

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2020-0110239
(43) 공개일자 2020년09월23일

- (51) 국제특허분류(Int. Cl.)
G06F 16/901 (2019.01) G06F 12/0871 (2016.01)
G06F 3/06 (2006.01)
- (52) CPC특허분류
G06F 16/9024 (2019.01)
G06F 12/0871 (2013.01)
- (21) 출원번호 10-2020-0031154
(22) 출원일자 2020년03월13일
심사청구일자 2020년03월13일
- (30) 우선권주장
1020190028971 2019년03월13일 대한민국(KR)
- (71) 출원인
한양대학교 산학협력단
서울특별시 성동구 왕십리로 222(행당동, 한양대학교내)
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
- (72) 발명자
서지원
서울특별시 성동구 왕십리로 222
노삼혁
울산광역시 울주군 언양읍 유니스트길 50, 106동 601-4호
이은재
울산광역시 울주군 언양읍 유니스트길 50, 106동 605호
- (74) 대리인
홍성욱, 심경식

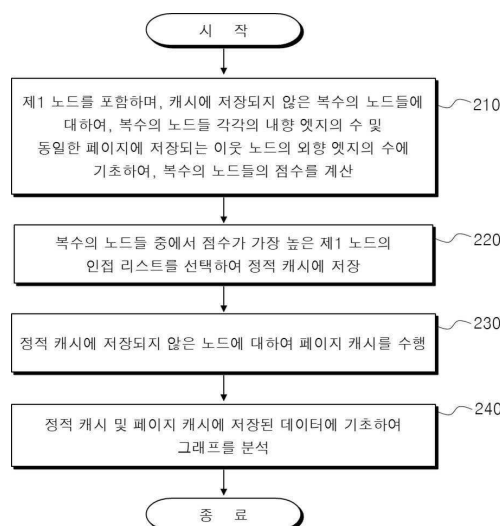
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 디스크 기반의 그래프 분석 방법 및 장치

(57) 요약

디스크 기반 그래프 분석 방법 및 장치를 개시한다. 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법은 제1 노드를 포함하며, 캐시에 저장되지 않은 복수의 노드들에 대하여, 복수의 노드들 각각의 내향 엣지(in-degree)의 수 및 동일한 페이지에 저장되는 이웃 노드의 외향 엣지(out-degree)의 수에 기초하여, 복수의 노드들의 점수를 계산하는 단계, 복수의 노드들 중에서 점수가 가장 높은 제1 노드의 인접 리스트를 선택하여 정적 캐시(static cache)에 저장하는 단계, 정적 캐시에 저장되지 않은 노드에 대하여 페이지 캐시를 수행하는 단계 및 정적 캐시 및 페이지 캐시에 저장된 데이터에 기초하여 그래프를 분석하는 단계를 포함한다.

대표도 - 도2



(52) CPC특허분류

G06F 3/0631 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2010013-2-00109
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	한국연구재단 부설 정보통신기획평가원 / 정보통신방송 연구개발사업 / 혁신성장동
력프로젝트(과기정통부)사업	-인공지능사업(이지바로)
연구과제명	(엑소브레인-2세부) WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기
술 개발	
기 여 율	1/2
과제수행기관명	(주)솔트룩스
연구기간	2019.01.01 ~ 2019.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711074130
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기술진흥센터
연구사업명	원천기술개발사업(차세대정보컴퓨팅기술개발사업)
연구과제명	기계학습 프레임워크를 활용한 Genome 분석 소프트웨어 개발
기 여 율	1/2
과제수행기관명	울산과학기술원
연구기간	2016.11.01 ~ 2020.12.31

명세서

청구범위

청구항 1

제1 노드를 포함하며, 캐시에 저장되지 않은 복수의 노드들에 대하여, 상기 복수의 노드들 각각의 내향 엣지(in-degree)의 수 및 동일한 페이지에 저장되는 이웃 노드의 외향 엣지(out-degree)의 수에 기초하여, 상기 복수의 노드들의 점수를 계산하는 단계;

상기 복수의 노드들 중에서 점수가 가장 높은 제1 노드의 인접 리스트를 선택하여 정적 캐시(static cache)에 저장하는 단계;

상기 정적 캐시에 저장되지 않은 노드에 대하여 페이지 캐시를 수행하는 단계; 및

상기 정적 캐시 및 페이지 캐시에 저장된 데이터에 기초하여 그래프를 분석하는 단계;

를 포함하는 것을 특징으로 하는 디스크 기반 그래프 분석 방법.

청구항 2

제1항에 있어서,

상기 복수의 노드들의 점수는,

상기 복수의 노드들 각각의 내향 엣지의 수가 감소할수록 증가하는 것을 특징으로 하는 디스크 기반 그래프 분석 방법.

청구항 3

제1항에 있어서,

상기 복수의 노드들의 점수는,

상기 복수의 노드들 각각의 상기 동일한 페이지에 저장되는 이웃 노드의 외향 엣지의 수가 감소할수록 증가하는 것을 특징으로 하는 디스크 기반 그래프 분석 방법.

청구항 4

제1항에 있어서,

상기 정적 캐시에 저장하는 단계는,

상기 복수의 노드들 중에서 점수가 가장 높은 노드가 다수인 경우, 상기 다수 노드들 중에서 외향 엣지의 수가 작은 상기 제1 노드의 인접 리스트를 선택하여 상기 정적 캐시에 저장하는 것을 특징으로 하는 디스크 기반 그래프 분석 방법.

청구항 5

제1항에 있어서,

상기 복수의 노드들의 점수를 계산하는 단계 및 상기 정적 캐시에 저장하는 단계는,

상기 정적 캐시의 공간을 모두 소비할 때까지 반복되는 것을 특징으로 하는 디스크 기반 그래프 분석 방법.

청구항 6

제1항에 있어서,

상기 복수의 노드들의 점수를 계산하는 단계 및 상기 정적 캐시에 저장하는 단계는,

수신되는 반복 횟수만큼 반복되는 것을 포함하는 것을 특징으로 하는 디스크 기반 그래프 분석 방법.

청구항 7

제1 노드를 포함하며, 캐시에 저장되지 않은 복수의 노드들에 대하여, 상기 복수의 노드들 각각의 내향 엣지(in-degree)의 수 및 동일한 페이지에 저장되는 이웃 노드의 외향 엣지(out-degree)의 수에 기초하여, 상기 복수의 노드들의 점수를 계산하는, 점수 계산부;

상기 복수의 노드들 중에서 점수가 가장 높은 제1 노드의 인접 리스트를 선택하여 정적 캐시(static cache)에 저장하는, 정적 캐시 할당부;

상기 정적 캐시에 저장되지 않은 노드에 대하여 페이지 캐시를 수행하는, 페이지 캐시 할당부; 및

상기 정적 캐시 및 페이지 캐시에 저장된 데이터에 기초하여 그래프를 분석하는, 그래프 분석부;

를 포함하는 것을 특징으로 하는 디스크 기반 그래프 분석 장치.

청구항 8

제7항에 있어서,

상기 복수의 노드들의 점수는,

상기 복수의 노드들 각각의 내향 엣지의 수가 감소할수록 증가하는 것을 특징으로 하는 디스크 기반 그래프 분석 장치.

청구항 9

제7항에 있어서,

상기 복수의 노드들의 점수는,

상기 복수의 노드들 각각의 상기 동일한 페이지에 저장되는 이웃 노드의 외향 엣지의 수가 감소할수록 증가하는 것을 특징으로 하는 디스크 기반 그래프 분석 장치.

청구항 10

제7항에 있어서,

상기 정적 캐시 할당부는,

상기 복수의 노드들 중에서 점수가 가장 높은 노드가 다수인 경우, 상기 다수 노드들 중에서 외향 엣지의 수가 작은 상기 제1 노드의 인접 리스트를 선택하여 상기 정적 캐시에 저장하는 것을 특징으로 하는 디스크 기반 그래프 분석 장치.

청구항 11

제7항에 있어서,

상기 점수 계산부 및 상기 정적 캐시 할당부는,

상기 정적 캐시의 공간을 모두 소비할 때까지 반복적으로 기능을 수행하는 것을 특징으로 하는 디스크 기반 그래프 분석 장치.

청구항 12

제7항에 있어서,

상기 점수 계산부 및 상기 정적 캐시 할당부는,

수신되는 반복 횟수만큼 반복적으로 기능을 수행하는 것을 포함하는 것을 특징으로 하는 디스크 기반 그래프 분석 장치.

청구항 13

컴퓨터 프로그램이 프로세서에 의해 실행될 때, 제1항 내지 제6항 중 어느 한 항에 따른 방법이 수행되는 컴퓨

터 프로그램을 저장한 컴퓨터-판독가능 저장 매체.

발명의 설명

기술 분야

[0001] 본 발명은 디스크 기반의 그래프 분석 방법 및 장치에 관한 것으로, 보다 상세하게는 정적 캐시 및 페이지 캐시를 동시에 사용함으로써 캐시에 저장된 데이터의 이용률을 향상시키는 그래프 분석 방법 및 장치에 관한 것이다.

배경 기술

[0002] 디스크 기반 그래프 분석 시스템은 디스크에 저장된 대용량 그래프를 적은 양의 메모리(DRAM)를 사용해서 분석하기 위하여, 그래프 알고리즘을 실행할 때 디스크에 저장된 그래프의 일부를 메모리에 임시로 저장(캐싱, Caching) 한다. 이때, 한번 읽은 파일의 내용을 페이지 캐시라는 영역에 저장시킨 후, 동일한 파일 접근이 발생하면 페이지 캐시로부터 파일을 제공하는 페이지 캐시 방식이 주로 사용된다.

[0003] 그러나 페이지 캐시의 경우, 페이지 단위(일반적으로 4KB 내지 1MB 크기)로 메모리 영역에 저장되기 때문에, 너비 우선 탐색 기반 알고리즘들 또는 BFS(Breadth First Search) 기반 알고리즘들에 사용되는 데이터 양에 비해 디스크에서 읽어오는 데이터가 훨씬 많게 되어, 페이지 캐시에 저장된 데이터의 이용률(utilization)이 낮은 경우가 생길 수 있다. 또한, 필요없는 데이터를 캐시에 저장하고 있으므로 필요한 데이터를 저장할 공간이 감소하는 문제가 생길 수 있다.

[0004] 한편, 페이지 캐시와 정적 캐시를 동시에 사용하면서, 정적 캐시에 저장될 데이터를 무작위로 선택하는 기법 역시 사용될 수 있다. 그러나 정적 캐시에 저장될 데이터가 무작위로 선택되므로 캐시에 저장된 데이터의 이용률이 저하될 수 있다.

[0005] 이에 따라 캐시에 저장된 데이터의 이용률을 증가시킬 수 있는 여러 기법들이 연구 중에 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 대한민국 등록특허공보 제10-1658885호(2016. 9. 22)

발명의 내용

해결하려는 과제

[0007] 본 발명은 디스크 기반의 그래프 분석 방법 및 장치에 관한 것으로, 캐시에 저장된 데이터의 이용률을 향상시킬 수 있도록, 정적 캐시에 저장되는 데이터를 결정하고, 정적 캐시 및 페이지 캐시를 동시에 사용함으로써 캐시에 저장된 데이터의 이용률을 향상시키는 그래프 분석 방법 및 장치를 제공하고자 한다.

[0008] 본 발명이 해결하고자 하는 과제는 이상에서 언급한 과제(들)로 제한되지 않으며, 언급되지 않은 또 다른 과제(들)은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법은 제1 노드를 포함하며, 캐시에 저장되지 않은 복수의 노드들에 대하여, 복수의 노드들 각각의 내향 엣지(in-degree)의 수 및 동일한 페이지에 저장되는 이웃 노드의 외향 엣지(out-degree)의 수에 기초하여, 복수의 노드들의 점수를 계산하는 단계, 복수의 노드들 중에서 점수가 가장 높은 제1 노드의 인접 리스트를 선택하여 정적 캐시(static cache)에 저장하는 단계, 정적 캐시에 저장되지 않은 노드에 대하여 페이지 캐시를 수행하는 단계 및 정적 캐시 및 페이지 캐시에 저장된 데이터에 기초하여 그래프를 분석하는 단계를 포함한다.

[0010] 바람직하게는, 복수의 노드들의 점수는 복수의 노드들 각각의 내향 엣지의 수가 감소할수록 증가한다.

- [0011] 바람직하게는, 복수의 노드들의 점수는 복수의 노드들 각각의 동일한 페이지에 저장되는 이웃 노드의 외향 엣지의 수가 감소할수록 증가한다.
- [0012] 바람직하게는, 정적 캐시에 저장하는 단계는 복수의 노드들 중에서 점수가 가장 높은 노드가 다수인 경우, 다수 노드들 중에서 외향 엣지의 수가 작은 제1 노드의 인접 리스트를 선택하여 정적 캐시에 저장하는 것일 수 있다.
- [0013] 바람직하게는, 복수의 노드들의 점수를 계산하는 단계 및 정적 캐시에 저장하는 단계는 정적 캐시의 공간을 모두 소비할 때까지 반복되는 것일 수 있다.
- [0014] 바람직하게는, 복수의 노드들의 점수를 계산하는 단계 및 정적 캐시에 저장하는 단계는 수신되는 반복 횟수만큼 반복되는 것일 수 있다.
- [0015] 또한, 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 장치는 제1 노드를 포함하며, 캐시에 저장되지 않은 복수의 노드들에 대하여, 복수의 노드들 각각의 내향 엣지(in-degree)의 수 및 동일한 페이지에 저장되는 이웃 노드의 외향 엣지(out-degree)의 수에 기초하여, 복수의 노드들의 점수를 계산하는 점수 계산부, 복수의 노드들 중에서 점수가 가장 높은 제1 노드의 인접 리스트를 선택하여 정적 캐시(static cache)에 저장하는 정적 캐시 할당부, 정적 캐시에 저장되지 않은 노드에 대하여 페이지 캐시를 수행하는, 페이지 캐시 할당부 및 정적 캐시 및 페이지 캐시에 저장된 데이터에 기초하여 그래프를 분석하는 그래프 분석부를 포함할 수 있다.
- [0016] 바람직하게는, 복수의 노드들의 점수는 복수의 노드들 각각의 내향 엣지의 수가 감소할수록 증가한다.
- [0017] 바람직하게는, 복수의 노드들의 점수는 복수의 노드들 각각의 동일한 페이지에 저장되는 이웃 노드의 외향 엣지의 수가 감소할수록 증가한다.
- [0018] 바람직하게는, 정적 캐시 할당부는 복수의 노드들 중에서 점수가 가장 높은 노드가 다수인 경우, 다수 노드들 중에서 외향 엣지의 수가 작은 제1 노드의 인접 리스트를 선택하여 정적 캐시에 저장하는 것일 수 있다.
- [0019] 바람직하게는, 점수 계산부 및 정적 캐시 할당부는 정적 캐시의 공간을 모두 소비할 때까지 반복적으로 기능을 수행하는 것일 수 있다.
- [0020] 바람직하게는, 점수 계산부 및 정적 캐시 할당부는 수신되는 반복 횟수만큼 반복적으로 기능을 수행하는 것일 수 있다.
- [0021] 본 발명은 컴퓨터 프로그램이 프로세서에 의해 실행될 때, 상술된 방법이 수행되는 컴퓨터 프로그램을 저장한 컴퓨터-판독가능 저장 매체를 제공한다.

발명의 효과

- [0022] 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법 및 장치는 캐시에 저장된 데이터의 이용률을 향상시킬 수 있도록 정적 캐시에 저장되는 데이터를 결정하기 때문에, 페이지 캐시만 이용하는 경우 및 정적 캐시에 저장되는 데이터를 무작위로 선택하는 경우에 비해 데이터 이용률이 향상될 수 있다.
- [0023] 또한, 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법 및 장치는 BFS 기반의 알고리즘들을 수행할 때, 동일한 양의 메모리를 사용하여 캐싱을 하는 경우 기존 방법에 비해, 캐시에 저장된 데이터의 이용률이 향상되기 때문에, 그래프 분석에 소요되는 실행 시간이 감소하는 효과가 있다.

도면의 간단한 설명

- [0024] 도 1은 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법을 설명하기 위한 도면이다.
- 도 2는 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법을 설명하기 위한 흐름도이다.
- 도 3은 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법에 관한 의사코드(pseudo code)를 나타낸 도면이다.
- 도 4 및 도 5는 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법 및 장치에 관한 실험 결과를 설명하기 위한 도면이다.
- 도 6은 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 장치를 설명하기 위한 블록도이다.

발명을 실시하기 위한 구체적인 내용

- [0025] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.
- [0026] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.
- [0027] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.
- [0028] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0029] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0030] 도 1은 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법을 설명하기 위한 도면이다.
- [0031] 도 1의 (a)의 각 원은 그래프의 노드를 의미하고, 화살표는 엣지를 나타낸다. 도 1의 (a)를 참고하면, u 노드에 x, y, z, v 노드가 연결되어 있고, x, y, z, v 노드 각각에도 인접한 노드의 리스트(이하 '인접 리스트'로 지칭됨)가 있다고 가정할 수 있다. 예를 들어, x 노드의 경우 내향 엣지(in-degree)는 u 노드로부터 1 이다. 그리고, x 노드의 외향 엣지(out-degree)는 4 이므로, x 노드의 인접 리스트의 크기는 4 일 수 있다. 또한, y 노드의 경우 내향 엣지는 u 노드로부터 1 이다. 그리고, y 노드의 외향 엣지(out-degree)가 3 이므로, y 노드의 인접 리스트의 크기는 3 일 수 있다. 이하에서는 u 노드의 인접 리스트 중에서 x, y, z, v 노드들이 접근 요청된 것을 전체로 설명한다. 한편, 디스크(SSD 또는 HDD)에는 그래프의 노드와 각 노드의 인접 리스트가 노드 번호 순서대로 저장될 수 있다.
- [0032] 도 1의 (b)는 도 1의 (a)의 그래프가 디스크에 저장된 경우를 설명하기 위한 도면이고, 도 1의 (c)는 도 1의 (a)의 그래프가 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법에 입력된 경우, 메모리 영역(캐시)에 저장될 데이터를 설명하기 위한 도면이다.
- [0033] 도 1의 (b)를 참고하면, 작은 사각형 각각은 메모리 영역이고, 큰 점선 사각형은 페이지 단위의 메모리 영역을 의미한다. 도 1의 (b)의 제1 페이지에는 x 노드의 4개의 인접 리스트, y 노드의 3개의 인접 리스트, z 노드의 1개의 인접 리스트들이 각각 저장되어 있고, 제2 페이지에는 v의 2개의 인접 리스트가 순서대로 저장된 것을 확인할 수 있다.
- [0034] 만약 페이지 캐시를 활용한 BFS 기반 알고리즘을 수행한다면 u 노드에 접근한 후, u 노드와 인접한 노드인 x, y, z, v의 인접 리스트가 저장된 제1 페이지 및 제2 페이지를 읽어와야 한다. 이때 제1 페이지를 읽어오면, 제1 페이지의 8/9에 해당하는 영역의 데이터를 실제 사용하게 되어 페이지 캐시의 이용률(utilization)이 좋다고 할 수 있다. 하지만 제2 페이지의 경우, 읽어들인 데이터의 2/9에 해당하는 v 노드의 인접 리스트가 저장된 데이터만 실제로 사용하게 되므로, 캐시에 저장된 전체 데이터의 이용률은 감소하게 된다.
- [0035] 따라서, 본 발명은 페이지 캐시와 함께 교체가 발생하지 않는 정적 캐시를 사용하고, 캐시에 저장된 데이터 이

용량을 향상시킬 수 있도록 정적 캐시에 저장할 데이터를 결정하는, 디스크 기반 그래프 분석 방법 및 장치를 개시한다.

- [0036] 도 1의 (c)를 참고하면, 페이지 캐시의 이용률을 향상시키기 위해 페이지 캐시(110)에는 x, y, z 노드의 인접 리스트가 저장된 제1 페이지만 저장하고, 교체가 발생하지 않는 정적 캐시(120)에는 v 노드의 인접 리스트를 저장할 수 있다. 이러한 경우 x, y, z 노드의 인접 리스트들에 접근하게 될 때 제1 페이지만을 디스크 또는 페이지 캐시 영역(110)에서 읽어오고, v 노드의 인접 리스트는 정적 캐시(120)로부터 읽어서 그래프 분석 알고리즘을 실행하기 때문에, 전체적으로 메모리의 효율이 높아질 수 있다.
- [0037] 이때 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법은 메모리 소모 공간 대비 높은 성능 이득을 얻을 수 있는 노드들의 점수를 집계하고, 점수가 높은 노드들의 인접 리스트를 정적 캐시 영역에 미리 읽어와 저장하며 이에 기초하여 그래프 분석 알고리즘을 수행한다.
- [0038] 따라서, 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법 및 장치는 기존 페이지 캐시를 사용하는 디스크 기반 그래프 분석 시스템에 적용 가능하며, 특히 그래프 분석 알고리즘을 노드 간의 메시지 전달 방식으로 구현하도록 인터페이스가 정의되고, 메시지 전달을 위한 글로벌 배리어(Global Barrier)로 구성된 그래프 분석 시스템에서 효과적으로 활용될 수 있다. 예를 들어, 사회연결망 (Social Network) 분석, 생명공학 네트워크 분석 등 대용량 그래프의 효율적인 분석이 필요한 분야에서 적용될 수 있다.
- [0039] 한편, BFS 기반 알고리즘은 그래프의 노드를 재귀적으로 방문하여 계산을 하는 모든 알고리즘을 의미하며 BFS, 최단경로 찾기 (하나의 노드 혹은 모든 노드로 부터의 최단경로 찾기), 그래프의 지름 계산, 연결 요소 (Connected Component) 찾기, 매개 중심성(Betweenness Centrality) 계산 등의 알고리즘이 포함될 수 있으므로, 본 발명의 디스크 기반 그래프 분석 방법 및 장치는 BFS 알고리즘을 수행하는 분야에 적용될 수 있다.
- [0040] 도 2는 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법을 설명하기 위한 흐름도이다.
- [0041] 단계 210에서, 캐시에 저장되지 않은 복수의 노드들에 대하여, 복수의 노드들 각각의 내향 엣지(in-degree)의 수 및 동일한 페이지에 저장되는 이웃 노드의 외향 엣지(out-degree)의 수에 기초하여, 복수의 노드들의 점수를 계산할 수 있다.
- [0042] 여기서 복수의 노드들의 점수는 복수의 노드들 각각의 내향 엣지의 수가 감소할수록 증가한다. 또한, 복수의 노드들의 점수는 복수의 노드들 각각의 동일한 페이지에 저장되는 이웃 노드의 외향 엣지의 수가 감소할수록 증가한다.
- [0043] 단계 220에서, 복수의 노드들 중에서 점수가 가장 높은 제1 노드의 인접 리스트를 선택하여 정적 캐시(static cache)에 저장할 수 있다.
- [0044] 단계 220에서, 복수의 노드들 중에서 점수가 가장 높은 노드가 다수인 경우, 다수 노드들 중에서 외향 엣지의 수가 작은 제1 노드의 인접 리스트를 선택하여 정적 캐시에 저장할 수 있다.
- [0045] 한편, 단계 210 및 단계 220은 정적 캐시의 공간을 모두 소비할 때까지 반복될 수 있다.
- [0046] 또한, 단계 210 및 단계 220은 수신되는 반복 횟수만큼 반복될 수 있다.
- [0047] 단계 230에서, 정적 캐시에 저장되지 않은 노드에 대하여 페이지 캐시를 수행할 수 있다.
- [0048] 단계 240에서, 정적 캐시 및 페이지 캐시에 저장된 데이터에 기초하여 그래프를 분석할 수 있다.
- [0049] 도 3은 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법에 관한 의사코드(pseudo code)를 나타낸 도면이다.
- [0050] 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법 및 장치는 그래프 노드마다 인접 리스트를 메모리에 저장할 때 필요한 메모리 소모량, 해당 인접 리스트가 포함된 페이지가 페이지 캐시에서 캐시 적중할 비율 그리고 각 페이지마다 캐시되지 않은 노드의 인접 리스트가 차지하는 비율 등을 고려해서 각 노드마다 점수를 계산한다. 이후, 메모리 내에 캐시 교체가 발생하지 않는 정적 캐시 영역을 할당받고, 점수가 높은 노드의 인접 리스트들을 할당된 영역에 캐싱한다. 그 다음으로 캐시되지 않은 노드의 리스트를 갱신하고, 캐시되지 않은 노드마다 점수를 다시 계산하여, 점수가 높은 노드의 인접 리스트를 정적 캐시 영역에 추가로 저장한다. 할당받은 캐시 영역을 모두 소비할 때까지 상술한 작업을 반복하며, 정적 캐시 영역을 모두 소비하면, 그래프 분석을 수

행한다.

- [0051] 도 3의 의사코드는 정적 캐시에 저장할 노드들을 선택하는 단계를 구체적으로 설명하기 위한 것이다.
- [0052] 도 3을 참고하면, 정적 캐시에 저장할 노드들을 선택하기 위해서 그래프 G, 정적인 캐시 크기 M, 알고리즘의 반복(iteration) 횟수 K를 입력받고, 캐시에 저장할 노드들을 의미하는 C 집합을 출력하는 것을 확인할 수 있다. 여기서 그래프 G의 변수인 V와 E는 각각 노드 집합과 엣지 집합을 의미한다.
- [0053] 도 3의 P(u_1)은 u_1 노드와 같은 페이지에 저장된 노드들의 집합이며 T는 각 노드들의 점수를 의미하는 집합이다. 예를 들어, 도 1의 (a)와 같은 그래프의 경우, u_1 노드가 x 노드이면, P(u_1)은 {y, z}일 수 있다. 또한 $deg_i(u_1)$ 와 $deg_o(u_1)$ 는 각각 u_1 노드의 내향 엣지의 수(in-degree, u_1 를 가리키는 노드들의 개수)와 외향 엣지의 수(out-degree, u_1 가 가리키는 노드들의 개수)를 의미한다.
- [0054] 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법 및 장치는 점수가 높으면서 캐시 영역을 적게 차지하는 노드들의 인접리스트를 선택하여 캐시에 저장할 수 있다.
- [0055] 도 3의 제1 부분(310)에서는 각 노드마다 점수를 계산한다. 구체적으로 제1 부분(310) 중, 제7행은 그래프 G에 속한 노드 집합 V의 원소인 v의 인접 리스트에 포함된 노드 중 C 집합에 포함되지 않은 하나를 u_1 으로 결정하고, 제9행에서 v의 인접 리스트에 포함된 노드 중 u_1 노드와 동일한 페이지에 인접 리스트가 저장되고, C 집합에 포함되지 않은 하나를 u_2 로 결정한다. 이하에서는 u_2 노드를 u_1 노드와 동일한 페이지에 저장되는 이웃 노드로 지칭하기로 한다.
- [0056] 이후 각 노드의 점수에 u_2 의 외향 엣지의 수를 더하고, 제11행의 수식에 따라 점수를 계산한다. 여기서 $r(u_1)$ 은 u_1 노드의 인접 리스트 저장된 페이지가 읽혀질 예상 횟수를 나타낸 것으로, u_1 노드의 인접 리스트가 저장된 페이지에 저장된 다른 인접 리스트들에 대응하는 노드의 수와 비례하는 것을 가정하였다. 따라서 제1 부분(310)의 제10행 및 제11행을 참고하면, 일 실시예에 따른 디스크 기반 그래프 분석 방법 및 장치는 해당 노드의 내향 엣지의 수, 동일한 페이지에 저장되는 이웃 노드의 외향 엣지의 수 및 $r(u_1)$ 에 기초하여 각 노드들의 점수를 계산하는 것을 확인할 수 있다.
- [0057] 또한 일 실시예에 따른 디스크 기반 그래프 분석 방법 및 장치는 도 3의 제2 부분(320)에서 점수가 높으면서 저장하는 인접 리스트를 저장하기 위해, 캐시 영역을 상대적으로 적게 드는 노드들을 C 집합에 저장한다. 이때 각 노드들의 외향 엣지의 수에 기초하여 C 집합에 저장될 노드들을 선택할 수 있다.
- [0058] 한편 일 실시예에 따른 디스크 기반 그래프 분석 방법 및 장치는 도 3의 알고리즘을 반복 횟수 K 만큼 반복하여 수행할 수 있고, 캐시 영역을 모두 소비할 때까지 수행할 수 있다.
- [0059] 반복 횟수 K가 증가하면 도 3의 알고리즘을 한번 수행할 때 마다 더 적은 노드들을 C 집합에 저장할 수 있으므로, 정적 캐시에 저장될 노드를 더 정확하게 결정할 수 있다. 그러나 K가 증가하면 정적 캐시에 저장될 노드를 선택하는 복잡도는 증가할 수 있다.
- [0060] 도 4 및 도 5는 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법 및 장치에 관한 실험 결과를 설명하기 위한 도면이다.
- [0061] 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법에 관한 실험에 이용된 데이터 셋은 아래의 표 1과 같다.
- [0062] [표 1]

그래프	V	E
Wikipedia(WK)	18 MB	172 MB
LiveJournal(LJ)	4.8 MB	68 MB
Twitter(TW)	53 MB	1.9 Bytes

- [0063]
- [0064] 다시 말해, 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법의 성능을 증명하기 위해 진행된 실험에

서 대용량 그래프 데이터인 Livejournal(LJ, 사회연결망 그래프), Wikipedia(WK, 웹 연결 그래프), Twitter(TW, 사회연결망 그래프)가 입력으로 사용되었다.

- [0065] 도 4는 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법을 Zheng, D., Mhembere, D., Burns, R., Vogelstein, J., Priebe, C. E., & Szalay, A. S. (2015) "FlashGraph: Processing billion-node graphs on an array of commodity SSDs", In 13th USENIX Conference on File and Storage Technologies (FAST 15) (pp. 45-58)에 개시된 FlashGraph라는 오픈소스 그래프 분석 시스템에 적용을 하였을 때의 성능을 나타내는 그래프이다.
- [0066] 도 4는 표 1의 데이터 셋을 입력으로 사용하고, 6 가지 BFS 기반 그래프 알고리즘 - BFS, DIAM(measuring diameter of graph, 그래프 지름 계산), BC (Betweenness Centrality), SP(shortest paths, 최단거리 계산), APSP(all-pair shortest paths, 최단거리 계산을 모든 노드로부터 하기), WCC(weakly connected components, 네트워크에서 서로 연결된 컴포넌트 찾기) - 을 실행한 결과이다. 한편 본 실험은 <http://konect.uni-koblenz.de/networks/> 에서 배포하는 데이터를 그래프 데이터로 하였다.
- [0067] 도 4는 가로 축의 6 가지 BFS 기반 그래프 알고리즘에 표 1의 데이터 셋을 입력하였을 때의 상대적인 실행시간을 나타낸 것으로, 가로 축의 10% 내지 40%는 입력 그래프 크기에 대한 캐시 크기의 비율을 나타낸다.
- [0068] 도 4를 참고하면, 페이지 캐시만 사용하였을 때에 비해 정적 캐시에 저장되는 데이터를 무작위로 선택하였을 경우 실행 시간은 느려지는 것을 확인할 수 있다. 이를 통해 정적 캐시에 저장되는 데이터를 무작위로 선택하는 경우 오히려 그래프 분석 성능이 저하될 수 있음을 알 수 있다.
- [0069] 이에 반해 일 실시예에 따른 디스크 기반 그래프 분석 방법의 경우, 페이지 캐시만 사용하였을 때에 비해 최대 33% 성능이 빨라지는 것을 확인할 수 있다. 이를 통해 일 실시예에 따른 디스크 기반 그래프 분석 방법은 페이지 캐시만을 이용하거나 정적 캐시에 저장되는 데이터를 무작위로 선택하는 경우보다 그래프 분석 성능이 향상될 수 있음을 알 수 있다.
- [0070] 도 5는 정적 캐시에 저장되는 데이터를 무작위로 선택한 경우와 본원 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법의 페이지 이용률을 나타내는 그래프이다.
- [0071] 도 5를 참고하면, 3가지 BFS 기반 그래프 알고리즘에 3가지 데이터 셋을 입력한 경우 모두에서 정적 캐시에 저장되는 데이터를 무작위로 선택한 경우와 본원 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법의 페이지 이용률이 높은 것을 확인할 수 있다. 특히 BC 그래프 알고리즘에 Twitter 데이터 셋을 입력한 경우, 본원 발명의 일 실시예에 따른 디스크 기반 그래프 분석 방법의 페이지 이용률은 정적 캐시에 저장되는 데이터를 무작위로 선택한 경우에 비해 약 22% 높은 것을 확인할 수 있다.
- [0072] 도 6은 본 발명의 일 실시예에 따른 디스크 기반 그래프 분석 장치를 설명하기 위한 블록도이다.
- [0073] 디스크 기반 그래프 분석 장치(600)는 점수 계산부(610), 정적 캐시 할당부(620), 페이지 캐시 할당부(630) 및 그래프 분석부(640)를 포함할 수 있다.
- [0074] 점수 계산부(610)는 제1 노드를 포함하며, 캐시에 저장되지 않은 복수의 노드들에 대하여, 복수의 노드들 각각의 내향 엣지(in-degree)의 수 및 동일한 페이지에 저장되는 이웃 노드의 외향 엣지(out-degree)의 수에 기초하여, 복수의 노드들의 점수를 계산할 수 있다.
- [0075] 이때 복수의 노드들의 점수는 복수의 노드들 각각의 내향 엣지의 수가 감소할수록 증가한다. 또한, 복수의 노드들의 점수는 복수의 노드들 각각의 동일한 페이지에 저장되는 이웃 노드의 외향 엣지의 수가 감소할수록 증가한다.
- [0076] 정적 캐시 할당부(620)는 복수의 노드들 중에서 점수가 가장 높은 제1 노드의 인접 리스트를 선택하여 정적 캐시(static cache)에 저장할 수 있다.
- [0077] 여기서 정적 캐시 할당부(620)는 복수의 노드들 중에서 점수가 가장 높은 노드가 다수인 경우, 다수 노드들 중에서 외향 엣지의 수가 작은 제1 노드의 인접 리스트를 선택하여 정적 캐시에 저장하는 것일 수 있다.
- [0078] 한편, 점수 계산부(610) 및 정적 캐시 할당부(620)는 정적 캐시의 공간을 모두 소비할 때까지 반복적으로 기능을 수행하는 것일 수 있다.
- [0079] 또한, 점수 계산부(610) 및 정적 캐시 할당부(620)는 수신되는 반복 횟수만큼 반복적으로 기능을 수행하는 것일

수 있다.

[0080] 페이지 캐시 할당부(630)는 정적 캐시에 저장되지 않은 노드에 대하여 페이지 캐시를 수행할 수 있다.

[0081] 그래프 분석부(640)는 정적 캐시 및 페이지 캐시에 저장된 데이터에 기초하여 그래프를 분석할 수 있다.

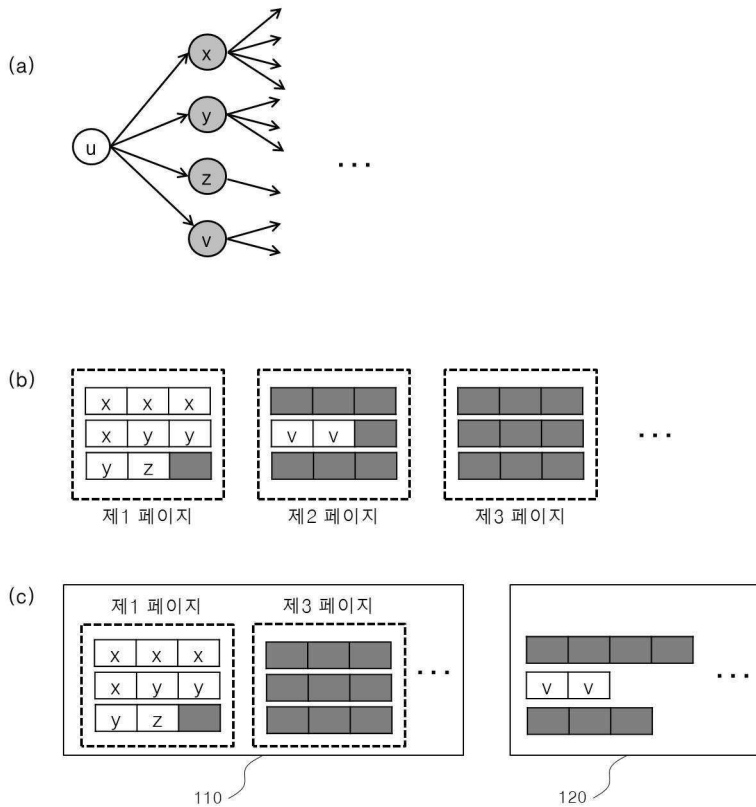
[0082] 상술한 본 발명의 실시예들은 컴퓨터에서 실행될 수 있는 프로그램으로 작성가능하고, 컴퓨터로 읽을 수 있는 기록매체를 이용하여 상기 프로그램을 동작시키는 범용 디지털 컴퓨터에서 구현될 수 있다.

[0083] 상기 컴퓨터로 읽을 수 있는 기록매체는 마그네틱 저장매체(예를 들면, 롬, 플로피 디스크, 하드디스크 등), 광학적 판독 매체(예를 들면, 시디롬, 디브이디 등)를 포함한다.

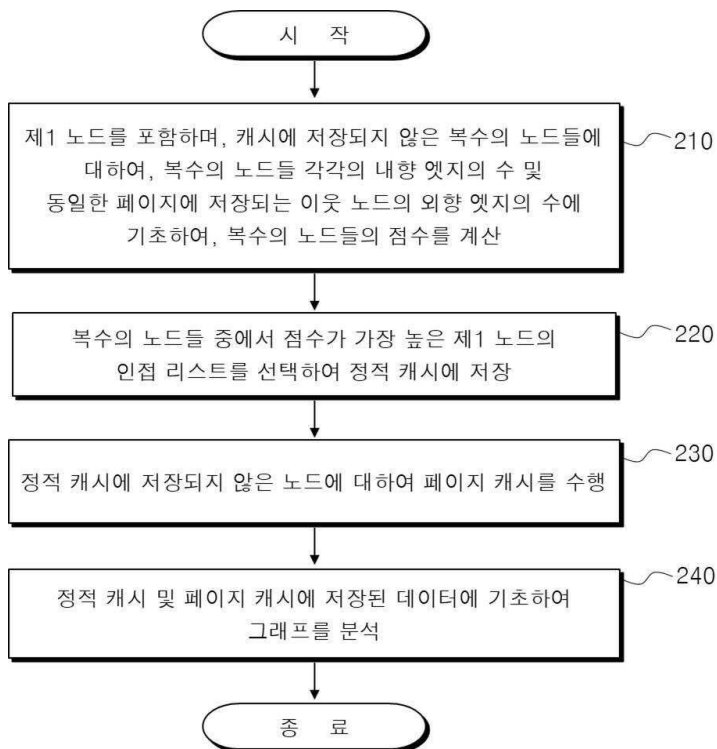
[0084] 이제까지 본 발명에 대하여 그 바람직한 실시예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

도면

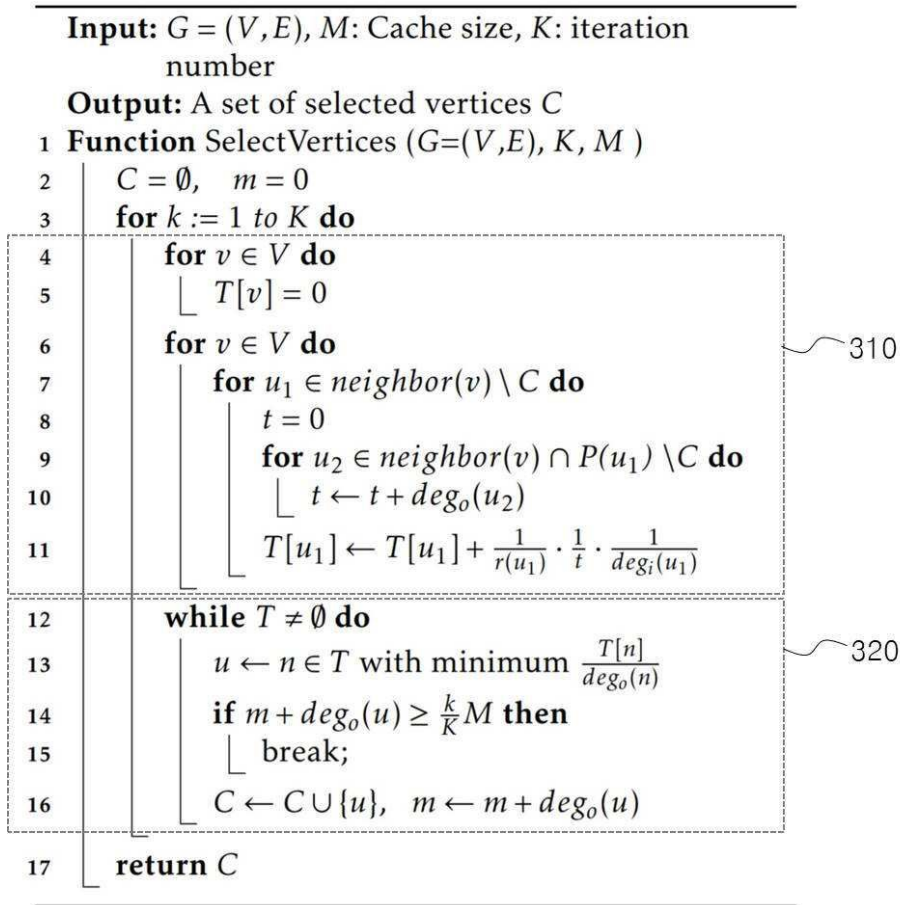
도면1



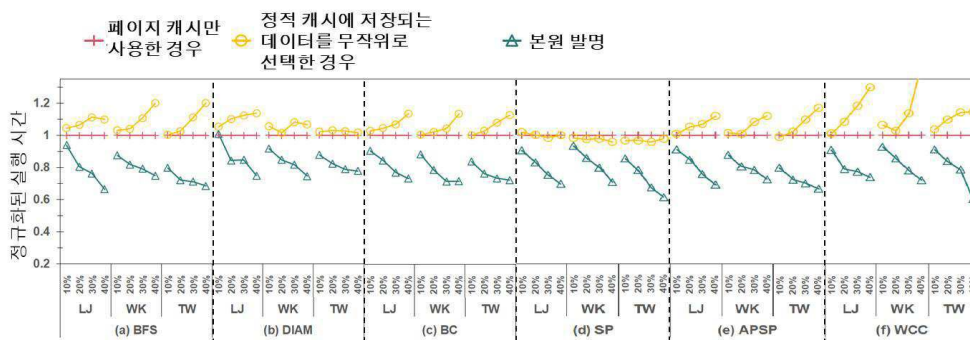
도면2



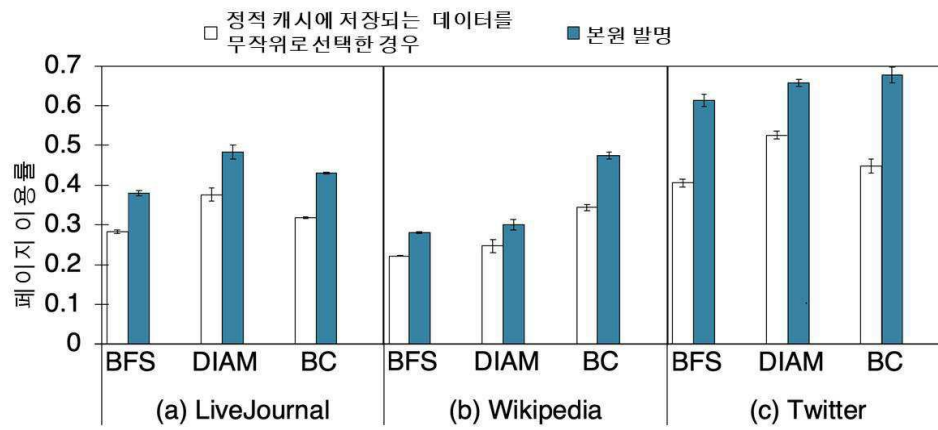
도면3



도면4



도면5



도면6

