유전자 알고리즘을 이용한 문서 클러스터링 연구

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

A Study on Clustering using Genetic Algorithm

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

요 약

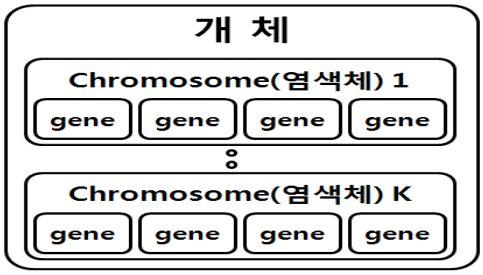
본 논문에서는 효율적인 인공지능 알고리즘인 유전자 알고리즘(GA)을 이용한 문서 클러스터링 시스템을 제안한다. 일반적으로 클러스터링 알고리즘에 가장 많이 사용되는 K-Means는 임의로 결정되는 초기 센트로이드 벡터에 따라 그 성능이 많이 달라지는 것을 볼 수 있다. 이에 본 논문에서는 유전자 알고리즘을 이용하여 안정적이면서도 높은 성능을 보여주는 클러스터링 알고리즘을 개발하였다. 제안한 클러스터링 알고리즘의 성능 평가를 위하여 HANTEC 2.0과 문서 범주화 집단 데이터 셋을 사용하였다. 제안된 방법은 효율적이고 빠른 K-Means를 이용한 클러스터링 알고리즘에 비하여 훨씬 뛰어난 성능을 보였다.

1. 서론

오늘날의 정보 검색 시스템은 날로 거대해지는 인터넷의 정보를 얼마나 효율적으로 분석하여 빠르고 정확하게 검색할 수 있느냐에 그 초점이 맞추어지고 있다[1]. 이러한 대용량의 정보를 효율적으로 처리하기 위하여 문서를 그룹화 하여 검색의 성능을 향상시키는 방법들이 연구되고 있는데 이러한 문서의 그룹화에는 문서 클러스터링과 문서 범주화가 있다[2]. 문서 클러스터링은 비슷한 특성을 지니는 문서들을 그룹화 하는 것으로 정해진 주제가 아닌 문서의 특성 자체를 바탕으로 그룹화의 기준을 정하는 특성이 있다. 문서 클러스터링에 가장 보편적이고 일반적인 알고리즘으로는 K-Means Algorithm이 있는데 그 속도가 빠르면서 성능이 비교적 좋다는 특징을 가지고 있다[3]. 하지만 초기의 센트로이드를 어떤 것으로 선택하느냐에 따라서 그 성능의 폭이 커 안정적이지는 못하다. 이에 본 논문에서는 유전자 알고리즘을 문서 클러스터링에 대입하여 안정적으로 더 좋은 성능을 가지는 클러스터링 시스템을 구축할 수 있는지 확인하였다.

2. Clustering using Genetic Algorithm

유전자 알고리즘(Genetic Algorithm)은 생물의 유전과 진화의 메카니즘을 공학적으로 모델화하는 것으로 생물이 갖는 환경에서의 적응 능력을 취급하여 자연 선택과 적자생존의 원리에 기반한 알고리즘이다[4]. 유전자 알고리즘은 개체, 염색체(Chromsome), 유전자(Gene)로 구분되는 내부 변수와 적합도(Fitness)라는 평가 규정을 바탕으로 구현된다. 개체는 염색체들에 의해 특징지어지는 자율적인 하나의 집단을 의미하고 유전자는 각 염색체의 형질을 규정하는 기본 구성 요소가 된다. 유전자 알고리즘의 중요한 인자중 하나인 적합도(Fitness)는 현재의 개체가 가지고 있는 특성, 특징이 우리가 원하는 그것에 얼마나 근접했는지를 나타내는 척도로서 전체의 성능을 평가하는데 중요한 역할을 하는 인자이다. 우리는 클러스터링에 유전자 알고리즘을 적용시키기 위하여 한 개의 개체를 Clustering Solution으로 하나의 염색체를 하나의 센트로이드로 그 센트로이드가 가지게 되는 문서의 특징 벡터의 값을 하나의 유전자(Gene)로 각각 대입하여 생각하였다. 그림1는 유전자 알고리즘을 클러스터링에 대입한 GA Clustering의 구성도를 나타낸다.



(그림 1) GA Clustering 구성도

그림1 에서 보는 바와 같이 하나의 개체는 문헌 집단의 클러스터링을 위한 하나의 단위가 된다. 우리는 이러한 개체를 여러 개 생성하게 되고 각각의 개체는 서로 다른 특징들을 지니게 된다. 다양한 개체들의 Fitness를 바탕으로 유전자 알고리즘의 선택, 교배, 돌연변이 특성을 이용하여 최종적인 클러스터를 정의하게 된다. 그림1 은 유전자 알고리즘 클러스터링의 순서도를 나타낸다.

1. K개의 클러스터 센트로이드를 선정
2. 가장 유사한 센트로이드를 갖는 클러스터로 나머지 문헌을 배치
3. 클러스터와 배치된 문서와의 Fitness 계산
4. Fitness를 바탕으로 집단을 선택, 선택된 집단에서 교배 돌연변이 알고리즘을 적용
5. 새로운 염색체 집단의 유전자 형질에 따른 클러스터 생성
6. Fitness의 변화가 없는 상태가 일정 기간동안 유지될 때까지 2, 3의 과정을 반복

(그림 2) GA Clustering 알고리즘

알고리즘의 진행과정 중 클러스터의 성능에 가장 영향을

미칠 수 있는 척도인 Fitness는 센트로이드와 배치된 문헌들과의 거리 값을 통하여 판단하게 되는데 거리 값이 작게 나올수록 유사도가 높고 클러스터링의 정확도가 높다는 의미가 된다. 적합도의 변화가 거의 없어질 때, 유전자 알고리즘을 통해서 더 나은 클러스터를 만들 수 없을 때 유전자 알고리즘에 의한 클러스터링은 마치게 된다.

3. 성능 평가

본 논문에서는 제안한 알고리즘의 성능을 평가하기 위하여 HANTEC 2.0, 한국일보(20000, 40075) 문서 범주화 실험 문서 집합의 데이터 셋을 이용하였다. 테스트에는 총 4가지의 Topic Set을 사용하였다. Topic 1 ~ 3은 서로 분명이 구분되는 주제를 가지고 Topic 4는 서로 비슷한 주제를 가지도록 선발하였다. 그림8 은 테스트에 사용된 Topic을 보여준다.

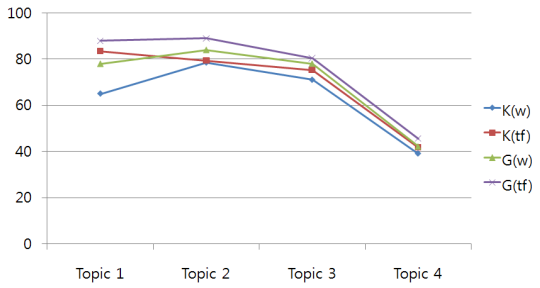
Topic 1
"여가생활_실외_여행관광", "정치_외교_외교(대중동)", "문화와 종교_스포츠_축구", "사회_사회질서_사건사고(화재)"
Topic 2
"여가생활_실내_바둑", "정치_외교_외교(대중)" "문화와 종교_생활_날씨 스트레스", "사회_사회질서_사건사고(폭력)"
Topic 3
"여가생활_실외_등산낚시", "경제_부동산_부동산", "문화와 종교_공연_영화", "산업_제조업_컴퓨터"
Topic 4
"경제_금융_은행(여신 대출 업무)", "경제_금융_은행(수신 예금 업무)", "경제_금융_은행(금융 통화 감시 사고)", "경제_금융_은행(금융 업계 동향)"

(그림 3) 테스트셋에 사용된 Topic

본 논문에서 제안한 유전자 클러스터링 알고리즘을 평가하기 위해서 우리는 정보검색분야의 평가 방법으로 가장 널리 사용되는 방법인 정확율(Precision)과 재현율(Recall)을 기반으로 평가하였다. 정확율과 재현율은 상호 보완적인 관계로 두 값에 따라 생성되는 클러스터를 평가할 수 있다. 이러한 관계성에 기반하여 두 값을 이용한 클러스터의 평가는 정확성과 재현율의 분기점(Break Even Point : BEP)인 F-measure로 이루어진다. BEP는

$$BEP = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

으로 정의된다.



(그림 4) 클러스터링 성능

그림 9는 각 Topic에 대한 클러스터링 성능을 나타내는 그래프 이다. 우리는 제안한 GA Clustering과 K-Means Clustering의 성능을 평가 비교하여 보았고 그래프에서 나타나듯이 일반적으로 GA Clustering이 K-Means Clustering보다 좋은 성능을 가지는 것을 알 수 있었고, w를 tf 자체로 정의한 것이 $tf_{ij} \cdot idf_i$로 정의한 것보다 성능이 더 좋은 것을 알 수 있었다. K-Means Clustering의 경우 매번의 알고리즘에 걸리는 시간이 적게 들고 결과값의 편차가 큰 반면 GA Clustering의 경우는 K-Means에 비하여 결과를 얻는데 시간이 많이 들고 결과 값의 편차가 적은 것으로 나타났다. 또한 Topic 4의 경우에는 성능이 현저하게 차이가 나는 것을 볼 수 있는데 비슷한 주제로 묶인 문서의 경우에는 나타나는 용어가 비슷하여 클러스터링의 정확도가 떨어지는 것을 볼 수 있었다.

4. 결론 및 향후 방향

대용량의 정보를 처리하여 빠르고 정확하게 정보를 검색하도록 도와주는 정보 검색 시스템의 성능을 향상시키기 위한 방법으로 자주 사용되는 문서 클러스터링을 효율적인 인공지능 알고리즘의 하나인 유전자 알고리즘을 이용하여 구현하여 보았다. 한글 문서의 테스트를 위하여 문서 범주화 실험집단을 이용하였고 알고리즘에 의하여 결정된 클러스터의 평가를 위하여 정보 검색 시스템에서 일반적으로 사용되는 정확율, 재현율을 이용한 BEP를 추출하였다. 일반적인 Topic에 대하여 K-Means Clustering에 비해 GA Clustering이 더 뛰어난 성능을 나타낸 것을 볼 수 있었다. 전혀 다른 주제에 대한 클러스터에는 GA Clustering이 K-Means Clustering에 비하여 월등한 성능을 보였지만 비슷한 주제에 대한 클러스터링에는 두 알고리즘 모두 성능이 좋지 않은 것을 볼 수 있다. 그 문서의 특징을 나타 낼 수 있는 용어의 구분이 어려운 경우의 클러스터링에 대한 연구가 필요하다.

Acknowledgement

This work was partially supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government(MOST) (No. 2009-0076721) and partially supported by the third stage of Brain Korea 21 project.

참고문헌

[1] B. Y. Ricardo and R. N. Berthier, *Modern information retrieval*, Addison Wesley, 1999.

[2] 한국 형태소 분석과 정보 검색, 강승식, 2002[3] 정보 검색 연구, 정영미, 2005

[3] S. Selim and M. Ismail, "K-means-type algorithm: generalized convergence theorem and characterization of local optimality," IEEE Trans. Pattern Anal. Mach Intell. 6, pp.81-87, 1994.

[4] S. Bandyopadhyay and U. Maulik, "Nonparametric genetic clustering: comparison of validity indices," IEEE Trans. Systems, Man and Cybernetics-C, Applications and Reviews, Vol.31, No.1, pp.120-125, 2001.