



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2010-0094263
(43) 공개일자 2010년08월26일

(51) Int. Cl.

G06F 17/30 (2006.01) G06F 17/00 (2006.01)

G06F 17/21 (2006.01)

(21) 출원번호 10-2009-0013597

(22) 출원일자 2009년02월18일

심사청구일자 2009년02월18일

(71) 출원인

고려대학교 산학협력단

서울 성북구 안암동5가 1

(72) 발명자

이상근

서울 동대문구 제기2동 153 안암골 벽산아파트
101동 203호

알람 히즈블

서울특별시 성북구 안암동5가 1번지 고려대학교
CJ I-HOUSE

하중우

서울특별시 성북구 안암동5가 149-1번지 103호

(74) 대리인

특허법인충현

전체 청구항 수 : 총 5 항

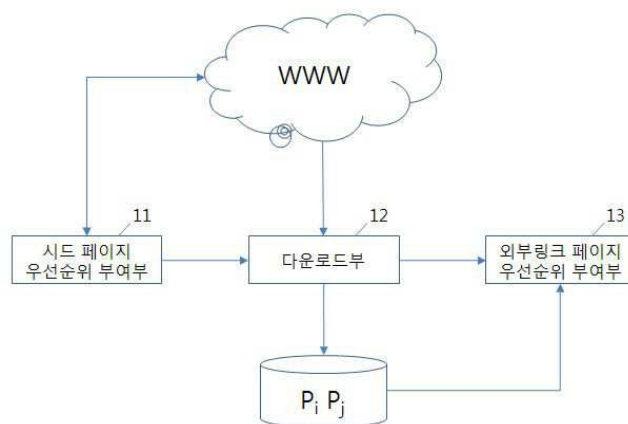
(54) 웹 크롤링 시스템 및 그 방법

(57) 요약

본 발명은 웹 크롤링에 소요되는 시간을 획기적으로 단축시킬 수 있는 웹 크롤링 시스템에 관한 것이다.

본 명세서에서 개시하는 웹 크롤링 시스템은 웹 크롤링을 위한 기준 웹 페이지들(시드 페이지들(seed pages))을 설정하고, 웹 크롤링을 통해 발견되는(Discovered) 상기 시드 페이지들의 각 시드 페이지(p_i)에의 접근 확률(중요도)을 산출하여 상기 각 시드 페이지(p_i)에 우선순위를 부여하는 시드 페이지 우선순위 부여부; 상기 부여된 각 시드 페이지(p_i)의 우선순위 중 가장 높은 순위를 갖는 시드 페이지($p_{i,max}$)를 추출하여 우선적으로 다운로드하되, 상기 시드 페이지($p_{i,max}$)에 링크된 외부링크(outlink) 페이지들도 일괄적으로 다운로드하는 다운로드부; 및 상기 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 상기 시드 페이지($p_{i,max}$)내에서의 접근 확률(중요도)을 산출하여, 상기 각 링크 페이지(p_j)에 우선순위를 부여하는 외부링크 페이지 우선순위 부여부를 포함하여 본 시스템 발명의 과제를 해결한다.

대표도 - 도1a



특허청구의 범위

청구항 1

웹 크롤링을 위한 기준 웹 페이지들(시드 페이지들(seed pages))을 설정하고, 웹 크롤링을 통해 발견되는(Discovered) 상기 시드 페이지들의 각 시드 페이지(p_i)에의 접근 확률(중요도)을 산출하여 상기 각 시드 페이지(p_i)에 우선순위를 부여하는 시드 페이지 우선순위 부여부;

상기 부여된 각 시드 페이지(p_i)의 우선순위 중 가장 높은 순위를 갖는 시드 페이지($p_{i,max}$)를 추출하여 우선적으로 다운로드하되, 상기 시드 페이지($p_{i,max}$)에 링크된 외부링크(out link) 페이지들도 일괄적으로 다운로드하는 다운로드부; 및

상기 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 상기 시드 페이지($p_{i,max}$)내에서의 접근 확률(중요도)을 산출하여, 상기 각 링크 페이지(p_j)에 우선순위를 부여하는 외부링크 페이지 우선순위 부여부를 포함하는 웹 크롤링 시스템.

청구항 2

웹 크롤링을 통해 이미 다운로드 되어있는 웹 페이지들의 각 웹 페이지(p_i)에의 접근 확률(중요도)을 산출하여 상기 각 웹 페이지(p_i)에 우선순위를 부여하는 페이지 우선순위 부여부;

상기 각 웹 페이지(p_i)에 부여된 우선순위 중 가장 높은 순위를 갖는 페이지($p_{i,max}$)를 추출하고, 상기 추출된 웹 페이지($p_{i,max}$)에 링크되어 있으나 탐색되지 아니한(unexplored) 외부링크 페이지들을 탐색하여 다운로드하는 다운로드부; 및

상기 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 상기 페이지($p_{i,max}$)내에서의 접근 확률(중요도)을 산출하고, 이 크기 순서에 따라 상기 각 링크 페이지(p_j)에 우선순위를 부여하는 외부링크 페이지 우선순위 부여부를 포함하는 웹 크롤링 시스템.

청구항 3

(a) 웹 크롤링을 위한 기준 웹 페이지들(시드 페이지들(seed pages))을 설정하고, 웹 크롤링을 통해 발견되는(Discovered) 상기 시드 페이지들의 각 시드 페이지(p_i)에의 접근 확률(중요도)을 산출하여 상기 각 시드 페이지(p_i)에 우선순위를 부여하는 단계;

(b) 상기 부여된 각 시드 페이지(p_i)의 우선순위 중 가장 높은 순위를 갖는 시드 페이지($p_{i,max}$)를 추출하여 우선적으로 다운로드하되, 상기 시드 페이지($p_{i,max}$)에 링크된 외부링크(out link) 페이지들도 동시에 일괄적으로 다운로드하는 단계; 및

(c) 상기 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 상기 시드 페이지($p_{i,max}$)내에서의 접근 확률(중요도)을 산출하여, 상기 각 링크 페이지(p_j)에 우선순위를 부여하는 단계를 포함하는 웹 크롤링 방법.

청구항 4

(a) 웹 크롤링을 통해 이미 다운로드 되어있는 웹 페이지들의 각 웹 페이지(p_i)에의 접근 확률(중요도)을 산출하여 상기 각 웹 페이지(p_i)에 우선순위를 부여하는 단계;

(b) 상기 각 웹 페이지(p_i)에 부여된 우선순위 중 가장 높은 순위를 갖는 페이지($p_{i,max}$)를 추출하고, 상기 추출된 웹 페이지($p_{i,max}$)에 링크되어 있으나 탐색되지 아니한(unexplored) 외부링크 페이지들을 탐색하여 다운로드하는 단계; 및

(c) 상기 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 상기 페이지($p_{i,max}$)내에서의 접근 확률(중

요도)을 산출하고, 이 크기 순서에 따라 상기 각 링크 페이지(p_j)에 우선순위를 부여하는 단계를 포함하는 웹 크롤링 방법.

청구항 5

제 3 항 내지 제 4 항의 어느 한 항의 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록 매체.

명세서

발명의 상세한 설명

기술 분야

[0001] 본 발명은 정보 검색(information retrieval) 방법에 관한 것으로, 보다 상세하게는 정보 검색 엔진에서 웹 페이지(웹 문서, 이하 웹 페이지의 고유 주소를 의미하는 Uniform Resource Locator(URL)와 동일한 의미를 나타내는 것으로 한다)들에 대한 크롤링(crawling)에 관한 것이며, 특히 웹 크롤링에 소요되는 시간을 획기적으로 단축시킬 수 있는 웹 크롤링 시스템 및 그 방법에 관한 것이다.

배경 기술

[0002] 웹 크롤링(web crawling)은 정보 검색 엔진에 흔히 쓰이는 것으로, 웹 크롤러(web crawler)가 새로운 웹 페이지 또는 갱신된 웹 페이지를 찾아 정보 검색 엔진에 다운로드하여 지정하는 것을 의미하는데, 웹 크롤링은 웹 페이지들(URLs)의 각 웹 페이지의 접근 확률(중요도)에 따라 각 웹 페이지에 우선순위를 부여하고(prioritizing), 보다 높은 접근 우선순위를 갖는 웹 페이지를 정보 검색 엔진에 우선적으로 다운로드하여 지정시켜 구현된다.

[0003] 웹 페이지들에 우선순위를 부여하는 것에 관해서는 많은 연구가 이루어졌고, 몇 가지 기법들은 이미 사용되고 있다. 우선순위를 부여하는 기존의 기법들은 가급적 많은 웹 페이지들에 우선순위를 부여하는 것에만 집중했지, 그 부여에 소요되는 시간 측면은 전혀 고려하지 아니하여 웹 페이지들에 우선순위를 부여함에 있어 지나치게 많은 시간이 소요되어 결국 웹 크롤링에 소요되는 시간이 지나치게 증가하는 문제가 있는데, 웹 페이지들이 폭발적으로 증가하고 있고 아울러 웹 페이지가 제공하는 정보 변화의 속도가 매우 빨라지는 요즈음의 추세를 감안하면 소요 시간 측면을 무시한 웹 크롤링의 의의는 반감될 수밖에 없다.

[0004] 기존의 우선순위 부여 기법 중에서 대표적으로 사용되고 있는 기법인 RankMass 기법에 의하면 8천만 개의 웹 페이지(1억 4천만 페이지를 외부링크(outlink)로 갖는)에 각 우선순위를 부여하여 정보 검색 엔진에 다운로드하여 지정하는 작업에 무려 222 시간이 소요된다. 이는 URL에 우선순위를 부여함에 있어 반복적인 작업(이미 우선순위가 부여된 페이지에 대해 소정의 상황에 따라 또 다시 우선순위를 부여)을 하기 때문에 수많은 우선순위 부여에 있어서의 오버헤드(overhead)가 발생하기 때문이다.

발명의 내용

해결 하고자하는 과제

[0005] 본 발명은 상기한 요즈음의 추세에 부응하고 기존의 웹 크롤링 기법들의 시간 과다 소요 문제를 해결하기 위해 창안된 것으로, 본 발명이 해결하고자 하는 과제는 웹 크롤링에 소요되는 시간을 획기적으로 단축시킬 수 있는 웹 크롤링 시스템 및 그 방법을 제공하는 것이다.

과제 해결수단

[0006] 상기와 같은 과제를 해결하기 위해 본 명세서에서 개시하는 웹 크롤링 시스템은 웹 크롤링을 위한 기준 웹 페이지들(시드 페이지들(seed pages))을 설정하고, 웹 크롤링을 통해 발견되는(Discovered) 상기 시드 페이지들의 각 시드 페이지(p_i)에의 접근 확률(중요도)을 산출하여 상기 각 시드 페이지(p_i)에 우선순위를 부여하는 시드 페이지 우선순위 부여부; 상기 부여된 각 시드 페이지(p_i)의 우선순위 중 가장 높은 순위를 갖는 시드 페이지($p_{i,max}$)를 추출하여 우선적으로 다운로드하되, 상기 시드 페이지($p_{i,max}$)에 링크된 외부링크(outlink) 페이지들도 일괄적으로 다운로드하는 다운로드부; 및 상기 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 상기

시드 페이지($p_{i,max}$)내에서의 접근 확률(중요도)을 산출하여, 상기 각 링크 페이지(p_j)에 우선순위를 부여하는 외부링크 페이지 우선순위 부여부를 포함하여 본 시스템 발명의 과제를 해결한다.

[0007] 상기와 같은 과제를 해결하기 위해 본 명세서에서 개시하는 또 다른 웹 크롤링 시스템은 이미 다운로드 되어있는 웹 페이지들의 각 웹 페이지(p_i)에의 접근 확률(중요도)을 산출하여 상기 각 웹 페이지(p_i)에 우선순위를 부여하는 페이지 우선순위 부여부; 상기 각 웹 페이지(p_i)에 부여된 우선순위 중 가장 높은 순위를 갖는 페이지($p_{i,max}$)를 추출하고, 상기 추출된 웹 페이지($p_{i,max}$)에 링크되어 있으나 탐색되지 아니한(unexplored) 외부링크 페이지들을 탐색하여 다운로드하는 다운로드부; 및 상기 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 상기 페이지($p_{i,max}$)내에서의 접근 확률(중요도)을 산출하고, 이 크기 순서에 따라 상기 각 링크 페이지(p_j)에 우선순위를 부여하는 외부링크 페이지 우선순위 부여부를 포함하여 본 시스템 발명의 과제를 해결한다.

[0008] 상기와 같은 과제를 해결하기 위해 본 명세서에서 개시하는 웹 크롤링 방법은 (a)웹 크롤링을 위한 기준 웹 페이지들(시드 페이지들(seed pages))을 설정하고, 웹 크롤링을 통해 발견되는(Discovered) 상기 시드 페이지들의 각 시드 페이지(p_i)에의 접근 확률(중요도)을 산출하여 상기 각 시드 페이지(p_i)에 우선순위를 부여하는 단계; (b)상기 부여된 각 시드 페이지(p_i)의 우선순위 중 가장 높은 순위를 갖는 시드 페이지($p_{i,max}$)를 추출하여 우선적으로 다운로드하되, 상기 시드 페이지($p_{i,max}$)에 링크된 외부링크(out link) 페이지들도 동시에 일괄적으로 다운로드하는 단계; 및 (c)상기 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 상기 시드 페이지($p_{i,max}$)내에서의 접근 확률(중요도)을 산출하여, 상기 각 링크 페이지(p_j)에 우선순위를 부여하는 단계를 포함하여 본 방법 발명의 과제를 해결한다.

[0009] 상기와 같은 과제를 해결하기 위해 본 명세서에서 개시하는 또 다른 웹 크롤링 방법은 (a)이미 다운로드 되어있는 웹 페이지들의 각 웹 페이지(p_i)에의 접근 확률(중요도)을 산출하여 상기 각 웹 페이지(p_i)에 우선순위를 부여하는 단계; (b)상기 각 웹 페이지(p_i)에 부여된 우선순위 중 가장 높은 순위를 갖는 페이지($p_{i,max}$)를 추출하고, 상기 추출된 웹 페이지($p_{i,max}$)에 링크되어 있으나 탐색되지 아니한(unexplored) 외부링크 페이지들을 탐색하여 다운로드하는 단계; 및 (c)상기 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 상기 페이지($p_{i,max}$)내에서의 접근 확률(중요도)을 산출하고, 이 크기 순서에 따라 상기 각 링크 페이지(p_j)에 우선순위를 부여하는 단계를 포함하여 본 방법 발명의 과제를 해결한다.

효 과

[0010] 본 발명에 의하면, 웹 크롤링 과정에서 발견되는 웹 페이지 중 접근 확률(중요도)이 가장 높은 웹 페이지를 우선적으로 다운로드하되, 이러한 웹 페이지를 다운로드 시에 웹 페이지에 링크된 외부링크(out link) 페이지들도 동시에 일괄적으로 다운로드하므로 웹 크롤링에 소요되는 시간을 획기적으로 줄일 수 있다.

[0011] 본 발명의 또 다른 태양에 의하면, 정보 검색 엔진에 이미 다운로드 되어 지정되어 있는 웹 페이지들의 접근 확률(중요도)을 구하여 각 웹 페이지에 우선순위를 부여하고, 이 우선순위에 근거하여 아직 탐색되지 않은 새로운 웹 페이지 또는 갱신된 웹 페이지를 탐색하므로(크롤링하므로) 웹 크롤링에 소요되는 시간을 획기적으로 줄일 수 있다.

발명의 실시를 위한 구체적인 내용

[0012] 이하, 본 발명이 해결하고자 하는 과제의 해결 방안을 명확하게 하기 위한 발명의 구성을 본 발명의 바람직한 실시예에 근거하여 첨부 도면을 참조하여 상세히 설명하되, 도면의 구성요소들에 참조번호를 부여함에 있어서 동일 구성요소에 대해서는 비록 다른 도면상에 있더라도 동일 참조번호를 부여하였으며 당해 도면에 대한 설명 시 필요한 경우 다른 도면의 구성요소를 인용할 수 있음을 미리 밝혀둔다. 아울러 본 발명과 관련된 공지 기능 혹은 구성에 대한 구체적인 설명 그리고 그 이외의 제반 사항이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우, 그 상세한 설명을 생략한다.

[0013] 상기와 같은 과제를 해결하기 위해 본 발명은 웹 페이지의 중요도를 평가하여 우선순위를 부여하기 위해 상기한 PageRank를 변형시켜 적용한다.

- [0014] 상기한 PageRank에 의하면, 웹 페이지에 도달하는 모든 에지 또는 경로(edge or path)의 확률이 계산되며 그 웹 페이지에 도달하는 모든 에지 또는 경로를 선택할 확률의 합이 바로 그 웹 페이지의 중요도(접근 확률)가 된다. 상기 모든 경로의 확률의 합을 계산하는 것은 PageRank lower bound로서 알려져 있는데, 이에 의하면 웹 페이지에 도달하는 새로운 경로가 발견되면 그 웹 페이지에 도달하는 모든 경로들의 접근 확률(중요도)을 다시 계산하게 된다. 그런데 새로운 경로는 수없이 많이 발견될 수 있으며, 따라서 PageRank lower bound를 이용하여 URL에 우선순위를 부여하는 것은 매우 많은 시간이 소요된다.
- [0015] 본 발명에서는 이러한 시간 과다 소요 문제를 해결하기 위해, 웹 페이지에 우선순위를 부여함에 있어, 각 웹 페이지의 접근 확률(중요도)을 산출하되 기존과 달리 접근 확률(중요도)이 산출된 웹 페이지에 대해서는 그 웹 페이지가 정보 검색 엔진에 다운로드되어 지정되는 작업이 다시 진행되지 않는 한 1회만 실시된다.
- [0016] 본 발명은 이렇게 산출되는 각 웹 페이지의 접근 확률(중요도) 중 가장 높은 접근 확률(중요도)을 가지는 웹 페이지를 최초로 다운로드하되 이러한 웹 페이지에 링크된 페이지(외부링크 페이지)도 정보 검색 엔진에 일괄적으로 다운로드하여 추가하거나, 정보 검색 엔진에 이미 다운로드되어 지정된 웹 페이지들 중 가장 높은 접근 확률(중요도)을 가지는 웹 페이지에 링크된 새로운(또는 갱신된) 웹 페이지를 우선적으로 탐색하여 정보 검색 엔진에 다운로드하여 지정한다.
- [0017] 본 시스템 발명의 제1 구성은 도 1a에 제시되어 있으며, 본 방법 발명의 제1 흐름은 도 1b에 제시되어 있다.
- [0018] 본 발명의 제1 구현예(제1 구성 및 제1 흐름)의 근간 개념은 웹 공간상의 다수의 웹 페이지 중 웹 크롤링을 위한 n개의 기준 웹 페이지들(신뢰하는 시드 페이지들(trusted seed pages))을 설정하고, 웹 크롤링을 통해 발견되는(Discovered) 이러한 시드 페이지들 중 접근 확률(중요도)이 가장 높은 시드 페이지를 우선적으로 다운로드하되, 이러한 시드 페이지에 링크된 외부링크(outlink) 페이지들도 아울러 다운로드하여 크롤링에 소요되는 시간을 감소시키는 것이다.
- [0019] 이를 위해 본 발명의 제1 구현예는 웹 크롤링을 통해 발견되는 시드 페이지들의 각 시드 페이지(p_i , i 는 1에서 n 까지의 자연수)의 접근 확률(중요도) D_i 를 산출하는 것으로 시작된다. 시드 페이지 우선순위 부여부(11)는 각 시드 페이지(p_i)의 접근 확률(중요도) D_i 를 아래의 식 1에 의해 산출 후, 산출된 D_i 의 크기 순서에 따라 각 시드 페이지(p_i)에 우선순위를 부여한다(s11).
- [0020] $D_i = (1-d)*t_i$ --- (식 1).
- [0021] 여기서 d 는 상기 PageRank에서 정의된 것으로, '랜덤 서퍼(random surfer)'가 현재 방문 중인 웹 페이지의 링크를 따라 다음 웹 페이지들을 연쇄적으로 서핑(surfing)할 확률을 의미한다. 특정 웹 페이지의 페이지랭크 값은 랜덤 서퍼가 그 웹 페이지를 방문할 확률로 정의할 수 있다. 따라서 랜덤 서퍼가 방문할 확률이 높을수록 중요한 웹 페이지라고 판단하는 것이다. 한편 상기 랜덤 서퍼는 사용자의 웹 서핑 패턴을 모델링한 것으로, 랜덤 서퍼가 현재 방문 중인 웹 페이지에서 다음으로 방문할 웹 페이지로 이동하는 경우의 수는 아래와 같이 2가지이다.
- [0022] 1) 사용자가 현재 방문 중인 웹 페이지 내의 링크를 따라간다. 이는 사용자가 페이지 현재 방문 중인 웹 페이지 내의 링크를 클릭하여 다음 페이지를 방문하는 것과 같다.
- [0023] 2) 사용자가 신뢰하는 웹 페이지들(상기한 시드 페이지들) 중 하나를 방문한다. 이는 사용자가 자신의 웹 브라우저에 등록해놓은 '즐겨찾기' 웹 페이지들 중 하나를 선택하여 방문하는 것과 같다.
- [0024] d 는 랜덤 서퍼가 1)의 경우를 선택하여 웹 서핑을 계속해 나갈 확률이며, 반대로, 랜덤 서퍼가 2)의 경우를 선택하여 웹 서핑을 계속해나갈 확률은 $(1-d)$ 가 된다. 랜덤 서퍼는 1)의 경우와 2)의 경우 중 반드시 하나를 선택하여 웹 서핑을 계속하기 때문에, 1)의 경우를 선택할 확률(d)과 2)의 경우를 선택할 확률($1-d$)를 합하면 1이 되는 것이다.
- [0025] t_i 는 사용자가 신뢰하는 웹 페이지(trusted seed page)들에 대하여, 그 신뢰 정도를 확률로 표현한 것이다. 예를 들어서, 사용자 자신이 신뢰하는 웹 페이지는 네이버(t_1)와 다음(t_2) 두 개의 웹 페이지뿐이며, 네이버를 9번 방문할 때 다음은 1번 정도 방문한다고 가정하면, 네이버의 t_i 값은 0.9, 다음의 t_i 값은 0.1이 된다(모든 trusted seed page 들의 t_i 값을 더하면 1이 된다). 이는 네이버를 다음보다 9배 더 신뢰한다고 생각할 수 있다.

- [0026] 여기서 시드 페이지는 일례로 사용자들로부터 접속 요청이 많은 웹 페이지이거나 외부링크 페이지를 많이 가진 웹 페이지들이 설정될 수 있는데, 예를 들면 네이버, 구글, 야후 등 포털 정보 검색 사이트들을 들 수 있다. 그러나 이러한 웹 페이지들을 반드시 시드 페이지로 정한다는 규칙은 없으며 어떤 웹 페이지를 시드 페이지로 설정하느냐는 웹 크롤러 구현 정책에 의한다. 다만 본 발명의 목적인 웹 크롤링 시간의 획기적 감소의 실현을 위해서는 이러한 페이지들을 시드 페이지로 설정하는 것이 바람직하다.
- [0027] 다운로드부(12)는 상기 부여된 각 시드 페이지(p_i)의 우선순위 중 가장 높은 순위를 갖는 시드 페이지($p_{i,max}$)를 추출하여 정보 검색 엔진에 우선적으로 다운로드하여 지정한다(s12). 이때 가장 높은 우선순위를 갖는 시드 페이지($p_{i,max}$)를 다운로드 시에 이러한 시드 페이지($p_{i,max}$)에 링크된 외부링크 페이지들도 일괄적으로 다운로드하게 된다. 이는 이러한 시드 페이지에 링크된 외부링크 페이지들은 상기 가장 높은 우선순위를 갖는 시드 페이지($p_{i,max}$)를 통해 접근될 확률이 높다고 볼 수 있기 때문에, 이러한 외부링크 페이지들을 가장 높은 우선순위를 갖는 시드 페이지($p_{i,max}$)를 다운로드시에 일괄적으로 다운로드함으로써 웹 크롤링에 소요되는 시간을 획기적으로 줄일 수 있게 한다.
- [0028] 외부링크 우선순위 부여부(13)는 상기 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 접근 확률(중요도) D_j 를 산출하고, D_j 의 크기 순서에 따라 각 링크 페이지(p_j)에 우선순위를 부여한다. D_j 의 산출은 아래의 식 2에 의해 산출한다(s13).
- [0029] $D_j = D_j + (d \cdot D_i) / o_i$ --- (식 2).
- [0030] 여기서 d 는 (식 1)에서의 d 와 동일하며, o_i 는 특정 웹 페이지 D_i 가 가지고 있는 아웃 링크(outlink)의 개수를 의미한다.
- [0031] (식 2)는 특정 웹 페이지 D_i 가 D_j 를 링크하고 있을 때, 랜덤 서퍼가 D_j 를 방문할 확률을 업데이트한다는 의미이다. D_i 로부터 D_j 로의 링크는 웹 크롤링 과정에서 새롭게 발견되었기 때문에, D_j 에 방문될(접근될) 확률은 높아져야 한다. 따라서 이전 D_j 값((식 2)의 우변의 D_j 값)에 추가된 확률값($(d \cdot D_i) / o_i$)을 더하여 좌변의 D_j 값 즉, p_j 에 대한 접근 확률을 산출하게 된다.
- [0032] 추가된 확률값($(d \cdot D_i) / o_i$)은 1) 랜덤 서퍼가 D_i 를 방문할 확률에 2) 랜덤 서퍼가 다음 방문할 웹 페이지로 D_j 를 선택할 확률을 곱하여 얻을 수 있고, 3) 2)의 확률은 랜덤 서퍼가 D_i 에서 다음 웹 페이지를 선택할 때, D_j 의 아웃링크 중 하나를 따라갈 확률인 d 에, D_i 내에 존재하는 총 o_i 개의 아웃링크 중 D_i 를 선택할 확률인 $1/o_i$ 를 곱함으로써 계산할 수 있다.
- [0033] 부여되는 각 링크 페이지(p_j)의 우선순위는 웹 공간 전체에서의 우선순위가 아니라, 가장 높은 우선순위를 갖는 시드 페이지($p_{i,max}$)내에서의 우선순위이다. 다시 말하면, 본 발명의 제1 구현예에 의한 각 링크 페이지(p_j)에의 우선순위 부여는 기존과는 달리 웹 공간 전체를 기준으로 하는 것이 아니라, 가장 높은 우선순위를 갖는 시드 페이지를 기준으로 이루어진다. 따라서 우선순위 부여에 소요되는 시간(웹 크롤링에 소요되는 시간)이 기존의 기법에 비해 현저히 감축된다.
- [0034] 외부링크 우선순위 부여부(13)는 상기 각 링크 페이지(p_j)에 우선순위의 부여를 끝내면 상기 $p_{i,max}$ 의 다음 우선순위를 갖는 시드 페이지에 대해 상기한 과정들이 이루어지게 한다.
- [0035] 본 시스템 발명의 제2 구성은 도 2a에 제시되어 있으며, 본 방법 발명의 제2 흐름은 도 2b에 제시되어 있다.
- [0036] 본 발명의 제2 구현예(제2 구성 및 제2 흐름)는 정보 검색 엔진에 이미 다운로드 되어 지정되어 있는 웹 페이지들의 접근 확률(중요도)을 구하여 각 웹 페이지에 우선순위를 부여하고, 이 우선순위를 이미 다운로드 되어 지정되어 있는 웹 페이지들에 링크되어 있으나 아직 탐색되지 않은(unexplored, 아직 미발견된) 새로운 웹 페이지 또는 갱신된 웹 페이지의 탐색(크롤링)에 활용하는 것이다.
- [0037] 이를 위해 본 발명의 제2 구현예는 상기 다운로드 되어 있는 n 개의 웹 페이지들의 각 웹 페이지(p_i , i 는 1에서 n 까지의 자연수)의 접근 확률(중요도) E_i 를 산출하는 것으로 시작된다. 페이지 우선순위 부여부(21)는 상기 각 페이지(p_i)의 접근 확률(중요도) E_i 를 아래의 식 3에 의해 산출 후, 산출된 E_i 의 크기의 순에 따라 각 웹 페이지

(p_i)에 우선순위를 부여한다(s21).

[0038] $E_i = (1-d) \cdot t_i$ --- (식 3).

[0039] 여기서 d , t_i 는 상기한 식들에서 언급한 바와 같다.

[0040] 다운로드부(22)는 상기 산출된 각 페이지(p_i)의 접근 확률(중요도) E_i 중 가장 높은 값을 갖는(가장 높은 우선순위를 갖는) 웹 페이지($p_{i,max}$)를 추출하고, 추출된 페이지에 링크되어 있으나 아직 탐색되지 아니한(unexplored) 외부링크 페이지들을 다운로드한다(s22).

[0041] 외부링크 우선순위 부여부(23)는 다운로드된 외부링크 페이지들의 각 링크 페이지(p_j)에 대한 접근 확률(중요도) E_j 를 산출하여 E_j 의 크기 순서에 따라 상기 각 링크 페이지(p_j)에 우선순위를 부여한다. E_j 의 산출은 아래의 식 4에 의해 산출한다(s23).

[0042] $E_j = E_i + (d \cdot E_i) / o_i$ --- (식 4).

[0043] (식 4)와 관련된 사항은 (식 2)와 관련된 사항과 동일하다.

[0044] 제2 구현예에 의하여 부여되는 각 링크 페이지(p_j)의 우선순위도, 상기 제1 구현예와 유사하게, 전체 웹 공간에서의 우선순위가 아니라, 최대의 E_i 를 갖는 이미 다운로드된 페이지내에서의 우선순위이다. 다시 말하면, 본 발명의 제2 구현예에 의한 외부링크에의 우선순위 부여는 기존과는 달리 웹 공간 전체를 기준으로 하는 것이 아니라, 최대의 E_i 를 갖는 페이지를 기준으로 이루어진다. 따라서 제2 구현예의 경우에도 우선순위 부여에 소요되는 시간(웹 크롤링에 소요되는 시간)이 기존의 기법에 비해 현저히 감축된다.

[0045] 외부링크 우선순위 부여부(23)는 각 링크 페이지(p_j)에 우선순위의 부여를 끝내면 상기 $p_{i,max}$ 에 링크된 웹 페이지들에 대한 탐색을 끝냈음을 알리는 마킹을 상기 $p_{i,max}$ 와 이에 링크된 각 웹 페이지(p_i)에 부여하고, 상기 $p_{i,max}$ 의 다음 우선순위를 갖는 웹 페이지에 대해 상기한 과정이 이루어지게 한다. 마킹은 아울러 $p_{i,max}$ 와 $p_{i,max}$ 의 다음 우선순위를 갖는 웹 페이지에 동시에 링크된 외부링크 페이지에 관해서는, $p_{i,max}$ 의 다음 우선순위를 갖는 웹 페이지에 대해 상기한 과정이 수행될 때, 탐색이 이루어지지 않게 하도록 한다.

[0046] 도 3은 본 발명의 성능을 기존의 기법과 비교하여 제시한 그래프이다.

[0047] 범례 중 첫 번째가 본 발명의 제1 구현예에 의한 것을(FPR Discovered Crawler), 두 번째가 본 발명의 제2 구현예에 의한 것이다. 세 번째와 네 번째는 기존의 기법에 의한 것이다.

[0048] 도 3의 그래프에 제시된 바와 같이, 본 발명에 의해 크롤링되는 웹 문서의 수(Number of Documents)는 기존의 기법에 비해 동일 시간 대비 현격히 차이가 나는 것을 알 수 있다.

[0049] 본 방법발명은 또한 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다.

[0050] 컴퓨터가 읽을 수 있는 기록매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피디스크, 광 데이터 저장장치 등이 있으며, 또한 캐리어 웨이브(예를 들어 유무선 네트워크를 통한 전송)의 형태로 구현되는 것도 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

[0051] 이제까지 본 발명에 대하여 그 바람직한 실시예를 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.

[0052] 그러므로 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 균등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

도면의 간단한 설명

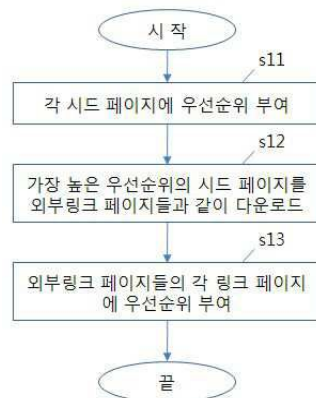
- [0053] 도 1a는 본 시스템 발명의 제1 구성을 제시한 도면이다.
- [0054] 도 1b는 본 방법 발명의 제1 흐름을 제시한 도면이다.
- [0055] 도 2a는 본 시스템 발명의 제2 구성을 제시한 도면이다.
- [0056] 도 2b는 본 방법 발명의 제2 흐름을 제시한 도면이다.
- [0057] 도 3은 본 발명의 성능을 기존의 기법과 비교하여 제시한 그래프이다.

도면

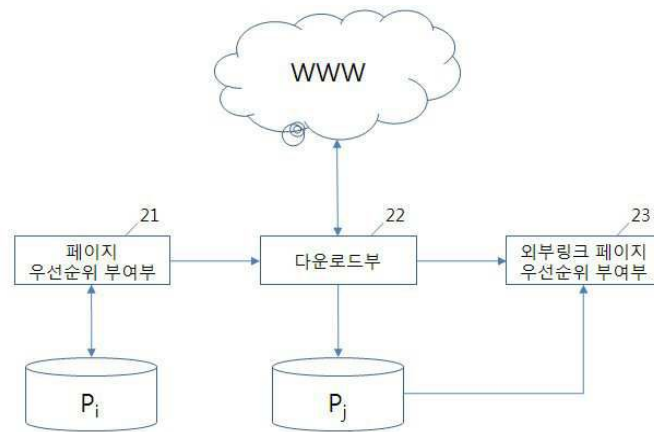
도면1a



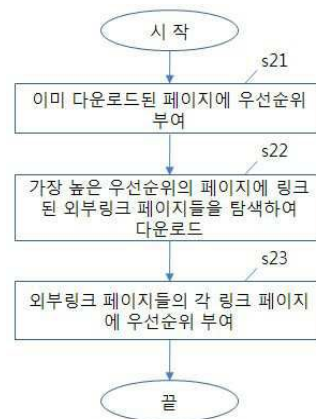
도면1b



도면2a



도면2b



도면3

