

# 다목적 유전자 알고리즘을 이용한 문서 클러스터링<sup>†</sup>

(The Document Clustering using Multi-Objective Genetic Algorithms)

이 정 송\*, 박 순 철\*\*

(Jung Song Lee and Soon Cheol Park)

**요 약** 본 논문에서는 텍스트 마이닝 분야에서 중요한 부분을 차지하고 있는 문서 클러스터링을 위하여 다목적 유전자 알고리즘을 제안한다. 문서 클러스터링에 있어 중요한 요소 중 하나는 유사한 문서를 그룹화 하는 클러스터링 알고리즘이다. 지금까지 문서 클러스터링에는 k-means 클러스터링, 유전자 알고리즘 등을 사용한 연구가 많이 진행되고 있다. 하지만 k-means 클러스터링은 초기 클러스터 중심에 따라 성능 차이가 크며 유전자 알고리즘은 목적 함수에 따라 지역 최적해에 쉽게 빠지는 단점을 갖고 있다. 본 논문에서는 이러한 단점을 보완하기 위하여 다목적 유전자 알고리즘을 문서 클러스터링에 적용해 보고, 기존의 알고리즘과 정확성을 비교 및 분석한다. 성능 시험을 통해 k-means 클러스터링(약 20%)과 기존의 유전자 알고리즘(약 17%)을 비교할 때 본 논문에서 제안한 다목적 유전자 알고리즘의 성능이 월등하게 향상됨을 보인다.

**핵심주제어** : 문서 클러스터링, k-means 클러스터링, 다목적 유전자 알고리즘, 유전자 알고리즘

**Abstract** In this paper, the multi-objective genetic algorithm is proposed for the document clustering which is important in the text mining field. The most important function in the document clustering algorithm is to group the similar documents in a corpus. So far, the k-means clustering and genetic algorithms are much in progress in this field. However, the k-means clustering depends too much on the initial centroid, the genetic algorithm has the disadvantage of coming off in the local optimal value easily according to the fitness function. In this paper, the multi-objective genetic algorithm is applied to the document clustering in order to complement these disadvantages while its accuracy is analyzed and compared to the existing algorithms. In our experimental results, the multi-objective genetic algorithm introduced in this paper shows the accuracy improvement which is superior to the k-means clustering(about 20 %) and the general genetic algorithm (about 17 %) for the document clustering.

**Key Words** : Document Clustering, k-means clustering, Multi-Objective Genetic Algorithm, Genetic Algorithm

<sup>†</sup> 이 논문은 2011년 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행되었음(No. 2011-0004389) 그리고 이 연구에 참여한 연구자는 2단계 BK21 사업의 지원비를 받았음

\* 전북대학교 컴퓨터공학부

\*\* 전북대학교 컴퓨터공학부

## 1. 서 론

클러스터링이란 대량의 데이터들을 의미 있는 그룹으로 분류하는 것을 의미한다. 클러스터링에서 문서 클러스터링은 정보 검색 분야에 중요한 부분을 차지하고 있다. 문서 클러스터링은 대량의 문서들을 유사한 문서들로 그룹화 하여 특정 주제에 따라 자동 분류하는 것을 뜻한다[1]. 문서 클러스터링은 사용자가 원하는 정보를 포함하고 있는 문서를 탐색하는데 소비되는 시간을 절약하며 검색 효율성을 높일 수가 있다[2]. 가장 일반적인 문서 클러스터링 알고리즘에는  $k$ -means 클러스터링[3]과 유전자 알고리즘(Genetic Algorithm)[4] 등이 있다.

$k$ -means 클러스터링은 클러스터링 문제를 해결하는 가장 간단한 알고리즘 중 하나로 구현의 용이성과 우수한 시간 효율성으로 문서 클러스터링에 많이 적용되고 있다. 하지만 초기 클러스터 중심에 따라 성능차가 커진다는 문제점을 가지고 있다.

또 다른 문서 클러스터링 방법은 현재 활발히 연구가 진행되고 있는 유전자 알고리즘을 통한 클러스터링 방법이다[5, 6]. 유전자 알고리즘을 이용한 문서 클러스터링은 유전자의 진화 개념을 이용하여 클러스터링에 필요한 요소를 개체, 염색체, 유전자에 대입한 알고리즘이다. 유전자 알고리즘을 이용한 문서 클러스터링은  $k$ -means 클러스터링보다 성능이 우수하나 세대를 평가하는 목적 함수를 정의하는데 문제점이 있다. 또한, 적절치 못한 목적 함수를 정의 하였을 때 지역 최적해(Local Optimal Solution)에 수렴 할 수 있다는 단점을 내포하고 있다[7].

본 논문에서는 현재까지 문서 클러스터링 연구에 있어 다목적 유전자 알고리즘을 적용한 사례가 알려지지 않는 가운데 처음으로 다목적 유전자 알고리즘을 문서 클러스터링에 적용해 보았다. 적용한 결과  $k$ -means 클러스터링과 유전자 알고리즘이 갖는 단점을 개선하여 성능이 향상됨을 알 수 있었다.

본 논문의 구성은 2장에서는 다목적 최적화 문제 및 다목적 유전자 알고리즘에 대하여 설명하고, 3장에서는 제안하는 다목적 유전자 알고리즘을 이용한 문서 클러스터링 방법에 대하여 논의한다. 마지막으로 문서 클러스터링에 있어 세 가지 알고리즘( $k$ -means 클러스터링, 유전자 알고리즘, 다목적 유전 알고리즘)을 비교 및 분석하였다.

## 2. 다목적 유전자 알고리즘

### 2.1 다목적 최적화 문제

다수의 목적 함수를 동시에 최적화 할 수 있는 해를 찾는 것을 다목적 최적화 문제(Multi-Objective Optimization Problem)라 한다[8].

다목적 최적화 문제에서 가장 이상적인 해는 모든 목적 함수들을 최적화하는 해이다. 하지만 각 목적 함수들은 서로 경합하기 때문에 어떤 목적 함수를 개선시키는 것은 다른 목적 함수를 악화시키는 것이 된다. 즉, 모든 목적 함수들을 동시에 최적화 하는 것은 일반적으로 불가능하다. 따라서 목적함수들 간의 타협점을 찾아 최적의 해를 찾아야 하며 이러한 해집합을 파레토 최적해(Pareto Optimal Solution)라 한다[9].

### 2.2 파레토 최적해

다목적 최적화 문제에서 최적의 해를 찾기 위해서는 목적 함수 공간의 해를 비교하여 얻을 수 있다. 두 해의 비교에는 지배(Dominance)관계를 사용하는데 지배 관계 정의는 식 (1)과 같다.

$$\begin{aligned} x, y \in S \\ \forall i \in \{1, 2, \dots, k\}: f_i(x) \leq f_i(y) \\ \text{and } \exists j \in \{1, 2, \dots, k\}: f_j(x) < f_j(y) \end{aligned} \quad (1)$$

최소화 문제로 가정 하였을 때, 식 (1)을 만족하면 해  $x$  는 해  $y$  보다 더 좋은 해로 평가받고  $x < y$  로 표현한다. 이를 해  $x$  가 해  $y$  를 지배(Dominate)한다고 하며,  $x$  를 비지배해(Non-Dominate Solution)라 한다. 또한, 두 해 중 지배관계를 명확하게 알 수 없는 경우가 있을 수 있다. 이 경우 두 해는 서로 비지배 관계에 놓여 있다고 표현하고 서로 동등한 수준의 해로 간주된다. 이러한 비지배해들의 집합을 파레토 최적해 또는 파레토 프론트(Pareto Front)라 한다.

### 2.3 다목적 유전자 알고리즘

다목적 최적화 문제를 풀기 위해 여러 가지 알고리즘이 제시되어 왔지만 최초 시작 위치(탐색 공간)에 의존적이며, 지역 최적해에서 벗어나지 못하고 여러

개의 다양한 해를 찾지 못한다는 단점을 가지고 있다. 이러한 단점을 해결할 수 있는 방법 중 하나로 주목 받고 있는 것이 유전자 알고리즘이다[10].

유전자 알고리즘은 다목적 최적화 문제에서 비교적 파레토 최적해에 근사한 해들을 만들어 낼 수 있으며, 다양한 해들을 제공한다는 장점을 가지고 있다.

현재 까지 제안된 다목적 유전자 알고리즘 기법은 파레토 랭킹 평가 기법과 다목적 함수 가중치 부가 평가기법이 있다.

최근에는 파레토 랭킹 평가 기법중 비우월 정렬(NSGA : Non-dominated Sorting Genetic Algorithm)을 확장한 NSGA-II가 선호되고 있다. NSGA는 공유(Sharing)기법을 적용하여 파레토 최적해의 다양성을 유지하도록 하였다. 하지만 계산량이 많고 최적의 해를 유지하는 엘리트 보존 정책이 적용되지 않아 진화 과정에서 최적의 해들의 손실이 야기되었다. 이를 개선하기 위해 고속 비우월 정렬(Fast Non-Dominated Sort), 밀집도 거리 할당(Crowding Distance Assignment) 등의 연산들을 추가한 NSGA-II가 제안되었다[11, 12].

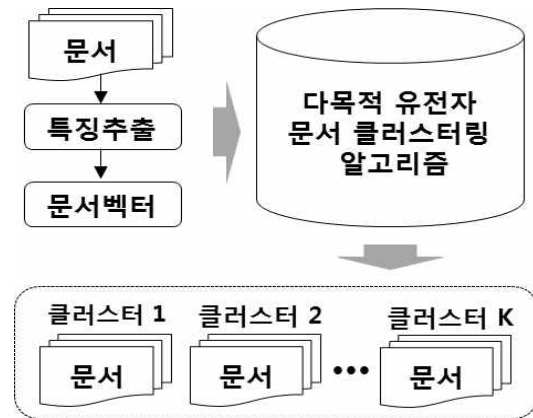
### 3. 다목적 유전자 알고리즘을 이용한 문서 클러스터링

유전자 알고리즘을 이용한 문서 클러스터링은 다른 클러스터링 알고리즘보다 성능이 우수하나 초기에 지역 최적해에 쉽게 수렴(조속 수렴 현상)하여 전체적인 성능 저하를 가져오게 된다.

본 논문에서는 다양한 목적 함수를 사용하는 다목적 유전자 알고리즘이 하나의 목적 함수에 의해 지역 최적해에 수렴 할 때 또 다른 목적 함수가 이를 방지할 수 있기 때문에 이를 문서 클러스터링에 적용해 보고 실험적으로 증명하고자 한다.

#### 3.1 전체 시스템 구조

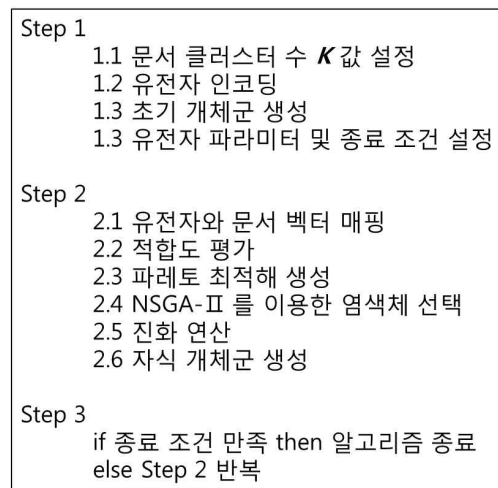
본 논문에서 제안하는 다목적 유전자 알고리즘을 이용한 문서 클러스터링은 문서 벡터를 구성하는 단계, 다목적 유전자 알고리즘을 이용하여 문서를 클러스터링하는 단계로 나누어지며 <그림 1>과 같다.



<그림 1> 문서 클러스터링 시스템 구조

문서 벡터를 구성하는 단계에서는 각 문서의 특징을 추출하고 이를 이용하여 문서 벡터를 구성한다. 다목적 유전자 문서 클러스터링 알고리즘에서는 문서 클러스터링에 적합한 유전자 구조와 진화 연산자 및 목적 함수를 사용하여 문서들을 클러스터링한다.

다목적 유전자 문서 클러스터링 알고리즘은 <그림 2>와 같다.



<그림 2> 다목적 유전자 문서 클러스터링 알고리즘

- 1) 문서 클러스터의 수  $K$  값을 설정하고 유전자 인코딩을 수행하여 초기 개체군을 생성한다. 또한, 유전자 파라미터와 종료 조건을 설정한다.
- 2) 유전자 인코딩 방식에 따라 문서 벡터와 매핑하여 적합도를 평가한다. 평가 후 파레토 최적해를 생성하고 NSGA-II 기법을 사용하여 염색체를

선택한다. 선택된 검색체들에 진화연산을 적용하여 자식 개체군을 생성한다.

- 3) 종료조건을 만족하면 알고리즘이 종료되어 현재 세대의 파레토 최적해를 최종해로 산정한다. 만족하지 않으면 Step 2를 반복한다.

### 3.2 다목적 유전자 문서 클러스터링 데이터 구조

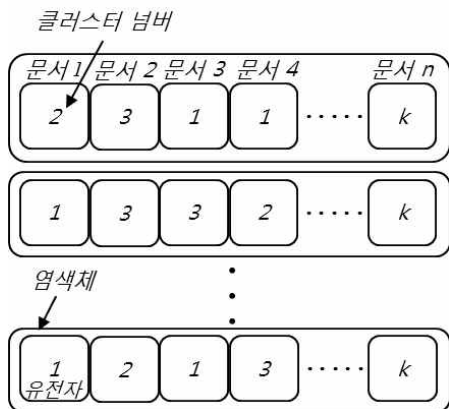
#### 3.2.1 특징 추출 및 문서 벡터 구성

본 논문에서는 여러 가지 문서 특징 추출 방법 중에서도 *Okapi's* 계산법[13]을 사용하여 문서의 용어들의 가중치를 계산하였다. 이 가중치를 이용하여 문서의 특징을 나타내는 문서 벡터를 구성하게 되며 식 (2)는  $t$  개의 용어 가중치를 가지는  $n$  번째 문서 벡터  $d$  를 나타낸다.

$$d_n = \langle w_{n,1}, \dots, w_{n,i}, \dots, w_{n,t} \rangle \quad (2)$$

#### 3.2.2 유전자 인코딩

유전자 알고리즘을 이용하기 위해서는 문제의 해가 되는 변수를 유전자와 비슷한 데이터 구조의 인코딩이 선행되어야 한다. 본 논문에서는 <그림 3>과 같이 초기 설정되는 클러스터의 수  $1 \sim K$  범위의 정수로 유전자를 인코딩하였다. 각 유전자는 문서를 나타내며 유전자의 값은 해당하는 문서가 포함되는 클러스터 번호를 나타낸다.



<그림 3> 유전자 인코딩 구조

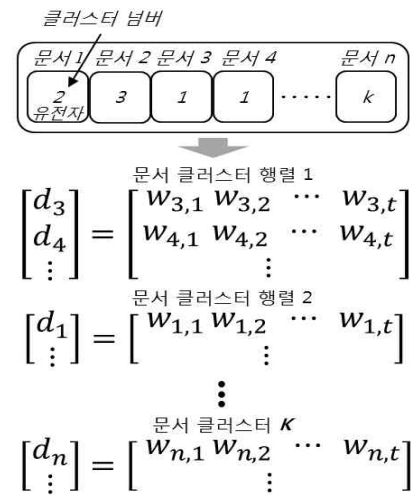
<그림 3>은  $n$  개의 문서들을  $K$  개의 그룹으로 클러스터링하고자 할 때 유전자 인코딩 구조이다. 첫 번째 검색체에서, 첫 번째 문서는 두 번째 클러스터, 두 번째 문서는 세 번째 클러스터, 세 번째 문서는 첫 번째 클러스터에 할당되어 있음을 나타낸다.

### 3.3 문서 클러스터 행렬

본 논문에서 제안하는 다목적 유전자 알고리즘을 이용한 문서 클러스터링에서는 검색체의 적합도를 평가하기 위해 문서 클러스터 행렬을 구성하는 단계가 필요하다.

구성 방법은 초기 설정되는 클러스터의 수  $K$ 개의 문서 클러스터 행렬들을 생성한다. 생성 후 유전자 인덱스 값과 동일한  $n$  번째 문서의 문서벡터를 클러스터 넘버와 같은 문서 클러스터 행렬에 추가한다.

예를 들어 <그림 4>에서 첫 번째 유전자의 인덱스는  $1$  이며 클러스터 넘버  $2$  를 가지므로 첫 번째 문서 벡터  $d_1$  을 문서 클러스터 행렬  $2$  에 삽입한다. 이와 같이  $n$  개의 모든 유전자에 적용하면 문서 클러스터 행렬의 구성이 완료된다.



<그림 4> 문서 클러스터 행렬

### 3.4 적합도 평가

적합도 평가에서는 일차적으로 검색체에 대하여 문서 클러스터 행렬이 구성된다. 구성된 후  $K$  개의 문서 클러스터 행렬에 목적 함수를 적용하여 생성된 문서 클러스터가 최적인지 평가하게 된다.

### 3.4.1 목적 함수

목적 함수는 현재 세대가 얼마나 해에 근접한지를 판단 할 수 있는 기준으로서 유전자 알고리즘의 전체 성능을 좌우한다.

본 논문에서 사용한 목적 함수는 클러스터링 결과를 평가하기 위해 사용되는 클러스터링 인덱스(Clustering Validity Index)를 사용하였다. 클러스터링 인덱스는 클러스터링 결과의 유효성을 평가하기 위해 사용되는 평가기준이며 최적의 클러스터가 생성되었을 경우 값은 최소값 또는 최대값을 가지게 된다.

문서 클러스터링을 위하여 본 논문에서 적용한 다목적 유전자 알고리즘인 NSGA-II의 시간 복잡도는  $O(MN^2)$ 이며  $M$ 은 목적 함수의 개수,  $N$ 은 개체수를 나타낸다. 이와 같이 클러스터링 인덱스를 목적 함수로 사용하는 다목적 유전자 알고리즘을 이용한 문서 클러스터링은 높은 계산 시간을 필요로 하므로 본 논문에서는 클러스터링 인덱스 중에 가장 간단한 CH(Calinski and Harabasz)[14]와 DB(Davis and Bouldin)[15]인덱스를 사용하였다.

CH 인덱스는 클러스터 그룹내 분산과 클러스터 그룹간 분산을 사용하며 최대값을 가질 때 적합한 클러스터라고 판단하며 식 (3)과 같다.

$$\frac{B}{W} \cdot \frac{n-k}{k-1} \quad (3)$$

$B$  는 클러스터 그룹간 분산(Between Cluster Sum of Squares),  $W$  는 클러스터 그룹내 분산(Within Cluster Sum of Squares),  $n$  은 문서의 수,  $k$  는 클러스터의 수를 나타낸다.

DB 인덱스는 그룹간 클러스터 중심 벡터의 유사도와 그룹내 클러스터 중심 벡터와 각 문서 벡터의 유사도를 사용한다. 여기서 유사도가 유클리디언 거리값이면 최소값, 코사인 유사도이면 최대값을 가질 때 적합한 클러스터라고 판단하며 식 (4)와 같다.

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i$$

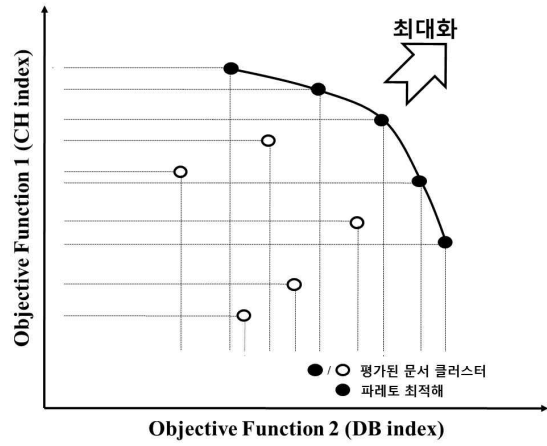
$$R_i = \max(R_{ij}), i \text{ and } j = 1 \dots n_c, i \neq j \quad (4)$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$n_c$  는 클러스터의 수,  $s$  는 그룹내 클러스터 중심 벡터와 문서 벡터와의 유사도,  $d$  는 그룹간 클러스터 중심 벡터의 유사도를 나타낸다.

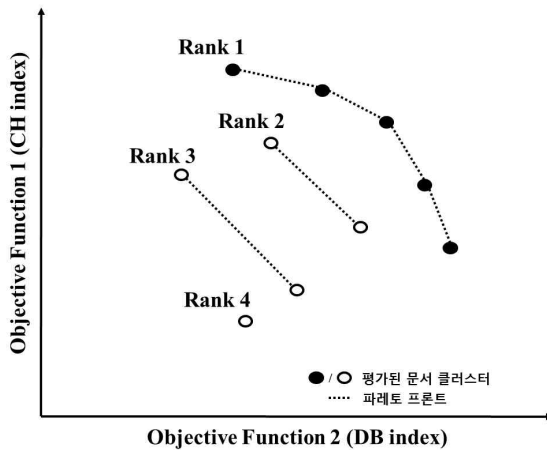
### 3.5 문서 클러스터링에서의 파레토 최적해 및 NSGA-II

각 세대의 염색체들 즉, 문서 클러스터들은 두 개의 목적 함수(CH, DB 인덱스)를 사용하여 적합도를 평가하고 목적 함수 공간으로 사상된다.



<그림 5> 문서 클러스터링에서의 목적 함수 공간

목적 함수 공간에서 두 개의 목적 함수 평가값이 최대일 때를 가정하면 <그림 5>와 같이 최대화 방향의 비지배해들을 선정하여 파레토 최적해를 생성한다.



<그림 6> 문서 클러스터링에서의 NSGA-II

다음에는 NSGA-II 기법을 사용하여 <그림 6>과 같이 같은 순위에 있는 비지배해들을 하나의 파레토 프론트에 위치시키고 밀집도를 비교하여 토너먼트 선택 연산을 수행한다. 마지막으로 선택된 염색체들은 진화 연산을 통해 다음 세대의 개체를 구성한다.

### 3.6 진화 연산자

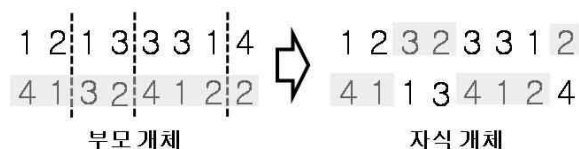
본 논문에서 진화 연산은 아래와 같이 선택 연산은 토너먼트 선택, 교배 연산은 다점 교배 연산, 돌연변이 연산은 확률적 돌연 변이 연산을 사용하였다[16].

#### 선택 연산자

NSGA-II에서는 불분명한 지배 관계에 놓인 개체들에 관하여 각 목적 함수에 따라 인접한 두 개체의 거리의 합으로 밀도를 측정한다. 이 밀도를 측정함으로써 불분명한 지배관계에 놓인 해들 사이의 순위를 결정한다. 순위를 결정하기 위해 선택 연산자는 기본적으로 토너먼트 선택을 사용한다.

#### 교배 연산자

교배 연산으로는 다점 교차 연산을 사용하였다. 이 점 교차 연산에서 확장된 다점 교차 연산은 <그림 7>과 같이 여러 점을 중심으로 나누어 부모 개체의 조합으로 자식 개체가 생성된다.



<그림 7> 다점 교차 연산

#### 돌연변이 연산자

돌연변이 연산자는 확률적 돌연변이 연산을 사용하였다. 선택되고 교배되어진 개체에서 각 유전자에 확률을 부여하여 임의의 값 이하의 유전자만 무작위로 변화시켜 새로운 개체를 생성한다.

## 4. 실험 및 결과 분석

본 논문에서는 *k*-means 클러스터링과 유전자 알고리즘, 그리고 다목적 유전자 알고리즘 사용하였을 때의 클러스터링 성능을 비교 및 분석하였다. 유전자 알고리즘과 다목적 유전자 알고리즘은 300개의 동일한 개체수와 세대수가 1,000이거나 성능 향상 없이 20번의 세대가 진행될 때 알고리즘이 종료되도록 설정하였다.

클러스터링 알고리즘의 성능을 비교하기 위해 데이터 셋은 한국일보-20000/한국일보-40075 문서 범주화 실험문서집합을 사용하였다. 총 3개의 테스트 셋을 만들어 각각 4개의 주제를 선정하였으며 <표 1>과 같다.

테스트 셋 1, 2는 비교적 구분이 명확한 주제로 선정하고 테스트 셋 3은 비슷한 주제로 선정하였다. 또한, *F-measure*[17]를 사용하여 각 클러스터링 알고리즘의 성능을 평가하고 객관성을 높이기 위해 100번을 반복하여 그 평균값을 성능 결과로 측정하였다.

<표 1> 실험 테스트 셋

테스트 셋	주제
테스트 셋 1	문화와 종교-생활-주거 결혼 정치-외교-외교(대북) 여가생활-실내 TV 사회-사회질서-사건사고(교통)
테스트 셋 2	여가생활-실내-공연 경제-가계 물가-가계 물가 문화와 종교-스포츠-야구 산업-제조업-비금속
테스트 셋 3	과학-자연과학-환경 과학-자연과학-화학 과학-자연과학-생물 과학-자연과학-물리

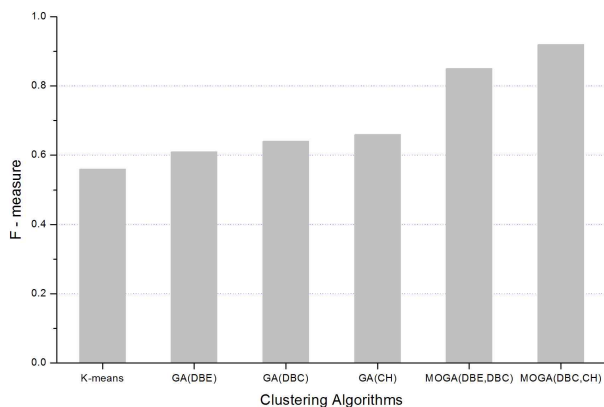
<그림 8, 9, 10>는 *F-measure*를 사용하여 각 테스트 셋에 대한 클러스터링 알고리즘의 성능 결과를 측정하는 것이며, 이것을 그래프로 표현한 것이다.

GA는 단일 목적 함수를 가지는 유전자 알고리즘, MOGA는 다수의 목적 함수를 가지는 다목적 유전자 알고리즘을 의미한다. DBE, DBC 그리고 CH는 적용한 목적 함수를 의미하며 DBE는 유클리디언 유사도

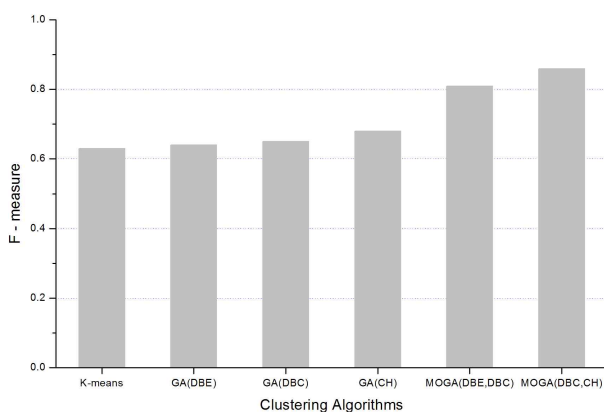
를 사용한 DB 인덱스, DBC는 코사인 유사도를 사용한 DB 인덱스, CH는 CH 인덱스를 나타낸다.

모든 테스트 셋에 대해 유전자 알고리즘이 *k*-means 클러스터링보다 성능이 우수하였고 특히, CH 인덱스를 적용하였을 때 약 10%의 가장 높은 성능 향상을 보였다. 또한, 다목적 유전자 알고리즘은 *k*-means 클러스터링과 유전자 알고리즘보다 성능이 우수하였으며, *k*-means 클러스터링보다 약 20%, 유전자 알고리즘보다 약 15%의 성능 향상을 보였다.

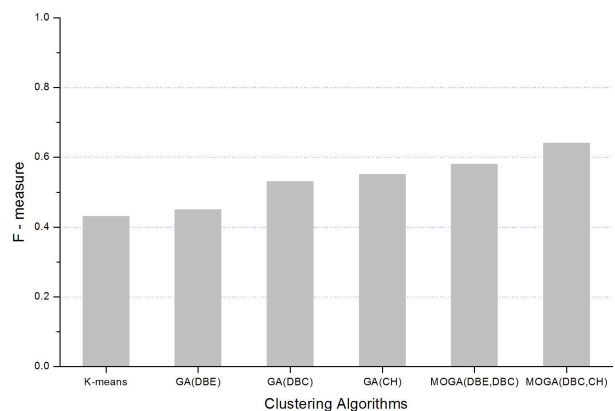
전체적으로 다른 클러스터링 알고리즘보다 다목적 유전자 알고리즘을 이용한 문서 클러스터링이 월등한 성능을 보였으며, DB 인덱스에서 코사인 유사도와 CH 인덱스를 목적 함수로 적용하였을 때가 가장 우수한 성능을 보였다.



<그림 8> 테스트 셋 1 클러스터링 성능 결과



<그림 9> 테스트 셋 2 클러스터링 성능 결과



<그림 10> 테스트 셋 3 클러스터링 성능 결과

## 5. 결론 및 향후 연구

본 논문에서는 *k*-means 클러스터링을 포함한 단일의 목적 함수를 갖는 유전자 알고리즘 그리고 다수의 목적 함수를 갖는 다목적 유전자 알고리즘을 문서 클러스터링에 적용하여 비교해 보았다. 유전자 알고리즘에서는 목적 함수로 사용된 클러스터링 인덱스에 따라 성능의 차이가 보였다. 다목적 유전자 알고리즘은 *k*-means 클러스터링, 유전자 알고리즘 보다 약 20% 이상의 성능 향상을 보였다. 특히, 코사인 유사도를 사용한 DB 인덱스와 CH 인덱스를 목적 함수로 사용하였을 때가 다른 알고리즘에 비해서 우수한 클러스터링 결과를 보였다.

문서 클러스터링에 있어 다목적 유전자 알고리즘을 적용한 사례가 아직까지 찾아지지 않고 있다. 향후 이와 관련된 연구를 통하여 본 논문에서 소개된 클러스터링 인덱스 이외의 다양한 인덱스를 파악하고 이를 목적 함수에 적용하며 성능을 향상시킬 계획이다. 또한, 클러스터링 인덱스들 사이의 관계성을 파악함으로써 클러스터링 성능에 영향력이 큰 요소를 찾을 수 있을 것이다.

## 참고 문헌

- [1] W. B. Croft, D. Metzler and T. Strohman, Search Engines Information Retrieval in Practice,

Addison Wesley, 2009.

[2] 정영미, 정보 검색 연구, 구미무역, 2005.

[3] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observation", Proc. 5th Berkeley Symp, vol. 1, pp. 281-297, 1967.

[4] W. Song and S.C Park, "Genetic algorithm for text clustering based on latent semantic indexing", Computers & Mathematics with Applications, vol. 57, pp.1901-1907, 2009.

[5] 최임천, 송웨이, 박순철, "개체 구조에 따른 유전자 알고리즘 기반의 문서 클러스터링 성능 비교", 한국산업정보학회논문지, 제16권, 3호, pp. 47-56, 2011.

[6] 김대희, 박상호, "분류시스템의 분류 규칙 발견을 위한 유전자 알고리즘", 한국산업정보학회논문지, 제9권, 4호, pp. 16-25, 2004.

[7] 차성민, 권기호, "다중 개체군 유전자 알고리즘의 새로운 이주 방식", 정보과학회논문지, 제28권, 1호, pp. 26-30, 2001.

[8] A. Osyczka, Multicriteria Optimization for Engineering Design, New York Academic Press, 1985.

[9] Censor. Y, "Pareto Optimality in Multiobjective Problems", Applied Mathematics & Optimization, vol. 4, pp. 41-59, 1977.

[10] 김갑환, 조정복, 고창성, 네트워크 모델과 다목적 GA, 한산, 2010.

[11] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast Elitist Multiobjective Genetic algorithm: NSGA-II", IEEE Transaction on Evolutionary Computation, vol. 6, no. 2, pp. 182-197, 2002.

[12] 박순규, 이수복, 이원철, "다목적 최적화를 위한 Goal-Pareto 기반의 NSGA-II 알고리즘", 한국통신학회논문지, 제32권, 11호, pp. 1079-1085, 2007.

[13] 최임천, 최경용, 박순철, "An Automatic Semantic Term-Network Constriction System", Computer Science and its Applications, pp. 48-51, 2009

[14] T. Calinski, and J. Harabasz, "A Dendrite Method for Cluster Analysis", Communications in Statistics, vol. 3, no. 1, 1974.

[15] D. L. Davies and D. W. Bouldin, "A Cluster

Separation measure", IEEE transactions on Pattern analysis and Machine Intelligene, vol. 1, no. 2, 1979.

[16] 문병로, 쉽게 배우는 유전 알고리즘-진화적 접근법, 한빛미디어, 2008.

[17] D. Fragoudis, D. Meretakis, and S. Likothanassi, "Best terms: an efficient feature-selection algorithm for text categorization", Knowledge and Information, vol. 8, pp. 16-33, 2005.



**이 정 송** (Jung Song Lee)

- 학생 회원
- 2011년 : 전북대학교 컴퓨터공학 (공학사)
- 2011년 3월 ~ 현재 : 전북대학교 컴퓨터공학 (석사 과정)
- 관심분야 : 정보 검색, 문서 클러스터링, 시멘틱 웹, 온톨로지



**박 순 철** (Soon Cheol Park)

- 평생 회원
- 1979년 2월 : 인하대학교 공과대학
- 1991년 12월 : 미국 루이지아나 주립대학 (전산학박사)
- 1991년 ~ 1993년 : 한국전자통신연구원 근무
- 1993년 ~ 현재 : 전북대학교 컴퓨터공학부 교수
- 관심분야 : 정보 검색, 시멘틱 웹, 온톨로지

논문 접수 일 : 2011년 12월 23일  
 1차수정완료일 : 2012년 01월 10일  
 2차수정완료일 : 2012년 01월 29일  
 게재 확정 일 : 2012년 02월 14일