Latent Ontological Feature Discovery for Text Clustering

Van T.T. Duong
Faculty of Information Technology & Applied Mathematics
Ton Duc Thang University
Ho Chi Minh City, Viet Nam

Tru H. Cao, Cuong K. Chau, Tho T. Quan
Faculty of Computer Science and Engineering
HCM City University of Technology
Ho Chi Minh City, Viet Nam

***Abstract*—The content of a text is mainly defined by keywords and named entities occurring in it. In particular for news articles, named entities are usually important to define their semantics. However, named entities have ontological features, namely, their aliases, types, and identifiers, which are hidden from their textual appearance. In this paper, we explore weighted combinations of those latent named entity features with keywords for text clustering. To that end, the traditional Vector Space Model is adapted with multiple vectors defined over spaces of entity names, types, name-type pairs, identifiers, and keywords. Clustering quality is evaluated by both of the self purity-separation type and the relative comparison type of measures. Hard and fuzzy clustering experiments of the proposed model on selected data subsets of Reuters-21578 are conducted and evaluated.**

Keywords—named entity; latent semantics; hard clustering; fuzzy clustering; 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 (cf. [13]). In particular, it helps to overcome the deficiencies of the query-list approach to showing search results by grouping returned documents into a hierarchy of meaningful thematic categories, providing better data views to users than sequential listings ([18], [14]). The idea has been realized in real-world application systems like Clusty ([6]) and Carrot2 ([4]).

Traditionally text clustering is only based on keywords (KW) occurring in the texts. Words include those that represent named entities (NE), which are referred to by names such as people, organizations, and locations ([15]). In particular, news articles usually contain such named entities, which are important for the news contents.

However, named entities in a text cover under their textual forms (i.e., names) ontological features that are significant to the semantics of the text. Firstly, it is the type of a named entity in the ontology of discourse, for which texts containing "*Ha Noi*", "*Paris*", and "*Tokyo*" may be grouped together as those about capital cities in the world. Clustering purely based on keywords fails to do that because it does not use the common latent type information of such named entities. Secondly, it is the identifier of a named entity, for which texts about "*U.S.*", "*USA*", "*United States*", and "*America*" may be grouped together as those about the same country *United States of America*. Keyword-based clustering also fails because it does not use the fact that an entity may exist under different aliases. These are just two basic ontological features of named entities.

In [16], the most significant entity name in a text was used as its label, based on an enhanced version of the *tf.idf* measure. Then the texts with labeling named entities of the same type were grouped together. As such, it was simply classification of texts 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 texts 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, [3] introduced a multi-vector space model on all the NE features, then explored and evaluated the information retrieval performance of various combinations of keywords and named entities.

In this paper, we employ the hybrid entity-keyword model in [3] for text clustering and show that using the latent ontological features of named entities is necessary for improving the clustering quality. Both hard clustering and fuzzy clustering are experimented. Researching efficient clustering algorithms is not the purpose of this work, so we apply the most popular hard clustering technique *k*-means [8] and its fuzzy counterpart fuzzy *c*-means ([2]). However, any technique could be used with our proposed model.

For clustering quality evaluation, measures like Overall Entropy (OE) in [9] for hard clustering and Xie-Beni index (XB) in [17] for fuzzy clustering are based on the purity and separation of the resulting partition itself. We view them as *objective measures*, for which a clustering result is not tested against a pre-constructed gold-standard one. In contrast, the so-called Variation of Information (VI) measure recently proposed in [11] quantifies how different two partitions are. 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. We apply both of these objective and subjective measures in this work.

Section II summarizes the basic notions and formulation of our multi-vector space model combining named entities and keywords. Section III recalls the OE, XB and VI measures, and proves the equivalence of OE and VI for hard clustering under a certain condition. Sections IV and V respectively present our experiments and evaluation on hard and fuzzy text clustering. Finally, Section VI draws concluding remarks and further work to be investigated.

II. An Entity-Keyword Multi-Vector Space Model

Despite having known disadvantages, VSM is still a popular model and a basis to develop other models for information processing, because it is simple, fast, and its similarity measure is in general either better or almost as good as a large variety of alternatives (cf. [1]). We recall that, in the keyword-based VSM, each text 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 text, called *tf*, and the inverse occurrence frequency of the term across all the existing texts, called *idf*. The similarity degree between a text and a query is then defined as the cosine of their representing vectors.

For formally representing a text 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:

1. Each text d is modelled as a subset of $(\boldsymbol{N}\cup\{*\})\times(\boldsymbol{T}\cup\{*\})\times(\boldsymbol{I}\cup\{*\})$, where '*' denotes an unspecified name, type, or identifier of a named entity in d, and
2. 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***×***T***, and ***I***.

For example, following is a text and its set of named entity features:

"*U.N. team survey of public opinion in North Borneo and Sarawak on the question of joining the federation of Malaysia*".

{(*U.N.*/*/*), (*North Borneo*/*Province*/*), (*Sarawak*/*Location*/*), (*Malaysia*/*Country*/*Country_T.MY*)}

Here, *Country_T.MY* is the identifier of the country *Malaysia* in the knowledge base of discourse. Meanwhile, the type of *U.N.* is presumably unrecognized, and *North Borneo* and *Sarawak* are only recognized as of the types *Province* and *Location*, respectively.

A feature of a named entity could be unspecified due to the incomplete information about that named entity in a text, or the inability of an employed NE recognition engine to fully recognize it. Each of the four component vectors introduced above for a text 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 vector weights:

1. The frequency of a name also counts identical entity aliases. That is, if a text contains an entity having an alias identical to that name, then it is assumed as if the name occurred in the text. For example, if a text refers to *Saigon City*, then each occurrence of that entity in the text is counted as one occurrence of the name *Ho Chi Minh City* too, because it is an alias of *Saigon City*.

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

The similarity degree of a text d and a text q, with respect to the named entity features, is then defined to be:

$$w_N.cosine(\vec{d}_N,\vec{q}_N) + w_T.cosine(\vec{d}_T,\vec{q}_T) + w_{NT}.cosine(\vec{d}_{NT},\vec{q}_{NT}) + w_I.cosine(\vec{d}_I,\vec{q}_I) \quad \text{(Eq. 1)}$$

where $w_N + w_T + w_{NT} + w_I = 1$.

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 $\vec{d}_{NT}$ 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 redundant.

Further, for combining ontological features with keywords, 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:

$$sim(\vec{d},\vec{q}) = \alpha.[w_N.cosine(\vec{d}_N,\vec{q}_N) + w_T.cosine(\vec{d}_T,\vec{q}_T) + w_{NT}.cosine(\vec{d}_{NT},\vec{q}_{NT}) + w_I.cosine(\vec{d}_I,\vec{q}_I)] + (1-\alpha).cosine(\vec{d}_{KW},\vec{q}_{KW}) \quad \text{(Eq. 2)}$$

where $w_N + w_T + w_{NT} + w_I = 1$ and $\alpha \in [0, 1]$. Here α represents the weight of the named entity component, and $(1 - \alpha)$ of the keyword component, in defining the similarity of the texts.

The proposed multi-vector space model can be used for clustering documents into a hierarchy via top-down phases each of which uses one of the four NE-based vectors presented above. 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.

Meanwhile, the KW-based vector is complementary to the NE-based vectors in representing the salient points in the content of a document. For instance, texts about tourist attraction places should contain both keywords related to tourist attraction and named entities being places. Optimal weighting of the

NE component and the KW component depends on the contents of the texts to be clustered. However, the point is that relying on keywords alone as in traditional techniques may not be satisfactory in practice, as shown by experimental results in our work.

There are still possible variations of the proposed model that are worth exploring, depending on whether entity names in a text are counted as keywords in constructing its KW-based vector or not. In other words, the entity name set and the keyword set of a text may or may not be considered as overlapping. We call these two alternative models NEKW_OVL and NEKW_NOVL respectively.

III. Measures of Clustering Quality

Traditionally, clustering quality is evaluated using two complementary measures: (1) *internal measure* that reflects the average semantic distance between data points within each cluster; the smaller the better for the cluster purity; and (2) *external measure* that reflects the average semantic distance between the clusters themselves; the larger the better for the cluster separation. In [9] *cluster entropy* and the *class entropy* are defined as the internal and external measures, respectively, and the Overall Entropy as their linear combination. The smaller the overall entropy is, the better clustering quality is.

Formally, 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 $\{l_1, l_2, \ldots, l_{k*}\}$. Let n_j be the total number of data points of label l_j in the dataset, and n_{ij} be the number of data points labeled l_j in cluster C_i. Then, the cluster entropy E_c, the class entropy E_l, and the overall entropy are defined as follows:

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

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

where $\beta \in [0, 1]$ is empirically determined. The smaller $E(C)$ is, the better clustering quality is. Ideally, all data points in each cluster have the same label, i.e., $E_c = 0$, and all data points of the same label reside in the same cluster, i.e., $E_l = 0$.

Meanwhile, for the Variation of Information measure ([11]), 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 $H(C \mid C^*)$ 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^*. Significantly, the following theorem states the equivalence of VI and OE, if the data point labels are as given by C^* and the cluster and the class entropies in OE have the same weight. The proof is presented in the Appendix.

Theorem 1. Assume that $C^* = C_1^* \cup C_2^* \cup \ldots \cup C_{k^*}^*$ is a partition on a set of data points and the label l_i of each cluster C_i^* is also the label of all the data points in it. Let $C = C_1 \cup C_2 \cup \ldots \cup C_k$ be an arbitrary partition on that same data point set. Then $VI(C, C^*) = 2E(C)$ if the cluster entropy and the class entropy in the computation of $E(C)$ have the same weight 0.5.

Since taking the equal weights for the cluster and the class entropies in the OE measure is natural and reasonable, the significance of the property proved above is that one can use either OE or VI for measuring clustering quality when all data points have pre-defined labels. Nevertheless, in practice, data point labels may not be pre-defined but generated as part of a clustering technique, which also affect the clustering quality in terms of OE. That is the case when VI is useful for testing a partition generated by that technique with respect to a subjectively constructed partition on the same dataset.

For fuzzy clustering, Xie-Beni index ([17]) is among the most popularly used ones, measuring the overall average purity and separation of a fuzzy partition by:

$$S = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}[\mu_i(x_k)]^m \| x_k - v_i \|^2}{n * \min_{i,j} \| v_i - v_j \|^2} \qquad \text{(Eq. 5)}$$

where n is the number of data points x_k's, m is the *fuzziness index*, v_i is the centroid of the i-th cluster, $\mu_i(x_k)$ is the membership value of x_k into the i-th cluster, and $\| x_k - v_i \|$ represents the distance between the data point x_k and the i-th cluster, which is usually calculated by Euclidian Distance. The smaller the value of the index, the better the fuzzy partition is. This index could be considered as a fuzzy counterpart of the OE measure.

For a subjective measure of fuzzy clustering quality, a gold standard partition can also be pre-constructed from a dataset like Reuters-21578. However, in such a dataset, although a text can belong to more than one cluster, its membership to a cluster can only be either 0 or 1. Therefore, for directly applying the VI measure to a fuzzy partition against a pre-constructed one, we introduce a threshold to an output of a fuzzy clustering algorithm to decide if a text is assigned to a cluster or not, i.e., the membership of a text in a cluster is coerced to be either 0 or 1.

IV. Hard Clustering Experiments

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:

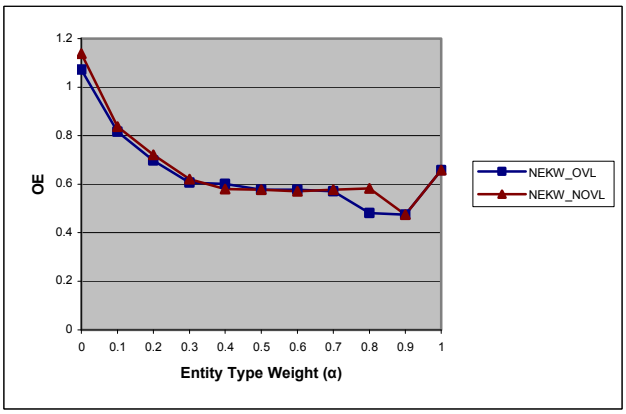
PLACES: 195 documents
PEOPLE: 105 documents
ORGS: 129 documents
EXCHANGES: 71 documents

Here we employ *k*-means for hard clustering. Basically, the *k*-means algorithm keeps relocating data points into *k* clusters until the following objective function stops decreasing:

$$f = \sum_{i=1}^{k} \sum_{x_j \in c_i} \left| x_j - \overline{c_i} \right| \qquad \text{(Eq. 6)}$$

where c_i is the *i*-th cluster and $\overline{c_i}$ is the average value of its data points x_j's, called the centroid. In practice, for obtaining the best clustering quality, the optimal value of *k* can be determined by experiments.

First, we run *k*-means on the constructed 500-document dataset with $k = 4$ and α varying from 0 to 1 on 0.1 incremental steps. Figure 1 illustrates the clustering quality of the NEKW_OVL and NEKW_NOVL models with respect to the OE and VI measures. For the OE measure, we take the equal weight for the cluster entropy and the class entropy, i.e., $\beta = 0.5$ for Equations 3. The corresponding data are presented in Table I. In accordance to Theorem 1, the corresponding OE and VI curves actually have the same shape. Second, we vary *k* from 2 to 10, take the best case for each value of *k*, and plot their OE and VI values as in Figure 2, from the obtained data in Table II. As expected, $k = 4$ is the optimal value for the testing dataset with 4 pre-defined clusters.



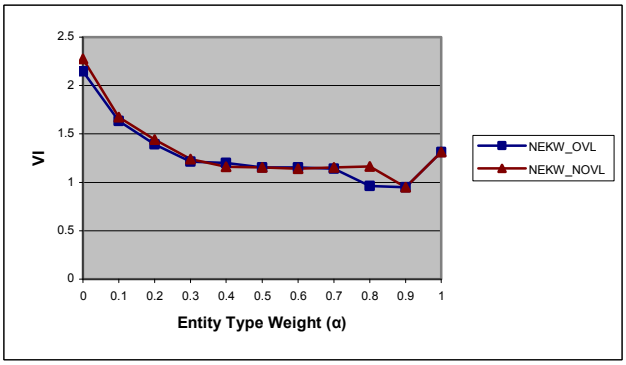
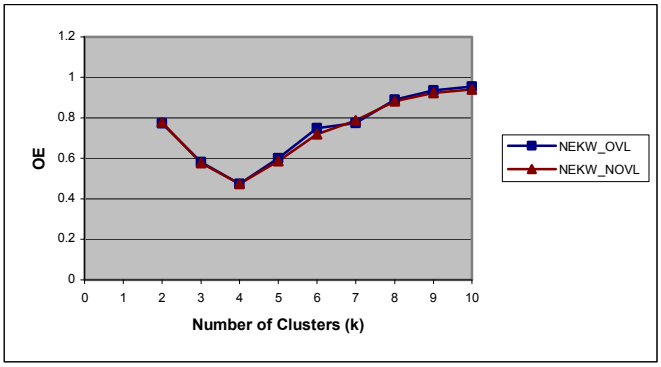


Figure 1. OE and VI diagrams for hard clustering with $k = 4$ and varied α

TABLE I
OE AND VI MEASURES FOR HARD CLUSTERING WITH $K = 4$ AND VARIED α

OE	**α=0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
OVL	**1.07**	0.82	0.7	0.61	0.6	0.58	0.58	0.57	0.48	**0.47**	0.66
NOVL	**1.14**	0.84	0.72	0.62	0.58	0.58	0.57	0.58	0.58	**0.47**	0.66

VI	**α=0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
OVL	**2.15**	1.63	1.39	1.21	1.2	1.15	1.15	1.14	0.96	**0.95**	1.31
NOVL	**2.28**	1.67	1.44	1.24	1.16	1.15	1.14	1.15	1.16	**0.95**	1.31



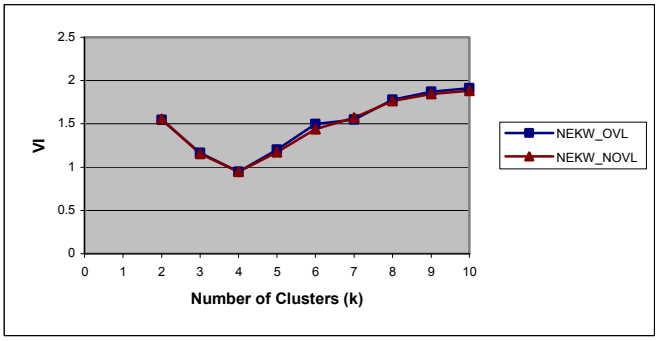


Figure 2. OE and VI diagrams for hard clustering with varied *k*

TABLE II
OE AND VI MEASURES FOR HARD CLUSTERING WITH VARIED *K*

OE	***k*=2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
OVL	0.77	0.58	**0.47**	0.6	0.75	0.78	0.89	0.94	0.96
NOVL	0.78	0.58	**0.47**	0.59	0.72	0.79	0.88	0.92	0.94

VI	***k*=2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
OVL	1.55	1.16	**0.95**	1.2	1.5	1.55	1.78	1.87	1.91
NOVL	1.56	1.15	**0.95**	1.17	1.44	1.57	1.76	1.85	1.88

The experimental results show that:

1. The NEKW_OVL and NEKW_NOVL models perform nearly the same. That is, counting or not counting entity names for KW-based vectors make little difference. It means that entity names themselves, i.e., only their textual forms, are not significant to assignment of named entity tags to documents in the Reuters-21578 dataset.
2. The clustering quality is improved by more than 100% with $\alpha = 0.9$ as compared with $\alpha = 0$ (OE = 0.47 vs. 1.07 for NEKW_OVL). We note that the NEKW_OVL model with $\alpha = 0$ is actually the traditional purely keyword-based VSM. So, the latent ontological features (e.g. entity types in these experiments) are important to the clustering results.
3. The best clustering quality is obtained when $k = 4$, which is the same as the number of clusters of the pre-constructed testing dataset. It implies that our proposed models represent well the contents of documents like those of the Reuters-21578 dataset for the clustering task.

V. FUZZY CLUSTERING EXPERIMENTS

We recall that, basically the fuzzy *c*-means algorithm keeps relocating data points into *c* clusters until the following objective function stops decreasing (cf. Equation 5):

$$J_m(P) = \sum_{k=1}^{n}\sum_{i=1}^{c}[\mu_i(x_k)]^m \| x_k - v_i \|^2 \qquad \text{(Eq. 7)}$$

where $P = \{\mu_1, \mu_2, \ldots, \mu_c\}$ is a fuzzy *c*-partition. The centroid of each cluster is computed by the following formula:

$$v_i = \frac{\sum_{k=1}^{n}[\mu_i(x_k)]^m x_k}{\sum_{k=1}^{n}[\mu_i(x_k)]^m} \qquad \text{(Eq. 8)}$$

At each iteration of the algorithm, after the cluster centroids are re-calculated, membership values $\mu_i(x_k)$'s are updated based on the data point x_k's and the cluster centroids:

$$\mu_i(x_k) = \frac{1}{\sum_{j=1}^{c}\left(\frac{\| x_k - v_i \|}{\| x_k - v_j \|}\right)^{\frac{2}{m-1}}} \qquad \text{(Eq. 9)}$$

The process is stopped when the maximum change of membership values between two consecutive iterations is less than a pre-defined threshold value.

We also use the Reuters-21578 dataset for fuzzy clustering experiments. First, about 7,500 documents containing one or more of the four tags PLACES, PEOPLE, ORGS, and EXCHANGES are automatically extracted. Next, we manually select from those documents 500 typical ones whose contents are clearly about their associated tags. For fuzzy clustering, the documents are selected so that some of them are about more than one named entity type, including:

200 documents containing only one NE tag
238 documents containing two NE tags
57 documents containing three NE tags
5 documents containing four NE tags

The distribution of the 500 documents across the four NE tags is as follows:

PLACES: 300 documents
PEOPLE: 200 documents
ORGS: 281 documents
EXCHANGES: 86 documents

We run fuzzy *c*-means on the constructed dataset, using both the NEKW_OVL and NEKW_NOVL and evaluating clustering quality with respect to both of the XB and VI measures. We set the fuzzy index $m = 2$ and the threshold value to stop the iterative process equal to 0.01. For evaluating the fuzzy clustering results, the membership threshold introduced in Section III is chosen by $1/c = 0.25$, in order to coerce a document to be or not to be in one of the four NE type clusters mentioned above.

As for hard clustering, first we fix the number of clusters $c = 4$ and vary α from 0 to 1 on 0.1 incremental steps. Since fuzzy *c*-means relies on the initial membership degrees of the documents to the projected clusters, which are initialized randomly, for each α we run the algorithm 10 times and take the average of the results for the XB or VI measures. Figure 3 illustrates the

resulting XB and VI diagrams of the NEKW_OVL and NEKW_NOVL models, based on the obtained data in Table III. Second, for finding the optimal number of clusters, we vary c from 2 to 10 and compute the XB and VI values of the best cases for each value of c as plotted in Figure 4, from the obtained data in Table IV. Again, it turns out that $c = 4$ is the optimal value for the testing dataset with 4 pre-defined clusters.

The results are consistent with those of hard clustering on the followings: (1) The overlapping and non-overlapping models make little difference on the performance; (2) The clustering quality is drastically improved when taking into account the latent named entity types, i.e., with $\alpha > 0$; and (3) That the proposed models represent the contents of Reuters-21578 documents properly is supported by the fact that the obtained optimal number of clusters coincides with that of the pre-constructed partition. Besides, comparing the XB and VI measures, one can observe that VI is much more stable than XB when the value of α or the number of clusters moves away from the optimal value.

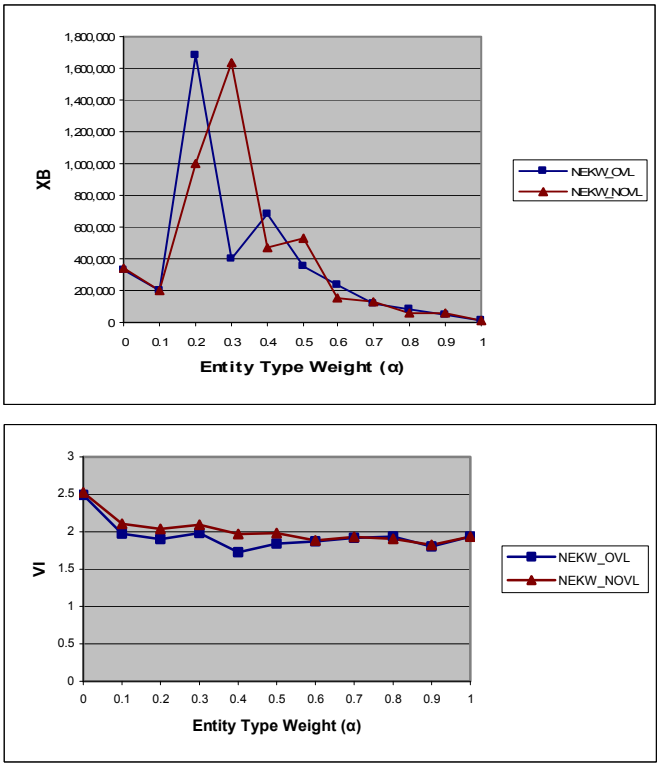


Figure 3. XB and VI diagrams for fuzzy clustering with c=4 and varied α

TABLE III
XB AND VI MEASURES FOR FUZZY CLUSTERING WITH $C = 4$ AND VARIED α

XB ×10,000	**α=0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
OVL	**33**	19.8	168.2	39.7	68	35	23.1	11.9	8.7	4.4	**1.7**
NOVL	**33.9**	20	99.9	163.9	47	52.4	15.6	12.8	5.4	5.3	**1.7**

VI	**α=0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
OVL	**2.49**	1.97	1.9	1.98	**1.73**	1.84	1.87	1.92	1.93	1.8	1.93
NOVL	**2.52**	2.11	2.04	2.09	1.97	1.98	1.88	1.93	1.91	**1.82**	1.93

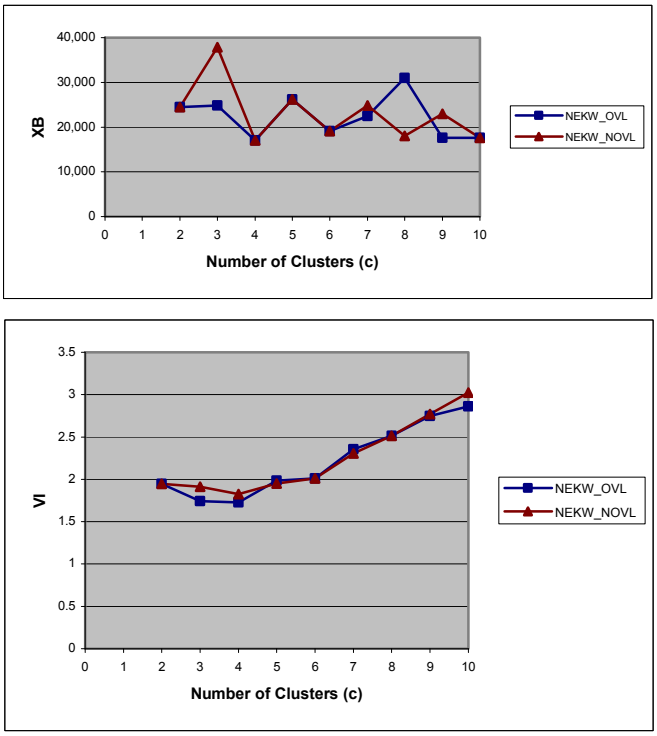


Figure 4. XB and VI diagrams for fuzzy clustering with varied c

TABLE IV
XB AND VI MEASURES FOR FUZZY CLUSTERING WITH VARIED C

XB ×10,000	***c*=2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
OVL	2.44	2.49	**1.7**	2.62	1.91	2.25	3.1	1.76	1.76
NOVL	2.44	3.78	**1.7**	2.62	1.91	2.48	1.81	2.3	1.76

VI	***c*=2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
OVL	1.95	1.74	**1.73**	1.98	2.01	2.35	2.51	2.75	2.86
NOVL	1.95	1.91	**1.82**	1.95	2.01	2.3	2.51	2.77	3.02

VI. CONCLUSION

We have presented a hybrid named entity-keyword-based multi-vector space model for information processing. Our experimental results using the introduced model for text clustering on the well-known Reuters-21578 dataset are two-fold. First, they show that the latent ontological features of named entities in a text are important to define its semantics. In particular, our model taking into account the types of named entities, which are covered under their textual forms, drastically outperforms the traditional purely keyword-based VSM for both hard and fuzzy clustering on the testing dataset. Second, they show that our model is suitable for representing the subjects of documents involving named entities like Reuters-21578 ones.

As another result of this research, we have proved the equivalence of the Variation of Information measure and the Overall Entropy measure whose class and cluster entropies have the same weight. The experiments have also show that, for fuzzy clustering, the adapted Variation of Information measure is more stably converge towards optimal points than a classical one like Xie-Beni index.

In general, all the four NE-based vectors in the introduced model could be employed for text clustering. The model also supports hierarchical clustering for which each layer uses a certain clustering objective corresponding to a NE feature. We are carrying out further experiments along this direction. On the other hand, since named entities are pervasive and play an important role in news articles, we are investigating the proposed model and method for knowledge discovery and integration on the Web.

APENDIX

Proof of Theorem 1.

Under the given condition, n_j and n_{ij} in Equations 3 in Section III are equal to $|C_j^*|$ and $|C_i \cap C_j^*|$, respectively. So, one obtains:

$$\begin{aligned}
E_c(C) &= -\sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{|C_i|} \\
&= -\sum_{i=1}^{k}\sum_{j=1}^{k^*}\left(\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} - \frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i|}{N}\right) \\
&= -\sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} \\
&\quad + \sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i|}{N} \\
&= -\sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} + \sum_{i=1}^{k}\frac{|C_i|}{N}\log\frac{|C_i|}{N}
\end{aligned}$$

$$\begin{aligned}
E_l(C) &= -\sum_{j=1}^{k^*}\sum_{i=1}^{k}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{|C_j^*|} \\
&= -\sum_{j=1}^{k^*}\sum_{i=1}^{k}\left(\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} - \frac{|C_i \cap C_j^*|}{N}\log\frac{|C_j^*|}{N}\right) \\
&= -\sum_{j=1}^{k^*}\sum_{i=1}^{k}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} \\
&\quad + \sum_{j=1}^{k^*}\sum_{i=1}^{k}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_j^*|}{N} \\
&= -\sum_{j=1}^{k^*}\sum_{i=1}^{k}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} + \sum_{j=1}^{k^*}\frac{|C_j^*|}{N}\log\frac{|C_j^*|}{N} \\
&= -\sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} + \sum_{j=1}^{k^*}\frac{|C_j^*|}{N}\log\frac{|C_j^*|}{N}
\end{aligned}$$

With $\beta = 0.5$, Equation 3 gives:

$$\begin{aligned}
E(C) &= -\sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} + \frac{1}{2}\sum_{i=1}^{k}\frac{|C_i|}{N}\log\frac{|C_i|}{N} \\
&\quad + \frac{1}{2}\sum_{j=1}^{k^*}\frac{|C_j^*|}{N}\log\frac{|C_j^*|}{N}
\end{aligned}$$

On the other hand, from Equations 4 one has:

$I(C,C^*)$

$$\begin{aligned}
&= \sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|/N}{(|C_i|/N).(|C_j^*|/N)} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k^*}\left(\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} - \frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i|}{N} - \frac{|C_i \cap C_j^*|}{N}\log\frac{|C_j^*|}{N}\right) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} - \sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i|}{N} \\
&\quad - \sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_j^*|}{N} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} - \sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i|}{N} \\
&\quad - \sum_{j=1}^{k^*}\sum_{i=1}^{k}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_j^*|}{N} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} - \sum_{i=1}^{k}\frac{|C_i|}{N}\log\frac{|C_i|}{N} \\
&\quad - \sum_{j=1}^{k^*}\frac{|C_j^*|}{N}\log\frac{|C_j^*|}{N}
\end{aligned}$$

Hence:

$VI(C, C^*)$

$$\begin{aligned}
&= -\sum_{i=1}^{k}\frac{|C_i|}{N}\log\frac{|C_i|}{N} - \sum_{j=1}^{k^*}\frac{|C_j^*|}{N}\log\frac{|C_j^*|}{N} \\
&\quad - 2\sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} \\
&\quad + 2\sum_{i=1}^{k}\frac{|C_i|}{N}\log\frac{|C_i|}{N} + 2\sum_{j=1}^{k^*}\frac{|C_j^*|}{N}\log\frac{|C_j^*|}{N} \\
&= -2\sum_{i=1}^{k}\sum_{j=1}^{k^*}\frac{|C_i \cap C_j^*|}{N}\log\frac{|C_i \cap C_j^*|}{N} + \sum_{i=1}^{k}\frac{|C_i|}{N}\log\frac{|C_i|}{N} \\
&\quad + \sum_{j=1}^{k^*}\frac{|C_j^*|}{N}\log\frac{|C_j^*|}{N} \\
&= 2E(C).
\end{aligned}$$

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

[3] T.H. Cao, K.C. Le, and V.M. Ngo, "Exploring combinations of ontological features and keywords for text retrieval," in Proc. of the 10th Pacific Rim Intl Conference on Artificial Intelligence, LNAI 5351. Springer-Verlag, 2008, pp. 603-613.

[4] Carrot2: Open Source Search Results Clustering Engine. Available at: http://project.carrot2.org/architecture.html.

[5] P. Castells, M. Fernández, and D. Vallet, "An adaptation of the vector-space model for ontology-based information

retrieval," IEEE Transactions on Knowledge and Data Engineering, vol. 19, 2006, pp. 261-272.

[6] Clusty Search: Clustering Search Engine. Available at: http://clusty.com.

[7] A. Gonçalves, J. Zhu, D. Song, V. Uren, and R. Pacheco, "LRD: Latent relation discovery for vector space expansion and information retrieval," in Proc. of the 7th International Conference on Web-Age Information Management, 2006.

[8] J. Hartigan and M. Wong, "Algorithm AS136: A K-means clustering algorithm," Applied Statistics, vol. 28, 1979, pp. 100-108.

[9] J. He, A. H. Tan, C. L. Tan, and S. Y. Sung, "On quantitative evaluation of clustering algorithms," in Clustering and Information Retrieval, Wu et al., Ed. Kluwer Academic, 2003, pp. 105-133.

[10] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," Journal of Web Semantics, vol. 2, 2005.

[11] M. Meilă, "Compare clusterings – an information based distance," Journal of Multivariate Analysis, 2007, pp. 873-895.

[12] Z.-Y. Niu, D.-H. Ji & Chew-Lim Tan, **"**Using cluster validation criterion to identify optimal feature subset and cluster number for document clustering," Information Processing and Management, vol. 43, 2007, pp. 730-739.

[13] J.V. Oliveira and W. Pedrycz, Ed., Advances in Fuzzy Clustering and its Application*s*. John Wiley & Sons, 2007.

[14] S. Osinki, "Improving quality of search results clustering with approximate matrix factorisations," in *Proc. 28th European Conference on Information Retrieval*, LNCS 3936. Springer-Verlag, 2006, pp. 167-178.

[15] S. Sekine, "Named entity: history and future," Proteus project report, 2004.

[16] H. Toda and R. Kataoka, "A search result clustering method using informatively named entities," in Proc. of the 7th ACM International Workshop on Web Information and Data Management, 2005, pp. 81-86.

[17] X.L. Xie and G. Beni, "A validity measure for fuzzy clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, 1991, pp. 841-847.

[18] D. Zhang and Y. Dong, "Semantic, hierarchical, online clustering of web search results," in Proc. 6th Asia-Pacific Web Conference, LNCS 3007. Springer-Verlag, 2004, pp. 69-78.