



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2018년11월14일
(11) 등록번호 10-1918208
(24) 등록일자 2018년11월07일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01)
(52) CPC특허분류
G06F 17/30522 (2013.01)
G06F 17/30864 (2013.01)
(21) 출원번호 10-2018-0017063
(22) 출원일자 2018년02월12일
심사청구일자 2018년02월12일
(56) 선행기술조사문헌
KR1020100068532 A*
(뒷면에 계속)

(73) 특허권자
김춘호
부산광역시 해운대구 양운로37번길 11, 105동 50
1호 (좌동, 해운대현대아파트)
부경대학교 산학협력단
부산광역시 남구 신선로 365 (용당동,
부경대학교)
(72) 발명자
이중화
울산광역시 남구 돌길로401번길 12, 112동 1305
호(삼산동, 세양청구아파트)
이현규
부산광역시 해운대구 대천로 225, 207동 601호(좌
동, 두산2차아파트)
김춘호
부산광역시 해운대구 양운로37번길 11, 105동 50
1호 (좌동, 해운대현대아파트)
(74) 대리인
리앤목특허법인

전체 청구항 수 : 총 3 항

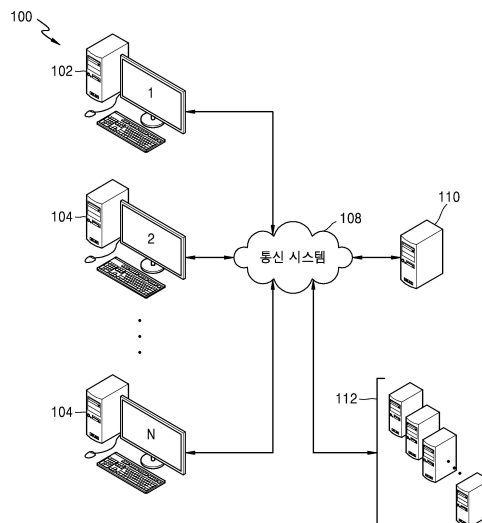
심사관 : 고재용

(54) 발명의 명칭 실시간으로 크롤링 데이터를 분석하는 분석 서버 및 분석 서버의 동작 방법

(57) 요약

본 실시예는 하나 이상의 프로세서 및 저장 매체를 포함하는 분석 서버에 있어서, 상기 저장 매체는 인터넷 망을 통해 크롤링되는 복수의 웹 페이지들에 포함된 표현에 포함된 키워드를 중심으로, 대상 및 대상에 대한 감정을 구조화하는 하나 이상의 제1 데이터 세트를 개별적으로 생성하고, 상기 제1 데이터 세트를 상기 제1 데이터 세트에 출현되는 하나 이상의 대상을 기준으로 군집화하고, 군집화된 각 대상의 감정 및 각 대상의 감정의 서열을 포함하는 제2 데이터 세트를 생성하는 사전 분석 처리 수단을 포함하는, 분석 서버를 개시한다.

대표도 - 도1



(56) 선행기술조사문헌

KR1020100098014 A*

KR101783476 B1*

JP2016201080 A*

KR1020120003834 A*

KR101602898 B1*

*는 심사관에 의하여 인용된 문헌

명세서

청구범위

청구항 1

웹 크롤링 기능을 분산적으로 수행하는 클라이언트들과 통신하며, 하나 이상의 프로세서 및 저장 매체를 포함하는 분석 서버에 있어서,

상기 저장 매체는

상기 클라이언트들에 의해 또는 자체적으로 크롤링된 데이터를 수집하되, 처리 능력을 기초로 크롤링된 데이터의 범위를 동적으로 변경하며,

상기 크롤링된 데이터로부터 웹 페이지, 포스팅에 포함된 표현 중에서, 의미 없는 텍스트들을 제거하고, 소셜 네트워크 상에 각 포스팅과 관련된 객체들 간의 관계를 기초로 관계 데이터 세트를 생성하고, 포스팅의 표현에 포함된 키워드를 중심으로, 객체의 감정, 특정 대상에 대한 감정, 관심을 가지는 대상을 구조화하여 제1 데이터 세트를 {객체, 대상, 감정}의 형태로 생성하는 과정을 반복하여 생성된 데이터 세트들을 통해 대상에 대한 감정의 집합을 생성하고,

상기 대상에 대한 감정의 집합을 통해 상기 대상에 대한 주된 감정, 주된 감정의 변화를 분석하고, 각 대상에 대한 주된 감정 또는 감정들의 서열을 추출하여 제2 데이터 세트를 생성하고, 상기 제1 데이터 세트 및 상기 제2 데이터 세트를 별도의 분석 데이터 저장 수단에 기록하는, 사전 분석 처리 수단;를 포함하고,

제1 사용자 단말기로부터 제1 대상에 대한 분석 데이터 구매 요청을 수신하는 경우, 상기 분석 데이터 구매 요청과 대응되는 분석 대상을 결정하고 하나 이상의 분석 대상과 관련된 데이터 세트들을 검색하고,

상기 분석 대상과 관련된 데이터 세트들의 용량이 기 설정된 임계 용량을 초과하는지 여부를 판단하여, 상기 분석 대상과 관련된 데이터 세트들의 용량이 기 설정된 임계 용량을 초과하는 경우 분석 대상과 관련된 데이터 세트들을 기초로 분석 대상에 대한 감정 데이터 세트를 취합하고, 분석 대상의 유효 감정을 결정하며- 상기 유효 감정은 상기 분석 대상과 관련되어 수집된 감정 데이터 세트에 포함된 감정들 중에서, 출현 빈도수가 기 설정된 임계 횟수 이상의 감정 데이터 세트의 감정임-,

상기 분석 대상과 관련된 감정 데이터 세트 및 유효 감정을 포함하는 결과 구조체를 생성하고, 상기 결과 구조체를 상기 제1 사용자 단말기로 전송하고, 상기 분석 대상과 관련된 데이터 세트의 용량이 기 설정된 임계 용량을 초과하지 않는 경우 상기 제1 대상과 관련된 데이터 세트를 추가적으로 크롤링하는 점을 특징으로 하는, 실시간으로 크롤링 데이터를 분석하는 분석 서버.

청구항 2

제1항에 있어서,

상기 제1 데이터 세트의 대상은

크롤링된 데이터들에 출현되는 키워드들 중에서, 출현 빈도의 비율이 1프로 이상의 키워드와 대응되는 것일 수 있는, 실시간으로 크롤링 데이터를 분석하는 분석 서버.

청구항 3

제1항에 있어서,

상기 제1 데이터 세트의 감정은

기 등록된 감정 단어 및 감정 단어의 표현 등급을 기초로 결정되는 점을 특징으로 하는, 실시간으로 크롤링 데이터를 분석하는 분석 서버.

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

삭제

청구항 8

삭제

청구항 9

삭제

청구항 10

삭제

발명의 설명

기술 분야

[0001] 본 명세서의 실시예는 실시간으로 크롤링 데이터를 분석하는 분석 서버 및 분석 서버의 동작 방법에 관한 것이다.

배경 기술

[0002] 최근에는 시장의 특성이나 구조가 빈번하게 변화하고 기술의 발전 주기가 급속히 단축되고 있으며 제품이 출시되고 사장될 때까지의 기간이 현저하게 감소되는 추세이기 때문에 사업을 하고자 하는 사업가는 지속적으로 새로운 기술을 개발하는 것도 중요하나, 특히 성공 가능성이 큰 분야에 투자 및 기술 개발을 해야 한다.

[0003] 특히 유사한 기술과 유사한 서비스가 많이 존재하는 사업 분야에서는 특히나 기존 서비스를 분석하여 보다 적은 노력과 투자를 통해 최적의 결과를 얻어낼 수 있는 서비스를 제공하는 것이 중요하다.

발명의 내용

해결하려는 과제

[0004] 본 발명의 종래의 문제점을 해결하기 위해서, 통신망을 통해 새롭게 생성되고 공유되는 웹 페이지들을 크롤링하여 데이터를 획득하고, 획득된 데이터를 분석하여 현재 시대를 반영한 트렌드 데이터를 생성할 수 있다.

과제의 해결 수단

[0005] 본 발명의 실시예들에 따른 하나 이상의 프로세서 및 저장 매체를 포함하는 분석 서버에 있어서, 상기 저장 매체는 인터넷 망을 통해 크롤링되는 복수의 웹 페이지들에 포함된 표현에 포함된 키워드를 중심으로, 대상 및 대상에 대한 감정을 구조화하는 하나 이상의 제1 데이터 세트를 개별적으로 생성하고, 상기 제1 데이터 세트를 상기 제1 데이터 세트에 출현되는 하나 이상의 대상을 기준으로 군집화하고, 군집화된 각 대상의 감정 및 각 대상의 감정의 서열을 포함하는 제2 데이터 세트를 생성하는 사전 분석 처리 수단을 포함하는, 실시간으로 크롤링 데이터를 분석할 수 있다.

[0006] 상기 제1 데이터 세트의 대상은 크롤링된 데이터들에 출현되는 키워드들 중에서, 출현 빈도의 비율이 1프로 이상의 키워드와 대응되는 것일 수 있다.

- [0007] 상기 제1 데이터 세트의 감정은 기 등록된 감정 단어 및 감정 단어의 표현 등급을 기초로 결정되는 점을 특징으로 할 수 있다.
- [0008] 상기 제2 데이터 세트는 사용자 단말기로부터의 분석 데이터 구매 요청의 응답하여 제공되며, 상기 제2 데이터 세트의 용량과 비례하는 가격을 가지는 점을 특징으로 할 수 있다.
- [0009] 상기 사전 분석 처리 수단은 장치의 처리 능력에 따라 크롤링되는 범위를 가변적으로 조절하는 점을 특징으로 할 수 있다.
- [0010] 본 발명의 실시예들에 따른 동작 방법에 있어서, 분석 서버가 인터넷 망을 통해 크롤링되는 복수의 웹 페이지들에 포함된 표현에 포함된 키워드를 중심으로, 대상 및 대상에 대한 감정을 구조화하는 하나 이상의 제1 데이터 세트를 개별적으로 생성하는 단계; 및 분석 서버가 상기 제1 데이터 세트를 상기 제1 데이터 세트에 출현되는 하나 이상의 대상을 기준으로 군집화하고, 군집화된 각 대상의 감정 및 각 대상의 감정의 서열을 포함하는 제2 데이터 세트를 생성하는 단계;를 포함할 수 있다.
- [0011] 본 발명의 실시예에 따른 컴퓨터 프로그램은 컴퓨터를 이용하여 본 발명의 실시예에 따른 데이터 분석 방법 중 어느 하나의 방법을 실행시키기 위하여 매체에 저장될 수 있다.
- [0012] 이 외에도, 본 발명을 구현하기 위한 다른 방법, 다른 시스템 및 상기 방법을 실행하기 위한 컴퓨터 프로그램을 기록하는 컴퓨터 판독 가능한 기록 매체가 더 제공된다.
- [0013] 전술한 것 외의 다른 측면, 특징, 이점이 이하의 도면, 특허청구범위 및 발명의 상세한 설명으로부터 명확해 질 것이다.

발명의 효과

- [0014] 본 발명은 통신망을 통해 새롭게 생성되고 공유되는 웹 페이지들을 크롤링하여 데이터를 획득하고, 획득된 데이터를 분석하여 현재 시대를 반영한 가공 데이터를 생성하는 효과가 있다.

도면의 간단한 설명

- [0015] 도 1은 본 발명의 일 측면에 따른 데이터 분석 시스템의 블록도이다.
- 도 2는 본 발명의 일 측면에 따른 데이터 분석 시스템의 다른 블록도이다.
- 도 3은 본 발명의 일 측면에 따른 데이터 분석 시스템의 또 다른 블록도이다.
- 도 4 내지 도 6은 본 발명의 실시예들에 따른 데이터 분석 방법의 흐름도이다.
- 도 7은 본 발명의 여러가지 측면들을 구현하기 위한 예시적인 시스템 환경의 도면이다.
- 도 8은 본 발명과 상호작용할 수 있는 샘플 컴퓨팅 환경의 다른 블록도이다.

발명을 실시하기 위한 구체적인 내용

- [0016] 본 발명은 다양한 변환을 가할 수 있고 여러 가지 실시예를 가질 수 있는바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 본 발명의 효과 및 특징, 그리고 그것들을 달성하는 방법은 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 다양한 형태로 구현될 수 있다.
- [0017] 본 출원에서 사용되는 바와 같이, 용어 "수단"은 컴퓨터 관련 개체, 즉 하드웨어, 하드웨어와 소프트웨어의 조합, 소프트웨어, 또는 실행 중인 소프트웨어를 지칭하기 위한 것이다. 예를 들어, 수단은 프로세서 상에서 실행되는 프로세스, 프로세서, 객체, 실행 파일, 실행 쓰레드, 프로그램, 및/또는 컴퓨터일 수 있지만, 이에 한정되는 것은 아니다. 예로서, 서버 상에서 실행되는 애플리케이션 및 그 서버는 컴퓨터 수단일 수 있다. 하나 이상의 수단이 프로세스 및/또는 실행 쓰레드 내에 존재할 수 있고 수단은 하나의 컴퓨터 상에 로컬화되어 있거나 및/또는 2개 이상의 컴퓨터 사이에 분산되어 있을 수 있다. "쓰레드"는 오퍼레이팅 시스템 커널이 실행을 위해 스케줄링하는 프로세스 내의 개체이다. 당업계에 공지된 바와 같이, 각각의 쓰레드는 쓰레드의 실행과 관련된 휘발성 데이터인 관련 "컨텍스트"를 갖는다. 쓰레드의 컨텍스트는 시스템 레지스터의 내용 및 쓰레드의 프로세스에 속하는 가상 주소를 포함한다. 따라서, 쓰레드의 컨텍스트를 포함하는 실제 데이터는 그 쓰레드가 실행됨에 따라 변한다.

- [0018] 본 발명의 실시예에 따른 데이터 분석 시스템은 서버 및 클라이언트를 이용하여 웹 페이지들(소셜 미디어, 포털, 인터넷 뉴스 페이지, 홈 페이지, 웹/모바일 물 등)을 크롤링할 수 있다. 본 발명의 실시예에 따른 데이터 분석 시스템은 크롤링된 데이터를 지능형 마이닝 알고리즘을 활용하여 감성 사건을 구축할 수 있다. 본 발명의 실시예에 따른 데이터 분석 시스템은 지능형 마이닝 알고리즘을 활용하여 감성 사건을 학습시킬 수 있다. 본 발명의 실시예에 따른 데이터 분석 시스템은 구축되고 학습되는 감성 사건을 이용하여 고객의 반응도 및 호감도를 도출할 수 있다. 도출된 반응도 및 호감도는 다양한 방식으로 시각화될 수 있다.
- [0019] 도 1에서, 본 발명의 일 측면에 따른 데이터 분석 시스템(100)의 블록도가 도시되어 있다.
- [0020] 본 발명의 이 예에서, 데이터 분석 시스템(100)은 1부터 "N"(단, N은 1부터 무한대까지의 임의의 수임)까지의 번호가 부여된 클라이언트(102-106), 통신 시스템(108), 분석 서버(110), 및 웹 페이지 서버(112)로 이루어져 있다. 클라이언트(102-106)는 분석 서버(110)를 위한 웹 페이지 정보의 일군의 "분산 자원"을 구성한다. 클라이언트는 일반적으로 새로운 URL 및 웹 페이지 변동 등을 통신 시스템(108)을 거쳐 분석 서버(110)에 제공하는 기능을 한다. 통신 시스템(108)은 인터넷 및/또는 인트라넷 등으로 이루어져 있다. 통신 시스템은 분석 서버(110)와 클라이언트(102-106) 사이의 통신 목적을 위한 액세스 수단을 제공한다. 통신 시스템은 또한 웹 페이지 정보를 수집하기 위해 클라이언트(102-106)와 다른 웹 페이지 서버(112) 사이 및/또는 분석 서버(110)와 다른 서버들 사이의 통신을 가능하게 해준다. 본질적으로, 웹 크롤러 기능은 분석 서버에서만 동작하지 않고 분석 서버(110) 및 클라이언트(102-106) 사이에 분산되어 있다. 분석 서버(110)는 그 자신의 정보를 재구성하는 것을 용이하게 하기 위해 클라이언트(102-106)를 이용하여 웹 페이지 서버(112)로부터 정보를 획득한다. 이 기능을 분산시킴으로써, 본 발명은 보다 최신이고 안정된 데이터 세트를 제공하며, 분석 서버는 그로부터의 데이터를 이용할 수 있다.
- [0021] 분석 서버(110)는 데이터 크롤링을 통해 데이터를 수집할 수 있다. 분석 서버(110)는 정형/비정형 알고리즘을 이용하여 크롤링된 데이터를 데이터베이스로 구축할 수 있다. 분석 서버(110)는 클라이언트 및 분석 서버 사이의 분산 분석 처리 기술을 이용하여 신속하게 데이터를 크롤링할 수 있다.
- [0022] 도 2로 가면, 본 발명의 일 측면에 따른 데이터 분석 시스템(200)의 다른 블록도가 도시되어 있다. 데이터 분석 시스템(200)은 서로 간에 상호작용하는 통신 수단을 갖는 클라이언트(202) 및 서버(204)로 이루어져 있다. 일반적인 동작 동안, 서버(204)는 웹 페이지를 호스팅하는 다른 서버를 찾기 위해 인터넷 등의 통신 네트워크를 검색하는 웹 크롤러를 호스팅한다. 크롤러는 웹 페이지에 관한 정보의 소스를 편집할 수도 있다. 클라이언트(202)는 서버(204)가 알지 못하는 웹 페이지를 검출할 수 있다. 이것에 의해 클라이언트(202)는 알고 있는 웹 페이지 및 알지 못하는 웹 페이지에 관한 새로운 정보 및/또는 변동/상태를 편집할 수 있게 된다. 이어서, 이 정보는 서버(204)로 전송된다. 서버(204)는 그 정보를 사용하여 그의 원래의 크롤러 웹 페이지 데이터를 재구성할 수 있다. 분산 자원을 가짐으로써, 서버(204)는 그 자신의 직접적인 자원(예를 들어, 프로세서 이용도, 저장 공간 등)에 부담을 주지 않고 그의 크롤러 기능을 확장시킨다. 게다가, 웹 크롤러는 일반적으로 자신이 액세스하는 각각의 서버에 자신의 신원을 밝히기 때문에, 웹 크롤러는 그 서버 상의 잘못된 데이터로 리디렉션될 위험이 있다. 서버는 또한 웹 크롤러가 서버의 자원에 부과할 수 있는 액세스 및 시간량을 제한할 수 있다. 서버에 액세스하는 클라이언트는 일반적으로 이러한 제약을 갖지 않으며 잘못된 데이터로 리디렉션되지 않는다. 따라서, 클라이언트의 웹 페이지 데이터는 웹 크롤러에 의해 편집된 잘못된 데이터를 정정하는 데 이용될 수 있다.
- [0023] 도 3을 참조하면, 본 발명의 일 측면에 따른 데이터 분석 시스템(300)의 또 다른 블록도가 도시되어 있다. 데이터 분석 시스템(300)은 서로 간에 상호 작용하는 통신 시스템(CS)(306)을 갖는 클라이언트(302) 및 서버(304)로 이루어져 있다.
- [0024] 본 발명의 이 예에서, 클라이언트(302)는 사용자 인터페이스 수단(308), 클라이언트 제어 수단(310), 및 저장 수단(312)로 이루어져 있다. 사용자 인터페이스 수단(308)은 일반적으로 이용되는 통신 시스템의 유형에 고유한 인터페이스를 사용자에게 제공한다. 이러한 인터페이스의 일례가 적어도 월드 와이드 웹을 통해 정보를 그래픽적으로 릴레이하는 데 이용되는 웹 브라우저이다. 웹 브라우저는 또한 단일 회사 내에서 서비스되는 웹 페이지 등의 인트라넷을 "서핑"하는 데 이용될 수 있다. 본 발명의 다른 예에서, 유사한 정보가 그래픽 사용자 인터페이스보다는 오히려 텍스트 기반 인터페이스 등을 이용하여 릴레이될 수 있다. 사용자 인터페이스 수단(308)은 통신 시스템(306)으로 정보를 전송하거나 및/또는 그로부터 수신한다. 클라이언트 제어 수단(310)은 웹 크롤링을 용이하게 하는 것과 관련하여 클라이언트의 제어를 제공한다. 클라이언트 제어 수단(310)은 웹 페이지 등과 같은 정보에 관련된 데이터를 수신 및/또는 전송한다. 이 수단(310)은 알고리즘을 처리하고, 데이터 변동 및 상태를 추적하며, 및/또는 데이터 분석 시스템(300) 내의 클라이언트에 대한 로컬 데이터 저장을 제어한다. 이 수

단(310)은 또한 차이, 변경, 변화 등을 결정하기 위해 웹 크롤러로부터 수신된 정보를 사용하여 사용자 인터페이스 수단(308)로부터의 정보를 분석할 수 있다. 클라이언트 제어 수단(310)은 클라이언트가 웹 크롤러 등에 대한 "분산 자원"으로서 참여할 수 있게 해준다. 이 수단(310)은 또한 저장된 데이터에 액세스하여 정보를 사용자 인터페이스 수단(308)에 제공할 수 있다. 본 발명의 일례에서, 사용자 인터페이스 수단(308)은 구현된 크롤러 데이터를 전송 및/또는 수신한다. 따라서, 클라이언트 제어 수단(310)은 이와 같이 구현된 크롤러 관련 데이터를 수신 및/또는 전송하기 위해 사용자 인터페이스 수단(308)과 인터페이스한다. 이와 마찬가지로, 이 수단(310)은 또한 동일한 방식으로 제어를 서버로 전송 및/또는 그로부터 수신할 수 있다. 본 발명의 다른 예에서, 클라이언트 제어 수단(310)은 서버와 유사하게 행동하여 피어-투-피어 방식으로 다른 클라이언트에 제어를 제공할 수 있다. 당업자라면 클라이언트 제어 수단(310) 및 사용자 인터페이스 수단의 기능이 단일 수단 내에 결합될 수 있음을 잘 알 것이다. 또한, 클라이언트는 사용자 인터페이스 수단(308)을 포함하지 않고 분산 자원으로 클라이언트를 이용하는 것도 가능하다. 본 발명의 이러한 경우의 일례는 다른 클라이언트를 릴레이 및/또는 제어하는 클라이언트를 포함할 수 있다. 데이터 저장 수단(312)은 예를 들어 서버로부터의 크롤러 데이터, 클라이언트로부터의 크롤러 데이터, 빈도 분석 데이터, 감성 분석 데이터, 관계 분석 데이터, 군집 분석 데이터, 웹 페이지 변동, 새로운 웹 페이지 데이터, 및 클라이언트 제어 파라미터 등을 저장하는 데 이용된다. 이 수단(312)은 이용되는 본 발명의 예에 따라 클라이언트 제어 수단(310) 및/또는 사용자 인터페이스 수단(308)과 직접 인터페이스(interface)할 수 있다. 데이터 저장 수단(312)은 또한 하드 드라이브, 랜덤 액세스 메모리, 판독 전용 메모리, 분리형 매체, 및 CD-ROM 등과 같은 데이터 저장 장치일 수 있다. 본 발명의 또 다른 예에서, 데이터 저장 수단(312) 상에 저장된 정보는 사용자 인터페이스 수단(308) 또는 클라이언트 제어 수단(310)과 인터페이스하지 않고 서버에 의해 직접 액세스(access)될 수 있다. 어떤 경우, 이것은 보다 빠른 데이터 검색을 가능하게 해준다.

[0025] 본 발명의 일례에서, 통신 시스템(306)은 "인터넷" 등의 인터넷이다. 통신 시스템(306)은 또한 광역 통신망(WAN) 및/또는 근거리 통신망(LAN) 등과 같은 인트라넷 시스템일 수 있다. 통신 시스템(306)은 또한 예를 들어 전화 시스템, 무선 시스템, 광 신호(광학) 시스템, 및 사운드 시스템 등과 같은 보다 전통적인 통신 수단을 이용할 수 있다. 당업자라면 다른 글로벌 및 로컬 네트워크 구조도 본 발명에 의해 통신 시스템(306)으로 이용될 수 있음을 잘 알 것이다.

[0026] 서버(304)는 사전 분석 처리 수단(314), 분산 처리 수단(316), 크롤러 수단(318), 데이터 저장 수단(320)로 이루어져 있다. 본 발명의 일례에서, 크롤러 수단(318)은 서버 및/또는 프록시 서버에 액세스하여 웹 페이지 콘텐츠, 나이, 크기, URL, 및 구현된 링크 등과 같은 웹 페이지 관련 정보를 획득하기 위해 통신 시스템(306)을 이용한다. 획득된 정보는 데이터 저장 수단(320)에 저장된다. 데이터 저장 수단(320)은 하드 드라이브, 랜덤 액세스 메모리, 판독 전용 메모리, 분리형 매체, 및 CD-ROM 등과 같은 데이터 저장 장치일 수 있다. 사전 분석 처리 수단(314)은 크롤러 수단(318)에 의해 찾아내어져 데이터 저장 수단(320)에 저장되는 모든 웹 페이지에 대한 처리 기능을 제공한다.

[0027] 사전 분석 처리 수단(314)은 클라이언트에 의해 또는 자체적으로 크롤링된 데이터를 수집할 수 있다. 사전 분석 처리 수단(314)에 의해 크롤링되는 데이터의 범위는 조절될 수 있다. 사전 분석 처리 수단(314)은 처리 능력을 기초로 크롤링되는 데이터의 범위를 넓혔다가 좁히는 등의 동적으로 변경할 수 있다. 사전 분석 처리 수단(314)은 웹 페이지, 포스팅 등에 포함된 표현을 기초로 제1 데이터 세트를 생성할 수 있다. 사전 분석 처리 수단(314)은 웹 페이지, 포스팅 등에 포함된 표현을 분석하고 상기 표현에 포함된 키워드를 중심으로 제1 데이터 세트를 생성할 수 있다. 여기서, 키워드는 포털, 인터넷 뉴스, SNS, 쇼핑몰, 웹 사이트 등에 출현 빈도수가 급증하는 콘텐츠의 전부 또는 일부 일 수 있다.

[0028] 사전 분석 처리 수단(314)은 웹 페이지에 포함된 표현을 정형/비정형 알고리즘을 통해 제1 데이터 세트를 생성할 수 있다. 사전 분석 처리 수단(314)은 웹 페이지에 포함된 표현 중에서, 의미 없는 텍스트들을 제거하고, 키워드의 축약어(abbreviation), 유사어(synonyms), 단수(singular), 및 복수(plural) 중 선택된 하나 이상에 해당하는 키워드를 중심으로 제1 데이터 세트를 생성할 수 있다.

[0029] 사전 분석 처리 수단(314)은 소셜 네트워크 상에 포스팅의 표현을 기초로 제1 데이터 세트를 생성할 수 있다. 사전 분석 처리 수단(314)은 소셜 네트워크 상에 각 포스팅과 관련된 객체들(팔로워, 팔로잉)간의 관계를 기초로 관계 데이터 세트를 생성할 수 있다. 사전 분석 처리 수단(314)은 포스팅의 표현을 요약한 제1 데이터 세트를 생성할 수 있다.

[0030] 사전 분석 처리 수단(314)은 포스팅의 표현에 포함된 키워드를 중심으로, 객체의 감정, 특정 대상에 대한 감정,

관심을 가지는 대상 등을 구조화된 제1 데이터 세트로 생성할 수 있다. 예를 들어, 사전 분석 처리 수단(314)은 {객체, 대상, 감정}을 하나의 데이터로 생성할 수 있다. 사전 분석 처리 수단(314)은 객체에 의해 입력된 대상을 대표 대상으로 변환하여 하나의 데이터로 생성할 수 있다. 사전 분석 처리 수단(314)은 페이지 또는 포스팅의 표현에 포함된 키워드를 중심으로, 대상, 및 대상에 대한 감정 등을 구조화된 제1 데이터 세트로 생성할 수 있다. 이때, 감정은 감정 사전에 포함된 감정 단어를 기초로 결정될 수 있다. 예를 들어, 페이지에 포함된 감정 단어들을 취합하고, 감정 단어들과 대응되는 감정을 군집화하게 되면, 페이지에 포함된 감정이 결정될 수 있다. 각 감정 단어는 사람의 감정들의 표현 등급으로 구분될 수 있다. 예를 들어, 감정 단어 'ㅎㄷㄷ'은 긍정이 40프로, 부정이 50프로의 표현 등급을 가질 수 있다. 표현 등급은 인간의 감정 종류(긍정, 부정, 선호, 비선호 등) 및 감정 종류의 표현 정도를 포함할 수 있다.

[0031] 페이지 또는 포스팅에 포함된, 키워드와 관련된 감정 단어는 표현 등급을 기초로 주된 감정 표현에 따라 긍정, 부정, 중립 중 적어도 하나로 분류될 수 있다. 감정 단어는 기 등록된 감정 사전에 포함된 키워드 중 하나일 수 있다. 감정 단어는 감정을 표현하는 단어를 말하며, 긍정, 부정, 호감, 비호감, 우울, 슬픔, 기쁨, 행복 등의 감정을 표현하는 단어를 포함할 수 있다. 감정 단어는 사전에 등재된 단어일 수도 있으나, 맞춤법이나 어법에 맞지 않은 텍스트 일 수도 있다. 예를 들어, OTL(좌절 금지), ㅋㅋㅋ(웃음), 놀란 표정의 이모티콘 등과 같이 사전으로 그 의미를 알 수 없는 단어도 감정 단어로 될 수 있다.

[0032] 감정 사전은 크롤링된 데이터를 기초로 갱신될 수 있다. 감정 사전은 크롤링된 데이터를 이용하여 새로운 감정 단어를 추가할 수 있고, 기존에 사용되는 감정 단어의 의미를 변경할 수 있다. 사전 분석 처리 수단(314)은 기존의 감정 단어 외에 새롭게 출현된 감정 단어를 포함하는 페이지들을 모르는 페이지로 수집할 수 있다. 이렇게 수집된 모르는 페이지들에 포함된 동일한 새로운 감정 단어의 출현 빈도가 기 등록된 임계 값 이상인 경우, 새로운 감정 단어를 감정 사전에 추가할 수 있다. 사전 분석 처리 수단(314)은 기존의 감정 단어의 감정의 변화를 감지한 경우, 기존의 감정 단어의 의미를 변경할 수 있다. 사전 분석 처리 수단(314)은 감정 단어와 함께 포함되는 이모티콘, 콘텐츠 등의 감정 등을 이용하여 감정 단어의 의미 변화를 검출할 수 있다. 긍정인 감정 단어가 부정의 이모티콘을 포함하거나, 부정인 감정 단어가 긍정의 이모티콘을 포함하는 데이터 세트의 출현 빈도가 증가할 때, 해당 감정 단어의 의미가 변경된 것으로 기록할 수 있다.

[0033] 사전 분석 처리 수단(314)은 하나의 포스팅 또는 페이지의 표현으로부터 하나 이상의 데이터를 생성할 수 있다. 사전 분석 처리 수단(314)은 데이터 세트를 통해 대상에 대한 감정의 집합을 생성할 수 있다. 이때, 대상은 상기의 키워드와 동일할 수 있다. 사전 분석 처리 수단(314)은 대상에 대한 감정의 집합을 통해 대상에 대한 주된 감정, 주된 감정의 변화 등을 분석할 수 있다. 대상에 대한 감정 등의 주관적인 정보를 획득하기 위해서, 사전 분석 처리 수단(314)은 주제에 대한 의견이나 평가 태도 감정 등과 같은 주관적인 정보를 식별하고 추출하는 알고리즘을 이용할 수 있다.

[0034] 예를 들어, 사전 분석 처리 수단(314)은 포스팅에 포함된 감정 단어 예를 들어, 긍정 단어, 부정 단어, 중립 단어, 호감 단어, 비호감 단어를 기초로 대상에 대한 감정을 결정할 수 있다. 이때, 감정은 포스팅에 대한 긍정(호감) 공감의 수, 부정(비호감) 공감의 수 등을 고려하여 조정될 수 있다. 사전 분석 처리 수단(314)은 포스팅에 포함된 감정 단어의 표현 등급을 고려하여, 포스팅 또는 웹 페이지로부터 추론되는 대상에 대한 감정을 결정할 수 있다.

[0035] 사전 분석 처리 수단(314)은 포스팅에 대한 파급력 지수(공유의 수, 댓글의 수, 좋아요 등과 같은 공감의 수, 조회수 등)를 기초로 대상에 대한 관심도를 결정할 수 있다. 파급력 지수는 공유의 수, 댓글의 수, 공감의 수, 조회수를 가중 평균하여 산출한 값일 수 있다. 각 수에 대한 가중치는 파급력의 정도에 따라 달라질 수 있다. 공유의 수의 가중치는 파급력이 가장 크기 때문에, 댓글의 수 및 공감의 수에 대한 가중치 보다 크다. 댓글의 수의 가중치는 공감의 수의 가중치보다 크다. 공감의 수의 가중치는 조회수의 가중치보다 크다.

[0036] 사전 분석 처리 수단(314)은 일반적인 웹 페이지와 관련된 웹 페이지 데이터를 이용하여 대상 및 감정이 연결된 데이터 세트를 추출할 수 있다. 사전 분석 처리 수단(314)은 웹 페이지의 표현을 기초로 대상 및 감정을 추출할 수 있다. 사전 분석 처리 수단(314)은 온라인 상에서의 콘텐츠, 페이지 등을 통해서 사회 구성원들의 관심 대상, 관심 대상에 대한 감정을 추출할 수 있다. 사전 분석 처리 수단(314)은 웹 페이지의 표현 중에서, 호감 단어 및 비호감 단어의 출현 빈도를 비교하여, 웹 페이지의 주된 대상에 대한 감정을 추출할 수 있다. 이때, 감정은 페이지에 대한 긍정 공감의 수, 부정 공감의 수 등을 고려하여 조정될 수 있다. 사전 분석 처리 수단(314)은 웹 페이지의 조회수, 공유의 수, 공감의 수, 댓글의 수 등을 기초로 대상에 대한 빈도 분석 데이터를 생성한다. 사전 분석 처리 수단(314)은 빈도 분석 데이터를 이용하여 대상에 대한 관심도를 결정할 수 있다. 사전

분석 처리 수단(314)은 대상에 대한 관심도 및 감정을 하나의 데이터에 포함되도록 생성할 수 있다.

- [0037] 사전 분석 처리 수단(314)은 대상에 대한 감정 및 관심도의 변화에 따라 대상과 관련된 데이터를 과거 데이터로 분류하여 저장할 수 있다.
- [0038] 사전 분석 처리 수단(314)은 크롤링을 통해서 획득된 대상들을 목록으로 만들고, 새롭게 대상 목록에 추가된 신규 대상을 생성할 수 있다.
- [0039] 사전 분석 처리 수단(314)은 크롤링을 통해 사이버 상의 웹 페이지, 포스팅 등을 통해 사회적으로 관심이 되는 대상들 및 대상들에 대한 관심도, 호감도를 실시간으로 모니터링하는 기능을 수행한다. 사전 분석 처리 수단(314)은 감정 사전을 자동적으로 갱신함으로써, 크롤링 및 분석이 순환적으로 수행될 수 있다.
- [0040] 사전 분석 처리 수단(314)은 관심이 되는 대상에 대한 주된 감정, 또는 감정들의 서열을 추출하여 제2 데이터 세트를 생성할 수 있다. 사전 분석 처리 수단(314)은 신규 대상이 새롭게 추가되는 경우, 신규 대상과 관련된 데이터를 생성하고, 제2 데이터 세트에 신규 대상과 관련된 데이터를 추가한다.
- [0041] 사전 분석 처리 수단(314)은 제2 데이터 세트에 이미 포함된 제1 대상에 대한 주된 감정 또는 감정의 서열 등이 변경되는 경우, 제1 대상에 대한 주된 감정 또는 감정의 서열 등을 변경할 수 있다.
- [0042] 사전 분석 처리 수단(314)은 제1 데이터 세트 및 제2 데이터 세트를 별도의 분석 데이터 저장 수단(324)로 기록한다.
- [0043] 분산 처리 수단(316)은 데이터 저장 수단(320)에 저장된 정보의 편집을 용이하게 해주며 그 정보가 보다 안정되고 최신이며 보다 포괄적인 것일 수 있게 해준다. 분산 처리 수단(316)은 예를 들어 단일의 분산 크롤러 또는 "클라이언트 기반 웹 크롤러"로서 상호작용하는 서버의 클라이언트들과 같은 분산 자원에 대한 제어를 제공한다. 분산 처리 수단(316)은 예를 들어 클라이언트(302) 등과 같은 분산 자원으로부터 수신된 데이터의 분석과, 함수 및 데이터의 할당 및 이들의 타이밍의 결정 등의 기능을 제공하고, 알고 있는 크롤러 데이터를 결정하고 데이터 갱신 및/또는 추가를 수신하고 데이터 갱신 및/또는 추가를 데이터 저장 수단(320)에 저장하기 위해 분산 자원에 알고리즘을 제공할 수 있다. 분산 처리 수단(316)은 분산 자원에 알고리즘을 이용하여 분산 자원의 최적화된 이용을 결정할 수 있다. 분산 처리 수단(316)은 카운트, 유형, 스푸핑 퍼센티지 및 소스 등과 같은 데이터 특성을 추적할 수 있다. 본 발명의 다른 예에서, 사전 분석 처리 수단(314)은 분산 처리 수단(316)이 직접 통신 시스템(306)에 액세스하기 보다는 분산 처리 수단(316)에 대한 정보를 전송 및/또는 수신한다.
- [0044] 상기와 같이 분석 서버를 통해서, 서비스 제공자들은 사람의 감정을 움직이는 감성 마케팅 전략을 수립할 수 있다.
- [0045] 당업자라면 각각의 수단이 독립적으로 기술되어 있지만 본 발명의 다른 예들에서의 수단은 다른 수단과 관련된 기능을 포함할 수 있음을 잘 알 것이다. 이와 마찬가지로, 어떤 수단들은 본 발명의 범위를 변경하지 않고 생략될 수 있다.
- [0046] 이상에 도시되고 기술된 예시적인 시스템을 살펴보면, 본 발명에 따라 구현될 수 있는 방법들이 도 4 내지 도 6의 플로우차트를 참조하여 더 잘 이해될 것이다. 설명의 간략화를 위해, 이 방법들이 일련의 블록들로 도시되고 기술되어 있지만, 본 발명이 블록들의 순서에 의해 제한되는 것이 아니며 본 발명에 따르면 어떤 블록들은 본 명세서에 도시되고 기술된 것과 다른 순서로 및/또는 다른 블록들과 동시에 일어날 수 있음을 잘 알 것이다. 게다가, 도시된 모든 블록들이 본 발명에 따른 방법들을 구현할 필요가 있는 것은 아니다.
- [0047] 본 발명은 하나 이상의 수단에 의해 실행되는 프로그램 모듈 등의 컴퓨터 실행가능 명령어의 일반적인 관점에서 기술될 수 있다. 일반적으로, 프로그램 모듈은 특정의 태스크를 수행하거나 특정의 추상 데이터 유형을 구현하는 루틴, 프로그램, 객체, 데이터 구조 등을 포함한다. 일반적으로, 프로그램 모듈의 기능은 여러가지 실시예들에서 원하는 바에 따라 결합 또는 분산될 수 있다.
- [0048] 도 4 내지 도 6은 본 발명의 실시예들에 따른 데이터 분석 방법의 흐름도이다.
- [0049] 도 4를 참조하면, S110에서는 서버가 클라이언트로부터 수집된 웹 페이지 관련 정보를 획득한다.
- [0050] S120에서는 서버가 웹 페이지 관련 정보를 데이터 저장 수단에 저장한다.
- [0051] S130에서는 서버가 웹 페이지 또는 포스팅 등에 포함된 표현을 기초로 제1 데이터 세트를 생성한다. 제1 데이터 세트의 생성과 관련하여서는 도 2의 사전 분석 처리 수단에 대한 설명과 중복되므로 생략한다. S140에서는 서버

가 제1 데이터 세트를 객체, 대상, 및 감정의 구조체를 포함하는 제2 데이터 세트로 변환한다.

[0052] 도 5를 참조하면, S210에서는 서버가 제1 사용자 단말기로부터 제1 대상에 대한 분석 데이터 구매 요청을 수신한다. S220에서는 서버가 분석 데이터 구매 요청과 대응되는 분석 대상을 결정하고, 하나 이상의 분석 대상과 관련된 데이터 세트를 검색한다. S230에서는 서버가 분석 대상과 관련된 데이터 세트의 용량이 기 설정된 임계 용량을 초과하는지 여부를 판단한다. S240에서는 판단 결과, 초과하는 경우 분석 대상과 관련된 데이터 세트를 기초로 분석 대상에 대한 감정 데이터 세트를 취합하고, 분석 대상의 유효 감정을 결정한다. 유효 감정은 분석 대상과 관련되어 수집된 감정 데이터 세트에 포함된 감정들 중에서, 출현 빈도수가 기 설정된 임계 횟수 이상의 감정 데이터 세트의 감정을 말한다. S250에서는 서버가 분석 대상과 관련된 감정 데이터 세트 및 유효 감정을 포함하는 결과 구조체를 생성할 수 있다. S260에서는 서버는 결과 구조체를 제1 사용자 단말기로 전송할 수 있다. S245에서는 초과하지 않는 경우, 제1 대상과 관련된 데이터 세트를 추가적으로 크롤링할 수 있다.

[0053] 또 다른 실시예에서, 서버는 수신된 분석 데이터 구매 요청의 응답하여, 제2 데이터 세트를 통해 획득된 특정 대상에 대한 감정 데이터를 포함하는 회신 신호를 생성하여 전송할 수 있다. 서버는 분석 데이터 구매 요청에 포함된 서비스의 난이도, 데이터 소스를 고려하여 분석 데이터 구매 요청과 대응되는 가격을 결정할 수 있다. 분석 데이터 구매 요청의 가격은 난이도가 높아질수록 또는 데이터 소스의 종류가 늘어날수록 증가할 수 있다. 또한, 분석 데이터 구매 요청의 회신 신호의 용량(분석 대상의 수, 최종 분석 데이터 세트의 용량 등)에 따라서 분석 데이터 구매 요청의 가격은 조정될 수 있다. 예를 들어, 데이터 소스의 종류가 1가지(소셜 미디어)인 경우에도 분석 데이터 구매 요청과 대응되는 대상의 수가 증가하는 경우, 분석 데이터 구매 요청의 가격은 올라갈 수 있다.

[0054] 도 6을 참조하면, S310에서는 서버가 웹 페이지 관련 정보를 획득한다.

[0055] S320에서는 서버가 웹 페이지 관련 정보를 데이터 저장 수단에 저장한다. S330에서는 서버가 웹 페이지에 포함된 표현으로부터 긍정, 부정, 중립을 각각 표현하는 감정 단어를 추출한다. 여기서, 감정 단어는 감정 사전을 활용하여 결정된다. S340에서는 서버가 긍정, 부정, 중립을 각각 표현하는 단어를 포함하는 문장 세트를 선택한다. S350에서는 서버가 긍정, 부정, 중립을 각각 표현하는 단어 정보의 수를 카운팅한다. S360에서는 서버가 웹 페이지 관련된 오피니언 정보를 사용자 단말기로 전달한다.

[0056] 서버는 웹 페이지에 포함된 표현들을 분석하는데 있어서 텍스트 마이닝 기술을 활용할 수 있다. 서버는 웹 페이지에 포함된 콘텐츠, 또는 텍스트를 우선 추출한다. 서버는 추출된 콘텐츠 또는 텍스트에 포함되는 단어/구/절에 해당하는 내용을 언어 분석 처리 과정을 통해 가공할 수 있는 데이터로 표현한다. 서버는 전처리된 데이터 중 의미 있는 정보를 선별하여 저장한다. 서버는 복수의 의미 있는 정보들을 단순화시킨다. 즉, 동일한 의미를 가지면서 표현이 일부 상이한 단어들을 동일한 단어, 또는 정보로 분류할 수 있다. 이런 방식으로 서버는 의미 있는 정보를 기반으로 문서를 자동으로 군집화하거나 분류할 수 있다. 서버는 시각화 툴로 군집화되고 분류된 데이터를 표현할 수 있다. 서버는 문서에 포함된 긍정, 부정, 중립을 표현하는 단어를 기초로 군집화할 수 있다. 군집화된 데이터를 요약하는 오피니언 정보를 생성할 수 있다.

[0057] 도 7을 참조하면, 본 발명의 여러가지 측면들을 구현하기 위한 예시적인 시스템 환경(700)은 프로세싱 유닛(704), 시스템 메모리(706), 및 시스템 메모리를 비롯한 여러가지 시스템 수단을 프로세싱 유닛(704)에 연결시키는 시스템 버스(708)를 포함하는 종래의 컴퓨터(702)를 포함한다. 프로세싱 유닛(704)은 임의의 상업적으로 이용가능한 또는 독자적인 프로세서일 수 있다. 게다가, 프로세싱 유닛은 병렬로 연결될 수도 있는 것과 같은 2개 이상의 프로세서로 이루어진 멀티프로세서로서 구현될 수 있다.

[0058] 시스템 버스(708)는 메모리 버스 또는 메모리 컨트롤러, 주변 버스, 및 몇가지 열거하자면 PCI, VESA, 마이크로 채널, ISA, 및 EISA 등의 각종의 종래의 버스 아키텍처 중 임의의 것을 사용하는 로컬 버스를 포함한 여러가지 유형의 버스 구조 중 임의의 것일 수 있다. 시스템 메모리(706)는 판독 전용 메모리(ROM)(710) 및 랜덤 액세스 메모리(RAM)(712)를 포함한다. 시동 중과 같은 때에 컴퓨터(702) 내의 구성요소들 사이에서 정보를 전송하는 것을 도와주는 기본 입출력 시스템(BIOS)(714)은 ROM(710)에 저장되어 있다.

[0059] 컴퓨터(702)는 또한 예를 들어 하드 디스크 드라이브(716), 예를 들어 분리형 디스크(720)로부터 판독하고 그에 기록하기 위한 자기 디스크 드라이브(718), 및 예를 들어 CD-ROM 디스크(724) 또는 기타의 광학 매체로부터 판독하고 그에 기록하기 위한 광학 디스크 드라이브(722)를 포함할 수 있다. 하드 디스크 드라이브(716), 자기 디스크 드라이브(718), 및 광학 디스크 드라이브(722)는 각각 하드 디스크 드라이브 인터페이스(726), 자기 디스크 드라이브 인터페이스(1328), 및 광학 드라이브 인터페이스(730)에 의해 시스템 버스(708)에 연결된다. 드라

이브(716-722) 및 그의 관련 컴퓨터 판독가능 매체는 컴퓨터(702)의 데이터, 데이터 구조, 컴퓨터 실행가능 명령어 등의 비휘발성 저장을 제공한다. 상기의 컴퓨터 판독가능 매체에 대한 설명이 하드 디스크, 분리형 자기 디스크 및 CD에 대한 것이지만, 당업자라면 자기 카세트, 플래쉬 메모리 카드, DVD, 베르누이 카트리지 등과 같은 컴퓨터에 의해 판독가능한 다른 유형의 매체도 역시 예시적인 오퍼레이팅 환경(700)에서 사용될 수 있으며 또한 임의의 이러한 매체가 본 발명의 방법들을 수행하기 위한 컴퓨터 실행가능 명령어들을 포함할 수 있음을 잘 알 것이다.

[0060] 오퍼레이팅 시스템(712), 하나 이상의 애플리케이션 프로그램(734), 기타 프로그램 모듈(736), 및 프로그램 데이터(738)를 비롯한 다수의 프로그램 모듈이 드라이브(716-722) 및 RAM(712)에 저장될 수 있다. 오퍼레이팅 시스템(712)은 임의의 적당한 오퍼레이팅 시스템 또는 오퍼레이팅 시스템들의 결합일 수 있다. 예로서, 애플리케이션 프로그램(734) 및 프로그램 모듈(736)은 본 발명의 일 측면에 따른 클라이언트 기반 웹 크롤링을 용이하게 하는 것을 포함할 수 있다.

[0061] 사용자는 키보드(1340) 및 포인팅 장치(예를 들어, 마우스(742)) 등의 하나 이상의 사용자 입력 장치를 통해 컴퓨터(702)에 명령 및 정보를 입력할 수 있다. 다른 입력 장치(도시 생략)로는 마이크로폰, 조이스틱, 게임패드, 위성 안테나, 무선 리모콘, 스캐너 등이 있을 수 있다. 이들 및 다른 입력 장치는 종종 시스템 버스(708)에 연결되어 있는 직렬 포트 인터페이스(744)를 통해 프로세싱 유닛(704)에 연결되지만, 병렬 포트, 게임 포트 또는 USB(유니버설 직렬 버스) 등의 다른 인터페이스에 의해 연결될 수 있다. 모니터(746) 또는 다른 유형의 디스플레이 장치도 역시 비디오 어댑터(748) 등의 인터페이스를 통해 시스템 버스(708)에 연결된다. 모니터(746) 이외에, 컴퓨터(702)는 스피커, 프린터 등과 같은 다른 주변 출력 장치(도시 생략)를 포함할 수 있다.

[0062] 컴퓨터(702)가 하나 이상의 원격 컴퓨터(760)로의 논리적 연결을 사용하여 네트워킹된 환경에서 동작할 수 있음을 잘 알 것이다. 원격 컴퓨터(760)는 워크스테이션, 서버 컴퓨터, 라우터, 피어 장치 또는 다른 통상의 네트워크 노드일 수 있으며, 비록 간략함을 위해 도 7에는 메모리 저장 장치(762)만이 도시되어 있지만 일반적으로 컴퓨터(702)와 관련하여 기술된 구성요소들의 대부분 또는 그 전부를 포함한다. 도 7에 도시된 논리적 연결은 근거리 통신망(LAN)(764) 및 광역 통신망(WAN)(766)을 포함할 수 있다. 이러한 네트워킹 환경은 사무실, 기업 규모 컴퓨터 네트워크, 인트라넷 및 인터넷에서 통상적인 것이다.

[0063] 예를 들어 LAN 네트워킹 환경에서 사용될 때, 컴퓨터(702)는 네트워크 인터페이스 또는 어댑터(768)를 통해 로컬 네트워크(764)에 연결된다. WAN 네트워킹 환경에서 사용될 때, 컴퓨터(702)는 일반적으로 모뎀(예를 들어, 전화, DSL, 케이블 등)(770)을 포함하거나, LAN을 통해 통신 서버에 연결되거나, 인터넷 등의 WAN(766)을 통해 통신을 설정하기 위한 다른 수단을 갖는다. 컴퓨터(702)에 대해 내장형 또는 외장형일 수 있는 모뎀(770)은 직렬 포트 인터페이스(744)를 통해 시스템 버스(708)에 연결된다. 네트워킹된 환경에서, 프로그램 모듈(애플리케이션 프로그램(734)를 포함함) 및/또는 프로그램 데이터(738)는 원격 메모리 저장 장치(762)에 저장될 수 있다. 도시된 네트워크 연결이 예시적인 것이며 컴퓨터들(702, 760) 사이의 통신 링크를 설정하는 다른 수단(예를 들어, 유선 또는 무선)이 본 발명의 일 측면을 수행할 때 사용될 수 있음을 잘 알 것이다.

[0064] 컴퓨터 프로그래밍 업계의 당업자의 실시예에 따라, 본 발명은 달리 언급하지 않는 한 컴퓨터(702) 또는 원격 컴퓨터(760) 등의 컴퓨터에 의해 수행되는 동작의 심볼 표현 및 작용과 관련하여 기술되어 있다. 이러한 작용 및 동작은 때로는 컴퓨터 실행되는 것으로 말해진다. 작용 및 심볼로 표현된 동작은 프로세싱 유닛(704)에 의해 데이터 비트를 나타내는 전기 신호를 처리하고 그 결과 전기 신호 표현의 변환 또는 감축을 가져오며 데이터 비트를 메모리 시스템(시스템 메모리(706), 하드 드라이브(716), 플로피 디스크(720), CD-ROM(724), 및 원격 메모리(762)를 포함함) 내의 메모리 장소에 유지하고 그에 따라 컴퓨터 시스템의 동작은 물론 신호의 다른 처리를 재구성 또는 달리 변경하는 것을 포함함을 잘 알 것이다. 이러한 데이터 비트가 유지되어 있는 메모리 장소는 그 데이터 비트에 대응하는 특정의 전기적, 자기적 또는 광학적 특성을 갖는 물리적 장소이다.

[0065] 도 8은 본 발명과 상호작용할 수 있는 샘플 컴퓨팅 환경(800)의 다른 블록도이다. 시스템(800)은 또한 하나 이상의 클라이언트(802)를 포함하는 시스템을 예시한 것이다. 클라이언트(802)는 하드웨어 및/또는 소프트웨어(예를 들어, 쓰레드, 프로세스, 컴퓨팅 장치)일 수 있다. 시스템(800)은 또한 하나 이상의 서버(804)를 포함한다. 서버(804)는 또한 하드웨어 및/또는 소프트웨어(예를 들어, 쓰레드, 프로세스, 컴퓨팅 장치)일 수 있다. 서버(804)는 예를 들어 본 발명을 이용하여 변환을 수행하기 위한 쓰레드를 하우징할 수 있다. 클라이언트(802)와 서버(804) 사이의 한가지 가능한 통신은 2개 이상의 컴퓨터 프로세스 사이에서 전송되도록 구성된 데이터 패킷의 형태일 수 있다. 시스템(800)은 클라이언트(802)와 서버(804) 사이의 통신을 용이하게 하기 위해 이용될 수 있는 통신 프레임워크(1408)를 포함한다. 클라이언트(802)는 클라이언트(802)에 로컬인 정보를 저장하는 데 이

용될 수 있는 하나 이상의 클라이언트 데이터 저장부(1410)에 연결되어 동작한다. 이와 마찬가지로, 서버(804)는 서버(804)에 로컬인 정보를 저장하는 데 이용될 수 있는 하나 이상의 서버 데이터 저장부(1406)에 연결되어 동작한다.

- [0066] 본 발명의 일례에서, 데이터 패킷은 웹 크롤링을 용이하게 하는 2개 이상의 컴퓨터 수단 사이에서 전송되며, 그 데이터 패킷은 적어도 부분적으로 웹 크롤링을 위한 분산 시스템을 적어도 부분적으로 이용하는 웹 크롤링에 관련된 정보로 이루어져 있다.
- [0067] 본 발명의 다른 예에서, 웹 크롤링을 용이하게 하기 위한 시스템의 컴퓨터 실행가능 수단을 저장하는 컴퓨터 판독가능 매체는 적어도 부분적으로 웹 크롤링을 위한 분산 시스템에 의해 편집된 웹 페이지에 관한 정보를 적어도 부분적으로 결정하는 웹 크롤링 시스템으로 이루어져 있다.
- [0068] 본 발명의 시스템 및/또는 방법이 컴퓨터 수단 및 비컴퓨터 관련 수단 둘다를 용이하게 하는 웹 크롤링 시스템에서 이용될 수 있음을 잘 알 것이다. 게다가, 당업자라면 본 발명의 시스템 및/또는 방법이 유선 및/또는 무선 등일 수 있는 컴퓨터, 서버 및/또는 핸드헬드 전자 장치 등을 포함하지만 이에 한정되는 것은 아닌 광범위한 전자 관련 기술에서 사용가능함을 잘 알 것이다.
- [0069] 당업자라면 또한 본 발명이 클라이언트 기반 크롤링 시스템에 대한 서버에는 물론 피어-투-피어 크롤링 시스템에 대해서도 이용될 수 있음을 잘 알 것이다. 또한, 클라이언트가 "서버" 동작과 일반적으로 관련된 태스킹을 수행할 수 있으며 따라서 본 발명의 어떤 실시예에서 서버와 관련된 어떤 특징을 클라이언트로 전송하는 것도 가능하다. 본 발명의 일 실시예의 예는 서버로 전송할 정보를 확인 및/또는 검색하기 위해 다른 클라이언트들로의 "서브크롤링(sub-crawl)"을 수행하는 클라이언트이다. 이 예는 예를 들어 어떤 클라이언트들과 서버 사이의 병목 현상을 갖는 네트워크에서 유익할 수 있다. 데이터는 서버에 대한 최상의 액세스 권한을 갖는 클라이언트로 전송될 수 있다. 본 발명의 다른 예들에서, 클라이언트는 인트라넷 시스템에서 서브크롤링을 개시함으로써 서버 행동을 보일 수 있으며, 따라서 인트라넷 상에 존재하는 단지 하나의 클라이언트 및/또는 실질적으로 감소된 수의 클라이언트로부터 서버로 정보를 보고한다. 이와 같이, 검색 서버는 그의 크롤링 자원을 확장하기 위해 클라이언트들 내의 다수의 서브크롤링을 개시할 수 있다.
- [0070] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 어플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.
- [0071] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.
- [0072] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media),

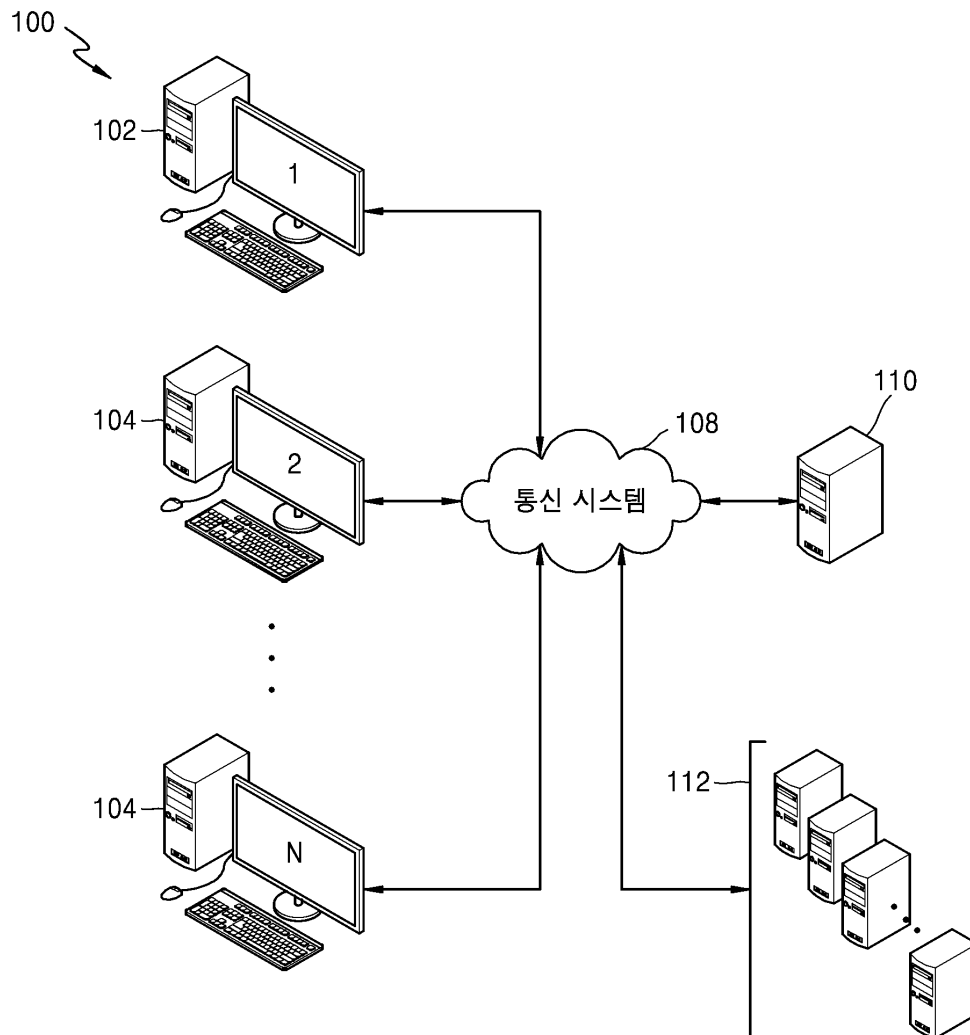
CD-ROM, DVD와 같은 광기록 매체(optical media), 플로피컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0073] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

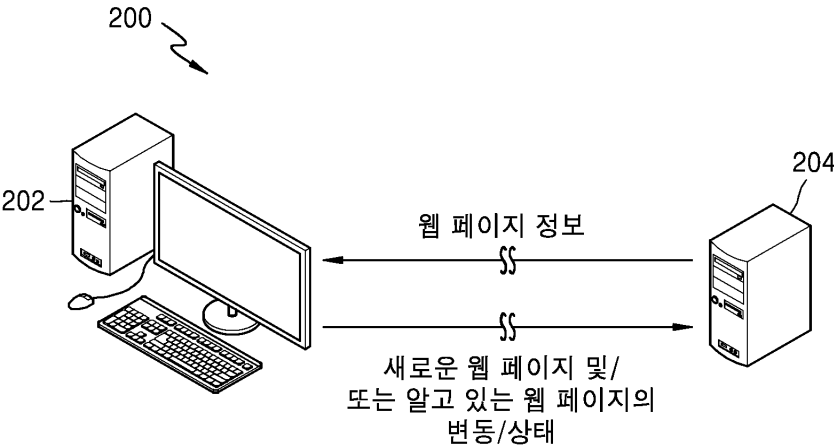
[0074] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

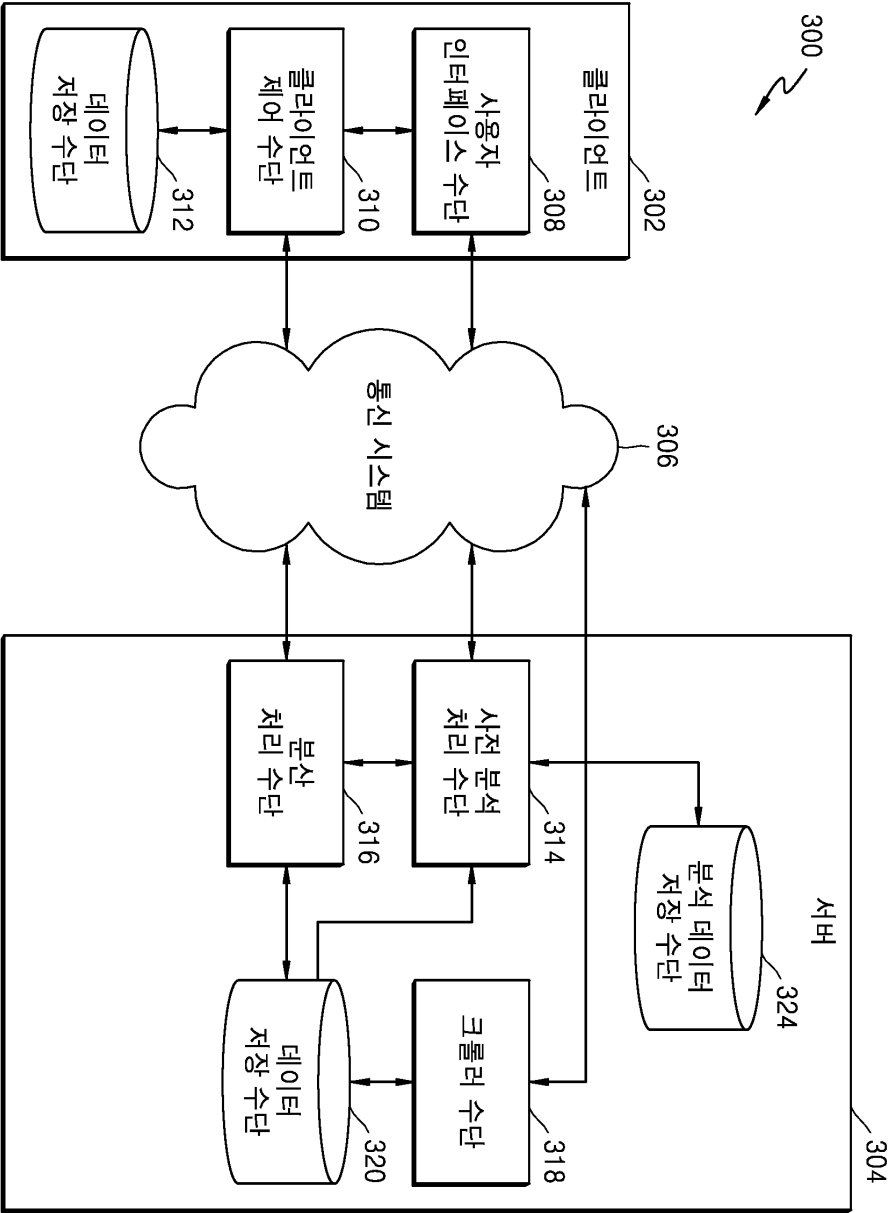
도면1



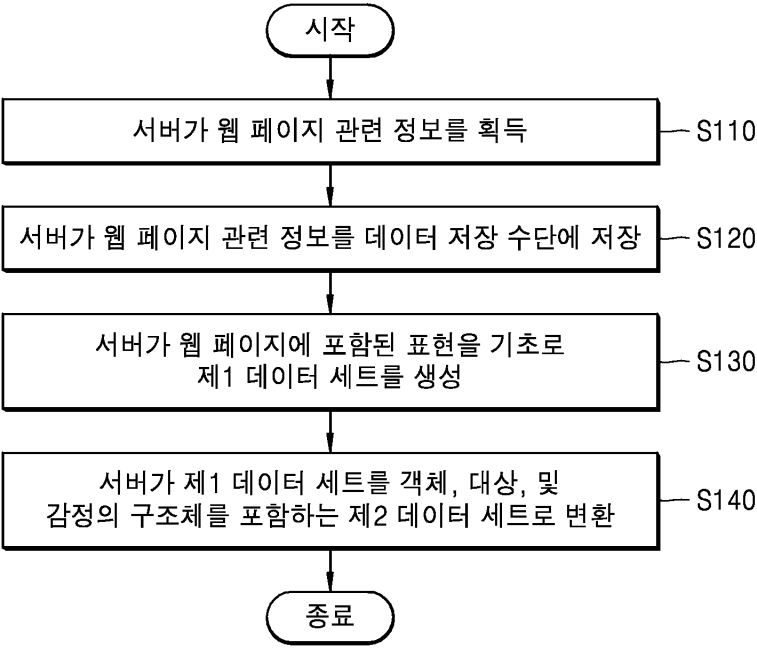
도면2



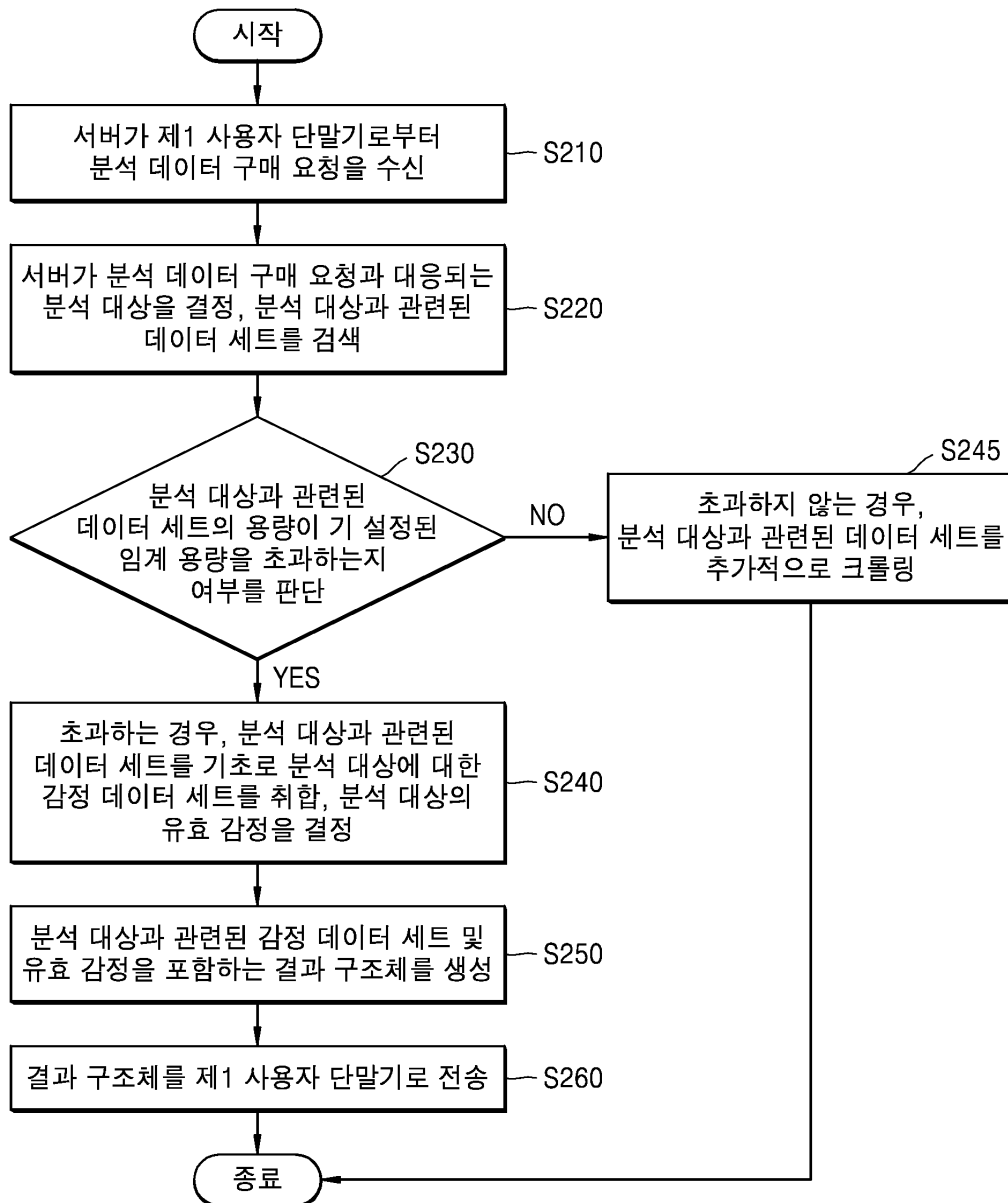
도면3



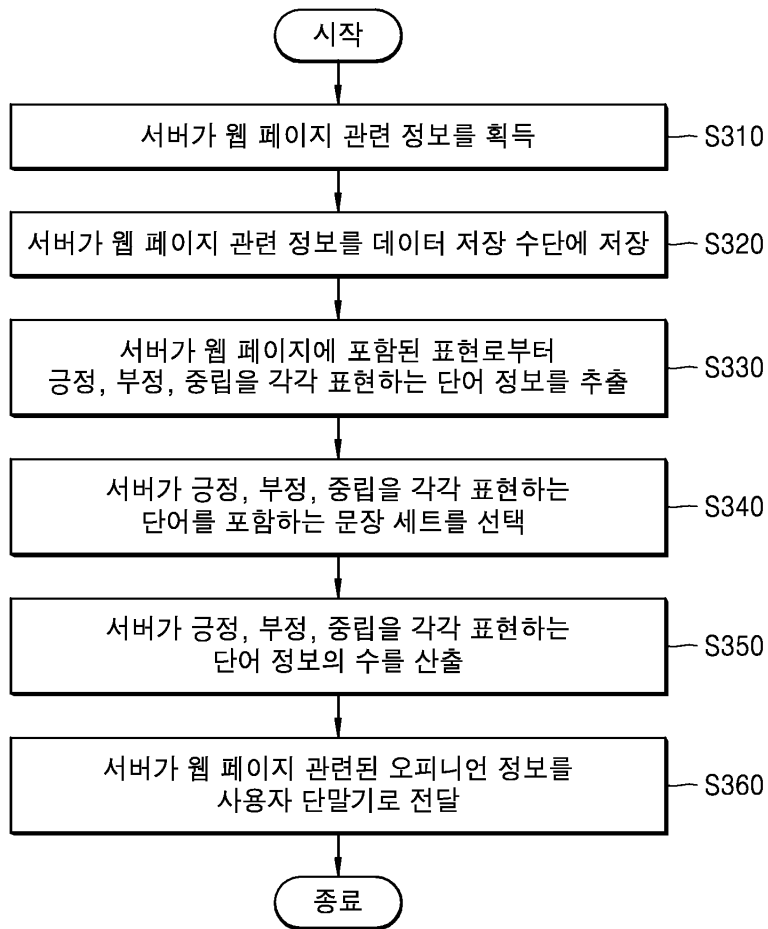
도면4



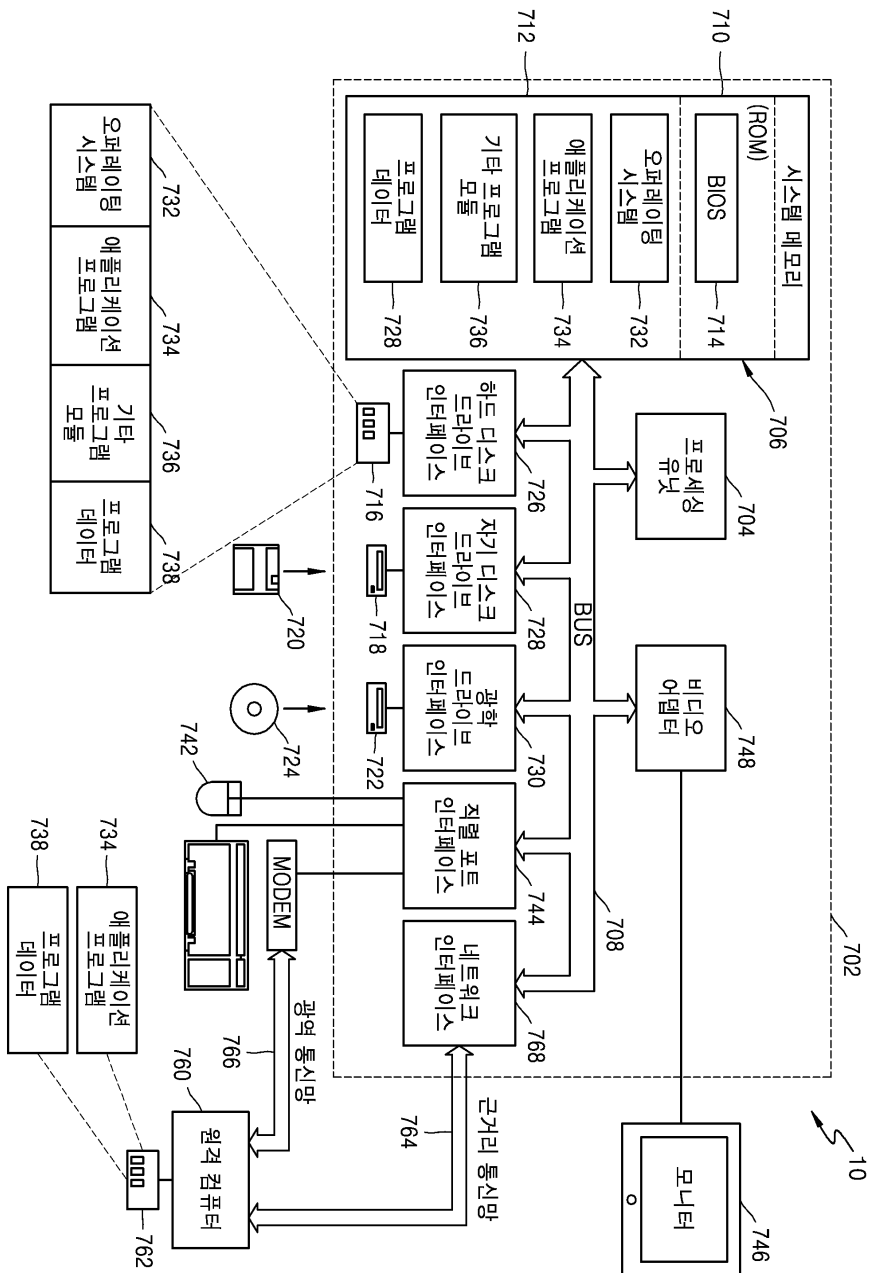
도면5



도면6



도면7



도면8

