고차원 데이터를 위한 관측값 기반의 자동화된 유사도 추론

김우철° 박상현
연세대학교
twelvepp@cs.yonsei.ac.kr, sanghyun@cs.yonsei.ac.kr

Inferring reliable similarity measure for high-dimensional data using observation

Woo-cheol Kim°, Sanghyun Park
Yonsei University

요 약

유사 검색은 데이터를 다루는 모든 연구 분야에서 가장 필수적인 연구 중 하나이다. 특히 데이터의 사용량이 늘어나고 데이터의 차원이 고차원이 될수록 유사 검색은 정확한 결과를 찾기 힘들기 때문에 더욱 더 많은 연구들이 진행되고 있다. 유사 검색의 연구 중에 가장 중요한 부분은 유사도를 나타내는 수식을 찾는 것이다. 현재 고차원 데이터를 대상으로 하는 많이 진행되고 있는 많은 연구들은 휴리스틱 기반으로 유사도를 정의한다. 하지만 휴리스틱 기반은 연구자에 따른 주관의 개입 및 과적응 문제가 있다. 휴리스틱 기반이 아닌 관측값을 이용하는 대표적인 연구에는 HMM이 있다. 하지만 HMM을 통한 경우 유사도를 유추해 낼 수 있지만 데이터의 어떤 속성이 유사도를 판단하는데 기여하는지 판단하기 힘들다. 따라서 본 논문에서는 고차원의 데이터에 대해서 관측값 기반의 자동화된 유사도를 추론방법을 제안하고 간단한 예비 실험을 통해서 검증한다.

1. 서 론

유사 검색(similarity search)은 데이터를 다루는 모든 연구 분야에서 가장 필수적인 연구 중 하나이다[1]. 특히 요즘과 같이 멀티미디어 데이터의 양이 늘어나고, 수많은 데이터들 중에 의미 있는 데이터를 찾아내려고 하는 데이터 마이닝 기술들이 발전함에 따라서 유사 검색 성능의 중요성은 더욱 높아가고 있다[2].

"유사하다"는 개념은 절대적인 결과로 나타나지 않는다. 그림 1의 a), b), c)과 같이 다양한 유사 검색 대상에 대해서 Q와 D 사이의 유사함의 정도는 쉽게 수치화된 절대값으로 나타낼 수 없다. 즉 "Q와 D가 유사한가?"라는 질문에 대해서는 쉽게 대답하기 힘들다. 특히 a)와 b)의 경우에는 더욱 더 어렵다. 하지만 c)의 경우에는 유사함의 정도를 유클리디언 거리를 이용한다면 수치화된 절대값으로 표현이 가능하다. 즉 유클리디언 거리가 3라고 한다면 대략적으로 "3만큼 유사하지 않다."라는 대답할 수 있다.

"유사하다"라는 개념은 상대적인 결과로 나타내는 것이 좀 더 일반적일 수 있다. 즉 유사검색을 위해 올바른 질문은 "그림 2에서 Q가 D1과 D2중 어느 것과 더 유사한가?"이다. 이런 질문에 대해서는 사람마다 일부 주관적인 판단이 포함될 수 있지만 모두 답을 할 수 있다. 즉 그림 2의 a)의 경우에는 대부분 "D2가 D1보다 유사하다."라고 답할 것이다.

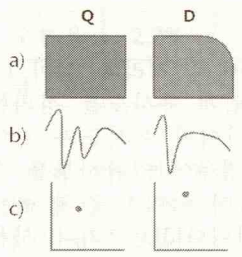

그림 1. 유사검색 1

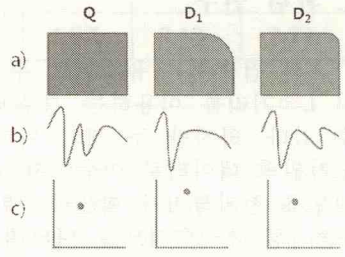

그림 2. 유사검색 2

하지만 실제로 다양한 연구 분야에서 사용 유사 검색을 위한 올바른 질의는 "Q가 D1과 D2중 어느 것과 더 유사한가?"보다는 "Q가 D1, D2,… .D1,000,000중 가장 유사한 것은?" 또는 "Q와 비슷한 순서대로 D1, D2,… .D1,000,000을 나열해라."라는 질문이다. 이러한 문제를 풀기 위해서는 유사도(similarity measure)가 필요하다. 유사도는 비교하려는 두 객체간의 유사 정도를 수치화된 값을 얻을 수 있는 척도 이다. 가장 많이 알려진 유사도는 L* 거리(L* distance)이다[3]. 즉 비교하려는 객체를 N차원 벡터공간의 객체로 표현한 다음에 그 객체 사이의 공간상의 거리를 이용한다.

따라서 대부분의 유사 검색을 다루는 연구 분야에서 가장 중요한 연구 내용은 유사 검색의 대상이 되는 객체간의 유사도를 나타내는 수식을 정의하는 것이다. 이렇게 수식을 정의하는 과정을 일반적으로 다음과 같은 2단계로 이루어진다. 1) 먼저 비교 객체에서 그 객체의 특

본 논문은 교육과학기술부 한국연구재단의 미래기반기술개발사업(2009-0083311)의 지원을 받아 수행되었습니다.