유사성 계수에 의한 문서 클러스터링 시스템 개발
Development of Similarity-Based Document Clustering System

우 훈식*, 임 동순**
Hoon-Shik Woo, Dong-Soon Yim

Abstract

Clustering of data is of a great interest in many data mining applications. In the field of document clustering, a document is represented as a data in a high dimensional space. Therefore, the document clustering can be accomplished with a general data clustering techniques. In this paper, we introduce a document clustering system based on similarity among documents. The developed system consists of three functions: 1) gatherings documents utilizing a search agent; 2) determining similarity coefficients between any two documents from term frequencies; 3) clustering documents with similarity coefficients. Especially, the document clustering is accomplished by a hybrid algorithm utilizing genetic and K-Means methods.

1. 서론

전통적인 클러스터링 방법들은 다수의 객체들을 어느 수 만큼의 그룹으로 묶어 같은 그룹에 속한 객체들은 서로 유사하고, 다른 그룹의 객체들 과는 서로 다르도록 한다. 다시 말하자면 클러스터링 문제는 어떤 최적화 기준에 의거하여 N개의 객체를 M개의 그룹으로 묶는 것이다. 이때 객체 간의 유사성을 어떻게 표현하느냐 하는 문제가 발생한다. 문서에 대한 클러스터링의 경우, 한 문서에 포함된 용어의 빈도 수를 특성벡터로 나타내고, 두 문서에 대한 유사도는 두 특성벡터의 코사인 값으로 표현된다[3].

이러한 문서 클러스터링은 다양한 응용 분야를 갖고 있다. 특히, World Wide Web상의 문서들을 조회, 필터링, 카테고라이징 하기 위한 많은 지능적 소프트웨어 에이전트에서 사용되어왔다. 또한, 클러스터링은 서로 연관된 웹 문서들의 중요한 특징을 추출하여 자동적으로 쿼리를 생성하거나 다른 비슷한 문서들을 찾는데 유용하게 이용된다.

본 연구에서는 웹 마이닝에서 중요한 분야 중의 하나인 문서들의 클러스터링을 위한 시스템을 개발하여 이를 소개하는데 그 목적이 있다. 특히, 클러스터링 문제를 정형화하고, 이를 해결하기 위한 방법으로 유전자 알고리듬과 k-means방법을 사용하였다.

* 대전대학교 컴퓨터정보통신공학부
** 한남대학교 산업시스템공학과

개발된 시스템은 web 상에서의 정보통신 관련 신문기사에 대한 클러스터링을 대상으로 하였다.

2. 문서 클러스터링 문제

본 연구에서의 문서 클러스터링 문제는 두 문서 간의 유사도를 이용하는 것을 대상으로 한다. 문서들의 특성벡터로부터 다음과 같은 유사성 행렬을 구하였다고 하자.

$$S = [s_{ij}] : N \times N \text{ 의 유사성 계수 행렬}$$

유사성 행렬을 이용한 한 최적화 문제를 설명하기 위하여 다음의 기호를 사용하기로 한다. G, N, M을 각각 전체 문서 집합, 전체 객체 수($|G|$), 클러스터 수라 하고, c_i와 G_k를 각각 i번째 객체가 속한 클러스터 번호와 k번째 클러스터의 객체 집합이라고 하자. 이러한 기호를 사용하여 우선 다음과 같은 정의를 하기로 한다.

- 객체와 클러스터 간의 유사성: i번째 객체와 t 번째 클러스터에 있는 객체간의 유사성 계수 합을 γ^{it}로 정의한다.

$$\gamma^{it} = \sum_{j \in G_t} s_{ij}$$

- 객체의 클러스터 내 평균 유사성: i번째 객체가 속한 클러스터 번호를 k라고 하자(즉, $c_i = k$). i번째 객체와 자신이 속한 클러스터에 있는 다른 객체 간의 유사성 평균을 i번째 객체의 클러스터내 평균 유사성이라고 부르고, 이를 α^i로 나타낸다.

$$\alpha^i = \frac{\gamma^{ik}}{|G_k| - 1}$$

- 객체의 클러스터 간 평균 유사성: i번째 객체와 서로 다른 클러스터에 속한 객체들 간의 유사성 평균을 i번째 객체의 클러스터 간 평균 유사성이라고 부르고, 이를 β^i로 나타낸다.

$$\beta^i = \frac{\sum_{l \neq k} \gamma^{it}}{N - |G_k|}$$

i번째 객체가 좋은 클러스터에 속하기 위하여는 클러스터 내 평균 유사성인 α^i가 커야 하는 반면, 클러스터간 평균 유사성인 β^i는 적어야 한다. 따라서, 다음과 같은 문제를 고려할 수 있다.

$$Max\ z = \sum_{i}^{N} [\lambda^i \alpha^i - (1 - \lambda^i)\beta^i$$

위의 문제에서 만약 $\lambda^i = \lambda = 1$라면, 단지 클러스터 내 유사성 평균 합을 최대화하는 문

제가 되어 클러스터 내의 흩어짐 정도를 최소로 하게 된다. 반대로, $\lambda_i = \lambda = 0$ 이라면, 클러스터 간의 유사성 평균 합을 최소화 하는 문제가 되어 클러스터 간의 차이를 최대로 하게 된다. 만약, $\lambda_i = \lambda = 0.5$ 이면 클러스터 내와 클러스터 간의 유사성 차이의 합을 최대화 하는 문제이다. 본 연구에서는 $\lambda_i = \lambda = 0.5$ 인 경우의 문제를 대상으로 한다.

3. 클러스터링 알고리듬

3.1 유전자 알고리듬

본 연구에서 사용하는 유전자 알고리듬의 염색체 표현 방법은 기본적으로 순열 표현 방법이다. 그러나, 순열로 표현된 하나의 염색체가 유효한 해를 나타낼 수 있도록 경험적 해석에서와 같은 특별한 복호화 작업이 필요하거나 또는 분리자가 요구된다[2]. 경험적 해석에 의한 방법은 각 문서가 어떤 그룹에 포함되어야 하는지를 결정하기 위하여 적응값을 구한다. 그러나, 이 작업은 일반적으로 많은 계산시간이 요구되어 효율성이 떨어진다. 분리자를 염색체 표현에 추가하는 방법은 단순함을 유지할 수 있으나, 복구 알고리듬이 요구된다. 본 연구에서는 순열로 표현된 하나의 염색체가 서로 인접한 문서끼리의 연관성 정보를 포함한다고 가정하여 가장 적은 유사성 계수를 갖는 문서간을 분리한다.

염색체의 순열 표현 방법에 적용될 수 있는 교배 연산자는 외판원 문제에 적용되는 path 표현방법에서의 교배 연산자인 PMX (Partially-mapped), OX (Order), CX (Cycle)등을 사용할 수 있다. PMX는 Goldberg 와 Lingle [1]에 의해 제안된 연산자로 한 부모로부터 연속된 부분 유전자를 유전 받고, 다른 부모로부터 가능한 많은 유전자를 상속받도록 한다. 본 연구에서는 교배 연산자로서 기존의 PMX방법을 고려한다.

순열 표현방법에서는 전형적인 돌연변이 연산자를 적용하기가 불가능하다. 대신에 전형적인 전위 연산자의 사용은 매우 간단히 적용될 수 있다. 이 연산자는 임의의 두 위치를 선택하여 이 위치의 수를 서로 교환시킨다.

3.2 K-Means 방법에 의한 유전자 알고리듬 해의 향상

유전자 알고리듬에 의한 해를 보다 좋은 해로 변환시키기 위하여 각 개체에 적용하거나, 또는 각 세대의 가장 좋은 해에 부분 최적화 알고리듬을 적용할 수 있다. 부분 최적화 알고리듬으로는 개체 분할 문제에 적용되는 최적화 알고리듬을 고려 할 수 있다. 본 연구에서는 유전자 알고리듬에 의한 해를 수정된 K-평균법 방법에 의해 보다 향상된 결과를 가져오도록 하였다.

원래의 K-Means 방법은 한 문서와 한 클러스터의 중심간의 거리 척도를 사용한다. 그러나, 이 연구에서 고려하는 클러스터링 문제에서는 클러스터의 중심이라는 개념을 사용할 수 없다. 따라서, 본 문제에 알맞게 K-Means 방법을 수정하여야 한다. K-Means 방법을 문서 클러스터링 문제에 적용하기 위하여 한 문서가 현재 속한 클러스터에서 다른 클러스터로의 이동을 했을 경우, 목적함수값의 변화량을 계산한다.

수정된 K-평균법 알고리듬

Input: (초기해)

c(i), i=1, ·,N　// I 번째 문서가 속해 있는 클러스터 번호

s(i,j) i=1, ·, N, j=1, ·, N // 두 문서 간의 유사성 계수

Output: (새로운 해)

c(i), i=1,·,N　// I번째 문서의 새로운 클러스터 번호

Process:

1. 현재 해에 대한 각 문서의 클러스터 내 평균 유사성과 클러스터 간 평균 유사성을 구하여 목적함수 값을 구한다.
2. 다음 절차를 문서의 이동이 불가능할 때 까지 반복한다.

2.1 각 문서(i)에 대하여

다른 각 클러스터로의 이동을 했다고 가정하여 목적함수의 순 변화량을 구한다.

2.2 각 클러스터로의 이동에 따른 순 변화량 중 가장 큰 것을 선택하여 이 값이 0 보다 크면 이동을 시키고, 각 문서에 대한 새로운 클러스터내 평균 유사성과 클러스터 간 평균 유사성을 구한다.

4. 시스템 개발

본 연구에서 개발된 문서 클러스터링 시스템을 구성하는 주요 모듈은 로봇, 용어 빈도수 계산, 문서 유사도 계산으로, 자바 언어를 이용하여 개발하였다. 주요 개발 내용은 다음과 같다:

4.1 로봇

로봇은 시스템에 지정된 URL 리스트로부터 해당 사이트 주소를 획득하여 웹 문서를 가져오는 역할을 담당한다. 본 연구에서는 테스트를 위하여 디지털 타임스[4]와 전자신문[5] 두 곳의 사이트를 대상으로 하였다. 로봇은 사용자로부터 해당 날짜를 입력 받아서 각 사이트에서 해당하는 날짜의 기사 URL을 가져 오는 것과 주어진 URL에서 신문 기사를 가져오는 두 가지 기능으로 구성되어 있다.

4.2 용어 빈도수 계산

용어 빈도수 계산은 로봇이 가지고 온 웹 문서에 대하여 각 용어 별로 빈도수를 계산하기 위한 모듈이다. 즉, 용어 빈도수는 문서 내에 존재하는 특정 용어의 발생 빈도를 측정한 것으로, 이 때 측정하고자 하는 용어는 관리자가 별도로 입력 및 수정할 수 있다. 본 연구에서는 용어 인덱스를 정보 통신과 관계된 용어를 선정하여 사용하였다.

4.3 문서 유사도 계산

본 연구에서는 문서 간의 유사성을 계산하기 위하여 vector space 모델[3]을 이용한다. 기본적으로 문서의 의미는 사용된 용어에 의한다고 가정하여 각 문서를 나타내는 특성 벡터를 구한다. 각 문서를 벡터로 표현하는 방법으로는 일반적으로 많이 사용되

는 tf-idf를 이용하였다. 두 문서 간의 특성벡터로부터 벡터 각의 코사인 값을 유사성 계수로 구하였다.

4.4 문서 클러스터링

문서간의 유사성 계수를 이용하여 유사한 문서끼리 그룹핑하기 위한 클러스터링을 수행한다. 클러스터링은 유전자 알고리듬과 k-means 방법을 혼용하였다. 기본적으로는 유전자 알고리듬에 의해 각 해를 구하나, 각 세대에서의 해에 k-means방법을 적용하여 해의 향상을 꾀하였다. [그림 1]은 클러스터링 수행 화면을 나타낸다.

4.5 클라이언트 인터페이스

[그림 2]는 본 연구의 결과 화면으로 일반 사용자가 직접 대면하는 클라이언트 화면으로 사용자 입력 용어에 대하여 문서 유사도를 중심으로 클러스터링된 카테고리 별 결과가 제시된다.

5. 결 론

본 연구에서는 웹의 정보를 지식 베이스로 변환하기 위해서 서치 에이전트를 이용하여 정보를 검색하고 검색된 결과를 유사도 측정을 이용하여 서로 연관된 웹 문서 간의 주요 특징을 추출하여 체계화하는 문서 유사도 기반의 문서 클러스터링 시스템을 개발하였다. 개발된 시스템은 현재 프로토타입 단계로 뉴스의 자동 분류, 인덱스 용어 사전 생성 등에 직접적으로 활용이 가능하여 한정된 범위의 지식 발견에 효과적으로 이용될 수 있을 것으로 예상한다.

본 연구에서는 클러스터링 알고리듬으로 유전자 알고리듬과 K-Means 방법을 혼용하였다. 추후 이러한 방법에 대한 효과성 및 효율성을 평가하여 보다 성능이 뛰어난 알고리듬에 대한 연구를 계속할 예정이다.

참고문헌

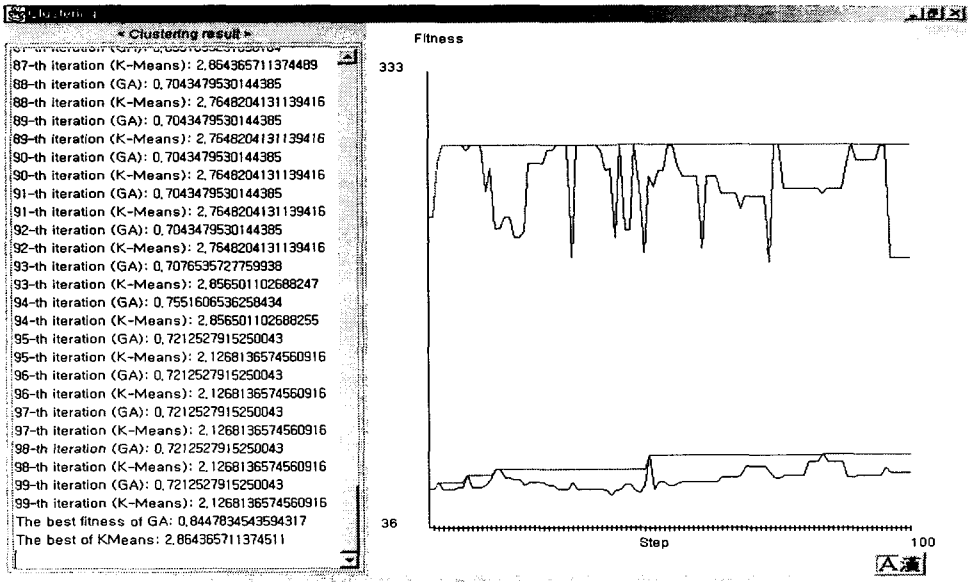
[1] Goldberg, D. E. and Lingle, R., Alleles, Loci, and the TSO, *Proceedings of the First International Conference on Genetic Algorithms*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1985, pp. 154-159.

[2] Jones, D.R. and Bertramo, M.A., Solving Partitioning Problems with Genetic Algorithms , *in Proceedings of the Fourth International Conference on Genetic Algorithms*, Morgan Kaufmann Publishers, Los Altos, CA, 1991, pp.442-449.

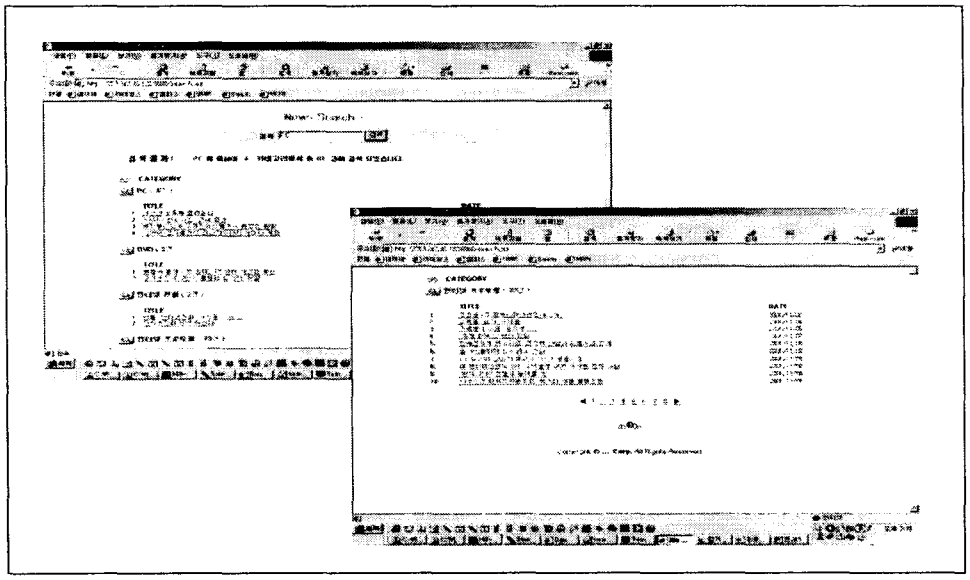
[3] Salton, G., Yang, C., and Wong, A., A vector-space model for automatic indexing , Communications of the ACM, Vol. 18, No. 11, pp. 613- 620, 1975

[4] http://www.dt.co.kr

[5] http://www.etimesi.com



[그림 1] 클러스터링 수행 화면



[그림 2] 문서 클러스터링 시스템 클라이언트 화면