Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

Paragraph-based K-Means Clustering by using Meaning-based Paragraph Division

Sa-Joon Park[1], Jae-Ho Kim[*2]

[1]*Faculty of Medical Industry Convergence, Daegu Haany University*
[2]*Department of Information Technology Engineering, Gangneung-Wonju National University*

ABSTRACT

As the number of electronic documents explosively increases, it becomes more and more difficult to retrieve information from them rapidly and accurately. To solve this problem, documents are clustered in various ways and generally K-Means algorithm is used to achieve it. K-Means algorithm is adequate to cluster so many documents rapidly and easily, but it does not consider the meaning of documents on clustering. In this research, we propose a document clustering technique of using meaning-based paragraphs. The proposed technique divides documents in a document set into meaning-based paragraphs by measuring similarity between sentences, chooses representative paragraphs having the maximum coherence value from each document, and then commits K-Means algorithm depending on them. In this paper, different from existing methods, we proposed a novel similarity function between two adjacent sentences by using WordNet as a ontology to calculate the similarity between words. And we introduced a method which can be used to calculate coherence of meaning-based paragraph by normalizing the sum of tf-idf value of words in the paragraph. We conducted experiments to prove the performance of the proposed technique by using the Reuter-21578 document set. The experimental result showed the document clustering technique of using meaning-based paragraphs improves the precision and the recall of document retrieval.

KEYWORDS: Document retrieval, Clustering, Meaning-based paragraph, K-means clustering method, Reuter-21578 document set, Precision, recall

ARTICLE INFO: Received 19 January 2017, Revised 9 February 2017, Accepted 10 February 2017.

[*]Corresponding author is with the Department of Information Technology Engineering, Gangneung-Wonju National University, 150 Namwon-ro Heungup-myon, Wonju, 220-711, KOREA.

E-mail address: kimjaeho@gwnu.ac.kr

1. 서 론

인터넷의 발달로 종이 문서 대신에 송수신이 가능하고 재활용성이 높은 전자 문서의 사용이 늘어남에 따라 생산되는 전자 문서의 양은 기하급수적으로 늘어나고 있다. 인터넷 검색을 통해 사용자가 원하는 정보를 빠르고 정확하게 찾는 것은 정보 검색의 중요한 기능이지만 검색 결과가 방대해질수록 원하는 정보를 찾는 데에는 시간과 노력이 더 소비되고 있다. 이런 문제를 해결하기 위해 정보 검색과 데이터 마이닝 분야에서는 문서 클러스터링 기법의 연구가 지속적으로 수행되고 있다[1,2].

정보 검색에서 문서간의 클러스터링 응집도가 높다는 의미는 관련성이 높은 문서들이 함께 검색의 결과로 도출될 가능성 높다는 의미이다. 유사성이 높은 문서들을 클러스터로 형성하고, 사용자 질의에 대해서 군집화된 클러스터의 관련도를 검사하여 관련성이 높은 클러스터에 있는 문서들을 추출해내면 정확성이 높은 결과를 유도할 수 있다[3,4]. 문서 클러스터링 연구의 목표는 검색된 결과의 정확율과 재현율 향상에 있다. 클러스터링하여 구축된 문서 집합은 사용자의 질의에 정확하고 신속한 결과를 도출한다.

문서 비교 시 문서에 포함되어 있는 단어들의 유사도와 사용 빈도가 높을수록 두 문서의 내용은 비슷하다고 볼 수 있다. 문서 클러스터를 만드는 과정에서 유사도가 높은 문서들은 우선적으로 하나의 클러스터로 형성되므로, 관련 있는 문서들은 같은 클러스터에 속하게 된다[4,5].

클러스터링을 수행 시 문서 단위로의 분석은 문서-문서간의 관련도를 측정하기에는 좋으나, 정보 검색에서 사용하는 질의는 통계적으로 의미 있는 빈도 벡터를 얻기에는 너무 적은 단어들로 구성되기 때문에 질의-클러스터의 관련도를 계산하기에는 적합하지 않다[4]. 또한, 문서들 사이의 관련도는 단어의 출현 빈도에 따라 결정이 되므로 의미적으로 유사성을 기반으로 한 문서 검색은 아니라는 문제점을 가지고 있다[6].

본 논문에서는 문서를 의미 기반 단위로 문단화하고, 문서의 유사도를 계산하고, 문서의 클러스터링을 수행하여, 검색 결과에 대한 정확성과 재현율을 높이는 클러스터링 기법을 소개한다. 문서를 의미 단위로 구성하면 다음과 같은 장점이 있다. 첫째, 문서를 의미 기반 문단 단위로 분류를 수행함으로써 의미에 중심을 둔 클러스터링이 가능하다. 둘째, 클러스터링을 수행하는 단위가 전체 문서에서 의미 기반 단위의 문단으로 축소되기 때문에 비교 연산을 수행하는 영역이 줄어든다. 이에 따라 검색 시스템의 검색 속도를 줄여주며 정확도를 높여준다. 또한, 외부 온톨로지를 활용하여 문단이 포함하고 있는 단어의 의미를 확장하여 검색이 가능하다. 셋째, 의미 기반 단위 문단 검색이 가능함으로써 사용자로 하여금 문서 내에서 특정 문단의 검색을 원활히 수행하게 해준다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 통해 주요 기술적인 면을 고찰하고 3장에서는 의미기반 문단 분할과 이를 활용한 문서 클러스터링 기법을 제시한다. 그리고 4장에서는 기존의 문서 클러스터링 기법과 의미기반 문단 분할 클러스터링의 비교 실험을 수행한다. 실험은 Reuter-21578 문서 집합을 사용하여 수행했다. 5장에서는 결론 및 향후 연구 방향을 제시한다.

2. 관련 연구

문서의 문단화 혹은 분할화 기법은 크게 유사도 기반, 그래픽 기반, 그리고 어휘 연결 기반으로 나뉠 수 있다[3].

유사도 기반 기법은 문장을 벡터 형태로 표현하

고 문장들 사이의 코사인 유사도 측정을 이용하여 유사도를 계산한다. c99 알고리즘[4]은 주제 문단을 분리하고 문장을 지역적으로 분류하기 위한 유사도 매트릭스를 사용하여 분리한다. 하지만 이는 문장 내에 유사어가 사용된 경우 문장 간의 유사도 측정이 올바로 되지 않는다는 문제점이 있다.

그래픽적 기법은 용어의 빈도수를 그래픽적으로 표현하고 이것을 이용하여 주제 문단을 파악하기 위해 사용한다. Dotplotting 알고리즘[7]은 문서 분할화의 그래픽적 접근법에 가장 일반적인 예이다.

어휘 연결 기반 기법은 용어의 다중 발생을 연결 한다 그리고 용어에 대해 두 개의 발생들이 너무 많은 문장들에 포함되어 있을 때 연결이 끊어진 것으로 분류하는 기법이다. Segmenter[8]는 어휘 연결 기반 방법을 사용하는 예이다.

상기의 방법들은 용어의 발생 빈도를 바탕으로 문단화가 이루어진다. 하지만 문서 작성자가 유사어를 활용한 경우 이러한 방법은 큰 성과를 기대하기 어렵다. [9]에서는 온톨로지를 이용하여 의미 기반의 문서 분할을 수행하는 기법을 제시했다.

[6]에서는 의미 흐름 기반으로 문단화 된 문서에 K-means 기법을 도입, 효율성과 성능을 개선시키는 방법을 제시했다.

3. 의미 기반 문서 클러스터링

문서를 의미를 중심으로 문단화 시키고, 문단화 된 문서들에 K-Means 기법을 활용, 의미 기반으로 문서들을 클러스터링 시키는 의미 기반 문서 클러스터링 기법을 소개한다.

3.1 의미 기반 문단 분할

문서를 문서 전체를 대상으로 클러스터링을 수행하기 보다는 문서는 문단으로 구성되어 있기 때문에 문서 보다는 문서의 내용을 함축하고 있는 문단을 찾아 문서를 대표한다. 이를 위해서 문서를 의미를 기반으로 하여 문단화 시킨다. 문단화 방법은 문서 내 인접한 문장들 간의 유사도를 측정하여 유사도 변화를 추적하여 수행한다. 인접한 문장들 간의 유사도의 급격한 변화는 문서 내 의미의 변화가 발생하였음을 알 수 있다. 즉, 인접한 문장들 간의 유사도를 비교하여 유사도가 급격히 떨어지는 부분에서 문단 분할을 수행한다.

인접한 문장들 간의 유사도 측정을 위해 Cosine 유사도 측정을 사용하면 문장 내에 유사어가 사용된 경우 문장 간의 유사도 측정이 올바로 수행되지 않는다. 이를 해결하기 위해 문장들내에서 사용된 유사어의 유사 정도를 파악하기 위해 WordNet을 온톨로지로 사용하고 문장 간의 유사도를 측정하기 위해 다음과 같이 변형 Cosine 유사도 측정 Cossim 방법을 사용한다.

$$
\begin{aligned}
&d = \{s_1, s_2, \cdots\cdots, s_n\}\\
&s_i = \{w_{i1}, w_{i2}, \cdots\cdots, w_{in}\}\\
&s_j = \{w_{j1}, w_{j2}, \cdots\cdots, w_{jn}\} \quad j = i+1\\
&S_i = \{W_{i1}, W_{i2}, \cdots\cdots, W_{in}\}\\
&S_j = \{W_{j1}, W_{j2}, \cdots\cdots, W_{jn}\} \quad j = i+1
\end{aligned}
$$

$$Cossim(s_i, s_j) = \frac{s_i \bullet s_j}{|s_i||s_j|} \tag{1}$$

$$MeanSim(s_i, s_j) = \frac{len(s_i)}{\sum_k short\_distance(W_{ik},\ W \in W_j)} \tag{2}$$

$$
\begin{aligned}
&0 \le MeanSim(s_i, s_{i+1}) \le 1\\
&Sim(s_i, s_{i+1}) = Cossim(s_i, s_j) * MeanSim(s_i, s_j)\\
&0 \le Sim(s_i, s_{i+1}) \le 1
\end{aligned} \tag{3}
$$

상기 식 (1)의 d는 문서 집합에 포함된 문서를

나타내며, 하나의 문서는 문장(s)들의 집합으로 표현된다. 문장(s)는 단어(w)들로 구성된다. W_{xy}는 S_x번 문장의 y번째 단어이다. 유사도 측정을 위해 우선 두 인접한 문장에 포함된 단어 W_{xy}의 발생 빈도수 w_{xy}를 바탕으로 단어를 포함한 문장 벡터 S와 단어의 빈도수를 포함한 s를 구축했다.

$Cossim$은 변형 Cosine 유사도 측정을 이용하여 두 인접한 문장 벡터 사이의 유사도를 측정한다. 기존의 Cosine 유사도를 확장, WordNet을 온톨로지로 활용하여 유사한 단어를 포함하여 인접한 문장 내에서 사용된 단어 간의 유사도를 측정한다.

상기 식 (2)에서 *short_distance* 함수는 WordNet에서 두 단어들 간의 최단 거리 값을 구하는 함수이다. 동일한 두 단어 간의 거리는 1이고 의미적으로 멀어질수록 그 값은 커진다. 문장의 길이를 인접한 두 문장을 구성하는 단어들 간의 최단거리의 총합으로 나눈 것을 문장 간의 의미 유사도 $MeanSim(s_i, s_j)$ 으로 측정했다. 이는 두 문장 내에 포함된 단어들이 가진 의미적 유사도가 높을수록 1의 값을 가지고 낮을수록 0에 가까운 값을 가지게 된다.

상기 식 (3)에서 이웃한 두 문장의 유사도 $Sim(s_i, s_{i+1})$)는 $Cossim$유사도와 두 문장의 의미 유사도 $MeanSim(s_i, s_j)$ 측정값을 곱해서 구한다. 두 문장 간의 유사도를 나타낸다.

문서 내의 문장들 간의 유사도 측정을 통하여 의미를 측정하였다. 인접 문장 간의 유사도 $Sim(s_i, s_{i+1})$ 을 활용하여 의미가 달라지는 문단을 구분하는 식은 다음과 같다.

$$MS(d) = \{Sim(s_1, s_2), Sim(s_2, s_3), \cdots\cdots, Sim(s_{n-1}, s_n)\}$$
$$Para(d) = \{i : SIM(i, i+1) \le AVR(MS(d)) - \alpha * STD(MS(d))\} \quad (4)$$

상기 식 (4)와 같이 문서내의 인접한 문장들 사이의 의미적 유사도를 구한다. 그리고 인접한 문장들 간의 유사도가 급격히 차이를 보이는 부분을 분단하여 문단화를 시킨다. 의미 유사도의 급격한 변화를 측정하기 위해 문서의 의미 유사도 *MS(d)*의 평균과 표준 편차를 활용한다. 즉, 인접한 문장 간의 의미 유사도가 문서내의 평균 문장 간의 의미 유사도 차이가 일정의 표준편차 범위 보다 큰 경우에 이를 의미의 급격한 변화로 판단하여 문단화 지점으로 선택한다. α 값은 사용자에 의해 지정될 수 있는 상수 값으로 실험에서는 1을 사용하였다.

3.2 의미기반 문단을 이용한 클러스터링

기존의 K-Means 클러스터링 기법은 문서 단위로 클러스터링을 수행하지만, 본 연구에서는 문장 간의 유사도를 측정하여 의미 기반으로 문단을 분할하여 문단 기반으로 K-Means 클러스터링 기법을 사용한다.

문단 기반으로 K-Means 클러스터링을 수행하기 위해는 문서에서 적절한 문단을 추출해야 한다. 이를 위해 가장 의미 응집성이 높은 문단을 추출한다. 문단의 의미 응집성은 추출된 문단에서 사용된 단어의 tf-idf[10,11] 값의 합을 문단의 크기로 정규화 하여서 구한다.

$$d_i = \{p_{i1}, p_{i2}, \cdots\cdots, p_{in}\}$$
$$tf(t,p) = \frac{f(t,p)}{\max f(w,p) : w \in p} \quad (5)$$

$$idf(t,d) = \log \frac{|d|}{|p \in d : t \in p|} \quad (6)$$

$$C(p_{ik}) = \frac{\sum \{tf(t,d_i) * idf(t,D) | t \in p_{ik}\}}{|p_{ik}|} \quad (7)$$

위의 식 (5)에서 d_i는 문서 집합 내 문서에 대해 3.1 절의 문단 분할 기법을 활용하여 분할된 문단(p_j)으로 구성된 문서이다. tf와 idf는 분할된 문단 내의 단어에 대해 적용하도록 수정되었다.

$C(p_{ik})$는 문단의 의미 응집도 산출식으로 문단 내의 단어들이 가지는 tf-idf 값의 총합을 문단의 크기 $|p_{ik}|$로 정규화한 값이다. 이때 $|p_{ik}|$는 문단 내에 존재하는 문장의 수를 의미한다. 따라서 적은 문장이 사용되면서도 의미적으로 중요한 단어가 사용된 문단이 해당 문서를 대표하는 문단으로 추출된다.

문서의 클러스터링은 문서 집합의 문서들에서 의미 응집도($C(p_{ik})$)가 가장 높은 문단을 추출하고 이에 대해 K-Means 기법을 이용하여 문서들을 클러스터링한다.

4. 실 험

의미 기반 문서 클러스터링의 성능을 검증하기 위해 실험을 수행하였다. 실험에서는 Reuter-21578 문서 집합을 이용하여. 문서 크기순으로 2,800개의 문서를 추출하였다. 추출한 후 본 논문에서 제안한 의미 기반 문서 클러스터링을 수행하기 위해 K 값은 50, 반복 횟수는 30으로 적용하였다. 5개의 질의어 - earn, acq, grain, wheat, corn - 을 이용하여 클러스터링의 성능을 측정하였다. 실험에서는 가장 기본적이면서도 널리 사용되는 재현율과 정확성을 선택하여 실험하였다.

<그림 1>은 문서 기반 K-Means 클러스터링과 의미 기반 문서 클러스터링 간의 재현율을 비교한 결과다.

의미 기반 문서 클러스터링의 재현율 실험 결과는 60~80%의 사이를 보였으며 earn, acq 질의어에 대해서는 문서 기반 K-Means 클러스터링보다 재현율에서 10% 이상 나타났으며 grain, wheat 질의어에 대해서는 5% 정도 좋은 결과를 보였다. 다만, corn에 대해서는 재현율이 낮게 나타났다.

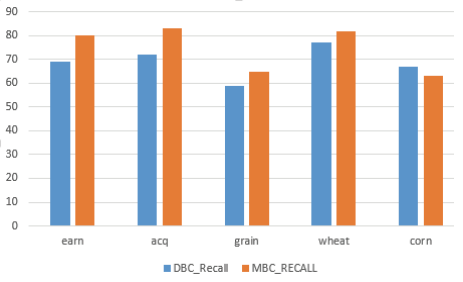

그림 1. 문서 기반 K-means 클러스터링 vs. 의미 기반 문서 클러스터링 재현율 비교
Figure 1. The comparison of recall between document-based K-means clustering and meaning-based document clustering

<그림 2>는 문서 기반 K-Means 클러스터링과 의미 기반 문서 클러스터링 간의 정확성을 비교한 결과다.

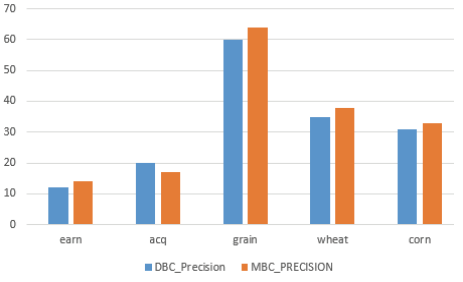

그림 2. 문서 기반 K-means 클러스터링 vs. 의미 기반 문서 클러스터링 정확성 비교
Figure 2. The comparison of precision between document-based K-Means clustering and meaning-based document clustering

정확성 비교의 실험결과에서 earn, grain, wheat, corn에 대해서는 문서 기반 K-Means 보다 5%이내에서 향상된 결과를 보였지만 acq 질의어에 대해서는 낮은 정확성 결과를 보였다.

5. 결 론

문서 클러스터링을 위해 문서를 문장 간의 유사도를 측정하여 의미 기반으로 문단화 했다. 문장 간의 유사도를 측정하기 위해서 WordNet의 단어 간의 유사도를 이용한 변형 Cosine 유사도 측정 방법을 제안했다. 문장 간의 유사도를 측정하여 유사도가 급격히 떨어지는 부분을 분리하여 문서 집합 내의 문서들을 의미 기반의 문단으로 분리한다. 각 문서는 문단으로 분리되면 분리된 문단으로부터 최대 응집도를 보이는 대표 문단을 선택한 후 이들을 이용하여 K-Means 기법을 Reuter-21578 문서 집합으로 클러스터링을 실험했다. 실험 결과에서 보듯이 의미 기반 문단 분할을 활용한 문단 기반 클러스터링이 문서 검색의 재현율과 정확성의 향상에 도움을 주었다. 따라서 문서 내에서 의미를 정확히 파악하고 의미에 따른 문단을 분할하는 것은 검색 시스템의 성능에 영향을 주었다.

본 시스템의 유사도 측정 기법 향상을 위해 향후 다음 연구를 수행할 예정이다. 대용량 문서의 경우 문서 내의 문단의 의미는 다양하게 나타난다. 의미를 정확히 파악하고 문단을 더 정확하게 분할할 수 있는 방법과 문서의 클러스터링 응집도 향상을 위한 클러스터링 방법 연구를 수행할 예정이다. 문서내의 모든 문단들을 기준으로 문서가 클러스터에 존재할 적합도를 계산하고 이를 바탕으로 클러스터링을 수행하는 것이다. 이는 문서의 전체적 의미성을 이용하므로 보다 나은 의미 기반 문서 클러스터링을 수행할 것이라 예상한다.

References

[1] T. Radecki, *A model of a document-clustering-based information retrieval system with a Boolean search request formulation,* SIGIR 1980, pp. 334-344, 1980.

[2] M. Steinbach, G. Karypis, and V. Kumar, *A comparison of document clustering techniques*, KDD Workshop on Text mining, 2000.

[3] A. Leusik. *Evaluating document clustering for interactive information retrieval*, CIKM 2001, pp. 33-40, 2001.

[4] K. S. Lee, *A document ranking model based on vector space retrieval and cluster analysis in information retrieval*, KAIST, Ph.D. dissertation, 2001.

[5] N. Y. Kim, H. J. Oh, D. U. An, and S. C. Park, *Document clustering analysis based on similarity calculation between cluster centroids*, Proceedings of the 2002 IEIE Autumn Conference, Vol. 25, No. 2, pp. 119-122, 2002.

[6] S. J. Park, *Meaning-flow based clustering for document retrieval in a large document set*, KKITS, Vol. 8, No. 4, pp. 37-42, 2013.

[7] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, *Scatter/Gather: A cluster-based approach to browsing large document collections*, SIGIR'92, pp. 318-329, 1992.

[8] H. J. Jain, M. S. Bewoor, and S. H. Patil, *Context senstive text summarization using K means clustering algorithm*, IJSCE, Vol. 2. Issue 2. pp. 301-304, 2012.

[9] A. Gelbukh, N. Kang, and S. Han, *Combining sources of evidence for recognition of relevant passages in texts*, LNCS 3563, pp. 283-290, 2005.

[10] G. Salton, *Automatic information organization and retrieval*, McGraw-Hill, New York. 1968.

[11] J. Spark, *A statistical interpretation of term*

specificity and its application in retrieval, Journal of Documentation, Vol. pp. 28, 11-21, 1972.

[12] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird, *Learning collection fusion strategies*, Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, pp. 172-179, 1995.

[13] H. J. Jain, M. S. Bewoor, S. H. Patil, *Context senstive text summarization using K means clustering algorithm*, IJSCE, Vol. 2, Issue 2, pp. 301-304, 2012.

[14] A. Goswami, N. R. Jin, and G. Agrawal, *Fast and exact out-of-core and distributed k-means clustering*, Knowlege and Information Systems, Vol. 10, pp. 17-40, 2006.

[15] G. Salton, A. Wong, and C. S. Yang, *A vector space model for automatic indexing,* Commun. ACM, Vol. 18, No. 11, pp. 613-620, 1975.

의미 기반 문단 분할을 활용한 문단 기반 K-Means 클러스터링

박사준[1], 김재호[2]

[1] *대구한의대학교, 의료산업융합학부*

[2] *강릉원주대학교, 정보기술공학과*

요 약

전자 문서의 수가 폭발적으로 늘어남에 따라, 문서들로부터 정보를 빠르고 정확하게 검색하는 것은 더욱더 어려워지고 있다. 이 문제를 해결하기 위해, 문서들은 다양한 방법으로 클러스터링 되는데 일반적으로 K-Means 알고리즘이 이용된다. K-Means 알고리즘은 많은 문서들을 빠르고 쉽게 클러스터링 하는데 적합하지만, 문서 클러스터링에 이용한 경우 문서가 가진 의미를 고려하여서 클러스터링 하지는 않는다. 본 연구에서는 의미를 기반으로 문서를 문단화하여 클러스터링하는 기법을 제안하다. 제안된 기법은 문장간의 유사도를 측정하여 문서 집합 내의 문서들을 의미 기반의 문단으로 분리하고, 각 문서로부터 최대 응집도를 보이는 대표 문단을 선택한 후 이들을 이용하여 K-Means 클러스터링을 수행한다. 본 연구는 다음의 점이 기존의 연구 방법과 차이가 있다. 문장간의 유사도 측정시에 단어간의 의미 유사도를 계산하기 위해 WordNet 온톨로지를 이용했다. 그리고 문단내 포함된 단어들의 tf-idf 값들의 합을 문단의 크기로 정규화함으로써 문단의 의미 응집도를 계산하는 방법을 제시했다. 제안된 기법의 성능을 증명하기 위해 Reuter-21578 문서 집합을 이용하여 실험을 수행했다. 실험의 결과는 의미 기반 문단을 사용하는 문서 클러스터링 기법이 문서 검색의 정확성과 재현율을 향상시킴을 보였다.

Sa Joon Park received the bachelor's degree, the M.S. degree and the Ph.D. degree in the Department of Computer Science and Engineering from Chung-Ang University in 1990, 1994, and 2004, respectively. He has been a professor in the Faculty of Medical Industry Convergence at Daegu Haany University since 2005. His current research interests include artificial intelligence, semantic web, and mobile contents. He is a life member of the KKITS.

E-mail address: phdjoon@dhu.ac.kr

Jae Ho Kim received the bachelor's degree, the M.S. degree and the Ph.D. degree in the Department of Computer Science and Engineering from Chung-Ang University in 1988, 1990, and 2004, respectively. He has been a professor in the Department of Information Technology Engineering at Gangneung-Wonju National University since 1997. His current research interests include image processing, machine learning, semantic web. He is a life member of the KKITS.

E-mail address: kimjaeho@gwnu.ac.kr