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.

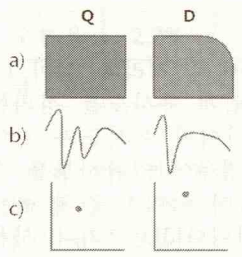

그림 1. 유사검색 1

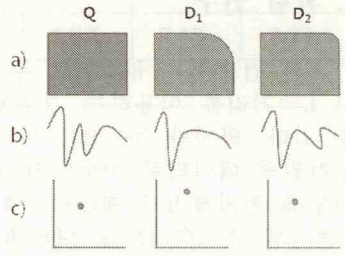

그림 2. 유사검색 2

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.

징을 잘 표현하는 차원을 추출하는 것이다. 2) 그 다음 추출된 차원들을 이용해서 검색 공간(search space)을 만들어서 유사도를 정의하는 것이다[1]. 물론 이러한 2단계의 과정은 독립적으로 분리되어 있지 않고 서로 영향을 미친다.

일반적으로 고차원의 데이터를 이용하는 경우에 유사도를 정의하는게 어렵다[4]. 유사도가 정의되기 어려운 점은 다음과 같다. 1) 먼저 비교 객체에서 속성값을 수치화된 값으로 추출하기 어렵다. 2) 추출된 속성값간의 의존관계를 확인하기 어렵다. 3) 추출된 속성을 벡터공간에 매핑하기 어렵다. 4) 마지막으로 벡터 공간에 매핑한 다음 거리를 정의하기가 어렵다. 따라서 많은 유사검색에서 객체의 속성을 추출하거나 벡터공간에 매핑하는 것은 대부분 휴리스틱 기반으로 처리되고 있다. 하지만 이러한 휴리스틱 기반의 유사도 정의는 수식을 정의하는 연구자의 주관이 개입될 수 있고 분석한 데이터의 특성에 과적응(over-fit)되는 경향이 크다는 문제가 있다[5].

본 논문에서는 연구자의 주관을 배제하고 주어진 데이터만을 이용해서 비교 객체간의 유사도를 자동으로 생성해 주는 방법에 대한 기초 연구를 제안한다. 그리고 간단한 예비 실험을 통해 제안한 방법의 효용성에 대해서 논의한다.

2. 관련 연구

유사 검색에서 유사도로 두 점사이의 거리와 같이 L2나 L∞거리를 이용하는 경우에는 쉽게 유사도를 정의할 수 있다. 하지만 두 개의 이미지 사이의 유사도와 같이 고차원의 데이터로 이루어져 있는 경우에는 유사도를 수식으로 정의하기가 힘들다. 예를 들어 이미지 간의 유사도의 정의는 수식으로 나타내기 어렵다[3]. 그러나 사용자는 두 개의 이미지간의 유사도를 어느 정도 관측할 수는 있다. 본 논문에서는 이렇게 관측을 통해서 얻은 유사도를 관측값(observation)이라고 정의 한다.

관측값을 기반으로 하는 유사검색에서 연구자가 주관적인 개입없이 이루어지는 연구는 히든 마르코프 모델(Hidden Markov Model: HMM)을 들 수 있다. 즉 관측값 S를 가지고 있는 두 개의 객체 OA와 OB를 이용해서 HMM의 모델의 훈련(training)시킨다. 이때 훈련에 사용된 관측값들이 많을수록 좋은 모델이 된다. 이렇게 만들어진 모델을 통해서 새로운 객체 OA와 OB를 입력으로 사용하면 두 객체의 유사 정도를 알 수 있다.

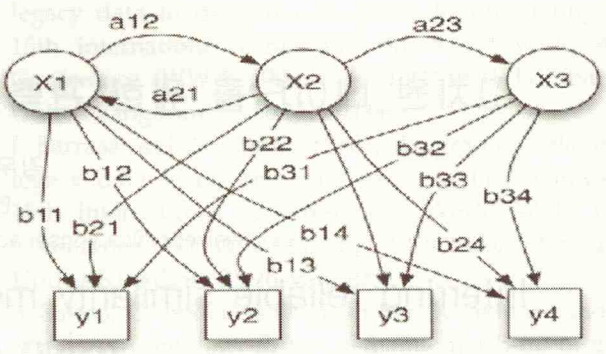

그림 4. HMM의 예: (1) 상태: x1, x2, x3; (2) 전이확율: b11, b12, ⋯, b34; (3) 관측값: y1, y2, y3, y4

HMM의 경우에는 확률 기반 모델로 그림 3과 같이 각 상태(state)와 각 상태간의 전이확율(transition probability)와 각 상태에서의 관측값(observation)들로 이루어져 있다. HMM은 사용자가 미리 정의한 상태와 관측값을 기반으로 훈련 과정을 통해서 각 전이확율값을 계산함으로써 구성된다.

유사 검색에서 HMM을 이용하는 경우의 가장 큰 단점을 유사도를 해석하기 어렵다는 것이다. 즉 훈련을 통해서 모델이 만들어지더라도 각 상태간의 전이 과정만 알 수 있고 객체의 어떠한 속성들이 객체간의 유사성을 판단하는데 어떤 영향을 미치는지 알 수 없다.

3. 관측값을 이용한 유사도

본 논문에서는 비교 대상이 되는 두 객체 사이의 유사도를 위한 공식을 자동 생성해 주는 알고리즘을 개발하려고 한다. 먼저 관측값들을 바탕으로 유사도를 유추하기 위해 필요한 가정을 하고 그러한 가정이 필요한 이유에 대해서 설명한다. 그런 다음 그 가정을 바탕으로 관측값 유추를 위한 방법에 대해서 설명한다.

3.1 유사도 유추를 위한 가정

관측값을 이용해서 유사도를 유추하기 위해서는 다음과 같은 2가지의 가정이 필요하다.

가정 1. 모든 객체간의 유사정도를 나타내는 관측값을 가지고 있다.

본 논문에서 사용하는 관측값은 유사도와는 다른 개념이다. 즉 유사도는 유클리디안 거리와 같이 공식에 의해서 명확하게 수치화된 값을 나타낸다. 하지만 관측값은 두 객체가 "유사하나", "유사하지 않다"와 같이 유사 정도를 관측해서

이러한 관측값은 2가지로 가질 수 있다. 첫 번째는 수치화된 관측값이다. 수치화되어 있다라는 의미는 문자 그대로 숫자로 표현된 값일 수도 있지만 "높음", "보통", "낮음"과 같이 단계적인 정도의 관측값일 수도 있다. 두 번째는 상대적인 관측값이다. 예를 들어 OA와 OB가

OA와 OC보다 더 유사와 같이 상대적으로 유사한 정도를 관측한 값일 수 있다.

가정 2. 유사도를 측정하기 위해 필요한 모든 변수의 값을 측정할 수 있다.

일반적으로 유사도는 다양한 변수간의 수식으로 정리된다. 예를 들어 L2-거리의 경우에는 각 객체가 2개의 변수 x, y로 구성되어 있고 각 객체의 값이 (x1, y1), (x2, y2)일 때 $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ 으로 정의된다. 따라서 올바를 유사도를 유추해내기 위해서는 유사도를 계산하는데 필요한 모든 변수를 유추해 낼 수 있어야 한다.

3.2 유사도 유추 알고리즘

3.1절의 가정을 기반으로 관측값들을 이용해서 유사도를 유추하는 가장 간단한 알고리즘은 모든 경우수를 다 해보는 것이다.

예를 들어 유사도를 유추하려는 객체가 2개의 변수(x, y)로 구성되어 있고 유사도에 쓰이는 연산자는 {+, -, ×, ÷}만 쓸 수 있고 단 1번의 연산만 허용한다고 하자. 이럴 때 가능한 수식은 { x+y, x-y, y-x, x×y, x÷y, y÷x }로 6가지이다.

이렇게 유사도를 찾는 경우 가장 큰 문제점은 검색공간(search space)이 넓다는 것이다. 유사도에 쓰일 수 있는 변수가 많을수록 허용하는 연산자의 개수가 많을수록 연산자로 이루어는 항의 개수가 많을수록 검색 공간은 기하급수적으로 커진다. 그에 비래해서 가장 적절한 유사도를 찾는 과정이 오래 걸린다. 따라서 이렇게 광범위한 검색공간에서 사용가능한 좀 더 효율적인 방법이 필요하다.

본 논문에서는 검색공간이 넓은 환경에서 좋은 효율을 보이는 유전자 알고리즘(genetic algorithm)을 적용했다. 단순한 유전자 알고리즘 적용은 예비 실험 단계에서 이용한다. 이 후 본 연구에서 좀 더 최적화된 방법을 제안하고자 한다.

4. 예비 실험

본 논문에서는 제안한 방법의 효율성과 정확성을 검증하기 위해서 예비 실험을 진행하였다.

정확성을 검증하기 위해서 이미 유사도를 정의할 수 있는 객체의 유사 검색을 진행하였다. 비교하려는 객체는 2차원 평면에 있는 점들이다. 객체의 값이 (x1, y1), (x2, y2)일 때, 점들 사이의 거리는 이미 $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ 로 정해져 있다. 이러한 점들을 이용해서 추론된 식이 $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ 과 얼마나 유사한지 검증한다.

실험에서는 변수를 {x1, y1, x2, y2} 4개만 사용하고 사용된 연산자는 {+, -, ×, ÷, √} 5개를 사용하였다. 마지막으로 연산에 사용된 항은 10개를 사용하였다.

훈련에 사용된 데이터의 개수는 100개를 사용하였고 추가적으로 1000개를 실험하였다.

예비 실험의 과정을 다음과 같다. 먼저 유전자 알고리즘을 이용해서 수식을 추론한 다음 추론된 수식과 실제 수식의 값들을 구해서 그 차이를 오류값을 계산하였다. 예를 들어 추론된 수식을 이용해서 값을 구했는데 89였고 실제 수식의 값이 100이였다면 오류율을 11%로 계산하였다.

먼저 훈련을 100개의 데이터 값을 이용하는 실험을 5회 실시하였다. 실험 결과는 표 1과 같다. 평균적으로 7.9%의 오류율을 보였다.

표 1. 100개의 관측값을 이용한 훈련결과

	1회	2회	3회	4회	5회
오류율	6.9%	9.1%	8.5%	7.9%	7.3%
평 균	7.9%				

두 번째는 훈련에 1000개의 데이터 값을 이용하는 실험을 5회 실시하였다. 실험 결과는 표 2과 같다. 평균적으로 2.4%의 오류율을 보였다.

표 2. 1000개의 관측값을 이용한 훈련결과

	1회	2회	3회	4회	5회
오류율	2.2%	3.1%	1.8%	2.1%	2.7%
평 균	2.4%				

비록 간단한 형태의 수식이기는 했지만 예비 실험의 결과를 보면 추론된 수식들이 원래 수식과 오류율이 크지 않다는 것을 알 수 있다. 앞으로 연구를 통해서 유사도 추론 과정에서 최적의 연산자 및 연산항의 수를 조절한다면 좀 더 정확한 수식을 추론할 수 있을 것이다.

5. 결론

멀티미디어 데이터와 같이 고차원의 데이터의 경우에 유사도를 정하기 쉽지 않다. 본 논문에서는 유사 검색에서 가장 중요한 유사도를 관측값들을 이용해서 유추하는 방법을 제시하였다. 기존에도 HMM을 이용하면 관측값으로 부터 블랙박스 방식으로 유사 검색에 이용될 수 있는 방법이 있다. 하지만 단순히 유사 검색 대상 객체들이 유사한 정도만 계산할 수 있을 뿐 어떠한 속성 때문에 유사하다고 판단되는지 알 수 없었다.

본 논문에서 관측값을 이용해서 유사도를 계산할 수 있는 수식을 제안함으로써 다음과 같은 장점을 지닌다. 먼저, 수식을 통해서 유사도를 바로 수치화된 값으로 나타낼 수 있다. 그리고 가장 중요한 장점은 수식을 통해서 각 객체의 어떠한 속성이 유사도를 판단하는데 중요한 영향을 미치는지 알 수 있다.

앞으로 계속적인 연구를 통해서 좀 더 많은 차원을 가

지고 있고 데이터들을 대상으로 유사도를 나타내는 공식을 유추해내는 일반화된 방법을 도출하고자 한다.

참고문헌

[1] Simone Santini, "Similarity measures", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 21, No. 9, September 1999.

[2] Anthony K. H. Tung, Rui Zhang, Nick Koudas, Beng Chin Ooi, "Similairty Search: A Matching Based Approach", In Proceedings of the 32th Very Large DataBase (VLDB'06), pp 631-642, September 2006.

[3] Simone Santini, Ramesh Jain, "Similarity is a Geometer", Multimedia Tools and Application, Vol. 5. pp 277-306, 1997.

[4] Aristides Gionis, Piotr Indyk, Rajeev Motwani, "Similarity Search in High Dimensions via Hashing", In Proceedings of the 25th Very Large DataBase (VLDB'99), pp 518-529, 1999.

[5] Ellen Spertus, Mehran Sahami, Orkut Buyukkokten, "Evaluating Similarity Measures: A Large-Scale Study In the Orkut Social Network", In Prodeedings of the Knowledge discovery in data mining (KDD'05), pp 678-684, 2005.