Chapter 10
Text Clustering with Named Entities: A Model, Experimentation and Realization

Tru H. Cao, Thao M. Tang, and Cuong K. Chau

Ho Chi Minh City University of Technology and
John von Neumann Institute VNU-HCM Vietnam
tru@cse.hcmut.edu.vn

Abstract. Named entities often occur in web pages, in particular news articles, and are important to what the web pages are about. They have ontological features, namely, their aliases, types, and identifiers, which are hidden from their textual appearance. In this chapter, for text searching and clustering, we propose an extended Vector Space Model with multiple vectors defined over spaces of entity names, types, name-type pairs, identifiers, and keywords. Both hard and fuzzy text clustering experiments of the proposed model on selected data subsets of Reuters-21578 are conducted and evaluated. The results prove that a weighted combination of named entities and keywords are significant to clustering quality. Implementation and demonstration of text clustering with named entities in a semantic search engine are also presented.

1 Introduction

Clustering, which is to partition and group data points of similar properties together, is not only an important technique for data mining and knowledge discovery, but also a useful technique for information processing in other application areas [21, 24]. Traditional text clustering is only based on keywords (KW) occurring in texts. Words include those that represent named entities (NE), which are referred to by names such as people, organizations, and locations [23]. In particular, news articles usually contain such named entities, which are important for the news contents. Indeed, in the top 10 search terms by YahooSearch[1] and GoogleSearch[2] in 2008, there are respectively 10 and 9 ones that are named entities. Besides, textual corpora, such as web pages and blogs, often contain named entities.

However, named entities in a document 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 documents 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.

[1] http://buzz.yahoo.com/yearinreview2008/top10/
[2] http://www.google.com/intl/en/press/zeitgeist2008/

D.E. Holmes, L.C. Jain (Eds.): Data Mining: Found. & Intell. Paradigms, ISRL 23, pp. 267–287.
springerlink.com © Springer-Verlag Berlin Heidelberg 2012

Secondly, it is the identifier of a named entity, for which documents 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 among the ontological features of named entities.

The ontology-based text clustering methods in [14] and [28] actually relied on an ontology of common concepts like WordNet rather than on named entities. In [25], 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 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 documents having close semantics regarding various named entities occurring in them.

Closely related to our work were [9] and [18]. In [9], a linear combination of one vector on proper names with their types and one vector on common words was used to represent a document. However, only proper names of the person, organization and location types were considered. In [18], each document was represented by three different vectors on named entities of each of the person, organization and location types. While [9] suggested that text clustering on only named entities was not good, [18] reported it was for multilingual news clustering.

Meanwhile, for text searching, in [6] 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. The latent semantics model proposed in [10] used both keywords and named entities as terms for a single vector space, but only entity names were taken into account. In contrast, [3] introduced a multi-vector space model on all of the NE features, then explored and evaluated the information retrieval performance of various combinations of keywords and named entities.

This paper contributes to text clustering using named entities in three aspects:

1. Our document representation model takes into account all types and all combined features of named entities.
2. Both hard clustering and fuzzy clustering are experimented. The results show that, for good clustering quality, the weights of the named entity and keyword components in the model depend on the actual contents of the documents to be clustered.
3. The model is realized and demonstrated in the semantic search engine called VN-KIM Search, for hierarchical clustering of resulting documents by keywords as well as different named entity features.

Section 2 summarizes the basic notions and formulation of our proposed multi-vector space model combining named entities and keywords. Section 3 recalls key measures of hard and fuzzy clustering quality. Sections 4 and 5 respectively present our experiments and evaluation on hard and fuzzy text clustering. Section 6 introduces VN-KIM Search with text clustering using named entities on search results. Finally, Section 7 draws concluding remarks and further work to be investigated.

2 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 document representation and 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, 16]). 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 two documents is then defined as the cosine of their representing vectors.

We represent each named entity by a triple (*name*/*type*/*identifier*) where *name*, *type*, and *identifier* are respectively the name, type, and identifier of that named entity. Let $\boldsymbol{N}$, $\boldsymbol{T}$, and $\boldsymbol{I}$ be respectively the sets of names, types, and identifiers of named entities in the ontology of discourse. Then:

1. Each document 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 $\boldsymbol{N}$, $\boldsymbol{T}$, $\boldsymbol{N}\times\boldsymbol{T}$, and $\boldsymbol{I}$.

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 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 vector weights:

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 the country *Georgia*, then each occurrence of that entity in the document is counted as one occurrence of the name *Gruzia*, because it is an alias of *Georgia*. Named entity aliases are specified in a knowledge base of discourse.

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 *Washington DC*, 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 type subsumption is defined by the type hierarchy of an ontology of discourse.

We then define the similarity degree of a document d and a document q, with respect to the named entity features, as follows:

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

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 developer-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 (e.g. the co-occurrence of an entity named *Georgia* and another country mention in a document does not necessarily refer to the country *Georgia*). Meanwhile, $\vec{d}_{NT}$ cannot replace $\vec{d}_I$ because there may be different entities of the same name and type (e.g. there are different cities named *Moscow* in the world). Also, since names and types of an entity are derivable from its identifier, products of ***I*** with ***N*** or ***C*** are not included.

Clearly, named entities alone are not adequate to represent a document. For instance, in the example text above, *opinion*, *joining,* and *federation* are keywords to be taken into account. Therefore, we propose to represent a document by one vector on keywords and four vectors on named entity features. Let $\vec{d}_{KW}$ and $\vec{q}_{KW}$ be respectively the vectors representing the keyword features of two documents 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 $\alpha \in [0, 1]$. The coefficient α weighs relative importance of the NE and KW components in document representation.

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, documents about tourist attraction places should contain both keywords related to tourist attraction and named entities being places. As shown in the experiments next, 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.

There are still possible variations of the proposed model that are worth exploring, depending on whether entity names in a document are counted as keywords in constructing its KW-based vector or not. For instance, in the example text above, *U.N*, *North Borneo*, *Sarawak*, and *Malaysia* could also be treated as keywords as usual. 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.

3 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 [13], for hard clustering, *cluster entropy* and the *class entropy* are defined as the internal and external measures, respectively, and the Overall Entropy (OE) 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:

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

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

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

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 (VI) measure [17], 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:

$$
\begin{aligned}
VI(C, C^*) &= H(C \mid C^*) + H(C^* \mid C) \\
&= H(C) + H(C^*) - 2I(C,C^*) \\
I(C,C^*) &= \sum_{i=1}^{k}\sum_{j=1}^{k^*} \frac{|C_i \cap C_j^*|}{N} \log \frac{|C_i \cap C_j^*| / N}{(|C_i| / N).(|C_j^*| / N)} \\
H(C) &= -\sum_{i=1}^{k} \frac{|C_i|}{N} \log \frac{|C_i|}{N} \\
H(C^*) &= -\sum_{j=1}^{k^*} \frac{|C_j^*|}{N} \log \frac{|C_j^*|}{N}
\end{aligned}
\qquad \text{(Eqs. 4)}
$$

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 was presented in [8].

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 [24, 26] 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.

Measures like OE for hard clustering and XB 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 VI measure 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.

4 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 documents 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. Figure 1 is an example of the header of a document in this dataset. It specifies that the document is about the topics *grain* and *wheat*, the places *USA* and *Australia*, and the people *Lyng* and *Yeutter*.

```
    <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>
```

Fig. 1. An example header of a document in Reuters-21578

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. [20]). 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, the selected documents are automatically annotated using the NE recognition engine of KIM [15], KIM PROTON ontology, and KIM World KB. The ontology consists of about 300 types and 100 relations, and the knowledge base contains over 77,000 named entities. The average precision and recall of the NE recognition engine are about 90% and 86%, respectively[3].

Then we obtain a testing dataset, denoted by D_h, 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 the most popular algorithm *k*-means [12] 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* is 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 2 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 1. 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 3, from the obtained data in Table 2. As expected, $k = 4$ is the optimal value for the testing dataset with 4 pre-defined clusters.

[3] It is reported at http://www.ontotext.com/kim/performance.html.

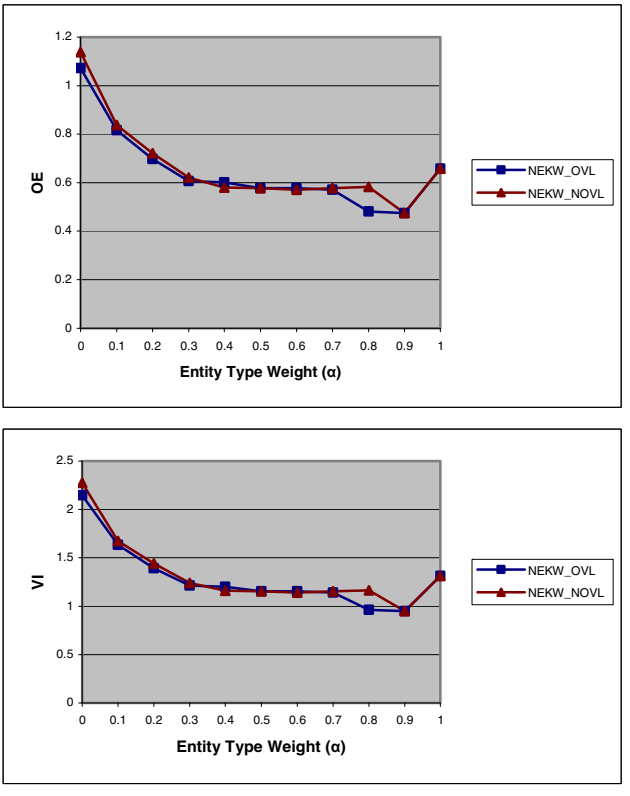

Fig. 2. OE and VI diagrams for hard clustering with $k = 4$ and varied α

Table 1. 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**
NEKW_OVL	1.07	0.82	0.7	0.61	0.6	0.58	0.58	0.57	0.48	**0.47**	0.66
NEKW_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**
NEKW_OVL	2.15	1.63	1.39	1.21	1.2	1.15	1.15	1.14	0.96	**0.95**	1.31
NEKW_NOVL	2.28	1.67	1.44	1.24	1.16	1.15	1.14	1.15	1.16	**0.95**	1.31

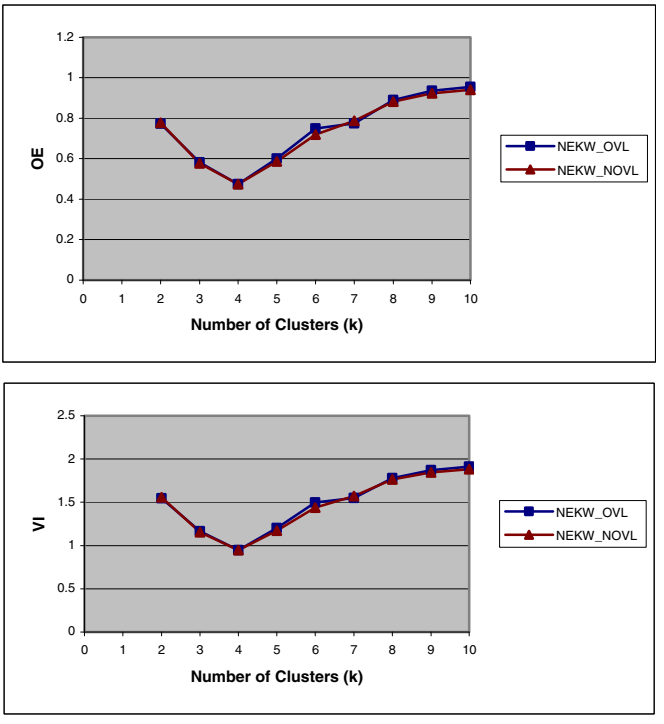

Fig. 3. OE and VI diagrams for hard clustering with varied *k*

Table 2. OE and VI measures for hard clustering with varied *k*

OE	***k*=2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
NEKW_OVL	0.77	0.58	**0.47**	0.6	0.75	0.78	0.89	0.94	0.96
NEKW_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**
NEKW_OVL	1.55	1.16	**0.95**	1.2	1.5	1.55	1.78	1.87	1.91
NEKW_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 for hard clustering. 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.

5 Fuzzy Clustering Experiments

The fuzzy counterpart of *k*-means is fuzzy *c*-means. 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):

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

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:

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

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:

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

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, constructing two testing datasets. For fuzzy clustering, the documents are selected so that some of them are about more than one named entity type or more than one document topic. One testing dataset consists of documents of only NE tags, while the other has documents with both NE and topic tags.

The first dataset, denoted by D_{f1}, comprises 500 documents with one or more of the four tags PLACES, PEOPLE, ORGS, and EXCHANGES, in which:

200 documents contain only one NE tag each
238 documents contain two NE tags each
57 documents contain three NE tags each
5 documents contain four NE tags each.

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

Details of the numbers of documents containing certain tags are given in Table 3.

The second dataset, denoted by D_{f2}, comprises 350 documents containing both NE tags, namely PLACES and PEOPLE, and topic tags, namely INTEREST and MONEY-FX, with the following distributions:

PLACES: 336 documents
PEOPLE: 136 documents
INTEREST: 148 documents
MONEY-FX: 174 documents

Details of the numbers of documents containing certain tags are given in Table 4.

Table 3. Document-tag distribution in the dataset D_{f1}

PLACES (300)	**PEOPLE** (200)	**ORGS** (281)	**EXCHANGES** (86)	**Number of Documents**
X				60
	X			29
		X		41
			X	70
X	X			57
X		X		120
X			X	2
	X	X		51
	X		X	1
		X	X	7
X	X	X		56
X	X		X	0
X		X	X	0
	X	X	X	1
X	X	X	X	5
			Total	500

We run fuzzy *c*-means on the two constructed datasets D_{f1} and D_{f2}, using both the NEKW_OVL and NEKW_NOVL models, and evaluating clustering quality with respect to the XB measure. The fuzzy index *m* is set to 2 and the threshold value to stop the iterative process is set to 0.01. In the experiments, we vary *c* from 2 to 10 and, for each value of *c*, 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 *c* and α we run the algorithm 10 times and take the average of the results for the XB measure.

Table 4. Document-tag distribution in the dataset D_{f2}

PLACES (300)	**PEOPLE** (200)	**INTEREST** (281)	**MONEY-FX** (86)	**Number of Documents**
X				50
	X			0
		X		6
			X	2
X	X			50
X		X		50
X			X	50
	X	X		0
	X		X	0
		X	X	6
X	X	X		20
X	X		X	50
X		X	X	50
	X	X	X	0
X	X	X	X	16
			Total	350

Table 5 and Table 6 present the XB values with varied *c* and α on the dataset D_{f1} for the models NEKW_OVL and NEKW_NOVL, respectively. For each value of *c*, there is an optimal value of α such that the XB measure is minimal, i.e., giving the best clustering quality. In order to evaluate the effect of α in average on clustering quality, we compute the average of the XB values for each common optimal value of α given certain values of *c*, as shown in the last rows of the two tables. It shows that the best values of α in average for NEKW_OVL and NEKW_NOVL on D_{f1} are respectively 0.9 and 0.7.

Table 5. The XB measure with varied c and α on the dataset D_{f1} for the model NEKW_OVL

XB ×1,000	α = 0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
c = 2	61040	8309	7747	566	42.3	76.6	44	181	61.8	13	**11.1**
3	398.8	35838	105	1464	768	237	168	58	93.5	68	**14.3**
4	265.5	181	6514	3037	726	334	234	99	36	28	**25.8**
5	296.2	166.9	76.6	52.9	803	457	127	92	75.8	32	**18.2**
6	314.5	169.6	93.4	38.3	243	311	**11**	90	128	37	20.5
7	240.2	168.5	82.9	35.1	23.1	**15.2**	116	41	367	22	23.2
8	228.7	146.7	88.4	43.9	17.5	18.1	10	42	28.1	**2.8**	22.8
9	206	146.8	82.7	47.5	20.6	12.9	6	**4.5**	23.4	11	18.1
10	214.7	139.1	58.5	38.9	23.9	17.5	6.8	5	3.44	**2.2**	12.2
Best Average						**15.2**	**11**	**4.5**		**2.5**	**17.3**

Table 6. The XB measure with varied c and α on the dataset D_{f1} for the model NEKW_NOVL

XB ×1,000	α = 0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
c = 2	29538	400.1	624	1605	917	518	117	212	108	**14**	22
3	382.3	232	3567	3224	1438	838	390	248	32.2	31	**11**
4	342.7	223.8	104	1495	1244	1089	448	157	58.3	97	**19**
5	271.8	205	105	55.7	523	346	222	147	81.8	86	**25**
6	258.4	188.2	80.5	856	1275	286	158	113	73	42	**17**
7	215.7	171	69.5	41.8	17.7	**11.1**	54	33	101	36	50
8	228.8	179.1	86.4	41.5	18.5	13	**8.6**	67	38.4	59	27
9	219.7	170.3	76.7	38.5	22.3	12.7	70	37	**4.2**	68	16
10	184.3	156.2	69.1	46.7	21.3	13.2	7.3	**4**	11.1	9.6	5.1
Best Average						**11.1**	**8.6**	**4**	**4.2**	**14**	**18**

The fact that the best value of α for NEKW_OVL is higher than that for NEKW_NOVL can be explained as follows. In the NEKW_OVL model, entity names are counted as keywords and may cause noises for fuzzy clustering with respect to NE tags. So, the weight for the KW component, i.e., $1 - \alpha$, should be decreased to reduce that noise effect. However, hard clustering as experimented above might not be effected by such noises.

One may also have another observation on the experimental results. That is, for each value of α, let us take the average XB measure on different values of c. Figure 4 plots that average XB measure with varied α on the dataset D_{f1}. It shows that, when α is big enough, e.g. from about 0.3 in this test, the performances of NEKW_OVL and NEKW_NOVL are almost the same. Probably, for that threshold α, counting entity names in the KW component of a document makes nearly no difference.

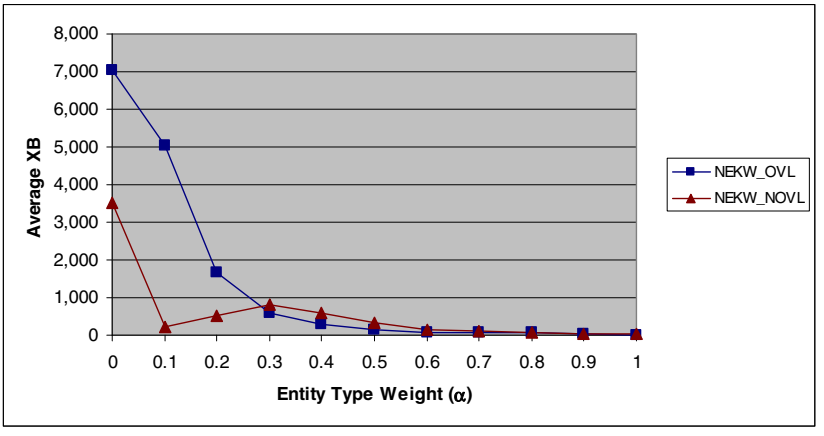

Fig. 4. Average XB with varied α on the dataset D_{f1}

Meanwhile, on the dataset D_{f2}, Table 7 and Table 8 show that the best values of α in average for NEKW_OVL and NEKW_NOVL are respectively 0.6 and 0.5. The lower best values of α as compared to those on D_{f1} are due to the documents containing not only NE tags but also topic tags, which rely on keywords. Figure 5. shows that, as for D_{f1}, NEKW_OVL and NEKW_NOVL perform nearly the same in terms of the average XB measure from a certain threshold α. Also, on both D_{f1} and D_{f2}, as for hard clustering, the fuzzy clustering quality is drastically improved when taking into account the latent named entity types, i.e., with $\alpha > 0$.

Table 7. The XB measure with varied c and α on the dataset D_{f2} for the model NEKW_OVL

XB ×1,000	**α = 0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
***c* = 2**	391.5	4258	996	227	797	170	46	48	**10.1**	24	24
3	22888	1624	925	650	698	235	82	44	169	**17**	45
4	6777	4820	825	711	343	235	74	178	26.5	**15**	21
5	1805	48.6	25.4	172	173	236	55	46	**20.2**	23	38
6	44.2	37.5	24.5	845	303	230	76	88	20.6	**11**	19
7	50.1	25.9	28.7	12.6	284	98	112	49	29.1	17	**9.5**
8	3613	30.8	397	20.8	153	**9.3**	64	52	31.3	17	13
9	51.2	38.2	30.5	18.4	7.5	6.4	**5.4**	50	33	37	8.7
10	41.5	39.9	24.4	10.4	8.5	**6.5**	28	77	9.3	33	11
Best Average						**7.9**	**5.4**		**15.1**	**14.3**	**9.5**

Table 8. The XB measure with varied c and α on the dataset D_{f2} for the model NEKW_NOVL

XB ×1,000	**α = 0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
***c* = 2**	90.3	5490	669	279	202	167	225	83.3	41	**16**	21
3	638.9	1187	1757	2143	566	321	61	62.4	20	**18**	68
4	3125	1486	826	549	465	126	115	35.4	37	21	**19**
5	59.8	67.9	654	1325	336	257	99	163	**17.4**	54	29
6	15563	37	1023	251	83	195	74	73.5	29	150	**16.9**
7	54.8	70.1	37	22	218	70	129	72	19	30	**14**
8	60.6	48	32.7	18.3	15	**9.1**	36	36.8	45	18	14.1
9	48.6	47.4	27.6	20	8.9	**6.4**	21	31.4	35	22	9
10	54.3	47.4	35.1	10.8	9.4	7.4	4	53	**2.7**	21	9.6
Best Average						**7.7**			**10**	**17**	**16.6**

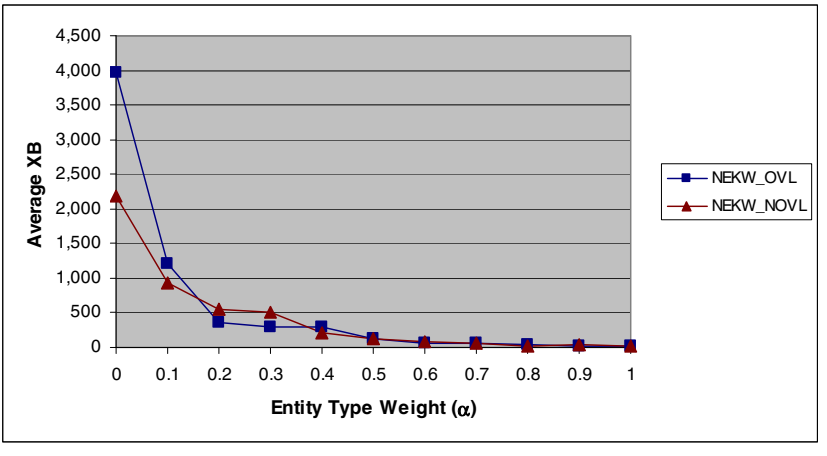

Fig. 5. Average XB with varied α on the dataset D_{f2}

6 Text Clustering in VN-KIM Search

Following KIM [15], we have developed a platform for Vietnamese Semantic Web called VN-KIM. It is firstly a knowledge-based system of popular named entities in Vietnam and the world. Currently VN-KIM ontology consists of 370 types and 115 relations. The knowledge base contains more than 210,000 selected named entities. It can automatically extract the type of a named entity in a web page written in Vietnamese and annotate that information in the web page, using the NE recognition engine for Vietnamese developed in [19].

For managing annotated web pages based on the combined entity-keyword VSM presented in Section 2, we have employed and modified Lucene [11], a general open source for storing, indexing and searching documents. In Lucene, a term is a character string and term occurrence frequency is computed by exact string matching. Here are our modifications for what we call S-Lucene:

1. Indexing documents over the four NE feature spaces corresponding to $\boldsymbol{N}$, $\boldsymbol{T}$, $\boldsymbol{N}\times\boldsymbol{T}$, and $\boldsymbol{I}$, besides the ordinary keyword space, to support the new model.
2. Modifying Lucene codes to compute dimensional weights for the vectors representing a document or a query, in accordance to the new model.
3. Modifying Lucene codes to compute the similarity degree between a document and a query, in accordance to the new model.

On the VN-KIM platform, we have implemented a semantic search engine called VN-KIM Search for text searching and clustering using named entities. The engine works on annotated Vietnamese web pages with the following essential features:

1. Its query syntax is designed to be similar to, and as expressive as, the Google's one.
2. However, being more powerful than a purely keyword-based search engine, its terms include both keywords and phrases representing named entities.
3. Moreover, it accepts named entity phrases that are not only simple entity names, but also complex constraints identifying named entities of user interest.
4. Besides, resulting web pages can be clustered with respect to the keywords and named entities that they contain.

VN-KIM Search has been then adapted for English and demonstrated using KIM ontology and NE recognition engine. As realized in real-world application systems like Clusty [7] and Carrot2 [5], clustering is used in VN-KIM Search 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 (cf. [22, 27]). However, it is ontology-based clustering as presented above instead of simply keyword-based clustering.

Figure 6 shows a screen interface of VN-KIM Search with the query "*peace (country of Asia)*" for searching documents tentatively about peace with countries in Asia. A phrase put in the parentheses is not a normal sequence of keywords, but represents named entities, which in this example are countries in Asia like *Israel* or *China*. Actually, such a query is first mapped to a conceptual graph to look up satisfying named entities in the knowledge base of discourse, using the processing method in [4]. Then the search engine retrieves documents containing those named entities.

The right window displays some top answer documents with queried named entities and keywords highlighted, e.g. *Israel* and *peace* for this example query. The left window displays hierarchical clusters of the answer documents. In this demonstration, the documents are clustered by two levels. The outer level is clustering by entity types and the inner level by entity names, combined with keywords. For instance, as indicated by the cluster labels, it shows that the dominant entities in the documents of the third outer cluster are of the type *Location*. Meanwhile, the four sub-clusters inside this cluster are more about *Israel*, *Cyprus*, *Pakistan*, or *Vietnam*, for instances. Figure 7 is a search result with highlighted named entities that are related to the query topic.

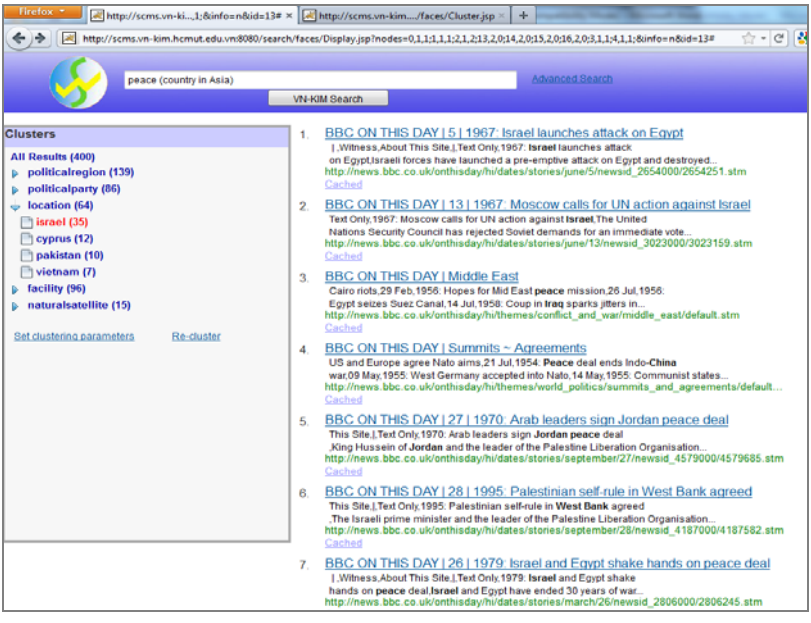

Fig. 6. Ontology-based searching and clustering in VN-KIM Search

Fig. 7. A resulting web page with highlighted named entities in VN-KIM Search

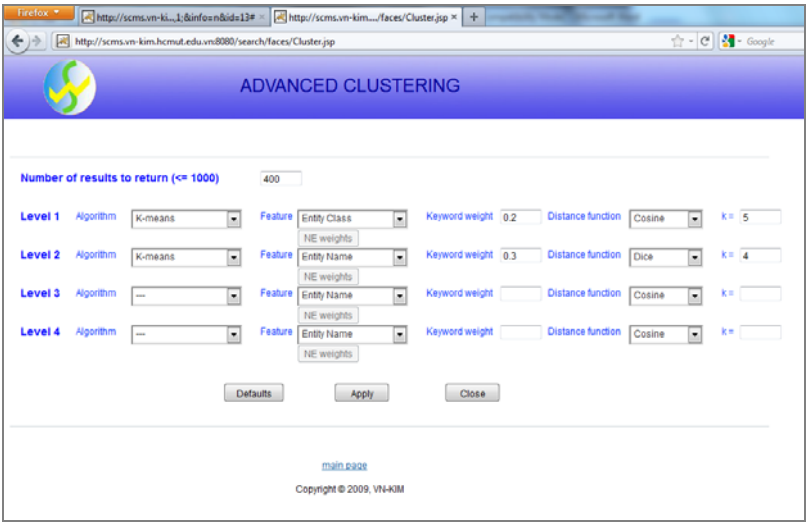

Fig. 8. Setting clustering parameters in VN-KIM Search

Figure 8 shows the interface to set the clustering parameters in VN-KIM Search. Answer documents could be clustered up to four levels. For each level, the user can choose a clustering algorithm (*k*-means or *c*-means), named entity features and their weights, a weight for the keyword component as expressed in Equation 2, a distance function (Cosine, Dice, Manhattan, or Euclidean), and a number of clusters. The current setting is for the clustering results in Figure 6.

7 Conclusion

We have presented a multi-vector space model for document representation, searching, and clustering. It is an extension of the VSM that represents a document as a linear combination of a vector on keywords and vectors on features of named entities occurring in the document. Our experimental results using the proposed 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 document are important to define its contents. In particular, taking into account named entity types, which are covered under their textual forms, drastically improves clustering quality as compared to the purely keyword-based VSM, for both hard and fuzzy clustering on the testing datasets. Second, they show that our model is suitable for representing the subjects of documents involving named entities like Reuters-21578 ones.

One can also observe from the experimental results that optimal weighting of the NE and KW components for clustering depends on document contents. For a dataset whose documents have only NE tags, e.g. D_{f1} in the experiments, the best value of α in average is close to 1, meaning that the NE component plays a major role. For a dataset whose documents have both NE and KW tags, e.g. D_{f2}, the best value of α in average is smaller. Besides, the overlapping and non-overlapping variations of the

proposed model have little difference in performance when the NE component weight is big enough.

The model also supports hierarchical clustering for which each layer uses a certain clustering objective corresponding to a NE feature. We have demonstrated that in the semantic search engine VN-KIM Search. For future work, 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.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading (1999)
2. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
3. Cao, T.H., Le, K.C., Ngo, V.M.: Exploring Combinations of Ontological Features and Keywords for Text Retrieval. In: Ho, T.-B., Zhou, Z.-H. (eds.) PRICAI 2008. LNCS (LNAI), vol. 5351, pp. 603–613. Springer, Heidelberg (2008)
4. Cao, T.H., Mai, A.H.: Ontology-Based Understanding of Natural Language Queries Using Nested Conceptual Graphs. In: Croitoru, M., Ferré, S., Lukose, D. (eds.) ICCS 2010. LNCS, vol. 6208, pp. 70–83. Springer, Heidelberg (2010)
5. Carrot2: Open Source Search Results Clustering Engine, `http://project.carrot2.org/architecture.html`
6. 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, 261–272 (2006)
7. Clusty Search: Clustering Search Engine, `http://clusty.com`
8. Duong, V.T.T., Cao, T.H., Chau, C.K., Quan, T.T.: Latent Ontological Feature Discovery for Text Clustering. In: Proceedings of the 7th IEEE International Conference on Research, Innovation and Vision for the Future - in Computing and Communication Technologies, pp. 264–271 (2009)
9. Friburger, N., Maurel, D., Giacometti, A.: Textual Similarity Based on Proper Names. In: Proceedings of the Workshop on Mathematical/Formal Methods in Information Retrieval at the 25th ACM SIGIR Conference, pp. 155–167 (2002)
10. 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)
11. Gospodnetic, O.: Parsing, Indexing, and Searching XML with Digester and Lucene. Journal of IBM DeveloperWorks (2003)
12. Hartigan, J., Wong, M.: Algorithm AS136: A K-Means Clustering Algorithm. Applied Statistics 28, 100–108 (1979)
13. He, J., Tan, A.H., Tan, C.L., Sung, S.Y.: On Quantitative Evaluation of Clustering Algorithms. In: Wu, et al. (eds.) Clustering and Information Retrieval, pp. 105–133. Kluwer Academic, Dordrecht (2003)
14. Hotho, A., Maedche, A., Maedche, E., Staab, S.: Ontology-based Text Document Clustering. KI 16, 48–54 (2002)
15. Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.: Semantic Annotation, Indexing, and Retrieval. Journal of Web Semantics 2 (2005)

16. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
17. Meilă, M.: Compare Clusterings – an Information Based Distance. Journal of Multivariate Analysis, 873–895 (2007)
18. Montalvo, S., Martínez, R., Casillas, A., Fresno, V.: Bilingual News Clustering Using Named Entities and Fuzzy Similarity. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 107–114. Springer, Heidelberg (2007)
19. Nguyen, V.T.T., Cao, T.H.: VN-KIM IE: Automatic Extraction of Vietnamese Named-Entities on the Web. Journal of New Generation Computing 25, 277–292 (2007)
20. Niu, Z.-Y., Ji, D.-H., Tan, C.-L.: Using Cluster Validation Criterion to Identify Optimal Feature Subset and Cluster Number for Document Clustering. Information Processing and Management 43, 730–739 (2007)
21. Oliveira, J.V., Pedrycz, W. (eds.): Advances in Fuzzy Clustering and its Applications. John Wiley & Sons, Chichester (2007)
22. Osinski, S.: Improving Quality of Search Results Clustering with Approximate Matrix Factorisations. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 167–178. Springer, Heidelberg (2006)
23. Sekine, S.: Named Entity: History and Future. Proteus Project Report (2004)
24. Theodoridis, S., Koutroumbas, K.: Patern Recognition. Academic Press, London (2008)
25. 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, pp. 81–86 (2005)
26. Xie, X.L., Beni, G.: A Validity Measure for Fuzzy Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 841–847 (1991)
27. Zhang, D., Dong, Y.: Semantic, Hierarchical, Online Clustering of Web Search Results. In: Yu, J.X., Lin, X., Lu, H., Zhang, Y. (eds.) APWeb 2004. LNCS, vol. 3007, pp. 69–78. Springer, Heidelberg (2004)
28. Zhang, X., Jing, L., Hu, X., Ng, M., Zhou, X.: A Comparative Study of Ontology Based Term Similarity Measures on PubMed Document Clustering. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 115–126. Springer, Heidelberg (2007)