論文
5-2-16

점근적 분석 모형에 기초한 유한개 레코드 정렬 알고리즘 효율성의 확률적 분석

(Probabilistic analysis of efficiencies for sorting algorithms with a finite number of records based on an asymptotic algorithm analysis)

김숙영(Suk-Young Kim)[1]

요 약

정렬 알고리즘 효율성을 분석하는 O 표기법은 자료 크기에 관한 모형을 구축하지 않고 자료 크기가 무한하게 증가될 때의 정렬 비교 횟수의 증가율에 관한 대략적인 정보만을 제공하는 점근적 알고리즘 분석 결과이다. 그러므로 제한된 유한개의 자료들만을 정렬하는 응용면에서도 정렬 알고리즘 효율성 검정이 필요하다. 9,000 개 이하의 수치 자료에 삽입 정렬과 퀵 정렬 알고리즘을 적용하여 자료 개수에 따른 정렬 시 필요한 원소 교환 횟수 관계 모형을 구축하였다. 효율성이 O(nlogn) 으로 분류되는 퀵 정렬의 경우 추정된 모형은 $S=0.9305N^{1.1339}$으로, $O(n^2)$ 으로 분류되는 퀵 정렬에서는 $S=0.12232N^{2.013}$으로 추정 되었다. 또한 모형의 적합도 검정 결과 정렬 시 자료 개수에 따른 원소 교환 횟수 관계가 추정된 모형들에 의하여 99% 이상이 설명될 수 있으며 적합성을 증명하는 강한 확률적 증거가 발견 되었다.

본 연구 결과들은 정렬 자료 개수가 적은 경우나 새로 개발된 정렬 알고리즘 효율성에 관한 검정의 필요성을 제시한다.

ABSTRACT

The Big O notation of a sorting algorithm analysis is an asymptotic algorithm analysis which gives information of a rough mathematical function with an infinite increase of a sample size, without any specification of a probabilistic model. Hence. in an application with a limited finite number of data, it is necessary to test efficiencies of sorting algorithms. I estimated probabilistic models which analyze the number of exchanges varying input sizes to sort. The estimated models to explain the relationship of sorting efficiency on the sample size(N of the sample size and S of the number of exchange of elements) are $S=0.9305N^{1.339}$ for Quick sort algorithm with O(nlogn) time complexity, and $S=0.2232N^{2.0130}$ for Insertion sort algorithm with $O(n^2)$ time complexity. Furthermore, there are strongly supports that more than 99% of the above relationship could be explained by the estimated models ($p<0.001$). These findings suggest it is necessary to analyze sorting algorithm efficiency in applications with a finite number of data or a newly developed sorting algorithm.

논문접수 : 2004. 2. 10.
심사완료 : 2004. 2. 18.

1) 정회원 : 안산 공과 대학 컴퓨터 정보과 부교수

1. 서론

체계적이고 효율적인 정보 관리 및 처리에 필수적인 기본 작업인 자료들을 오름차순 이나 내림 차순 으로 재배열 하는 정렬의 효율적인 수행을 위하여 기억 장소와 비교 및 교환 횟수를 최소화 하기 위한 최적의 알고리즘의 선택이 필수적이다. 그러므로 정렬된 자료 양 및 필요한 기억 장소 등의 요인들을 고려한 알고리즘 비교에 관한 많은 정보들이 제공 되어야 한다.

이론상으로 정렬 방식에 관한 알고리즘 평가는 각 방법에 있어 입력 크기가 커지거나 또는 한계에 도달할 때의 비교 횟수 또는 실행 시간을 평가함 으로써 구체적인 모형을 구축하지 않고 증가율 만을 분석하는 점진적 알고리즘 (Asymptotic Algorithm Analysis) 이다. 그러나 실제 자료를 정렬하는데는 이보다 훨씬 많은 횟수의 비교가 필요하다는 보고가 있으며 제한된 유한개의 레코드를 정렬하는 실생활 응용 소프트웨어 평가 에서는 점근적 분석에서 무시되는 비교 횟수 식의 상수가 결정적 요인이 될 수 있다 [1]. 일반적으로 정렬 알고리즘 분석을 위하여는 사용 시스템 이나 데이터 유형과 독립적이며 알고리즘 수행 시간과 밀접하게 관련있는 레코드 교환 횟수 측정이 가장 적합한 방법으로 알려지고 있다 [2]. 즉 정렬 자료 양에 따른 교환 횟수의 관계를 설명하는 모형을 구축하면 실제 응용 프로그램 개발 시 자료 양에 따라 가장 적합한 정렬 알고리즘을 선택할 수 있다.

그러므로 본 연구 목표는 점근적 분석 결과에 기초하여 유한 개의 레코드 정렬 알고리즘 효율성 분석 모형을 구축하고 적합도 여부를 확률적으로 분석 함이다.
이러한 정렬 알고리즘 효율성의 확률적 접근은 새로 개발되는 정렬 알고리즘 효율성을 평가하는 유용한 도구가 될 것이다.

2. 방법

2.1 정렬 알고리즘 효율성 평가 모형 구축을 위한 자료 생성

0 에서 16,383 범위에서 중복되지 않은 정수형 난수 9,000 개를 생성하여 삽입 정렬과 퀵 정렬 알고리즘을 적용하여 정렬 시 발생하는 원소의 교환 횟수들을 측정 하였다 (표 1) [3].

<표 1> 정렬 원소 교환 횟수 측정 계획

<Table 1> Design of measuring numbers of exchanges for sorting

자료수	정렬에 필요한 교환 횟수	
	삽입 정렬	퀵 정렬
100		
200		
300		
. . . 8900 9000		

2.2. 정렬 알고리즘 효율성 평가 모형 구축

자료 양 과 정렬에 필요한 교환 횟수 간의 관계를 설명하기 위한 모형을 구축하기 위하여 정렬 알고리즘의 점근적 분석 결과에기초하여 회귀 모형을 선택 하였다.
이론적으로 삽입 정렬은 $O(n^2)$ 이고, 퀵 정렬은 $O(n \cdot \log_2 n)$ 이므로 고정된 자료 개수 (N)를 독립 변수로, 정렬에 필요한 교환 횟수(S) 를 종속 변수로 하는

삽입 정렬의 경우 $S = \beta_1 N^2$ 모형을
퀵 정렬의 경우 $S = \beta_1 N \log(N)$
모형을 선택하고 계수를 측정 한다 [4].

회귀 모형의 계수들은 EXCEL 의 차트 메뉴 에서 추세선 옵션 기능을 활용하여 손쉽게 얻었다 [5].

2.3. 모형의 적합도 검정

추정된 회귀 모형이 자료 양과 정렬에 필요한 교환 횟수간의 관계를 설명하는데 적합한 가의 여부를 검정 한다.
정렬 자료 양에 따른 원소 교환 횟수 S_i 에 대한 모형에 의한 추정치 $\hat{S}_i$, 및 교환 횟수 평균 $\bar{S}$ 의 값 들로 부터

$$R = \frac{\text{모형에의한분산}}{\text{오차에의한분산}} = \frac{\sum(\hat{S}_i - \bar{S})^2}{\sum(S_i - \hat{S}_i)^2}$$

식에 의하여 검정 통계량을 계산한다.

2.4. 확률값 계산

회귀 모형의 적합도를 확률적으로 증명하기 위하여 $F(1,n-1)$ 분포에서 $P(F \geq R)$ 인 확률값을 계산한다.
확률 값이 작을수록 추정된 회귀 모형이 정렬 자료양에 따른 원소 교환 횟수 관계를 설명할 수 있는 확률적 근거가 강함으로 평가한다 [5,6].

3. 결과

3.1. 자료 기술

자료 개수와 정렬에 필요한 교환 횟수가 알고리즘 별로 그림 1 에 기술 되었다.
고정된 자료 크기에서 삽입 정렬에 필요한 교환 횟수는 퀵 정렬에 필요한 교환 횟수에 비하여 1,000 배 이상이며, 퀵 정렬은 자료 개수의 증가에 따라 교환 횟수가 직선형 으로 증가하는 것처럼 보이나, 삽입 정렬 에서는 자료 개수의 증가에 따라 교환 횟수가 2차 곡선 형태로 증가하는 경향을 보였다.

3.2. 모형 추정

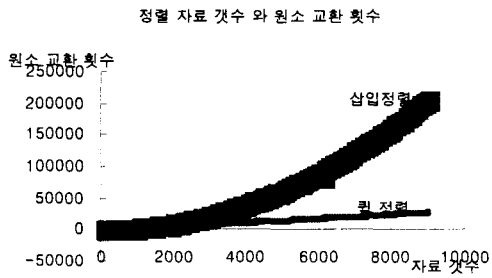
정렬 자료 개수에 따른 교환 횟수 관계를 선택한 회귀 모형 으로 추정한 결과,
퀵 정렬 알고리즘 적용 시 두 변수 관계는

$S=0.9305N^{1.1339}$ 함수로,

삽입 정렬 알고리즘 적용 시 에는

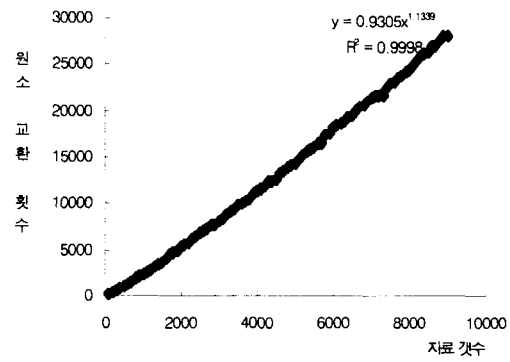
$S=0.2232N^{2.0130}$

함수 로 추정 되었다. (그림 2)



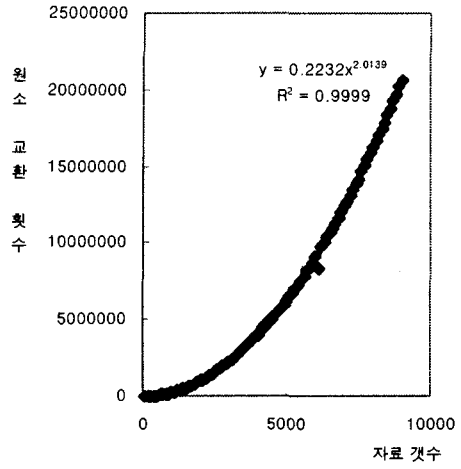
[그림 1] 정렬 알고리즘 별 자료 개수에 따른 원소 교환 횟수

[Fig 1] Plot of input size vs. number of elements exchange by sorting algorithms



[그림 2] 퀵 정렬 회귀 모형 추정

[Fig 2] Estimation of a regression model for Quick sort algorithm



[그림 3] 삽입 정렬 회귀 모형 추정

[Fig 3] Estimation of a regression model for Insertion sort algorithm

3.3. 모형 적합성 통계 분석

퀵 정렬과 삽입 정렬에서 자료 개수에 따른 원소 교환 횟수 관계가 추정된 회귀 모형에 의하여 설명되는 정도와 적합성을 확률적으로 분석한 결과가 표 2 에 있다.

퀵 정렬 및 삽입 정렬 적용 시 자료 개수에 따른 원소 교환 횟수의 관계를 설명하기 위한 적합한 모형은 앞에서 추정된 모형이며 각각의 모형에 의하여 두 변수들의 관계가 99% 이상이 설명될 수 있음이 확률적으로 증명 되었다 ($p<0.0001$).

<표 2> 추정된 모형의 적합성 검정 결과

<Table 2> Goodness of Fit test of the estimated models

정렬 알고리즘	퀵 정렬	삽입 정렬
추정된 모형	$S=0.9305N^{1.1339}$	$S=0.2232N^{2.0130}$
결정 계수*	0.9999	0.9998
p-value**	<0.0001	<0.0001

* 변수 관계들이 추정된 회귀 모형에 의해 설명될 수 있는 정도

** 두 변수 관계를 설명하는 모형으로 적합한 가의 검정 (작은 값 일수록 모형의 적합성을 증명하는 강한 근거가 존재함)

4. 고찰 및 결론

정렬은 다양한 분야의 응용 프로그램 개발 등을 포함한 정보 관리에 필수적인 도구 이므로 본 연구와 같은 알고리즘 효율성에 관한 실험적 연구가 필요하다.

O 표기법은 정렬 자료 개수의 증가에 따른비교 횟수 또는 원소 교환 횟수의 증가 형태 정보만을 제공할 뿐 변수들 관계를 설명하는 모형을 추정하지 못한다.

따라서 본 연구 에서는 자료 개수가 유한 개인 정렬 알고리즘의 효율성을 평가하기위하여 9,000 개 이하의 다양한 표본 수를 가진 음이 아닌 정수들을 정렬하여 자료 개수에 따른 원소 교환 횟수 모형을 구축 하였다.

점근적 정렬 알고리즘 분석 결과에 의하면 평균 시간 복잡도는 $O(n^2)$에서 O(nlogn) 형태를 취하므로 본 연구에서는 $O(n^2)$ 시간 복잡도를 가지는 삽입 정렬과 O(nlogn) 복잡도를 가지는 퀵 정렬 알고리즘을 적용하였다.

삽입 정렬 적용 시 자료 개수에 따른 원소 교환 횟수 관계 설명에 추정된 모형은 $S=0.2232N^{2.0130}$ 이었고 퀵 정렬 적용 시 에는 $S=0.9305N^{1.1339}$으로 추정 되었다.

그림 2 의 퀵 정렬 시 자료 개수를 X 축으로, 원소 교환 횟수를 Y 축으로 2차원 평면 상에서 자료들을 plot 하여 자료들을 지나는 곡선을 추정한 결과 99% 이상의 자료들이 곡선 상에 있었다.

삽입 정렬 에서도 2차원 평면에 plot 된 자료들을 지나는 곡선을 추정한 결과 99% 이상의 자료들이 추정된 곡선 상에 있었다.또한 추정된 모형들이 자료 개수와 원소 교환 횟수 관계를 설명하기에 적합함이 확률적으로 증명 되었다. ($p<0.001$) 퀵 정렬의 점근적 분석 결과는 평균 O(nlogn) 이나 본 실험 결과에서는 $O(n^{1.1339})$ 로 추정되었다.

수학적으로 $n \to \infty$ 일수록 $\log n < n^x$ ($x>0$) 이므로 정렬 자료 개수가 9,000 개 이하일 때 본 연구의 정렬 자료 개수에 따른 원소 교환 횟수 관계가 점근적 분석 결과 보다 약간 과대 추정 되었다. 실제 자료 정렬 시에는

이론적인 점근적 분석 모형에서 제시되는 횟수보다 훨씬 많은 원소의 비교 또는 교환 횟수가 필요하다는 보고 [1] 와 본 연구 결과가 일치한다.

삽입 정렬에서는 2차 모형의 계수가 1 미만으로 추정되었다. 즉 9,000 개 이하의 자료 정렬에서 자료 개수에 따른 원소 교환 횟수가 2차적 형태로 증가하나 증가율은 매우 적은 것으로 평가된다. 50,000 개 이하의 자료 개수를 가진 정렬 알고리즘 효율성 분석 실험[7,8]에서도 자료 개수가 10,000 개 이하일 때 삽입 정렬과 같은 $O(n^2)$ 으로 평가되는 정렬에서는 완만한 곡선 형태를 이루고 있었으므로 본 연구 결과와 일치한다.

결론적으로 본 연구에서는 9,000 개 이하의 자료 정렬 시 삽입 정렬과 퀵 정렬 알고리즘 효율성의 점근적 분석 결과가 적용됨이 확률적으로 증명 되었다. 그러나 500개 이하의 자료를 가진 정렬 알고리즘 실험 연구[9]에서는 퀵 정렬의 모형 적합성이 확률적으로 증명되지 못하였다.

따라서 소량의 유한개의 자료 정렬 시 효율적인 알고리즘 선택을 위한 검정이 필요함을 제시한다.

5. 참고 문헌

[1] GH Gonnet, RB Yates, Handbook of algorithm and Data Structures, Addison-Wesley Publishing Co., p.413, 2002

[2] "Quicksort", in http://clips.ee.uwa.edu.au /~morris /Year2 /PLD210 /niemann /s-man.html,

[3] "Sorting Algorithms" in http://cs.smith.edu/~thiebaut/java/sort/demo.html, 1997.

[4] BW Rust, Fitting Natures Basic Functions, Computing Science Engineering, Vol 4 No 4 p.72-77, 2002

[5] FE Harrel, Regression Modeling Strategies; with applications to Linear Models, Logistic Regression and Survival analysis. Springer New York, p. 103-105, 2002.

[6] KD Concannon, WJ Thompson Regression Lines, Computing Science Engineering vol 2, No 4, p.78-81, 2000.

[7] Comparisons of sorting algorithm, in http://linux.wku.edu/~lamonmol/algor/sort/sort.html

[8] DE Knuth, Sorting and Searching, Art of Computer Programming, Vol 3, p.113-115, Addison Wesley Professional, 1998.

[9] 김숙영, 회귀 분석과 분산 분석을 적용한 정렬 알고리즘 효율성 분석, 제2권, p.211-213. 안산 공과 대학 논문집, 1996.

김숙영

미국 오하이오 주립 대학교 컴퓨터 과학과 졸업

미국 오하이오 주립 대학교 대학원 통계학과 졸업 (응용 통계학 석사)

1995 - 현재 안산 공과 대학 컴퓨터 정보과 부교수

관심 분야 : 전산 통계 (데이터 분석). 자료 구조, 수치해석,