Research Journal of Applied Sciences, Engineering and Technology 12(2): 214-222, 2016
DOI:10.19026/rjaset.12.2323
ISSN:2040-7459; e-ISSN: 2040-7467

Submitted: July 2, 2015 Accepted: August 2, 2015 Published: January 20, 2016

Research Article

Document Similarity Measure Based on the Earth Mover's Distance Utilizing Latent Dirichlet Allocation

[1]Min-Hee Jang, [1]Tae-Hwan Eom, [1]Sang-Wook Kim and [2]Young-Sup Hwang
[1]Department of Computer and Software, Hanyang University, 17 Haengdang-Dong, Seongdong-Gu, Seoul 133-791,
[2]Department ofComputer Science and Engineering, Sun Moon University, Sunmoonro 221-70 Tangjoong-Myoon, Asan, Chungnam, 336-708, Korea

Abstract: Document similarity is used to search for such documents similar to a query document given. Text-based document similarity is computed by comparing the words in documents. The cosine similarity is the most popular text-based document similarity measure and computes the similarity of two documents based on their common word frequencies. It counts the exactly same words only, so cannot reflect semantic similarity between similar words having the same meaning. We propose a new document similarity measure to solve this problem by using the Earth Mover's Distance (EMD). The EMD enables to compute the semantic similarity of documents. To apply the EMD to the similarity measure, we need to solve the high computational complexity and to define the distance between attributes. The high computational complexity comes from the large number of words in documents. Thus, we extract the topics from documents by using Latent Dirichlet Allocation (LDA), a document generating model. Since the number of topics is much smaller than that of words, the LDA helps reduce the computational complexity. We define the distance between topics using the cosine similarity. The experimental results on real-world document databases show that the proposed measure finds similar documents more accurately than the cosine similarity owing to reflecting semantic similarity

Keywords: Cosine similairty, document similarity, earth mover's distance, latent dirichlet allocation, semantic similarity

INTRODUCTION

Document similarity is used to search for those documents similar to a given query document. Document search is widely used in such applications as duplicate document detection, technical paper searches and relevant document recommendations (Han and Kamber, 2006; Berry, 2003; Cao *et al*., 2006). For example, PubMed, the largest search site of medical documents, provides a service that recommends documents similar to each of the documents selected by a user (NCBI, 2009). To provide such a document search service, a method to compute the similarity between two documents correctly is essential. Document similarity can be computed by using the words in the documents as features. This similarity approach is called text-based similarity (Baeza-Yates and Ribeiro-Neto, 1999; Iosif and Potamianos, 2010; Robertson and Jones, 1976). In this study, we focus our attention on this text-based similarity. The representative text-based similarity is the cosine similarity (Salton *et al*., 1976; Steinbach *et al*., 2000), which is computed based on the frequency of the common words used in two documents. It decides two documents more similar when they have more common words inside. However, since it computes the similarity by only using the words that match exactly between the two documents and thus could not take into account such words that have a similar meaning but do not match exactly. To solve this problem, we propose a novel document similarity measure based on the Earth Mover's Distance (EMD) (Rubner *et al*., 2000). The EMD is a distance function used in various multimedia applications such as image, video and music search and has been known to provide good search results (Xu *et al*., 2010; Wichterich *et al*., 2008; Assent *et al*., 2006).

The EMD uses histograms with fixed-size bins to compare two data. The EMD computes the minimum work needed to transform one histogram into the other. The work is defined as the multiplication of the weight and the distance of the moved histogram bins. Since it uses the distance between the bins when computing the similarity, it can consider the similarity between the bins at different positions. If we assume that a document corresponds to a histogram and each word in a document does a bin of a histogram, we can compute

Corresponding Author: Young-Sup Hwang, Department of Computer Science and Engineering, Sun Moon,University, Sunmoonro 221-70 Tangjoong-Myoon, Asan, Chungnam, 336-708, Korea

the semantic similarity by considering the different words but having similar meanings by using the EMD.

However, there are some problems in computing the similarity by using the EMD. First, the computational complexity of the EMD is very high. For example, with n features, the complexity is O ($n^3 logn$) (Rubner *et al*., 2000; Jang *et al*., 2011). Since the number of words in a document is very large in general, the computation of the similarity by the EMD is too time-consuming. Second, the distance between words is needed to compute a document similarity by the EMD but is difficult to define. However, since the distance between words is difficult to measure as an exact number, computing the EMD itself is impossible.

To solve these problems, we propose an EMD-based approach to measure document similarity using the Latent Dirichlet Allocation (LDA) (Blei *et al*., 2003). The LDA is a probability model that analyzes topics in a document database in machine learning. When a user inputs the number of topics (m), LDA first extracts m topics latent in the document database and then computes the probability that each word in the database belongs to the extracted topics (Blei, 2004; Blei and Lafferty, 2006). By using the LDA, the proposed approach changes the features of a document from n words to m topics ($n \ll m$). The features of a document can be represented as m topics if we use the frequency of each word in the document and the probability that the word belongs to m topics. This has the benefit of reducing the time required to compute document similarity since the number of the features is substantially reduced using this method. The topic features are the result of considering all the meanings of the words in a document database by using the LDA (Wang and Grimson, 2007).

To calculate the EMD-based document similarity using the topic features, a distance measure between topics is needed. We calculate the distance using the probability that n words in a document database belongs to m topics. The proposed approach represents each topic as an n dimensional vector using the probability that the n word belongs to a topic and computes the distance between the topics in a document database by using the cosine similarity.

Since the proposed approach reduces the computational cost and measures the distance between topics, we can compute the document similarity based on the EMD. Even if the two compared documents are composed of different topics, the proposed approach has a high accuracy of document search because it can compute the semantic similarity between topics by using the topic distance and the EMD. Experimental results showed that the proposed approach had more accuracy and computed the similarity faster than the cosine similarity.

Table 1: Frequencies of words in document *A* and *B*

	Unix	Linux	Graphic	Multi-media
Document *A*	1	0	2	0
Document *B*	0	1	0	2

MOTIVATION

Cosine similarity: The cosine similarity represents each document in a document database as a vector to compute the document similarity. For each document, the number of words in a document database represents the dimensions of the vector and the frequency of each word represents an element of a vector. The cosine similarity between the two documents is represented as Eq. (1). In Eq. (1), A and B are the vectors and A_i and B_i is the *i*-th element of each vector:

$$\cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

The cosine similarity calculates the document similarity only using words matched exactly between two documents as seen in Eq. (1). That is, the cosine similarity cannot consider the word similarity for two words that do not match. This is a serious drawback because the document similarity cannot be computed correctly if words that have different forms but a similar or the same meaning are excluded.

For example, we assume that two documents *A* and *B* include the words in Table 1. The document *A* has ’Unix’ and ’graphic’ and the document *B* has ’Linux’ and ’multimedia’. ’Unix’ and ’Linux’ are similar in meaning as the OS. ’Graphic’ and ’Multimedia’ are similar also because they are related to computer graphics. That is, *A* and *B* are similar documents though they don’t have any exactly matching words. However, cosine similarity calculates the document similarity between *A* and *B* as 0 because these words do not exactly match.

Earth mover’s distance: In this study, we compute the document similarity by using the Earth Mover’s Distance (EMD) in order to consider the word similarity. The EMD represents the data as a histogram to compute the similarity between the two data. Let $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{ q_1, q_2, \ldots, q_n \}$ be two histograms.

p_i and q_i are the weight of *i*-th bin in each histogram. We assume that the total weights of P and Q are the same. The distance between P and Q can be measured by computing the minimum work, the cost to move the histogram of P to the histogram of Q. A work is defined as the multiplication of the ground distance d (a distance between bins) and the flow f (an amount of the weights of bins moved from one histogram to the other histogram). The ground distance is the standard distance such as the Euclidean distance or L_1 distance. Eq. (2) shows the EMD:

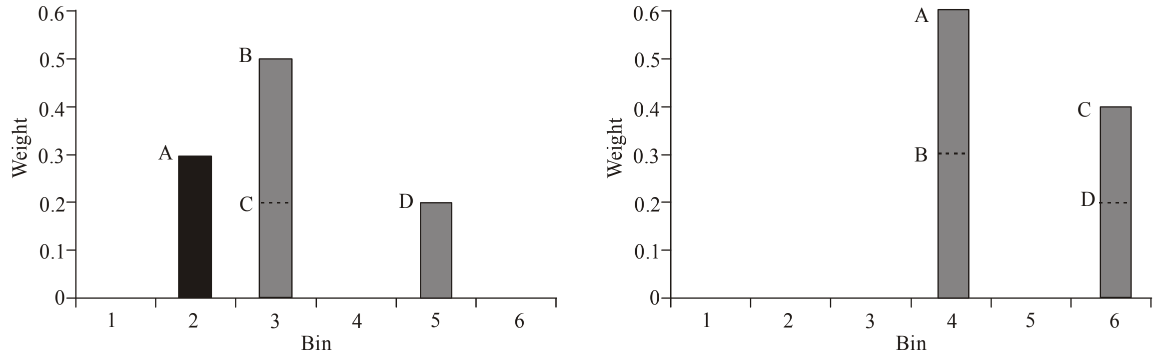


Fig. 1: The EMD between two histogram

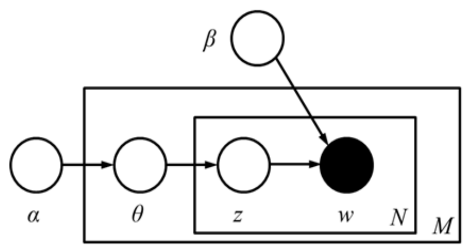


Fig. 2: Document generation in the LDA

$$\text{EMD}(P, Q) = \sum_{i=1}^{n} \sum_{j=1}^{n} \min\{d_{ij} \cdot f_{ij}\} \quad (2)$$

Figure 1 shows a computation example of the EMD. The x axis is the bins of a histogram and the y axis is the weight of each bin. Although there are various cases in transporting the histogram P to the histogram Q, the minimum work between the two histograms is computed when we move the bins in P to the same alphabets in Q:

(A part: (0.32) + B part: (0.3×1) + C part: (0.2×3) + D part:(0.2×1))

Since the EMD uses the distance between bins, it can compute the similarity between bins that are not matched exactly. If we represent a document as histograms, we can consider the word similarity by computing the EMD-based document similarity.

Applying the EMD to document similarity is difficult for two reasons:

- Defining the distance between words is difficult. A document must be represented as histograms in order to compute the similarity by the EMD. Each word in a document should be represented as a bin in a histogram and the frequency of a word as a weight. And the minimum work based on the ground distance between two documents should be calculated. However, the distance between words is difficult to measure by the exact numbers in contrast to the Euclidean distance or the L_1 distance. Therefore, computing the EMD itself is impossible because it is hard to define the ground distance.
- The computational complexity of the EMD is very high, $O(n^3 \log n)$ as stated earlier. Even though the ground distance between words exists, it is still difficult to compute the document similarity by the EMD because the number of n words in a document is very large in general. It has been shown that as the number of features increases 100 times, the time to compute the EMD increases by about 10,000 times (Rubner *et al*., 2000).

The proposed approach: In this section, we propose an approach to compute document similarity using the EMD. First, to reduce the number of features (n words), we extract m topics, a new feature of a document, by applying Latent Dirichlet Allocation (LDA) which is a generative model of a document ($n \gg m$). The LDA is a generative probabilistic model for a discrete data set such as documents and is used to analyze the topics in a document database (Blei *et al*., 2003). The LDA finds topics in a document database by analyzing the semantic relations between words in the document by statistical methods. The LDA is premised on the following two concepts. First, a document can have various latent topics. Second, each topic can be represented as the distribution of words. The LDA assumes that a document is generated as follows. When an author writes a document, they determine first which topics the document will include. After that, one topic among the choices of topics are selected. They then selects a word that has high probability that it belongs to the topic. The LDA assumes that a document is written from repetition of this process. Figure 2 illustrates the document generation process in the LDA. The rectangle N is the number of words in a document. The rectangle M is the number of documents in a document database. Circles is variables used in the LDA. The black circle is the variable which exists in the database and the white circle is the latent variable which does not exist in the database. The variable w denotes a word stored in the document database and z denotes the probability that this word belongs to a specific topic. The distribution of latent topics in the document is denoted by Θ and α denotes the distribution of latent topics in the document database. β

Table 2: Topics and representative words of each topic extracted by the LDA

Topic 1 (politics)		Topic 2 (health)		Topic 3 (medical)		Topic 4 (christian)	
Moralist	0.041	Food	0.010	Patient	0.010	God Ada	0.029
Human	0.024	Healthy	0.009	Doctor	0.008	Jesuit	0.016
Objectbuild	0.021	Cancer	0.008	Disectomy	0.007	Christian	0.013
Peoplenet	0.017	Disectomy	0.007	Vitiello	0.007	People	0.008
Donahue	0.017	Medicar	0.007	Candidacy	0.007	Christaian	0.008
System	0.015	Msg	0.006	Medicar	0.006	Biblical	0.007
Absolutist	0.013	Aider	0.006	Trebise	0.006	Lovecraft	0.006
Personal	0.012	Druggie	0.006	Subjective	0.006	Lord	0.006
Write	0.012	Infection	0.005	Articulate	0.005	Lifeblood	0.006
Claimant	0.011	Patient	0.005	Yeasteryear	0.005	Include	0.005

Table 3: Topic vectors of documents

	Topic 1	Topic 2	Topic 3	Topic 4
Document 1	0.00175	0.00058	0.95606	0.00006
Document 2	0.00092	0.84965	0.00078	0.11173
Document 3	0.95333	0.00117	0.00104	0.00117
Document 4	0.00052	0.00056	0.00063	0.95491
Document 5	0.23231	0.00050	0.72529	0.00056

denotes the distribution of the words of each topic in the document database.

The LDA extracts topics latent in a document database based on this document generative model. However, the information stored in a database are only words w. Thus, the LDA estimates z, Θ, α and β backwards from w. Among the various estimation methods, the LDA use the Expectation Maximization (EM) to estimates the topics. The EM tries to determine the maximum likelihood of the parameters in a probability model dependent on the latent variables. Details of the estimation in the EM can be found in Blei *et al*. (2003). The extracted topic consists of n words in a document database and the probability that each word belongs to the topic. Table 2 shows an example of 4 topics extracted from a document database used in Rennie (2008) by the LDA and the top 10 words that have a high probability for a word to belong to each topic. All examples shown in this study are real-world data.

In Table 2, the topics are labeled according to the distribution of words in each topic. The extracted topics are politics (topic 1), health (topic 2), medical (topic 3) and Christian (topic 4). Each topic has words of a similar meaning. Also, since the topic is labeled from the distribution of words, many of the same words exist between similar topics and they have also higher probability. For example, 'discectomy', 'Medicare' and 'patient' have a higher probability in both the health and medical topics. The proposed approach extracts m topics by the LDA and represents each document in a database as an m dimensional vector. The vector is computed by the multiplication of the frequencies of words in each document and the probabilities that the words belong to each topic. In other words, the distribution of topics in a document is analyzed from the frequencies of words in the document and the probabilities that the words belong to the topic. Since the frequencies of the words used in each document are different from the other documents, we should normalize them.

Table 3 is an example of the extracted topics by the LDA and the computed vectors of real-documents. A column is a topic and a row is a document. Since the proposed approach changes the features of a document from the number of n words to the number of m topics, we can reduce greatly the computational cost of the EMD ($n \gg m$). Also, these topic vectors consider the meaning of words in the document database, since these vectors are extracted from the frequency of the word in each document and the probability of the word belongs to the topic (Blei *et al*., 2003).

Although we reduce the number of features in a document, it is still not possible to compute document similarity using the EMD since the ground distance between the features does not exist. To solve the problem, we calculate the ground distance between the topics by using the cosine similarity. The proposed approach first represents each topic in a document database as an n-dimensional topic vector based on the n-words in a database. For each topic, the number of n-words in a document database represents the dimension of the vector and the probability that each word belongs to the topic represents an element of a vector.

The proposed approach applies the cosine similarity Eq. (1) to all pairs of the topic vectors. After calculating the cosine similarity, the distance between the topics is calculated as (1-cosine similarity). Since the proposed approach uses the probability that a word belongs to each topic, it can compute the accurate distance even if we use the cosine similarity. The proposed approach reduces the number of features in the document from n words to m topics ($n \gg m$) and calculates the distance between the features based on the relation between the words and the topics; thus, it can compute a document similarity by using the EMD.

As stated in above section, the EMD considers the similarity between the bins in different positions by using the ground distance between the bins. In our approach, a topic corresponds to the bin of a histogram and the ground distance between the topics exists. Therefore, the proposed approach can compute the semantic similarity between the documents even if the two documents consistof different topics.

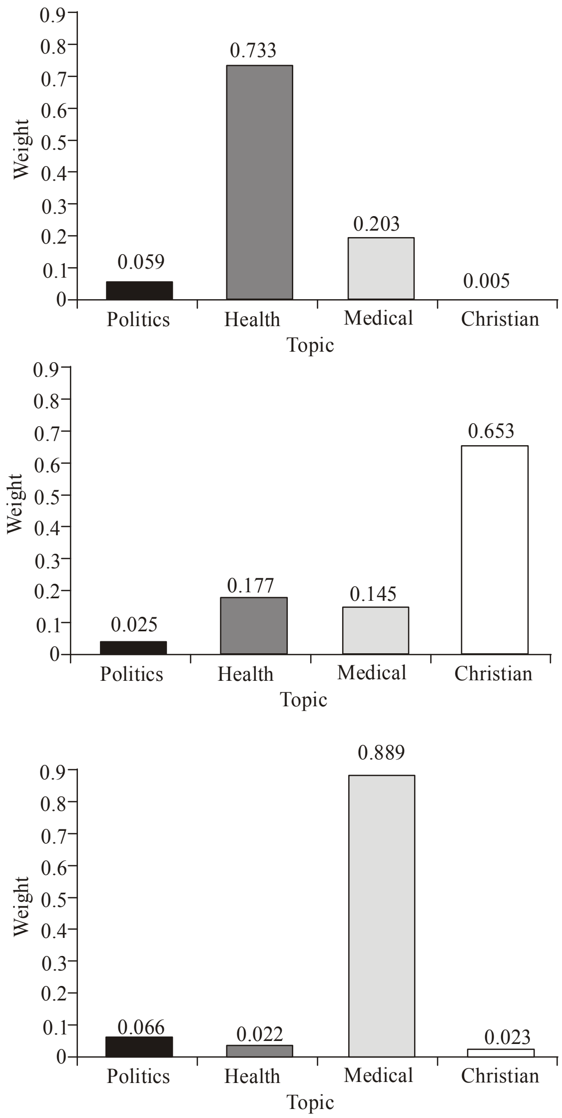


Fig. 3: Distribution of topics in documents

Table 4: Distance between topics

	Politics	Health	Medical	Christian
Politics	0.00	0.80	0.73	0.77
Health	0.80	0.00	0.50	0.84
Medical	0.73	0.50	0.00	0.79
Christian	0.77	0.84	0.79	0.00

We use Fig. 3 and Table 4 to show an example of the computation of the semantic similarity between the topics. Figure 3 shows 4 topics extracted by the proposed approach. In Fig. 3, the *x* axis is the extracted topics and the *y* axis is the weight. Table 4 shows the distances between the 4 topics calculated by the proposed approach. In Fig. 3, document *A* has a high weight on the topic ‘health’, document *B* on ‘Christian’ and document *C* on ‘medical’. Since ‘health’ and ’medical’ are topics related to health-care, document *A* is more similar with document *C* than document *B*. However the cosine similarity calculates the similarity between document *A* and *B* as 0.587, a similarity between document *A* and *C* as 0.433. The similarity between document *A* and *B* is measured higher than the similarity between document *A* and *C*.

This result comes from the problem that the cosine similarity is computed based only on the frequency of the common words used in twothe documents. The proposed approach solves this problem. The similarity of *A* is calculated by the proposed method between document *A* and *B* is 0.463 and the similarity between document *A* and *C* is 0.528, then document *A* is more similar to document *C* than *B*. As shown by this result, the proposed approach computes a document similarity more accurately since it considers the semantic similarity between the topics.

EXPERIMENTS

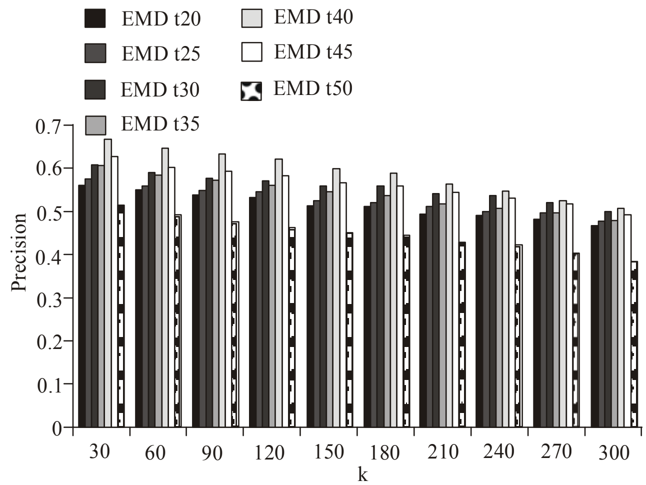
In this Section, we compare the accuracy and the similarity computation time of the proposed approach with those of the cosine similarity by various experiments.

Experimental environment: We used Newsgroup 20 for experiments.Newsgroup 20 is a document database widely used in document classification and clustering (Rennie, 2008; Bisson and Hussain, 2008). The Newsgroup 20 consists of 20,000 newsgroup articles and is divided into 20 topics. And these 20 topics are classified into 7 main topics. 4.5% documents in the Newsgroup 20 belong to more than two topics. Table 5 shows the topics in the Newsgroup 20. We randomly sampled 300 newsgroup articles for each topic; thus, the number of documents used was 6,000.

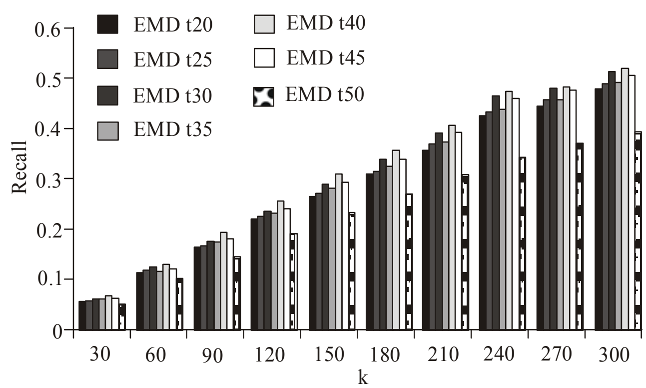
To verify the performance of the proposed approach, we measured the accuracy (precision and recall) of the *k*-nearest neighbor search and the similarity computation time. The correct answer set to a query document is the 300 documents in the same topic which the query document belongs to. The detailed process to measure the accuracy is as follows:

1) The similarities between all pairs of documents are measured by computing the proposed approach and the cosine similarity.
2) We searched the top *k* documents similar to a query document. The number of searched documents, *k*, was from 30 (10%) to 300 (100%).
3) We measured the accuracy by computing the precision and recall rate.
4) The processes in 2) and 3) were iterated over all the documents in the Newsgroup 20 and the results were averaged.

Experimental results: In the first experiment, we measured the precision and recall of the proposed approach according to the number of topics in the LDA. As stated in above section, users should give the



(a) Precision



(b) Recall

Fig.4: Accuracy of the proposed approach with the number of topics

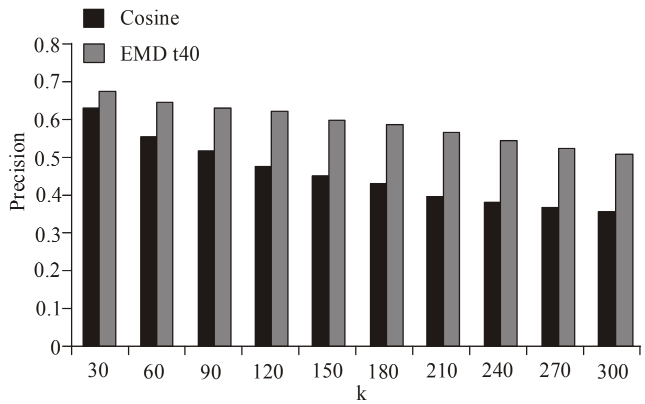
Table 5: Topics in the newsgroup 20

Computer-related topics	Contentious issues	Scientific subjects	Recreation
Graphics	Guns	Crypt	Autos
Ms-windows	Mideast	Electronics	Motorcycles
Ibm-hardware	Religion	Med	Baseball
Mac-hardware	Misc	Space	Hockey
Windows-x			
Miscellaneous	Social issues	Atheism	
Forsale	Christian	Atheism	

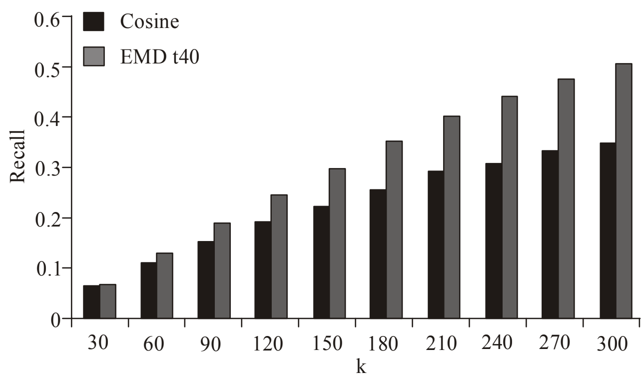
number of topics in the LDA. The accuracy of the proposed approach can be changed according to the number of topics. Figure 4 shows the measured accuracy of the proposed approach according to the number of topics. The *x* axis is the number of selected documents and the *y* axis is the precision and recall rate. The EMD t20, EMD t30, EMD t40 and EMD t50 mean that 20, 30, 40 and 50 topics are extracted from the document database. These are the same for all the experiments.

Figure 4 shows the experimental results. In Fig. 4, the accuracy increases as the number of topics increases but decreases after the EMD t40. When the number of extracted topics is 20, the same number of topics in the Newsgroup 20, the experimental result showed low accuracy. It is because the LDA uses a statistical method based on probabilities to extract the topics. Therefore, the extracted topics by the LDA do not exactly match the true topics. This makes the accuracy low at the EMD t20, even though the number of extracted topics is the same as the number of true topics.

The highest accuracy is at the EMD t40. Since a greater number than the true number of topics is extracted at the EMD t40, the possibility of the dissimilar documents tied to the same topics is lower than the EMD t20. On the other hand, though the possibility of the similar documents tied to different topics is high, the EMD can compensate for the similarity between these documents. As a result, the accuracy becomes high at the EMD t40. The reason why the accuracy falls greatly when the number of topic is over 40 is that the number of extracted topics is



(a) Precision



(b) Recall

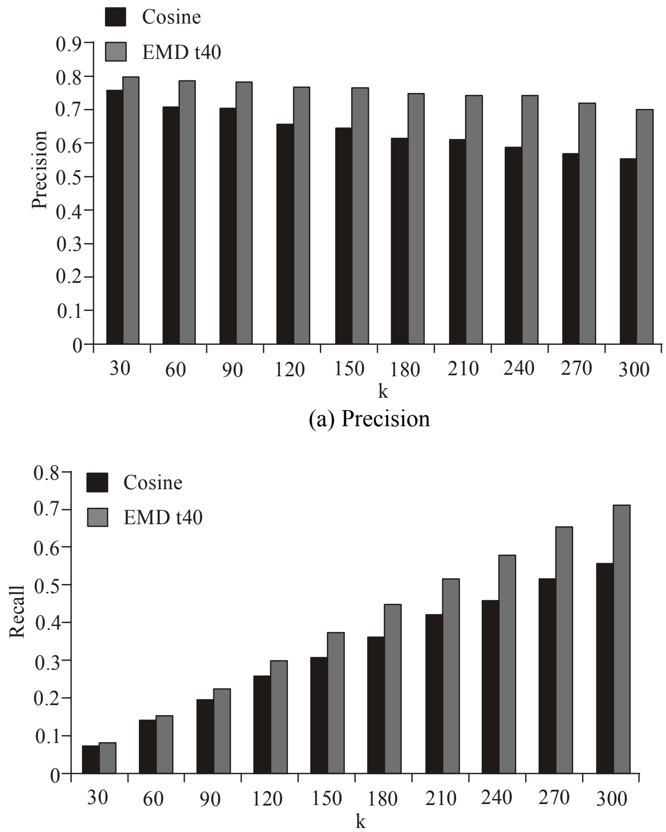
Fig.5: Comparison of the results between the proposed approach and the cosine similarity

much greater than the true number of topics. Since the similar documents are classified into various different topics, the EMD could not compensate for the similarity between these documents. In the following experiments, we fix the number of topics as 40.

The second experiment verifies the accuracy of the proposed approach with the cosine similarity by measuring the precision and recall. Figure 5 shows the results where the proposed approach had 32% higher accuracy on average in term of the precision and the recall. The cosine similarity computes the similarity only for the words that match exactly between the two documents. In comparison, the proposed approach considers the semantic similarity between the topics by the LDA and the EMD; thus, it shows higher accuracy. For such a reason, while the precision of the cosine similarity decreases greatly as the number of k increases, the precision of the proposed approach decreases relatively small. Likewise, the growth of the recall in the proposed method with the increase of the number of k is higher than that of the cosine similarity.

In the third experiment, we extended the correct answer set from the 20 topics to the 7 main topics and measured the precision and recall of the proposed approach and those of the cosine similarity. Figure 6 is the experimental results. As shown in the figure, the accuracy of both methods is higher than before because the answer set is extended. However, the overall tendency is almost the same as with the second experiment. On average, the precision and recall rate of the proposed approach are 19% higher than those of the cosine similarity.

In the final experiment, we measured the computation time of the proposed approach and the cosine similarity. The computation times of the proposed method are measured according to the number of topics. Figure 7 shows the results. The x axis denotes the method and the y axis denotes the computation time. As shown in the figure, the proposed approach showed better performance than the cosine similarity regardless of the number of topics. This is because the proposed approach changed the feature of a document from n dimensional word vector to m dimensional topic vector ($n \gg m$). Therefore, the proposed method has better performance even though the EMD has high computational complexity. As the number of topics increases, the computation time in the proposed approach increases, but it is still two times faster than the cosine similarity even at EMD t40 which has the highest accuracy.



(a) Precision

(b) Recall

Fig. 6: Comparison of the accuracy between the proposed approach and the cosine similarity on the main topics

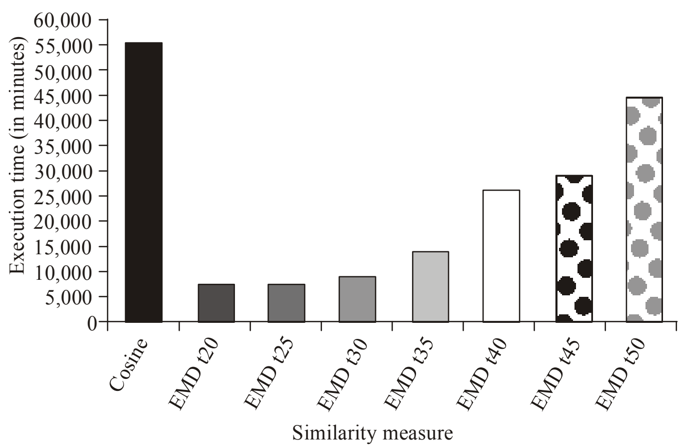


Fig. 7: Comparison of the computation time of the proposed approach and the cosine similarity

CONCLUSION

The cosine similarity, one of the most popular document similarity measures, computes similarity by counting the same words in two documents. It cannot consider the semantic similarity between words. In this study, we proposed a new similarity measure that can reflect semantic similarity by using the EMD. Since applying the EMD to word similarity suffers from the high computational complexity, we extracted topics from documents by using the LDA, a document generating model. The LDA solves the high computational complexity since the number of topics is much smaller than that of words. We also proposed a notion of a distance among topics based on the cosine similarity. The proposed method can search for documents more accurately by reflecting the semantic similarity. The experimental results on real-world document databases showed that the proposed method is not only more accurate but also faster than the original cosine similarity.

ACKNOWLEDGMENT

This study was supported by (1) the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2014R1A2A1A10054151) and (2) the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2015-H8501-15-1013) supervised by the IITP (Institute for Information & communication Technology Promotion).

REFERENCES

Assent, I., A. Wenning and T. Seidl, 2006. Approximation techniques for indexing the earth mover's distance in multimedia databases. Proceeding of the IEEE International Conference on Data Engineering(ICDE, 2006), pp: 1-12.

Baeza-Yates, R. and B. Ribeiro-Neto, 1999. Modern Information Retrieval. Addison-Wesley, Boston, MA, USA.

Berry, M., 2003. Survey of Text Mining: Clustering, Classification, and Retrieval. Springer-Verlag, New York, USA.

Bisson, G. and F. Hussain, 2008. Chi-Sim: A new similarity measure for the co-clustering task. Proceeding of the 7th International Conference on Machine Learning and Applications (ICMLA, 2008), pp: 211-217.

Blei, D., 2004. Probabilistic models for text and images. Ph.D. Thesis, U.C., Berkeley.

Blei, D. and J. Lafferty, 2006. Correlated topic models. Adv. Neur. In., 1: 147-154.

Blei, D., A. Ng and M. Jordan, 2003. Latent dirichlet allocation. J. Mach. Learn. Res., 3: 993-1022.

Cao, Y., J. Xu, T.Y. Liu, H. Li, Y. Huang and H.W. Hon, 2006. Adapting ranking SVM to document retrieval. Proceeding of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp: 186-193.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann, San Francisco, USA.

Iosif, E. and A. Potamianos, 2010. Unsupervised semantic similarity computation between terms using web documents. IEEE T. Knowl. Data En., 22: 1637-1647.

Jang, M.H., S.W. Kim, C. Faloutsos and S. Park, 2011. A linear-time approximation of the earth mover's distance. Proceeding of the 20th ACM International Conference on Information and Knowledge Management (CIKM, 2011), pp: 505-514.

NCBI (National Center for Biotechnology Information), 2009. PubMed, Retrieved from: http://www.ncbi.nlm.nih.gov/sites/entrez/.

Rennie, J., 2008. The 20 Newsgroups Data Set. Retrieved from: http://people.csail.mit.edu/jrennie/20Newsgroups/.

Robertson, S.E. and K.S. Jones, 1976. Relevance weighting of search terms. J. Am. Soc. Inform. Sci., 27: 129-146.

Rubner, Y., C. Tomasi and L. Guibas, 2000. The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vision, 40: 99-121.

Salton, G., A. Wong and C. Yang, 1976. A vector space model for automatic indexing. Commun. ACM, 18: 613-620.

Steinbach, M., G. Karypis and V. Kumar, 2000. A comparison of document clustering techniques. Proceeding of the ACM International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD, 400: 525-526.

Wang, X. and E. Grimson, 2007. Spatial latent dirichlet allocation. Adv. Neur. In., 20: 1-8.

Wichterich, M., I. Assent, P. Kranen and T. Seidl, 2008. Efficient EMD-based similarity search inmultimedia databases via flexible dimensionality reduction. Proceeding of the ACM SIGMOD International Conference on Management of Data(SIGMOD, 2008), pp: 199-212.

Xu, J., Z. Zhang, A.K.H. Tung and G. Yu, 2010. Efficient and effective similarity search over probabilistic data based on earth mover's distance. Proceeding of the VLDB Endowment, 3(1): 758-769.