TEXT CATEGORIZATION USING EARTH MOVER'S DISTANCE AS SIMILARITY MEASURE

Hidekazu Yanagimoto and Sigeru Omatu
Osaka Prefecture University, 1-1, Sakai, Osaka, Japan

Keywords: Text Categorization, Earth Mover's Distance.

Abstract: We propose a text categorization system using Earth Mover's Distance (EMD) as similarity measure between documents. Many text categorization systems adopt the Vector Space Model and use cosine similarity as similarity measure between documents. There is an assumption that each of words included in documents is uncorrelated because of an orthogonal vector space. However, the assumption is not desirable when a document includes a lot of synonyms and polysemic words. The EMD does not demand the assumption because it is computed as a solution of a transportation problem. To compute the EMD in consideration of dependency among words, we define the distance between words, which needs to compute the EMD, using a co-occurrence frequency between the words. We evaluate the proposing method with ModApte split of Reuters-21578 text categorization test collection and confirm that the proposing method improves a precision rate for text categorization.

1 INTRODUCTION

Text categorization systems automatically categorize informations using human-labeled documents. The system uses the similarity measure between an unlabeled document and the labeled documents. The documents are represented as vectors with the Vector Space Model (VSM) and the cosine similarity is used in text categorization. When the cosine similarity is computed, we assume that the vector space is am orthogonal vector space. However, this assumption is not always fulfilled because a document includes a lot of synonyms and polysemic words. Hence, we need to propose new similarity measure without the assumption that each of words is uncorrelated.

We propose a text categorization system using the Earth Mover's Distance (EMD) as a similarity measure between documents. The EMD needs no assumption that each of words is uncorrelated. Alternatively, it demands a distance between the words. We define the distance according to relationship between the words. To capture the relationship between the words, we use the co-occurrence frequency of the words. The co-occurrence is represented as the co-occurrence probability and the distance is computed according to the probability. To evaluate the proposing method, we carry out an experiment that ModApte split of Reuters-21578 text categorization test collection is categorized with the proposing method and a conventional method. We confirm that the proposing method is superior to a conventional method in the view of a precision rate.

2 PREVIOUS WORK

Text categorization is the activity of labeling natural language texts with predefined categories. To realize the text categorization, various machine learning algorithms are applied to it(Sebastiani, 2002). Expert Network(Yang and Chute, 1994), which is one of the text categorization systems using k-nearest neighborhood(Mitchell, 1997), achieves high recall and precision. The Expert Network uses a cosine measure between an input document and a labeled document.

Text categorization systems usually deal with texts represented as vectors by the Vector Space Model (VSM)(Salton et al., 1975). In the VSM, documents

Yanagimoto H. and Omatu S. (2007).
TEXT CATEGORIZATION USING EARTH MOVER'S DISTANCE AS SIMILARITY MEASURE.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - ISAS*, pages 632-635
DOI: 10.5220/0002406606320635

are represented as real-value vectors whose elements are weights for indexing words. When a retrieval system retrieves documents which relate to a query, a similarity measure between a document and the query is calculated by a cosine measure. In this case we assume that the vector space consists of orthogonal basis vectors. However, this assumption does not always fulfill because of synonyms, polysemic words, co-occurrence of words. Dependency of words distorts the similarity measure between a document.

Wan et al.(Wan and Peng, 2005) propose a similarity measure regarding dependency between words. They use the Earth Mover's Distance (EMD)(Rubner et al., 2000) to calculate the similarity measure between documents. They use the electronic lexical database - WordNet(Miller et al., 1990) and define the distance depending on semantical vicinity. However, it is difficult to digitize the semantical vicinity since the vicinity is defined in linguistics. A well-known problem with thesaurus-based method is that general-purpose thesauri do not have sufficient vocabulary coverage crossing different applications. The thesaurus-based method cannot cover neologisms since almost all thesauruses are often maintained with man power. Hence, we tackle these problems by improvement of similarity measure.

3 TEXT CATEGORIZATION USING EMD

3.1 Vector Space Model

In the VSM, all documents written with natural languages is represented as vectors to deal with them on computers. An element of the document vector denotes a weight of an indexing word included in the document. The weight for the indexing words is calculated by tf*idf.

$$w^i_j = \mathrm{tf}^i_j \ \log \frac{N}{\mathrm{df}_j} \tag{1}$$

where w^i_j is a weight of a term T_j in a document Doc_i, tf^i_j is a term frequency of T_j in Doc_i, df_j is a document frequency of T_j and N is the total number of the documents. A document Doc_i is represented as the document vector $\mathbf{d}_i = (w^i_1, w^i_2, \cdots, w^i_V)$ where V denotes the number of vocabulary for the corpus.

The cosine measure is often used as a similarity measure between documents. The cosine measure $\mathrm{sim}_{\cos}(\mathbf{d}_i, \mathbf{d}_j)$ between between documents; Doc_i and Doc_j, is calculated below.

$$\mathrm{sim}_{\cos}(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \mathbf{d}_j^{\mathrm{T}}}{\| \mathbf{d}_i \|_2 \| \mathbf{d}_j \|_2} \tag{2}$$

where T denotes transposition of a matrix and $\| \mathbf{d}_i \|_2$ denotes a quadratic norm of $\mathbf{d}_i$. Since we make the norms of all documents be equal to 1 in an experiment, Equation (2) is transformed into below.

$$\mathrm{sim}_{\cos}(\mathbf{d}_i, \mathbf{d}_j) = \mathbf{d}_i \mathbf{d}_j^{\mathrm{T}} \tag{3}$$

3.2 Earth Mover's Distance

The EMD is a method to evaluate similarity between two multi-dimensional distributions in a feature space where a distance between features can be defined. The distributions are represented as signatures which have feature vectors and weights for the features. A multi-dimensional distribution P which is represented as the signature is $P = \{(\mathbf{p}_1, w_{\mathbf{p}_1}), (\mathbf{p}_2, w_{\mathbf{p}_2}), \cdots, (\mathbf{p}_m, w_{\mathbf{p}_m})\}$ where $\mathbf{p}_i$ is a feature vector and $w_{\mathbf{p}_i}$ is a weight for the feature vector. Now, let $Q = \{(\mathbf{q}_1, w_{\mathbf{q}_1}), (\mathbf{q}_2, w_{\mathbf{q}_2}), \cdots, (\mathbf{q}_n, w_{\mathbf{q}_n})\}$ be the second distribution.

A distance between the features is defined in the feature space and is called a ground distance. Let $\mathbf{D} = [d_{ij}]$ be a ground distance matrix where d_{ij} is the ground distance between the features; $\mathbf{p}_i$ and $\mathbf{q}_j$. Let $\mathbf{F} = [f_{ij}]$ be a flow matrix where f_{ij} is the flow between $\mathbf{p}_i$ and $\mathbf{q}_j$. Here, we want to find an optimal flow $\mathbf{F}^*$ where makes a following cost function be a minimum.

$$\mathrm{WORK}(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij} \tag{4}$$

The cost function is minimized under the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq n \tag{5}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{\mathbf{p}_i} \qquad 1 \leq i \leq m \tag{6}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{\mathbf{q}_j} \qquad 1 \leq j \leq n \tag{7}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min(\sum_{i=1}^{m} w_{\mathbf{p}_i}, \sum_{j=1}^{n} w_{\mathbf{q}_j}) \tag{8}$$

Using the optimal flow $\mathbf{F}^*$, the earth mover's distance $\mathrm{EMD}(P, Q)$ is defined.

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}{}^*}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}{}^*} \tag{9}$$

Since the resulting work $\mathrm{WORK}(P, Q, \mathbf{F}^*)$ depends on the size of a signature, we need to normalize the resulting work. If the resulting work is not normalized, the smaller signature is favorable and this similarity measure is not desirable.

3.3 Earth Mover's Distance for Texts

Documents need to be represented as the signatures to compute similarity between the documents using the EMD. Let a feature vector $\mathbf{p}_i$ be an indexing word and let a weight for the feature be a weight for the indexing word. A document Doc$_i$ is represented as a signature D_i = $\{(\mathrm{T}_1, w_1^i), (\mathrm{T}_2, w_2^i), \cdots, (\mathrm{T}_m, w_m^i)\}$ which includes all indexing words in only the document Doc$_i$.

We need to define the ground distance between indexing words to compute the EMD. We define the ground distance between the indexing words depending on relationship between the indexing words. The relationship between the indexing words is defined based on the co-occurrence of the indexing words. Now let co-occurrence frequency of T_i and T_j be $\mathrm{occur}(\mathrm{T}_i, \mathrm{T}_j)$. A conditional probability $P(\mathrm{T}_i|\mathrm{T}_i)$ which shows a probability that T_j occurs in a sentence including T_i is defined below.

$$P(\mathrm{T}_j|\mathrm{T}_i) = \frac{\mathrm{occur}(\mathrm{T}_i, \mathrm{T}_j)}{\sum_j \mathrm{occur}(\mathrm{T}_i, \mathrm{T}_j))} \tag{10}$$

The ground distance d_{ij} from T_i to T_j is defined below using the conditional probability $P(\mathrm{T}_i|\mathrm{T}_i)$.

$$d_{ij} = 1 - P(\mathrm{T}_j|\mathrm{T}_i) \tag{11}$$

Since the EMD is not similarity measure, the similarity measure between the documents is computed using the EMD. When the sum of the weights for the indexing words in a signature is 1, the next formula consists.

$$\begin{aligned} &\mathrm{EMD}(D_k, D_l) \leq 1 && (12) \\ &\quad \text{if } \textstyle\sum_i w_i^k = 1 \text{ and } \sum_i w_i^l = 1 \end{aligned}$$

We define the similarity measure $\mathrm{sim}_{\mathrm{EMD}}(D_k, D_l)$ between Doc$_k$ and Doc$_l$.

$$\mathrm{sim}_{\mathrm{EMD}}(D_k, D_l) = 1 - \mathrm{EMD}(D_k, D_l) \tag{13}$$

3.4 Text Categorization

The proposing system uses architecture of the Expert Network (ExpNet)(Yang and Chute, 1994). The ExpNet consists of two steps: a similarity computing step and a category rank step. In the similarity computing step, the ExpNet computes cosine measure between an unlabeled document and a labeled document. It selects top K cosine measure labeled documents and uses them to decide a category for the unlabeled document. Our system uses the EMD to select K labeled documents.

In the category rank step, the ExpNet uses the K labeled documents to decide a category for the unlabeled document. In the ExpNet$P(c_k|\mathbf{d}_i)$ is defined as a conditional probability of a category c_k related to a document Doc$_i$ judged by human previously. Given the labeled documents, the conditional probability is estimated as

$$P(c_k|\mathbf{d}_i) = \frac{\text{frequency of category } c_k \text{ for } \mathbf{d}_i}{\text{frequency of } \mathbf{d}_i \text{ in labeled documents}} \tag{14}$$

Relevance measure of each category $\mathrm{rel}(c_k|\mathbf{d}_i)$ is defined as a weighted sum of the cosine similarity.

$$\mathrm{rel}_{\cos}(c_k|\mathbf{x}) = \sum_{j=1}^{K} P(c_k|\mathbf{d}_j)\mathrm{sim}_{\cos}(\mathbf{x}, \mathbf{d}_j) \tag{15}$$

The $\mathbf{x}$ denotes a document vector of the unlabeled document. Since our method uses the EMD, $\mathrm{rel}(c_k|X)$ is defined as

$$\mathrm{rel}_{\mathrm{EMD}}(c_k|X) = \sum_{j=1}^{K} P(c_k|D_j)\mathrm{sim}_{\mathrm{EMD}}(X, D_j) \tag{16}$$

where X denotes a signature for the unlabeled document.

Table 1: Content of Reuters-21578 using experiments.

	documents	unique words
Training	7,733	17,488
Test	3,008	10,731

4 EXPRIMENTS

We carried out text categorization for Reuters-21578 text categorization test collection. We used ModAPte split of the Reuters-21578. The ModApte split consists of 9,603 training documents and 3,299 test documents. In the experiment we used 7,733 training documents and 3,008 test documents including more than one indexing word after we removed SMART stoplist and applied Porter algorithm(Porter, 1980). On average a training document and a test document are labeled with 1.2 categories. 115 different categories exist in the training documents and 93 different categories exist in the test documents. And 3 categories in the test documents do not exit in training documents. Table 1 shows the content of the training documents and the test documents. Since 3,407 indexing words out of 10,731 unique words in the test documents was not included in the training documents, we used only 7,324 words to compute the similarity between a training document and a test document.

The experiment was that a test documents was labeled one category $\hat{c}$ which has the maximum $\mathrm{rel}_{\cos}(c_k|\mathbf{x})$ or the maximum $\mathrm{rel}_{\mathrm{EMD}}(c_k|X)$ although

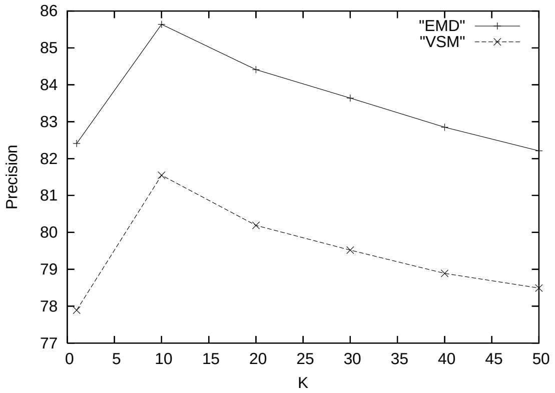

Figure 1: Precision rate on K=1, 10, 20, 30, 40, 50.

Table 2: The number of correct documents and error documents for VSM and EMD.

		VSM	
		correct	error
EMD	correct	2,359	217
	error	94	338

test documents were labeled with several categories. Hence, when an estimated category $\hat{c}$ is included in categories labeled in a test document as correct categories, we consider that that the text categorization algorithm can label the document correctly.

To evaluate each method based on previous idea we used precision rate for the test documents.

$$\text{precision} = \frac{\text{the number of correct labeled documents}}{\text{the number of all test documents}} \tag{17}$$

The precision rate depends on a value of K in Equation (15) and Equation (16). Figure 1 shows precision rates for cosine similarity (VSM) and our proposing method (EMD) on $K = 1, 10, 20, 30, 40, 50$. Our proposing method, EMD was superior to a conventional method, VSM on every K-values. The precision rate on $K = 10$ was the maximum value for VSM and EMD. The precision rate of VSM is 81.6% and the one of EMD is 85.6%. Hence, the difference of the precision rates is about 4.0%.

To discuss text categorization ability Table 2 shows the number of error documents and correct documents for VSM and EMD. The error documents for EMD was smaller than the one for VSM because of improvement of the precision rate.

In figure 1 we confirmed that it could improve the precision rate to regard the dependency of indexing words. In table 2 we think that we can improve our method. Our method makes a word related to too many words or contextually unrelated words. Hence, our method could not label the documents which VSM could label correctly. To make a word related to appropriate words increases the similarity between documents which the VSM can not evaluate. Hence, the number of error documents in VSM decreases. On the other hand, to make a word related to too many words boosts the similarity between documents which are unrelated. This cause the number of error documents, which can not exist in the VSM, to increase. We need to discuss how to define the distance between the indexing words beside the conditional probability $P(\mathrm{T}_i|\mathrm{T}_j)$.

5 CONCLUSION

We proposed a text categorization method using Earth Mover's Distance as a similarity measure. We realized to compute similarity between documents regarding the dependency of words using the Earth Mover's Distance. The distance between the words is defined with the conditional probability that one word occur with the other word in the same sentence. We confirm that the proposing method is superior to a conventional method using cosine similarity with Reuters-21578 text categorization test collection.

We will discuss how to define the distance between words beside the conditional probability and improve our proposing method.

REFERENCES

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An online lexical database. *International Journal of Lexicography*, 3(4):235–312.

Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill, New York, US.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Rubner, Y., Tomasi, C., and Guibas, L. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Wan, X. and Peng, Y. (2005). The earth mover's distance as a semantic measure for document similarity. In *the 14th ACM International Conference on Information and Knowledge Management*, pages 301–302. ACM Press.

Yang, Y. and Chute, C. G. (1994). An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems*, 12(3):252–277.