Castells, P., Fernández, M., Vallet, D.: An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. IEEE Transactions on Knowledge and Data Engineering **19** (2006) 261-272.
6. Dill, S. et al.: SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. In: Proceedings of the 12th Int. Conference on the WWW (2003).
7. Gonçalves, A., Zhu, J., Song, D., Uren, V., Pacheco, R.: LRD: Latent Relation Discovery for Vector Space Expansion and Information Retrieval. In: Proceedings of the 7th International Conference on Web-Age Information Management (2006).
8. Hartigan, J., Wong, M.: Algorithm AS136: A K-means Clustering Algorithm. Applied Statistics **28** (1979) 100-108.
9. He, J., Tan, A.-H., Tan, C.-L., Sung, S.-Y.: On Quantitative Evaluation of Clustering Algorithms. In: Wu, W. et al. (eds.): Clustering and Information Retrieval. Kluwer Academic (2003) 105-133.
10. Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.: Semantic Annotation, Indexing, and Retrieval. Journal of Web Semantics **2** (2005).
11. Meilă, M.: Compare Clusterings - An Information Based Distance. Journal of Multivariate Analysis (2007) 873-895.
12. Sekine, S.: Named Entity: History and Future. Proteus Project Report (2004).Toda, H., Kataoka, R.: A Search Result Clustering Method Using Informatively Named Entities. In: Proceedings of the 7th ACM International Workshop on Web Information and Data Management (2005) 81-86.