집합(Aggregation)개념을 이용한 클래스의 클러스터링 방법

임 근
서울보건대학 전산정보처리과
e-mail:LK04@www.shjc.ac.kr

Clustering Method of Class using Aggregation Concepts

Lim, Keun
Dept of Computer Information Process, Seoul Health Junior College

요 약

본 논문에서는 클러스터 정의시 사용하게 되는 특성으로 노드간 패스 수에 기반한 K-edge 컴포넌트의 그래픽 정의 방법과 노드를 클러스터화 하는 집합화(Aggregation) 방법을 제시하였다. 집단화된 하이퍼텍스트 분리를 통해 이전 결과를 개선할 수 있으며, 집단내의 노드간 관련성을 가시화하여 비교할 수 있다.

1. 서론

본 논문에서는 Term 클러스터링과 구문 문법형태를 통한 자연어 문장을 번역하기 위한 방법을 제시하고, 또한 클러스터 정의에 사용하게 되는 특성으로 노드간 패스 수에 기반한 K-edge 컴포넌트의 그래픽 정의를 이용하여, 노드를 클러스터화 하는 집합화(Aggregation) 방법을 제시한다. 2가지 방법은 모두 도큐멘트 검색에 대한 텍스트 표현을 통해서 검색의 질적인 면을 향상시켜보려는 목적을 가지고 있다. 또한 집단화 방법을 적용함으로서 클래스들의 클러스터화를 실현해 보고자 한다.

본 논문의 구성은 2장에서 관련 연구로 클러스터링 개념과 조건 등을 설명하며, 3장에서 클러스터링 메소드, 4장에서 적용 예와 결론을 맺는다.

2. 관련 연구

2.1 Term 클러스터링 개념

Term 클러스터링은 기존에 텍스트 표현으로부터 추출한 Term을 그룹화하기 위한 클러스터 분석중에 하나의 응용분야로 특성추출 방법이라고 할수 있으며 패턴분류에 유용하다. Term 클러스터는 비교적 적은 빈도로 유사한 Term에 의해 제한을 받게되며 Term 대치의 의미보다는 보충의 개념으로 이용된다. 클러스터링은 관련성 있는 도큐멘트의 분리여부에 따라 효과가 결정된다.

2.2 클러스터링 알고리즘

클러스터링 알고리즘은 독립 개체의 자연적인 클러스터를 찾기 위한 시도로서 개체의 집합은 내부적 응집력과 외부적 결합성을 통해서 특성을 설명할 수 있다.

- Clustering Criterion : N개의 표본,

즉 X_1, X_2, X_3 , . . . X_N 으로 분류한다고 가정하자. 이러한 벡터들은 임의 벡터로 표시하지 않는다. 그 이유는 벡터들이 고정적인 값이고 일반화되어 있는 요소라고 가정하기 때문이다. 각 표본은 L개 클래스(W_1, W_2 . . . W_N)로 표기한다. 편의상 K_i 의 값은 1과 L 사이의 정수값 이라고 가정한다. 하나의 분류요소(Classification) Ω 는 벡터로서 W_{ki} 라고 하며 Configuration X^*는 X_i 로 나타내며 다음과 같이 정리한다.

$\Omega = [W_{k1}, W_{k2} , . . . W_{kn}]^T$

$X^* = [X_1^T , X_2^T , , . . X_N^T]^T$

Clustering Criterion J 는 Ω 와 X^* 의 함수로 다음과 같이 표시할 수 있다.

$J = J (W_{k1}, W_{k2} , .. W_{kn} : X_1^T , X_2^T , . X_N^T)$

$= J(\Omega_n : X^*)$

정의에 의하여 Classification Ω 는 다음을 만족한다.

$J(\Omega_n : X^*) = \max_{\Omega}$ 또는 $\min_{\Omega} J(\Omega : X^*)$

- Clustering algorithm : 주어진 클러스터링 문제에 대하여 Configuration X^*는 고정적인 값이고 Classification Ω 는 변수이다. 일반적인 클러스터링 기법을 통한 방법은 실제 검색에 적용하기에는 부적합하다. 그 이유는 Classification Ω 가 이산적이고, 비순서적이기 때문이다. 그러나 Classification Ω 에서 분산에 관한 J상의 반복적 탐색 알고리즘 정의는 가능하다. 1번 반복한 Classification은 Classification Ω (1) 이라고 가정하자.

$\Omega(1) = [W_{k1(1)}, W_{k2(1)}, \ldots, W_{kn(1)}]^T$

만일 1개의 표본이 현재 클래스 Ki(1)에서 클래스 j로 Clustering Criterion 은 전체합 Δ J (i, j, 1)에 의해 변환된다. 이때 Δ J (i, j, 1)=$J(W_{k1(1)}, W_{ki-1(1)}, \ldots W_{kn(1)} : X^*) - J(\Omega(1):X^*)$ 만일 Δ J (i, j, 1)이 음수값이면, 클래스 j로의 1개 표본에 대한 재분류는 하나의 Classification을 만들게 되고, 이것은 다음 알고리즘에 기반한다.

① 초기 Classificaion Ω (0)을 선택한다.

② 1번째 Classificaion Ω (1)이 주어지고, Δ J (i, j, 1)은 계산을 통해 구한다.(j=1, 2 ..,L : i=1, 2 .., N)

③ i=1, 2 .., N 에 대하여, 1개의 표본을 클래스로 재분류한다. $\Delta J(i, j, 1) = min_j \Delta J(i, j, 1)$ 현재 Classification에 의해 j = ki(1)를 포함하고, 이 단계는 Ω (i+1)의 형태를 취한다.

④ 만일 Ω (i+1) <- Ω (1) 이면,② 단계로 간다.

이 알고리즘은 Clustering Criterion 에 기반한 Classification rule의 단순한 반복 응용이라고 할 수 있다.

3. 클러스터링 분석방법

바람직한 클러스터링은 자연스럽고, 세부사항에 독립적이며, 유사한 객체를 동일한 클러스터에 분류하는 것이다. 또한 적당한 크기의 클러스터를 생성하도록 한다. 이때 각 어플리케이션은 각 요소들에 대한 정의를 필요로 한다. 방향성 그래프와 같이 표현된 하이퍼텍스트에서는 노드사이의 독립 패스 수에 의해 표시되는 연결강도(strenth of connection)를 통해 노드간 관련성을 간접적으로 가시화 한다. 이 경우에도 하나의 링크는 노드간 관련성을 보일 필요가 있다. 만일 도큐멘트의 크기가 변화한다거나, 노드의 링크 위치를 변경했지만 현재 위치를 포인팅하고 있다면 클러스터링은 동일하다. 그러나 문맥내의 변화는 결정적인 요소이기 때문에 명확성을 고려해야 한다. 또한 이러한 변화요소를 분석할 수 있는 방법이 필요하다고 인식하여 분석 방법을 고려하였다.

3.1 클러스터 수

클러스터의 개수 L을 결정하는 문제를 다중분리지금까지 임의적으로 지정하였다. 그러나 L 값을 결정하고 적절하게 클래스를 할당할 필요가 있다. 클러스터링 과정이 변화된 이후에 주어진 L 값에 대한 최적 표준으로 J^*_L 이라고 한다. 만일 L 이 증가함에 따라 J^*_L 감소하거나, 최소값에 도달한다면 클래스의 적절한 수로 L_0가 결정될 수 있다. 그러나 잘 알려져 있는 표준들이 이러한 특성을 가지고 있지 않기 때문에 외부적인 방법이 필요하다.

3.2 다중분리

클러스터링 표준 J에 전적으로 의존하는 접근 방법의 경우에 적용했을 때 가정 만족할 만한 방법이다. 이 경우 다음과 같은 범주를 가지게 된다. L=2 라고 가정할 때, 여기에는 몇 가지 차별화된 classification이 존재하고 최소값에 가장 근접한 J를 만들게 된다. 클래스 L을 포함하는 표본의 모집단이 K 이분법이라고 가정하자. 각 이분법은 2개 그룹으로 표본을 분리한다. 여기에 S_{ij}는 I번 이분법에 의한 그룹 j에 할당되는 모든 표본의 집합이라고 하고 (i = 1, 2, . . k , j = 1, 2, . . , n)다음과 같은 2개 조건을 갖는다.

조건1) 하나의 이분은 하나의 그룹 상 각 클래스가 위치한다. 즉 클래스들은 이분에 의해 분리되지 않는다. □

조건2) 클래스 쌍은 적어도 하나의 이분이 있으며, 같은 그룹에 2개의 클래스를 할당하지 않는다. K분법에 각각에서 하나의 그룹을 선택하고 K 그룹의 교차로서 subset C를 구한다. □

3.3 집합(Aggregation)정의

Def 1

V' ⊂ V(G)라 하고 G는 노드의 집합
V'에 의해 유도된 서브그래프<V'>는 벡터의 집합 V'를 갖는다 모든 edge는 V'에서 양방향 포인트를 갖는다. ■

Def 2
하나의 그래프 G_i는최대값으로 서브그래프 G_j 가 없다면 $G_i \subset G_j$ 로 G_i는동일 특성을 갖는다. ■

Def 3
하나의 그래프 G의 K-component는 최대값으로서 서브그래프 <A>를 유도하며, <A>의 모든 분할 A_1 과 A_2 의 특성을 갖는다. 또한 적어도 <A>의 K개 링크는 A_1 이나 A_2 의 각 객체를 갖는 일련의 이벤트이다. ■

Def 4
집합(Aggregation)을 찾기 위해서 다음과 같은 과정을 거친다. ■
①그룹의 모든 K-component를 정의한다.
②노드를 집합화 한다. 이것은 강 컴포넌트(strong component)를 의미한다.
예) 노드(d,e,f,g,h,i)는 4-component안에 있으며, 다른 강 컴포넌트는 없다.
③강 컴포넌트는 제거하고, 각 부분들을 모아서 또 다른 약 컴포넌트에 분배한다.

주론

그래프 G의 K-component <A>의 별도의 객체인 모든 쌍 v, u ∈ A는 서브그래프 <A>의 k-link disjoint 패스로 연결되며, <A>는 이들 특성을 갖는 최대값이다.

4. 적용 예와 결론

본 논문에서는 하이퍼텍스트로 구성된 검색시스템에 클러스터를 정의하는 방법을 제시하였다. 즉 링크간 독립 패스 수에 있어서 클러스터는 임의적인 특성이 아닌 하이퍼텍스트의 근본적인 특성에 종속적이다. 또한 K-component를 기반으로 하여, 집합화에 의한 클러스터링 방법을 제시하였다. Fig. 1과 2에서와 같이 하이퍼텍스트 상태도는 주어진 크기, 즉 요구하는 대상이 방대할 경우 적용하기 어려운 문제점이 있다. 따라서 전역적인 것보다 국지적인 상태도를 보이게 된다. 다만 본 논문에서는 각 집합의 원시의 그래프보다 축소해서 보이므로 복잡하고, 크기가 큰 전체 모습을 보이는데 효과적이다. 또한 집합은 하이퍼텍스트상에서 구조적 질의에 대한 응답에 편리함을 줄수 있으며, 사용자가 여러개 노드를 검색하고자 할 경우, 2개 노드의 관련성을 비교해 볼수 있다.

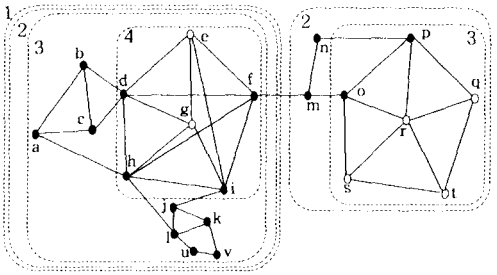

Fig.1 K-component 그래프

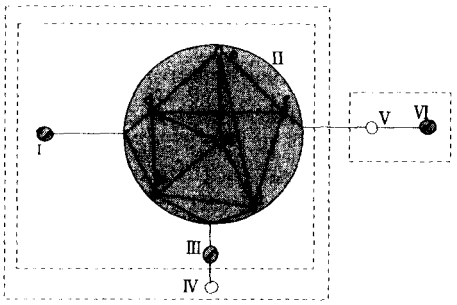

Fig. 2 K-component 클러스터링 그래프

5. 참고문헌

[1] F. Harary, Graph Theory. Addison-Wesley, Reading, 1989.
[2] 이경환, 소프트웨어 재사용을 위한 객체모델링 기법, 교학사, 1993
[3] Gellersen HW, Gaedke. M , object-oriented Web application development, IEEE Computer socity V, v N.1 60-68, 1999.
[4] D. Budgen and G. Friel, Augmenting the Design Process: Transformations from Abstract Design Representation, CAiSE-92, Springer-verlag, 1992
[5] Jeffrey M .voas, The challengers of Using COTS SW in Component-based Development, V,31 N 6, 44-45, 1998
[6] Lisa spicknall Fruth and James M. purtilo, journal of system and software V.32 N.3, 1996.