클러스터 측정과 유전자 알고리즘을 이용한 문서 클러스터링

최임천*, 박순철*
*전북대학교 컴퓨터 공학과
e-mail:cis1124@chonbuk.ac.kr, scpark@chonbuk.ac.kr

Document Clustering using Generic Algorithm and Cluster Measurement

Lim Cheon Choi*, Soon Cheol Park*
*Dept of Computer Engineering, Chonbuk National University

요 약

본 논문에서는 클러스터 측정(Cluster Measurement)과 유전자 알고리즘을 이용한 문서 클러스링 알고리즘을 제안한다. 유전자 알고리즘의 요소를 클러스터링에 대입하고 클러스터 측정을 적합도 함수에 대입하여 문서 클러스터링을 구현하였다. 성능 평가를 위하여 한국일보-20000/한국일보-40075 문서범주화 실험문서집합의 데이터 셋을 이용하였다. 클러스터링 성능 평가 결과 AS Index가 DB Index, RS Index 보다 좋은 성능을 보여준다. 또한 제안한 알고리즘이 K-means 클러스터링 알고리즘에 비교해 안정적으로 좋은 성능을 보여준다.

1. 서론

대용량의 정보를 빠르고 정확하게 검색하기 위해서 문서 클러스터링에 대한 연구가 활발히 진행되고 있다.[1,3] 문서 클러스터링은 미리 정의된 분류 주제 없이 주어진 문서 집합에서 문서와 문서 사이의 유사도에 근거하여 문서를 그룹화 하는 비교사학습(unsupervised-learning) 기법이다.[1,2,3] 대표적인 클러스터링 알고리즘인 K-means 클러스터링 알고리즘은 그 속도가 빠르면서 그 성능이 비교적 좋아 많이 사용된다. 하지만 임의로 정의되는 초기 센트로이드 벡터에 따라 그 성능의 편차가 커 안정적이지 못한 단점이 있다.[4] 본 논문에서는 클러스터링 측정법을 기반으로 유전자 알고리즘을 이용하여 최적의 해를 구해 안정적으로 좋은 성능을 가지는 클러스터링 알고리즘을 제안하였다.

본 논문은 2장에서 클러스터링 측정법에 대하여 살펴보고, 3장에서 유전자 알고리즘과 문서 클러스터링에 유전자 알고리즘을 적용하는 과정을, 4장에서 클러스터링 알고리즘의 실험 평가 및 결과 분석에 대해 살펴본다. 끝으로 5장에서 결론을 맺는다.

2. Cluster Measurement

클러스터링은 대부분 비교사학습 방법을 통해 진행 된다. 그렇기 때문에 클러스터링 알고리즘의 평가는 매우 중요한 부분이다.[5,6] 클러스터링에는 미리 정의된 정보가 없기 때문에 주어진 정보로 클러스터링 알고리즘을 평가할 수 있는 방법이 필요하다. 클러스터 측정법은 주어진 정보를 기반으로 현재 클러스터링 알고리즘의 성능을 평가할 수 있는 측정법이다. 표 1은 본 논문에서 살펴볼 클러스터 측정법에서 공통으로 사용되는 기호와 기호의 의미이다.

<표 1> 클러스터 측정법에 사용된 기호

기호	의미
n_c	총 클러스터의 개수
$d(x,y)$	x,y 사이의 거리(cosine similarity)
v_i	i번째 클러스터의 센트로이드 벡터
c_i	i번째 클러스터
$\| c_i \|$	i번째 클러스터에 속하는 문서 수

2.1 Davies Bouldin Index

DB Index(Davies Bouldin Index)는 클러스터 사이의의 유사도 평가 (R_{ij})를 중심으로 측정한다. R_{ij}는 클러스터 내부의 분산 평가(s_i)와 클러스터와 클러스터 사이의 비유사도(d_{ij})로 이루어져 있다. R_{ij}는 자유롭게 정의될 수 있으나 일반적으로 다음과 같이 정의된다.[5,6]

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

$$d_{ij} = d(v_i, v_j),\ s_i = \frac{1}{\| c_i \|} \sum_{x \in c_i} d(x, v_i) \qquad (1)$$

위의 R_{ij}에 따라 DB Index는 다음과 같이 정의된다.[5,6]

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

$$R_i = \max(R_{ij}), i \text{ and } j = 1...n_c, i \neq j \quad (2)$$

위의 식에 따라 DB Index는 클러스터와 클러스터에 속한 문서 사이의 유사도와 클러스터와 클러스터 사이의 유사도에 의하여 결과 값이 도출되게 된다. 결과적으로 더 높은 값을 가지는 클러스터가 더 좋은 클러스터로 평가된다.

2.2 RS(R Squared) Index

RS(R Squared) Index는 공식적으로 클러스터 간의 균질성을 측정하여 클러스터를 평가한다. RS Index는 다음과 같이 정의된다.[5,6]

$$RS = \frac{SS_t - SS_w}{SS_t}, where$$

$$SS_t = \sum_{j=1}^{d}\sum_{k=1}^{n_j}(x_k - \overline{x_j})^2, SS_w = \sum_{\substack{i=1...n_c \\ j=1...d}}\sum_{k=1}^{n_{ij}}(x_k - \overline{x_j})^2 \quad (3)$$

위의 식에 따라 RS Index는 전체 문서 집합의 분산 값과 클러스터의 분산 값을 바탕으로 클러스터를 측정한다. RS는 0~1 사이의 값을 가지게 되며 값이 1에 가까울수록 더 좋은 클러스터라고 평가한다.

2.3 Average Similarity Index

본 논문에서는 기존의 클러스터 측정법이 클러스터링 알고리즘에 직접적으로 사용되는데 부족하다는 판단에 AS(Average Similarity) Index라는 클러스터 측정법을 제안한다. AS Index는 다음과 같이 정의된다.

$$SA = \frac{1}{n_c}\sum_{i=1}^{n_c} S_i, where$$

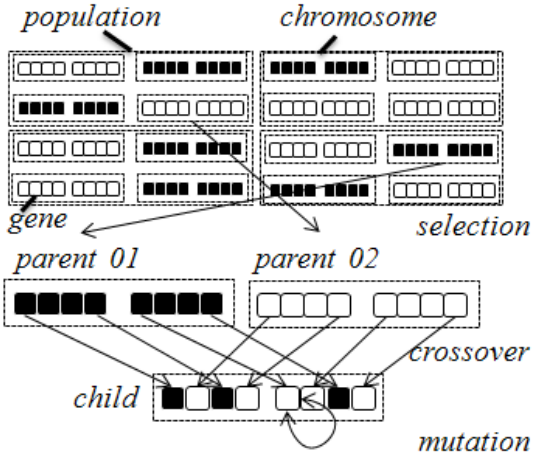
$$S_i = \sum_{j=1}^{\| c_i \| - 1}\sum_{k=j+1}^{\| c_i \|} d(x_j, x_k) \quad (4)$$

위의 식에 따라 AS Index는 각각의 클러스터에 속하는 문서와 문서 사이의 유사도 평균값을 나타낸다. 결국 각각의 클러스터에 속하는 문서와 문서 사이의 유사도가 높을 수록 더 좋은 클러스터라고 평가한다.

3. 문서 클러스터링을 위한 유전자 알고리즘

유전자 알고리즘

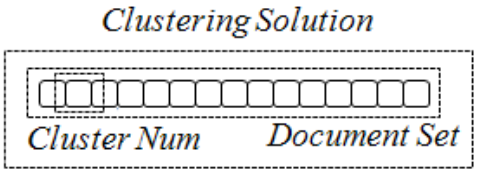
생물의 유전과 진화의 매카니즘을 공학적으로 모델화한 유전자 알고리즘은 자연선택과 적자생존의 원리에 의한 최적화 알고리즘이다. 개체(population), 유전자(gene) 염색체(chromosome)로 표현되는 내부 변수와 적합도 함수라는 내부 평가 규정으로 구성되고 선택(selection), 교배(crossover), 돌연변이(mutation)라는 기본 연산을 통하여 최적의 해를 찾는다.[7,8] 그림 1은 유전자 알고리즘의 기본 구성요소를 그림으로 표현한 것이다.



(그림 1) 유전자 알고리즘의 기본 구성 요소

유전자 알고리즘 적용(문서 클러스터링)

유전자 알고리즘을 문서 클러스터링에 적용[8]시키기 위하여 개체는 클러스터 솔루션(Cluster Solution)으로 염색체는 전체 문서 집합(Document Set)으로 유전자는 각 문서가 속하는 클러스터 번호(Cluster Num)를 할당하였다. 본 논문에서는 유전자 알고리즘과 클러스터 측정법을 이용한 문서 클러스터링 알고리즘을 GA(CM) 문서 클러스터링이라고 명명한다. 그림 2는 GA(CM) 문서 클러스터링의 구조도를 나타낸 것이다.



(그림 2) GA(CM) 문서 클러스터링 구조도

개체 초기화

GA(CM) 문서 클러스터링은 초기 100개의 개체를 가진다. 개체의 수는 GA(CM) 문서 클러스터링 결과의 정확도(성능)와 알고리즘 수행 시간에 영향을 미친다. 개체는 각 문서가 속하는 클러스터의 정보를 가지고 있고 초기에는 임의의 클러스터 번호가 할당된다.

적합도 평가

생성된 개체는 적합도 함수에 따라 각 개체의 적합도

를 평가받게 된다. GA(CM) 문서 클러스터링에서는 클러스터링 측정법에 의하여 적합도를 평가한다. 클러스터 측정법에 따라서 GA(CM)은 다음과 치환하기로 한다.

GA(DB) : Davies Bouldin Index
GA(RS) : RS(R Squared) Index
GA(AS) : Average Similarity Index

선택 연산

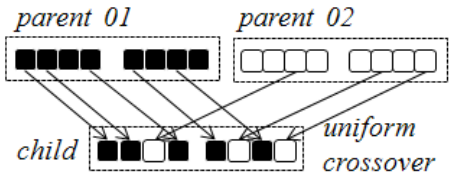
GA(CM) 클러스터링에서는 품질 비례 룰렛 휠 선택을 이용하여 선택 연산을 구현하였다. 가장 대표적인 선택 연산으로 각 해의 적합도를 평가하여 가장 좋은 적합도의 개체가 가장 나쁜 적합도의 개체보다 t배 높은 선택 확률을 가지게 되는 연산이다. 일반적으로 사용되는 t=3을 적용한다. 다음은 품질 비례 룰렛 휠 선택 알고리즘을 정의한 것이다.

$$C_i = (f_w - f_i) + (f_w - f_b)/(t-1),\ t > 1 \qquad (5)$$

C_i는 현재 개체가 선택될 확률을 f_i는 현재 개체의 적합도를 f_w는 전체 적합도 중 최하위를 f_b는 전체 적합도 중 최상위를 나타낸다. 모든 개체에 대한 C를 구하여 합하고, 그 값까지의 범위에서 Random한 숫자를 출력 그 숫자가 포함되어 있는 범위에 있는 개체를 선택한다.

교배 연산

교배 연산으로는 균등 교차 연산(uniform crossover)을 사용하였다. 일반적으로 사용되는 교차 연산으로 임계값 P를 중심으로 임계값 이상이면 parent 01로부터 그렇지 않으면 parent 02로부터 특징을 물려받는 연산이다. 본 논문에서는 0.5를 임계값으로 삼아 같은 확률로 두 부모를 선택하도록 하였다.



(그림 2) 균등 교차 연산

돌연변이 연산

돌연변이 연산을 통하여 우리는 개체가 가지지 못한 특성을 물려받을 수도 있는데 가끔 발생하는 성공적인 돌연변이는 클러스터의 성능을 향상시키고 대부분의 실패적인 돌연변이는 자연적으로 도태되어 사라진다. GA(CM) 문서 클러스터링에서 사용한 돌연변이 연산은 전형적 변이로 0-1 범위의 난수를 발생시키고 임계값 미만의 수가 나오면 돌연변이를 시키는 연산으로 본 논문에서는 일반적으로 사용되는 0.015를 임계값으로 사용하였다.

GA(CM) 문서 클러스터링

GA(CM) 문서 클러스터링은 개체 초기화, 적합도 함수, 선택, 교배, 돌연변이 연산을 이용하여 문서를 클러스터링 한다. 그림 3은 GA(CM) 문서 클러스터링 알고리즘을 나타낸다.

1. 초기 설정
   - P 개체 생성, N 문서 K 클러스터에 배치, 적합도 평가
2. 선택, 교배, 돌연변이 연산
   - 새로운 P개체 생성
3. 적합도 평가
4. if 클러스터 & 적합도 안정 then End
   else 2, 3 반복

(그림 3) GA(CM) 문서 클러스터링 알고리즘

4. 실험 및 결과 분석

본 논문에서는 클러스터링 측정법과 유전자 알고리즘을 이용하여 클러스터링 알고리즘을 제안하고 구현하였다. 제안된 클러스터링 알고리즘의 성능을 평가하기 위하여 한국일보-20000/한국일보-40075 문서범주화 실험문서집합의 데이터 셋을 이용하였다.[9] 총 3개의 Topic Set을 만들어 실험하였고, 각각의 Topic Set은 4개의 주제를 할당하였다. 각 주제별로 주제에 속하는 문서 중 임의의 50개씩의 문서를 선택하여 한 Topic Set당 200개의 문서로 이루어져 있다. 그림 4는 Topic Set에 포함되는 주제를 보여준다.

Topic 01
  여가생활_실내_바둑, 문화와 종교_생활_날씨 스트레스
  정치_외교_외교(대중), 사회_사회질서_사건사고(폭력)
Topic 02
  건강과의학_건강_영양 식품 식사, 문화와종교_공연_음악
  경제_국가_재정 경기전망, 산업_제조업_전자부품
Topic 03
  여가생활_실내_바둑, 경제_기업_도산
  문화와 종교_공연_음악, 산업_농축산수산_근해원양

(그림 4) Topic Set

그림에서 확인할 수 있는 것과 같이 Topic Set에 포함되어 있는 4개의 주제들은 문서 집합의 대분류에서 추출한 것으로 주제별 구분이 명확한 것을 확인할 수 있다.

클러스터링 성능 평가는 $F-measure$ 척도를 사용하였다. $Precision$과 $Recall$의 분기점으로 이루어지는 $F-measure$는 다음과 같이 정의된다.

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (6)$$

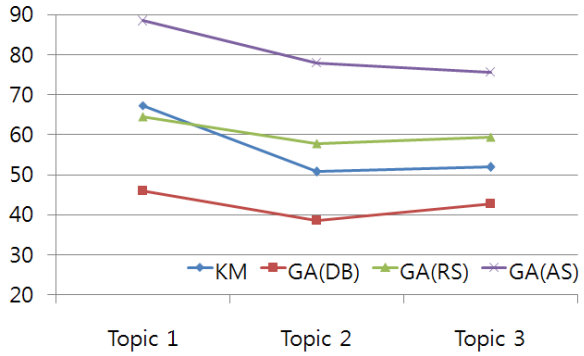
$F-measure$에 사용되는 $Precision$과 $Recall$은

$$Precision = \frac{Relevant \cap Clustered}{Clustered} \quad (7)$$

$$Recall = \frac{Relevant \cap Clustered}{Relevant} \quad (8)$$

로 정의된다.

그림 5는 본 논문에서 실험한 클러스터링 알고리즘의 성능을 Topic별로 표현한 것이다.



(그림 5) Topic별 클러스터링 알고리즘 성능

위의 그림을 통해서 우리는 클러스터 평가도와 유전자 알고리즘을 이용한 클러스터링의 성능을 볼 수 있다. 기존의 클러스터 평가도는 클러스터를 생성하는데 사용하기는 부적합한 것으로 판단되었다. 특히 DB Index의 경우는 K-means 클러스터링 알고리즘 보다 더 낮은 성능을 나타내었다. DB Index의 경우 알고리즘이 가지는 특성 상 클러스터링에는 부적합 하였다. 또한 제안한 AS Index의 경우는 클러스터의 평가에도 그리고 클러스터를 생성하는데 매우 적합하였다. K-means 클러스터링 알고리즘에 비해서 3가지 Topic Set에서 모두 약 20%정도의 성능향상을 보였다.

5. 결론 및 향후 방향

본 논문에서는 다양한 클러스터링 측정법과 유전자 알고리즘을 이용하여 안정적이면서도 높은 성능을 보이는 클러스터링 알고리즘을 구현하였다. 클러스터링의 성능 평가를 위하여 한국일보-20000/한국일보-40075 문서범주화 실험문서집합의 데이터 셋을 이용하였다. 클러스터링 결과의 평가 척도로는 정확률, 재현율에 의한 F1-Measure를 사용하였다. 기존 연구된 클러스터링 측정법(DB Index, RS Index)은 현재 클러스터의 적합도를 판별하기에는 좋은 성능을 보이나, 클러스터를 만드는데 사용하기에는 조금 부족한 것으로 확인되었다. 제안한 AS Index는 클러스터의 적합도를 판별하기에도, 또한 클러스터를 만드는데 사용하기에도 적합하였다.

GA(CM) 문서 클러스터링의 객관적인 성능 평가를 위하여 K-means 클러스터링 알고리즘과의 성능을 비교하였다. 모든 Topic Set에 대하여 제안한 GA(AS) 문서 클러스터링 알고리즘이 GA(DB), GA(SR) 클러스터링 알고리즘이나, K-Means 클러스터링 알고리즘에 비하여 클러스터링 결과에 훨씬 좋은 성능을 나타냈다. GA(AS) 문서 클러스터링 알고리즘이 성능에 있어서 편차가 큰 것을 수정, 보완한다면 더 좋은 클러스터링 알고리즘으로 발전할 것이다.

Acknowledgement

본 논문은 BK21 프로젝트와 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행되었습니다.(2010-0011997)

참고문헌

[1] B. Y. Ricardo and R. N. Berthier, Modern information retrieval, Addison Wesley, 1999.
[2] 강승식, 한국 형태소 분석과 정보 검색, 홍릉과학출판사, 2002
[3] 정영미, 정보 검색 연구, 구미무역, 2005
[4] S. Selim and M. Ismail, "K-means-type algorithm generalized convergence theorem and characterization of local optimality", IEEE Trans. Pattern Anal. Mach Intell. vol. 6, pp. 81-87, 1984.
[5] David L. Davies, Donald W. Bouldin, "A Cluster Separation Measure", IEEE Transactions on In Pattern Analysis and Machine Intelligence, Vol. 1, No. 2. pp. 224-227.
[6] Csaba Legany, Sandor Juhasz, Attila Babos, "Cluster validity measurement techniques", "Knowledge Engineering and Data Bases", Vol 5, pp. 388-393, 2006
[7] L. D. Davis, "Handbook of Genetic Algorithms", Van Nostrand Reinhold, 1991.
[8] U. Maulik, S. Bandyopadhyay, "Genetic algorithm-based clustering technique", Patten Recognition. vol. 33, pp. 1455-1465, 2000
[9] http://www.kristalinfo.com/TestCollections/