김일성종합대학학보
(자연과학)
주체106(2017)년 제63권 제3호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 63 No. 3 JUCHE106(2017).

유클리드거리에 의한 안전한 자료발굴의 한가지 방법

김 훈

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《정보산업을 빨리 발전시키고 인민경제의 모든 부문을 정보화하여야 합니다.》

(《**김정일**선집》 증보판 제20권 380페지)

오늘날 자료의 폭발적인 증가로 하여 지식추출을 위한 자료의 정확하고 효률적이며 고속인 분석이 필요하며 이로 하여 자료의 저장과 분석을 분리시키고 제3자가 자료분석을 진행하도록 할것이 요구된다.

제3자에 의한 안전한 자료발굴을 위하여 우연섭동, 우연사영[1, 2]과 같은 기술들이 연구되였지만 이러한 방법들은 자료점들사이 유클리드거리를 변화시키는것으로 하여 오늘날 자료발굴에서 널리 리용되고있는 유클리드거리에 의한 자료발굴에는 적합치 않다.

론문에서는 기밀성자료를 숨기며 유클리드거리를 높은 정확도로 보존하면서도 제3자에게 전송하는 자료량을 줄이기 위해 잘 알려진 푸리에변환의 에네르기압축성질을 리용하는 방법을 론의한다.

1. 안전한 자료발굴을 위한 푸리에변환의 리용

푸리에관련변환들은 푸리에토대집합을 리용하여 자료를 원래의 령역으로부터 다른 령역으로 에네르기를 보존하며 넘긴다.

이러한 푸리에관련변환의 일종으로서 리산코시누스변환은 에네르기압축능력이 강하고 실수자료에 적용할수 있으며 실수렬 $x_0, x_1, \cdots, x_{n-1}$을 실수결수렬 $f_0, f_1, \cdots, f_{n-1}$로 넘기는 변환으로서 다음과 같이 표시된다.

$$f_k = \sqrt{\frac{2}{n}} \sum_{j=0}^{n-1} \lambda_j x_j \cos\left(\frac{(2j+1)k\pi}{2n}\right)$$

여기서 λ_j는 $k=0$일 때 $1/\sqrt{2}$, 기타는 1인 수이다. 그리고 렬 x_j의 에네르기는

$$e = \frac{1}{n} \sum_{j=0}^{n-1} {x_j}^2$$

으로 정의된다.

선행연구[3]로부터 $x_0, x_1, \cdots, x_{n-1}$과 $f_0, f_1, \cdots, f_{n-1}$의 에네르기는 같으며 두 자료렬벡토르 $x_0, x_1, \cdots, x_{n-1}$과 $y_0, y_1, \cdots, y_{n-1}$사이의 유클리드거리는 푸리에변환후에도 보존된다. 즉 자료집합의 푸리에변환결과를 유클리드거리에 의한 자료발굴에 리용할수 있으며 저에네르기결수억제에 의하여 자료발굴의 안전성과 전송자료의 압축을 보장할수 있다.

2. 안전한 자료발굴을 위한 조작

자료집합의 매 행(record)을 하나의 리산신호렬로 보고 푸리에변환을 진행한 후 대다수 행들에서 큰 에네르기를 유지하는 결수들만을 선택하여 자료발굴자에게 전송하는것이 목표이다. 따라서 론문에서는 다음과 같은 조작을 진행하였다.

먼저 결수들의 순서치환을 진행한다.

속성의 수와 결수들의 배렬순서를 숨김으로써 거꿀푸리에변환에 의한 초기자료의 회복을 매우 어렵게 할수 있다.

만일 선택된 결수의 수가 μ, 최대속성수를 N이라고 가정하면 불순한 목적을 가진 제3자(자료발굴자)는 $\sum_{i=\mu}^{N}\frac{i!}{1-\mu}$개의 가능한 순렬을 조사해야 한다. 즉 자료발굴자에게 결수들의 순서를 치환하여 전송함으로써 초기자료값에 대한 안전성을 매우 높일수 있다.

다음 에네르기가 큰 결수를 선택한다.

선행연구[1]에서 제안한 에네르기가 큰 결수의 선택방법은 복잡도가 대단히 크다.

따라서 론문에서는 에네르기기여률을 정의하고 그것을 리용한 결수선택방법을 제안한다.

정의 에네르기기여률과 에네르기기여행렬

자료렬 $x_1, x_1, \cdots, x_m$이 주어졌을 때 $x_{12}/E(x_1, x_1, \cdots, x_m)$을 렬의 에네르기에 대한 $x_i(1\le i\le m)$의 에네르기기여률이라고 한다. 그리고 X를 $n\times m$차원자료집합,

$$A_{ij} = X_{ij2}E(X_{i1}, X_{i2}, \cdots, X_{im})$$

을 i째 행에 대한 j째 속성의 에네르기기여률이라고 할 때 행렬 A_{ij}를 에네르기기여행렬이라고 한다. 또한 $C_j = \frac{1}{n}\sum_{i=1}^{n}A_{ij}$를 X에 대한 j째 속성의 평균에네르기기여률이라고 한다.

기여행렬에 의한 결수선택알고리듬은 다음과 같다.

걸음 1 X에 대하여 행에 관한 푸리에변환결과 H를 얻는다.

걸음 2 H에 대하여 에네르기기여행렬 A_{ij}를 얻는다.

걸음 3 A_{ij}를 리용하여 평균에네르기기여률 C_j를 구한다.

걸음 4 C_j들을 그 크기에 따라 내림순서로 정렬한다.

걸음 5 $\sum_{j=1}^{\mu}C_j \ge \varsigma$를 만족시키는 최소인 μ를 구한다.

걸음 6 C_j값의 크기순서로 μ개의 결수를 선택하여 그 순서를 치환한 후 자료발굴자에게 전송한다.

알고리듬에서 X는 초기자료집합, ς는 자료발굴의 정확성에 대한 요구를 반영하는 (0, 1) 구간의 수이다.

이 알고리듬의 복잡도는 $O(mn\log m)$로서 선행연구[1]에서 제안한 방법에 따르는 알고리듬의 복잡도보다 훨씬 작으면서도 효률적이다.

3. 실험결과 및 분석

실험은 Matlab 7.0으로 진행하였으며 실험자료는 행의 개수가 320이고 렬의 개수가 16인 자료집합을 리용하였다.

실험에서는 우연사영방법과 우연결수선택방법, 기여행렬을 리용한 결수선택방법에 대한 안전성을 평가하였다. 실험을 위해 선행한 방법[1]에서 리용한 믿음구간과 F－척도를 리용하였으며 반복회수를 20으로 하였다.

실험결과 세가지 방법들이 모두 높은 안전성을 보여주었는데 제안한 방법인 기여행렬을 리용한 결수선택변환방법은 결수의 수가 9일 때 자료전송량을 거의 0.5배로 줄이면서도 안전성을 보장하였다.

맺 는 말

기여행렬에 의한 푸리에결수선택방법을 제안하고 성능평가를 진행하였다.

실험결과는 기여행렬을 리용한 결수선택방법이 선행한 방법들보다 보다 효과적이라는 것을 보여주었다.

참 고 문 헌

[1] Zhiyuan Chen; VLDB, 15, 4, 293, 2006.

[2] C. Aggarwal; EDBT, 12, 3, 183, 2004.

[3] M. Debeljak et al.; Ecological Indicators, 41 30, 2014.

주체105(2016)년 11월 5일 원고접수

A Method of Secure Data Mining by Euclide Distance

Kim Hun

We proposed a selection method of Fourier coefficient using the contribution matrix. This method reduces the transmitting data and also is secure. So it is efficient in data mining.

Key words: Fourier transform, data mining, Euclide distance