



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2011년12월08일
(11) 등록번호 10-1090549
(24) 등록일자 2011년11월30일

(51) Int. Cl.

G06F 17/30 (2006.01)

(21) 출원번호 10-2011-0042412

(22) 출원일자 2011년05월04일

심사청구일자 2011년05월04일

(56) 선행기술조사문헌

정보과학회논문지(2009.01)*

*는 심사관에 의하여 인용된 문헌

(73) 특허권자

인하대학교 산학협력단

인천 남구 용현동 253 인하대학교

(72) 발명자

심정섭

인천광역시 남동구 간석2동 금호어울림마을 110동 1101호

윤현철

인천광역시 남구 용현4동 84-6번지 신유성빌라 가동 102호

조석현

인천광역시 남구 학익동 풍림아이원아파트 105동 501호

(74) 대리인

특허법인무한

전체 청구항 수 : 총 3 항

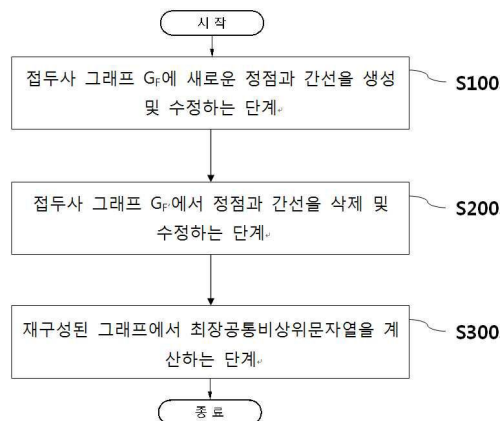
심사관 : 이명진

(54) 접두사 그래프 모델에 기반한 동적 최장공통비상위문자열 검색 방법

(57) 요약

접두사 그래프 모델에 기반한 동적 최장공통비상위문자열 검색 방법이 개시된다. 동적 최장공통비상위문자열 검색 방법은, 금지문자열 집합에 의해 생성된 접두사 그래프에, 금지문자열 집합에 추가된 문자열에 대응하는 정점과 간선을 생성하여 접두사 그래프를 재구성 하거나, 금지문자열 집합에서 삭제된 문자열에 대응하는 정점과 간선을 삭제하여 접두사 그래프를 재구성 하는 단계; 및 재구성 된 접두사 그래프를 이용하여 가장 긴 문자열에 해당되는 최장공통비상위문자열을 검색하는 단계를 포함하는 동적 최장공통비상위문자열 검색 방법이 제공된다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 2010-0028134

부처명 교육과학기술부

연구관리전문기관 한국연구재단

연구사업명 기본연구

연구과제명 문자열 포함 및 불포함 문제해결 알고리즘 및 응용 기술개발

기여율 1/1

주관기관 인하대학교 산학협력단

연구기간 2010.09.01 ~ 2011.08.31

특허청구의 범위

청구항 1

접두사 그래프 모델에 기반하여 금지문자열인 최장공통비상위문자열을 검출하는 침입탐지시스템에서의 동적 최장공통비상위문자열 검색 방법에 있어서,

상기 동적 최장공통비상위문자열 검색 방법은,

상기 침입탐지시스템에서, 금지문자열 집합에 의해 생성된 접두사 그래프에, 상기 금지문자열 집합에 추가된 문자열에 대응하는 정점과 간선을 생성하여 상기 접두사 그래프를 재구성 하는 단계; 및

상기 침입탐지시스템에서, 상기 재구성 된 접두사 그래프를 이용하여 가장 긴 문자열에 해당되는 최장공통비상위문자열을 검색하는 단계

를 포함하고,

상기 금지문자열 집합에 추가된 문자열에 대응하는 정점과 간선을 생성하여 상기 접두사 그래프를 재구성 하는 단계는,

상기 침입탐지시스템에서, 상기 접두사 그래프에서 상기 추가된 문자열에 대응하는 정점을 검색하는 단계; 및

상기 침입탐지시스템에서, 상기 추가된 문자열에 대응하는 정점이 상기 접두사 그래프에 존재하지 않으면 상기 추가된 문자열에 대응하는 정점과 간선을 상기 접두사 그래프에 생성하는 단계

를 포함하는, 동적 최장공통비상위문자열 검색 방법.

청구항 2

삭제

청구항 3

접두사 그래프 모델에 기반하여 금지문자열인 최장공통비상위문자열을 검출하는 침입탐지시스템에서의 동적 최장공통비상위문자열 검색 방법에 있어서,

상기 동적 최장공통비상위문자열 검색 방법은,

상기 침입탐지시스템에서, 금지문자열 집합에 의해 생성된 접두사 그래프에서, 상기 금지문자열 집합에서 삭제된 문자열에 대응하는 정점과 간선을 삭제하여 상기 접두사 그래프를 재구성 하는 단계; 및

상기 침입탐지시스템에서, 상기 재구성 된 접두사 그래프를 이용하여 가장 긴 문자열에 해당되는 최장공통비상위문자열을 검색하는 단계

를 포함하고,

상기 금지문자열 집합에 삭제된 문자열에 대응하는 정점과 간선을 삭제하여 상기 접두사 그래프를 재구성 하는 단계는,

상기 침입탐지시스템에서, 상기 접두사 그래프에서 상기 삭제된 문자열에 대응하는 정점을 검색하는 단계; 및

상기 침입탐지시스템에서, 상기 삭제된 문자열에 대응하는 정점이 상기 접두사 그래프에 존재하면 상기 삭제된 문자열에 대응하는 정점과 간선을 상기 접두사 그래프에서 삭제하는 단계

를 포함하는, 동적 최장공통비상위문자열 검색 방법.

청구항 4

삭제

청구항 5

제1항 또는 제3항에 있어서,

상기 최장공통비상위문자열을 검색하는 단계는,

상기 침입탐지시스템에서, 상기 재구성 된 접두사 그래프에서 정점들을 순회하는 최장 경로를 계산하는 단계;
및

상기 침입탐지시스템에서, 상기 계산된 최장 경로를 이용하여 상기 최장공통비상위문자열을 계산하는 단계
를 포함하는 동적 최장공통비상위문자열 검색 방법.

명세서

기술 분야

- [0001] 본 발명의 실시예들은 접두사 그래프 모델에 기반한 동적 최장공통비상위문자열 검색 방법에 관한 것으로서, 특히 문자열 불포함 검색을 위한 방향 그래프 수정 방법 및 그에 대한 이용에 관한 것이다.

배경 기술

- [0002] 주어진 문자열 집합에 대한 포함/불포함 문제가 압축 알고리즘, 분자 생물학, 컴퓨터 보안 등 다양한 분야에서 필요성이 대두되어 진행되고 있다.
- [0003] 문자열 포함 문제는 최장공통부분서열(Longest Common Subsequence), 최단공통상위서열(Shortest Common Supersequence), 최장공통부분문자열 (Longest Common Substring)과 최단공통상위문자열(Shortest Common Superstring) 등이 연구되어 왔고, 문자열 불포함 문제는 최장공통비상위서열(Longest Common Non-Supersequence), 최단공통비부분서열(Shortest Common Non-Subsequence), 최장공통비상위문자열(Longest Common Non-Superstring)과 최단공통비부분문자열(Shortest Common Non-Substring) 등이 연구되어 왔다.
- [0004] 이 중 최장공통비상위문자열에 관하여 상세히 설명하면 다음과 같다.
- [0005] 문자열 집합 F 가 주어졌을 때, F 의 모든 문자열을 포함하지 않는 공통비상위문자열(Common Non-Superstring, CNSS)들의 집합을 $CNSS_F$ 라 하고, 이들 중 가장 긴 문자열을 최장공통비상위문자열(Longest Common Non-Superstring, LCNSS)이라 한다. 집합 F 의 원소는 $CNSS_F$ 가 포함하지 않아야 할 문자열이므로 금지문자열이라 하고, 이에 따라 집합 F 를 금지문자열 집합이라 한다. 최장공통비상위문자열은 접미사 트리 또는 접두사 그래프 모델을 이용하여 구할 수 있음이 널리 알려져 있다.
- [0006] 최장공통비상위문자열 문제는 컴퓨터보안 분야에서 패킷 내의 악성패턴 등을 검출하는 침입탐지시스템에 활용될 수 있다. 악성패턴으로 알려진 문자열들을 F 로 정의하고 네트워크 상의 패킷 내에 악성패턴(문자열)이 존재하면 이를 검출하는 것이다. 그런데 이러한 악성패턴들은 계속해서 변화하고 있으며 이에 따라 F 의 원소들도 변경할 필요가 있다. 새로운 악성패턴들이 발견되면 F 에 새로운 문자열들을 추가해야 하며 더 이상 위협적이지 않은 패턴들은 F 에서 제거되어야 한다. 이러한 경우와 같이 F 에 한 문자열이라도 추가되거나 삭제될 경우 최장공통비상위문자열을 구하기 위해서 접두사 그래프를 재생성 하게 되는데, 실제 구현했을 때 그래프를 생성하는데 상당한 시간이 필요하므로 효율적이라 할 수 없다.

발명의 내용

해결하려는 과제

- [0007] 금지문자열 집합에서 수정된 문자열만을 금지문자열 집합으로 생성된 접두사 그래프에 부분 적용시켜 수정된 금지문자열 집합에 대해 능률적으로 최장공통비상위문자열을 찾을 수 있는 동적 최장공통비상위문자열 검색 방법이 제공된다.

과제의 해결 수단

- [0008] 금지문자열 집합에 의해 생성된 접두사 그래프에, 금지문자열 집합에 추가된 문자열에 대응하는 정점과 간선을 생성하여 접두사 그래프를 재구성 하는 단계; 및 재구성 된 접두사 그래프를 이용하여 가장 긴 문자열에 해당되는 최장공통비상위문자열을 검색하는 단계를 포함하는 동적 최장공통비상위문자열 검색 방법이 제공된다.

[0009] 일측에 따르면, 금지문자열 집합에 추가된 문자열에 대응하는 정점과 간선을 생성하여 접두사 그래프를 재구성하는 단계는, 접두사 그래프에서 추가된 문자열에 대응하는 정점을 검색하는 단계; 및 추가된 문자열에 대응하는 정점이 접두사 그래프에 존재하지 않으면 추가된 문자열에 대응하는 정점과 간선을 접두사 그래프에 생성하는 단계를 포함한다. 이때, 최장공통비상위문자열을 검색하는 단계는, 재구성된 접두사 그래프에서 정점들을 순회하는 최장 경로를 계산한 후 최장 경로를 이용하여 최장공통비상위문자열을 계산할 수 있다.

[0010] 금지문자열 집합에 의해 생성된 접두사 그래프에서, 금지문자열 집합에서 삭제된 문자열에 대응하는 정점과 간선을 삭제하여 접두사 그래프를 재구성하는 단계; 및 재구성된 접두사 그래프를 이용하여 가장 긴 문자열에 해당되는 최장공통비상위문자열을 검색하는 단계를 포함하는 동적 최장공통비상위문자열 검색 방법이 제공된다.

[0011] 일측에 따르면, 금지문자열 집합에 삭제된 문자열에 대응하는 정점과 간선을 삭제하여 접두사 그래프를 재구성하는 단계는, 접두사 그래프에서 삭제된 문자열에 대응하는 정점을 검색하는 단계; 및 삭제된 문자열에 대응하는 정점이 접두사 그래프에 존재하면 삭제된 문자열에 대응하는 정점과 간선을 접두사 그래프에서 삭제하는 단계를 포함한다. 이때, 최장공통비상위문자열을 검색하는 단계는, 재구성된 접두사 그래프에서 정점들을 순회하는 최장 경로를 계산한 후 최장 경로를 이용하여 최장공통비상위문자열을 계산할 수 있다.

발명의 효과

[0012] 금지문자열 집합 F에 새로운 문자열이 추가되거나 금지문자열 집합 F에서 문자열이 삭제될 경우, 접두사 그래프 전체를 재생성하는 것이 아니라 추가 또는 삭제된 문자열에 대응하는 정점과 간선만을 접두사 그래프에서 수정하는 방식을 이용함으로써 보다 신속하고 능률적으로 최장공통비상위문자열을 검색할 수 있다.

도면의 간단한 설명

[0013] 도 1은 본 발명의 일실시예에 있어서, 접두사 그래프 모델에 기반한 동적 최장공통비상위문자열 검색 방법을 도시한 순서도이다.

도 2는 본 발명의 일실시예에 있어서, 접두사 그래프에 새로운 정점과 간선을 추가하는 과정을 설명하기 위한 순서도이다.

도 3은 문자 집합 $\Sigma=\{a,b,c\}$ 에 대한 금지문자열 집합 $F=\{aaa, aabaab, abababa, abababcb, ababca, bbb, cc\}$ 를 접두사 그래프 모델로 생성한 그래프를 도시한 것이다.

도 4는 금지문자열 집합 $F=\{aaa, aabaab, abababa, abababcb, ababca, bbb, cc\}$ 에 새로운 문자열 abcaa를 추가하였을 때 접두사 그래프 변화의 중간과정을 설명하기 위한 도면이다.

도 5는 금지문자열 집합 $F=\{aaa, aabaab, abababa, abababcb, ababca, bbb, cc\}$ 에 새로운 문자열 abcaa를 추가한 후 완성된 접두사 그래프를 도시한 것이다.

도 6는 본 발명의 일실시예에 있어서, 접두사 그래프에서 정점과 간선을 삭제하는 과정을 설명하기 위한 순서도이다.

도 7은 문자열 abcaa를 추가한 금지문자열 집합 $F'=\{aaa, aabaab, abababa, abababcb, ababca, abcaa, bbb, cc\}$ 에서 문자열 abcaa를 삭제하였을 때 접두사 그래프 변화의 중간과정을 설명하기 위한 도면이다.

도 8은 문자열 abcaa에서 문자 c를 삭제한 후의 그래프 변화를 도시한 것이다.

도 9는 본 발명의 일실시예에 있어서, 최장공통비상위문자열의 계산 과정을 설명하기 위한 도면이다.

도 10 내지 도 13은 본 발명에서의 금지문자열 집합의 수정에 따른 수행시간을 기존의 알고리즘과 비교한 그래프로 도시한 것이다.

발명을 실시하기 위한 구체적인 내용

[0014] 이하, 본 발명의 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0015] 본 발명의 실시예들은 접두사 그래프 모델을 이용하여 최장공통비상위문자열을 구하는 방법에 관한 것이다. 특히, 본 발명의 일실시예에 따른 최장공통비상위문자열 검색 방법은 컴퓨터보안 분야에서 패킷 내의 악성패턴(즉, 금지문자열)을 검출하는 침입탐지시스템에 적용될 수 있다.

- [0016] 본 명세서에서는 금지문자열 집합을 F로 표기하고 금지문자열 집합 F로 생성된 접두사 그래프를 G_F 로 표기하기로 한다.
- [0017] 도 1은 본 발명의 일실시예에 있어서, 접두사 그래프 모델에 기반한 동적 최장공통비상위문자열 검색 방법을 도시한 순서도이다.
- [0018] 일실시예에 따른 동적 최장공통비상위문자열 검색 방법은, 도 1에 도시된 바와 같이, 금지문자열 집합 F에 추가된 문자열에 대응하는 새로운 정점과 간선을 접두사 그래프 G_F 에 생성 및 수정하는 단계(S100), 금지문자열 집합 F에 삭제된 문자열에 대응하는 정점과 간선을 수정된 접두사 그래프 G_F 에서 삭제 및 수정하는 단계(S200), 단계(S100)과 단계(S200)을 통해 최종 수정된 접두사 그래프 G_F 에서 최장공통비상위문자열을 계산하는 단계(S300)를 포함하여 구성될 수 있다.
- [0019] 도 2는 도 1에 도시한 단계(S100)의 세부 과정을 도시한 것이다. 도 2에 도시된 바와 같이, 금지문자열 집합 F에 추가된 문자열에 대응하는 새로운 정점과 간선을 접두사 그래프 G_F 에 생성 및 수정하는 단계(S100)는 접두사 그래프 G_F 에 추가된 문자열에 대응하는 정점을 검색하는 단계(S110)와, 접두사 그래프 G_F 에 추가된 문자열에 대응하는 정점이 존재하지 않을 경우 해당 정점과 간선을 생성 및 수정하는 단계(S120)로 구성될 수 있다.
- [0020] 도 3은 단계(S100)에서 금지문자열 집합 F로 생성된 접두사 그래프 G_F 를 개략적으로 나타낸 도면이고, 도 4는 금지문자열 집합 F에 새로 추가된 문자열에 대응하는 정점과 간선들을 생성 및 수정하는 단계(S120)에서의 중간 과정을 나타내는 도면이고, 도 5는 본 발명의 일실시예에 따른 금지문자열 집합 F에 새로 추가된 문자열에 대응하는 정점과 간선들을 생성 및 수정하는 단계(S120)가 완료된 후 완성된 접두사 그래프 G_F 를 개략적으로 나타낸 도면이다. 도 5는 금지문자열 집합 F에서 삭제된 문자열에 대응하는 정점과 간선을 단계(S100)에서 수정된 접두사 그래프 G_F 에서 삭제 및 수정하는 단계(S200)의 입력 그래프로도 사용된다.
- [0021] 도 2에 도시된, 금지문자열 집합 F에 추가된 문자열에 대응하는 새로운 정점과 간선을 접두사 그래프 G_F 에 생성 및 수정하는 과정을 상세히 설명하면 다음과 같다.
- [0022] 먼저, 금지문자열 집합 F에 있는 모든 문자열의 진접두사 집합을 P라고 표기한다. 그리고, 접두사 그래프 G_F 의 정점 v에 대해 $\text{str}(v)$ 의 진접미사 중 진접두사 집합 P에 속하는 최장문자열에 해당하는 정점을 $\text{SV}(v)$ 라 표기하고, $\text{str}(v)$ 의 진접두사 중 진접두사 집합 P에 속하는 최장문자열에 해당하는 정점을 $\text{PV}(v)$ 라 표기하고, $\text{str}(v)$ 의 마지막 문자를 $\text{LC}(v)$ 라 표기한다. 또한, $\text{SV}(w) = v$ 인 정점 w들의 집합을 $\text{ISV}(v)$ 라 표기하고, $v \oplus \sigma \in P \cup F$ 일 때 정점 $\text{ver}(v \oplus \sigma)$ 를 $\text{DL}(v, \sigma)$ 로, $v \oplus \sigma \in P \cup F$ 일 때 $v \oplus \sigma$ 의 진접미사 중 진접두사 집합 P에 속하는 최장문자열에 해당하는 정점을 $\text{FL}(v, \sigma)$ 로 표기한다.
- [0023] 도 3을 참조하여 예를 들어보면, $F = \{\text{aaa}, \text{aabaab}, \text{abababa}, \text{abababcb}, \text{ababca}, \text{bbb}, \text{cc}\}$ 일 때, $P = \{\lambda, a, aa, aab, aaba, aabaa, ab, aba, abab, ababa, ababab, ababc, b, bb, c\}$ 이다. $\text{SW}(w)$ 를 계산해 보면, $\text{str}(w) = \text{abab}$ 이고, abab의 진접미사 집합은 $\{\lambda, bab, ab, b\}$ 이므로 이들 중 P에 속하는 최장문자열은 ab이고 따라서 $\text{ver}(ab) = v$ 이다. 유사하게 $\text{PV}(y)$ 를 계산해 보면, $\text{str}(y) = \text{aab}$ 이고, aab의 진접두사 집합은 $\{\lambda, a, aa\}$ 이므로 이들 중 P에 속하는 최장문자열은 aa이고 따라서 $\text{ver}(aa) = y'$ 이다. 또한, $\text{LC}(v)$ 는 $\text{str}(v) = \text{ab}$ 의 마지막문자 이므로 b이고, $\text{SV}(y)$ 와 $\text{SV}(w)$ 가 v를 가리키므로, $\text{ISV}(v)$ 는 $\{y, w\}$ 이다. $\text{DL}(y', b)$ 는 $y' \oplus b = \text{aab}$ 이고 $\text{aab} \in P \cup F$ 이므로, $\text{ver}(\text{aab})$ 인 정점 y 이고, $v \oplus b = \text{abb}$ 이고 $\text{abb} \in P \cup F$ 이므로, abb의 진접미사집합 $\{\lambda, b, bb\}$ 중 P에 속하는 최장문자열인 bb에 해당하는 정점 v' 이다.
- [0024] F에 새로 삽입되는 문자열을 a_i 라 표기한다. a_i 가 삽입 되었을 경우 먼저 a_i 와 대응하는 정점을 G_F 에서 검색한다(S110). 이 때 검색은 G_F 의 λ 정점부터 시작하게 되며, 진행은 한 문자씩 일치되는 간선(DL 함수값)을 따라간다. 탐색이 실패하는 순간 즉, DL 함수값을 따라갈 수 없을 경우 그래프를 갱신하게 된다(S120). 탐색이 실패하게 되는 정점과 그때 a_i 의 문자를 각각 v와 σ 라 하면, $\text{DL}(v, \sigma)$ 는 존재하지 않고 $\text{FL}(v, \sigma)$ 는 존재하게 된다. 정점 v에서의 그래프 갱신과정은 다음과 같다. 먼저, $\text{FL}(v, \sigma) = \alpha$ 라 하자. 새로운 정점 α' 을 생성하고 $\text{DL}(v, \sigma) = \alpha'$ 으로 설정하고 $\text{FL}(v, \sigma)$ 를 삭제한다. 다음으로 v의 $\text{ISV}(v)$ 에 대해 FL, SV 함수값들을 수정할 필요가

있는지를 확인하여 처리한다.

[0025] 도 3에 제시된 그래프에서 문자열 $a_i = abcaa$ 가 추가되는 경우를 예로 들어 설명한다. $DL(\lambda, a) = \beta$, $DL(\beta, b) = v$ 인 S110단계를 완료하여 정점 v 까지 진행한다. 이후의 S120단계는 도 4에서 제시된 그래프를 예로 설명한다. 정점 v 에서 $DL(v, c)$ 는 존재하지 않고 $FL(v, c) = a$ 이다. 따라서 새로운 정점 a' 을 생성한 뒤 $FL(v, c)$ 값은 삭제하고 $DL(v, c) = a'$ 으로 설정한다. 다음으로 FL 함수값과 SV 함수값이 a 인 정점들을 찾아 이들의 함수값들을 a' 으로 변경해야 한다. 이는 정점 v 의 ISV를 이용한다. $ISV(v) = \{y, w\}$ 이므로 y 와 w 에 대해 처리한다. 먼저 $FL(y, c)$ 는 $FL(SV(y), c)$ 또는 $DL(SV(y), c)$ 이므로, $FL(y, c) = a'$ 으로 변경한다. 정점 y 의 함수값이 변했기 때문에 $ISV(y)$ 에 속한 정점들의 함수값도 변경될 수 있으므로 이를 확인하는 과정이 필요하다. $DL(y, c)$ 가 존재하지 않고, $ISV(y) = \{z\}$ 이므로 $FL(z, c)$ 를 확인한다. 이는 y 의 경우와 같이 $FL(z, c) = a$ 이므로 $FL(z, c) = a'$ 으로 변경한다. $ISV(z)$ 는 공집합이므로 더 이상 변경할 정점은 없다. 한편, $FL(w, c)$ 는 존재하지 않고 $DL(w, c)$ 가 존재하기 때문에 $SV(w') = a'$ 으로 변경한다. 이는 $str(w')$ 즉, $ababc$ 의 진접미사 중 P 에 속하는 가장 긴 문자열이 c 에서 abc 로 변경되었기 때문에, $SV(w')$ 가 가리키는 정점도 정의에 의해 a 에서 a' 으로 변경되어야 하기 때문이다. 정점 w 이후로 더 이상의 변경을 진행하지 않는 이유는 다음과 같다. 도 4에서 w 에 대해 처리를 끝내고, $ISV(w)$ 를 따라 x 로 진행한다. $DL(x, c) = x'$ 을 방문하여 $SV(x')$ 을 확인하면 a 가 아닌 w' 을 향하고 있으며 이는 a 와 a' 과는 관련이 없으므로 이후로는 변화되지 않는다. 도 5의 그래프는 도 3의 그래프에 $abcaa$ 의 삽입이 완료된 후의 그래프이다.

[0026] 도 6은 도 1에 도시한 단계(S200)의 세부 과정을 도시한 것이다. 도 6에 도시된 바와 같이, 금지문자열 집합 F 에 삭제된 문자열에 대응하는 정점과 간선을 수정된 접두사 그래프 G_F 에서 삭제 및 수정하는 단계(S200)는 접두사 그래프 G_F 에서 삭제된 문자열에 대응하는 마지막 정점을 검색하는 단계(S210)와, 접두사 그래프 G_F 에서 추가된 문자열에 대응하는 정점과 간선을 제거 및 수정하는 단계(S220)로 구성될 수 있다.

[0027] 도 7은 본 발명의 일실시예에 따른 금지문자열 집합 F 에서 삭제된 문자열에 대응하는 정점과 간선들을 삭제 및 수정하는 단계(S220)에서의 중간과정을 나타내는 도면이고, 도 8은 본 발명의 일실시예에 따른 금지문자열 집합 F 에서 삭제된 문자열에 대응하는 정점과 간선들을 삭제 및 수정하는 단계(S220)가 완료된 후 완성된 접두사 그래프 G_F 를 개략적으로 나타낸 도면이다.

[0028] 도 6에 도시된, 금지문자열 집합 F 에 삭제된 문자열에 대응하는 정점과 간선을 접두사 그래프 G_F 에서 삭제 및 수정하는 과정을 상세히 설명하면 다음과 같다.

[0029] 여기서, 금지문자열 F 에서 삭제되는 문자열을 d_i 라 표기한다. 먼저 d_i 의 마지막 문자와 대응하는 정점을 접두사 그래프 G_F 에서 검색하며(S210), 진행 방향은 PV 함수를 이용하여 역으로 진행한다. d_i 의 각 문자 σ 와 대응되는 정점을 v 라 하자. 정점 v 에서의 그래프 갱신과정(S220)은 다음과 같다.

[0030] 먼저, v 가 삭제될 정점인지를 판단한다. 만약 v 에서 DL 함수값이 하나도 정의되지 않는다면 v 는 삭제될 정점이다. 반대로, DL 함수값이 하나라도 정의되어 있다면 v 는 삭제할 수 없다. 정점 v 에 DL 함수값이 하나 이상 존재한다는 것은 $str(v)$ 를 포함하는 다른 문자열에 해당하는 정점이 그래프 상에 존재한다는 것을 의미하기 때문에 v 를 삭제할 경우 그래프 모델의 정의에 모순이 발생한다. 삭제할 정점임이 확인되면, v 를 삭제하기 전에 먼저 v 와 관련된 FL , SV 함수들을 $ISV(PV(v))$ 를 이용하여 수정한다. 이후 v 를 삭제하고 재귀적으로 $PV(v)$ 에 대해서 그래프 수정단계를 진행한다. 만약 v 가 삭제되지 않는 정점이라면 S220단계는 종료된다.

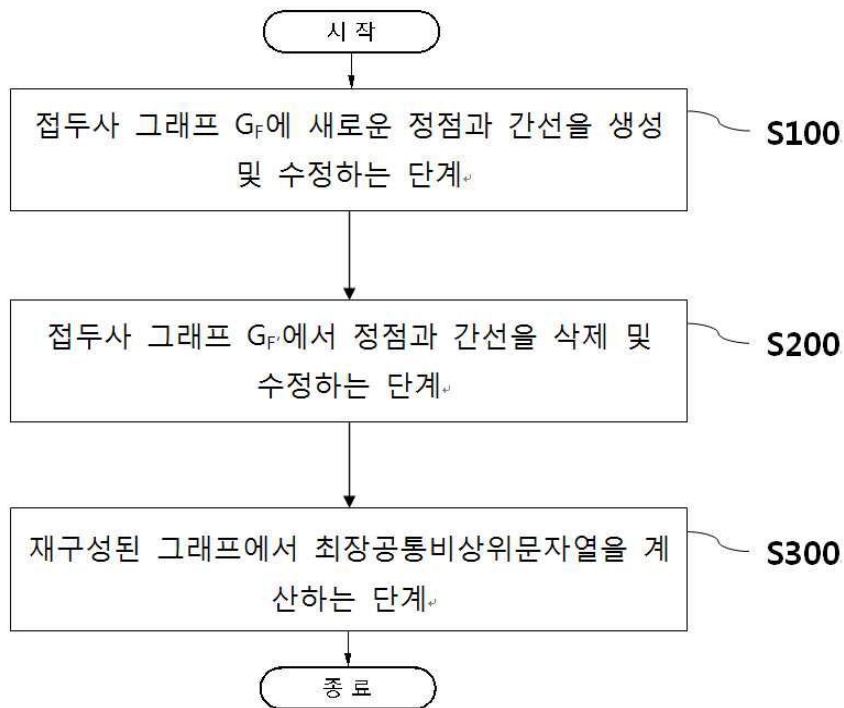
[0031] 앞서 설명한 도 5에서 제시된 그래프에서 문자열 $d_i = abcaa$ 가 삭제되는 경우를 예로 들어 설명한다. 마지막 문자 a 와 그 이전 문자 a 를 삭제하는 것은 간단한 연산들만으로 해결이 가능하므로 생략하고, 세 번째 문자인 c 를 처리하는 과정을 설명한다. 문자열 abc 를 나타내는 a' 이 삭제된다면 먼저 a' 과 관련된 간선들을 처리해 주어야 한다. 만약 a' 이 삭제되지 않는다면 S220을 종료하면 된다. 모든 $\sigma \in \Sigma$ 에 대해 $DL(a', \sigma)$ 가 하나도 정의되지 않으므로 a' 이 삭제되어야 한다. $PV(a') = v$ 이므로 먼저 $SV(PV(v))$ 를 이용하여 $DL(v, c) = a'$ 을 삭제하고 $FL(v, c) = ver(SV(PV(v)))^{\oplus} c$ 를 정의한다. 이후 a' 으로 향했던 SV , FL 들을 a 로 변경한다. 이는 S120단계에서 ISV 를 이용한 것과 같은 방식으로 진행한다. 정점 a' 과 관련된 간선들을 모두 처리한 후 a' 을 삭제한다. 다음으로 b 에 해당하는 정점은 v 이고 $DL(v, a)$ 가 정의되어 있으므로 S220 단계를 종료한다. 도 3의 그래프는 도 5의 그래프에서 문자열 $abcaa$ 와 대응하는 정점의 삭제와 간선의 삭제 및 수정이 완료된 후의 그래프 도

면 이다.

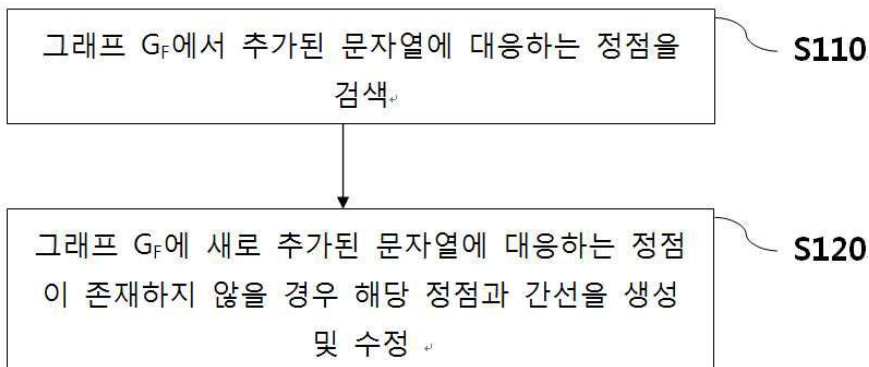
- [0032] 본 발명의 일실예에서는 앞서 설명한 과정을 거쳐 최종으로 재 구성된 접두사 그래프 G_F 에서 최장공통비상위문자열을 계산하게 된다. 도 9에 도시한 바와 같이, 최종 수정된 접두사 그래프 G_F 에서 최장공통비상위문자열을 계산하는 단계(S300)는 단계(S100)과 단계(S200)를 거쳐 수정된 접두사 그래프 G_F 에서 가장 긴 경로를 계산하고(S310), 가장 긴 경로를 이용하여 최장공통비상위문자열을 계산한다(S320). 즉, 금지문자열 집합에서 수정된 문자열만을 적용시켜 재 구성된 접두사 그래프 G_F 를 이용하여 가장 긴 문자열을 가진 최장공통비상위문자열을 찾을 수 있다.
- [0033] 도 10은 금지문자열 집합 F의 크기가 1000이고, 길이 300인 문자열 a_i 를 100개부터 2000개까지 100씩 증가시키며 F에 추가하였을 때 수행시간의 변화를 분석한 그래프 도면이고, 도 11은 a_i 의 길이는 300으로, 삽입 문자열 집합 A의 크기를 각각 1, 10, 100으로 고정한 후 F의 크기를 변화시키며 F에 A를 추가하였을 때 수행시간의 변화를 분석한 그래프 도면이다. 도 12는 집합 F의 크기가 3000이고, 길이 300인 문자열 d_i 를 100개부터 2000개까지 100씩 증가시키며 F에서 삭제하였을 때 수행시간의 변화를 분석한 그래프 도면이고, 도 13은 d_i 의 길이는 300으로, 삭제 문자열 집합 D의 크기를 각각 1, 10, 100으로 고정한 후 F의 크기를 변화시키며 F에서 D를 삭제하였을 때 수행시간의 변화를 분석한 그래프 도면 이다.
- [0034] 도 10은 약 1400개를 연속적으로 삽입 하였을 때까지 본 발명에서 제시한 방법이 기존 방식에 비하여 더 빨리 수행됨을 알 수 있다. 일반적으로, 연속적으로 삽입되는 문자열 집합 A의 크기는 집합 F의 크기보다 작을 것으로 예상되므로, A의 크기를 F의 크기의 약 20%정도로 가정할 경우, F의 크기가 1000 일 때 A의 크기는 200정도 이고, 수행시간은 기존의 방법이 0.535초, 본 발명에서 제시한 방법은 0.223초로 약 2.4배의 성능 향상을 확인할 수 있다.
- [0035] 도 11을 통하여 집합 F에 삽입되는 문자열 집합 A의 크기가 커질수록 수행시간이 오래 걸리는 것을 확인할 수 있으며, 이 또한 기존의 방법보다 본 발명의 방법이 약 4.1~7.6배의 성능 향상을 가져오는 것을 확인할 수 있다.
- [0036] 도 12는 삭제할 문자열 집합 D의 크기가 증가할수록 기존의 방법을 활용하였을 때는 집합 F의 크기가 감소하므로 수행시간이 빨라지지만, 약 1500개를 연속적으로 삽입 하였을 때까지 본 발명에서 제시한 방법이 기존 방식에 비해 더 빨리 수행됨을 알 수 있다. 삭제 역시 삽입의 경우와 마찬가지로 D의 크기를 F의 크기의 약 20%정도로 가정할 경우, F의 크기가 1000 일 때 D의 크기는 200정도 이고, 수행시간은 기존의 방법이 1.236초, 본 발명에서 제시한 방법은 0.250초로 약 4.9배의 성능 향상을 확인할 수 있다.
- [0037] 도 13을 통하여 집합 F에서 삭제되는 문자열 집합 D의 크기가 커질수록 수행시간이 오래 걸리는 것을 확인할 수 있으며, 이 또한 기존의 방법보다 본 발명의 방법이 약 5.7~7.5배의 성능 향상을 가져오는 것을 확인할 수 있다.
- [0038] 이와 같이, 본 발명의 실시예에 따르면, 금지문자열 집합 F에 새로운 문자열이 추가되거나 금지문자열 집합 F에서 문자열이 삭제될 경우, 기존의 방법을 사용하여 접두사 그래프 전체를 재생성 하는 것 보다 추가 또는 삭제된 문자열에 대응하는 정점과 간선만을 접두사 그래프에서 수정하는 방식을 이용하면 보다 효율적으로 최장공통비상위문자열을 구할 수 있다.
- [0039] 이상과 같이 본 발명은 비록 한정된 실시예와 도면에 의해 설명되었으나, 본 발명은 상기의 실시예에 한정되는 것은 아니며, 본 발명이 속하는 분야에서 통상의 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다.
- [0040] 그러므로, 본 발명의 범위는 설명된 실시예에 국한되어 정해져서는 아니 되며, 후술하는 특허청구범위뿐 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

도면

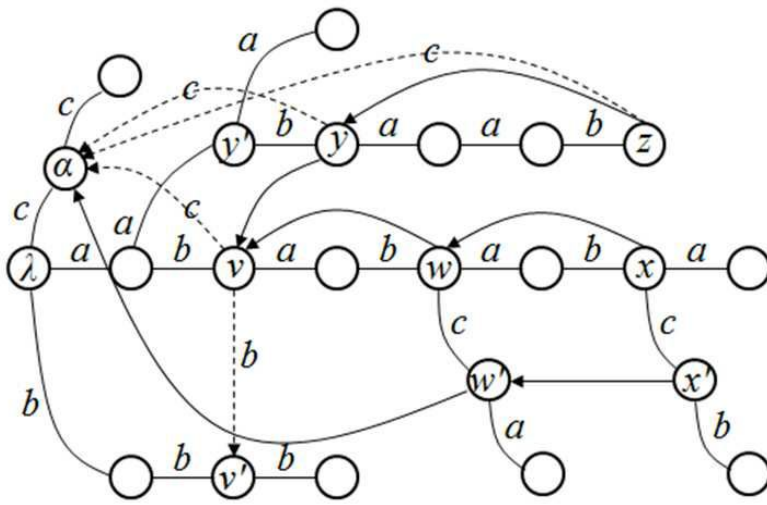
도면1



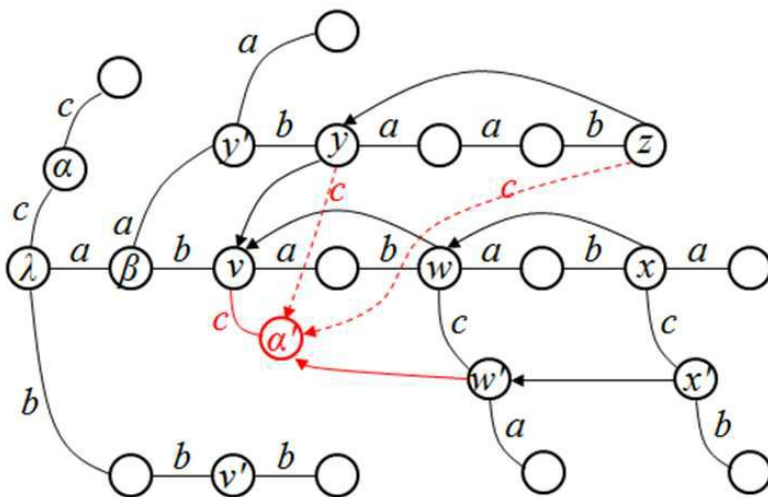
도면2



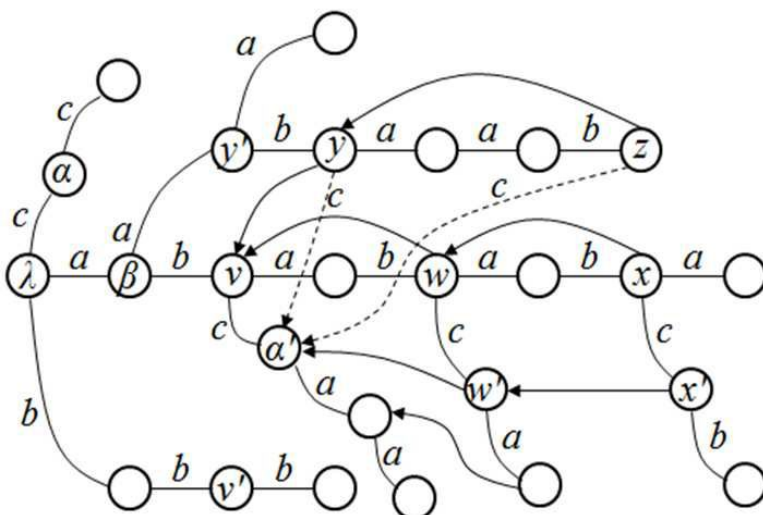
도면3



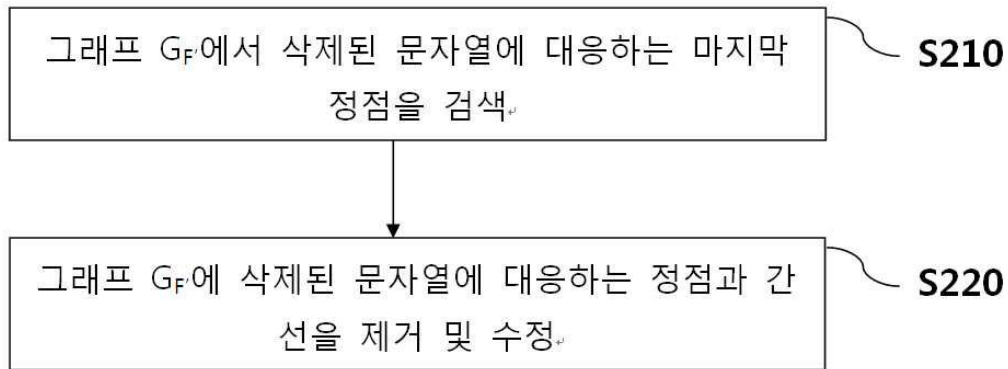
도면4



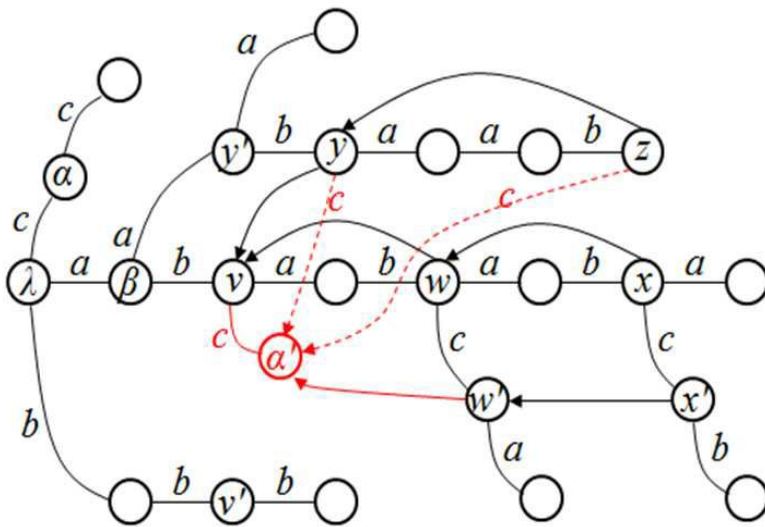
도면5



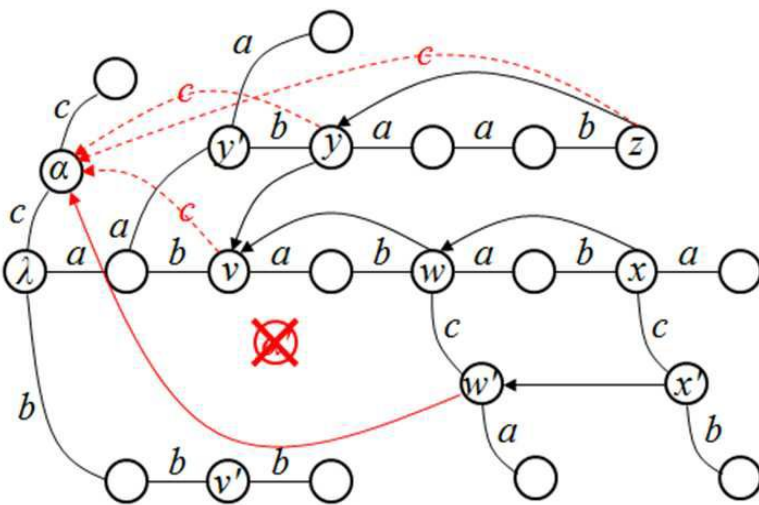
도면6



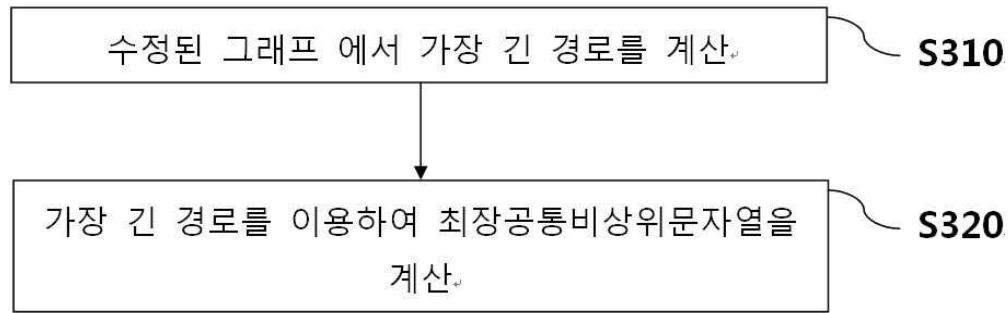
도면7



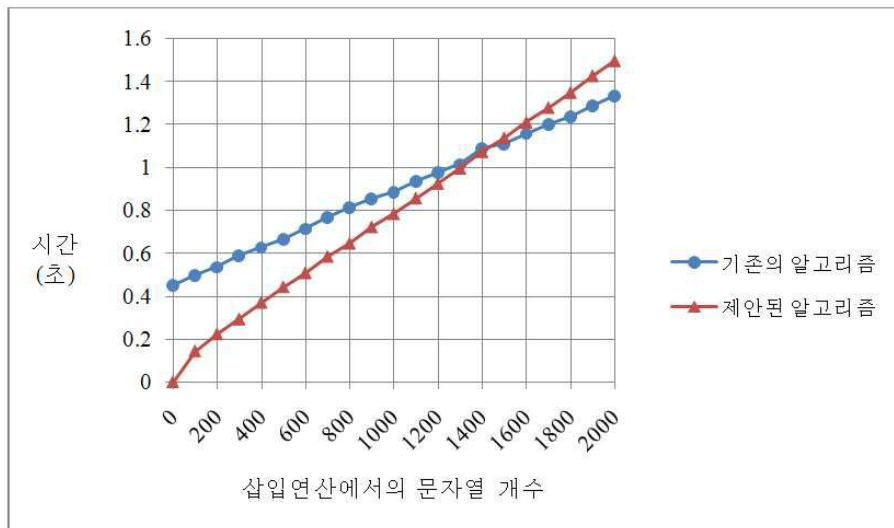
도면8



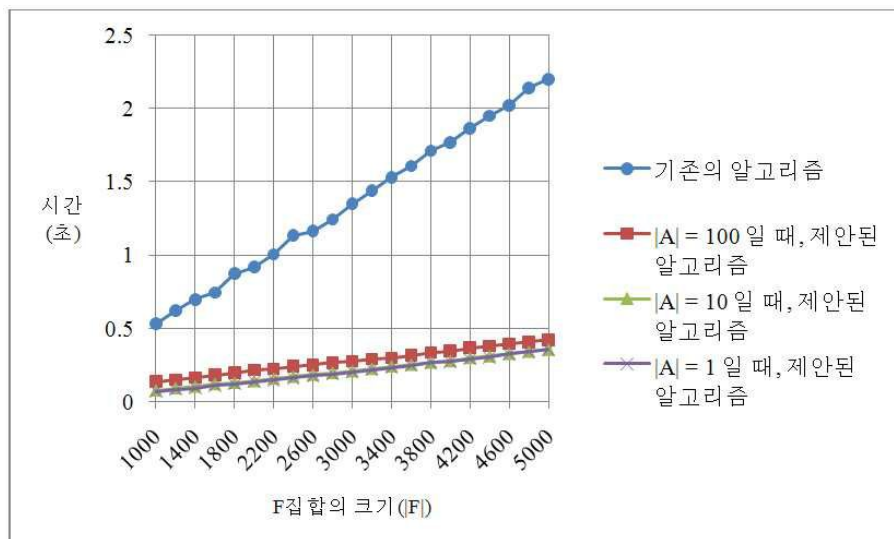
도면9



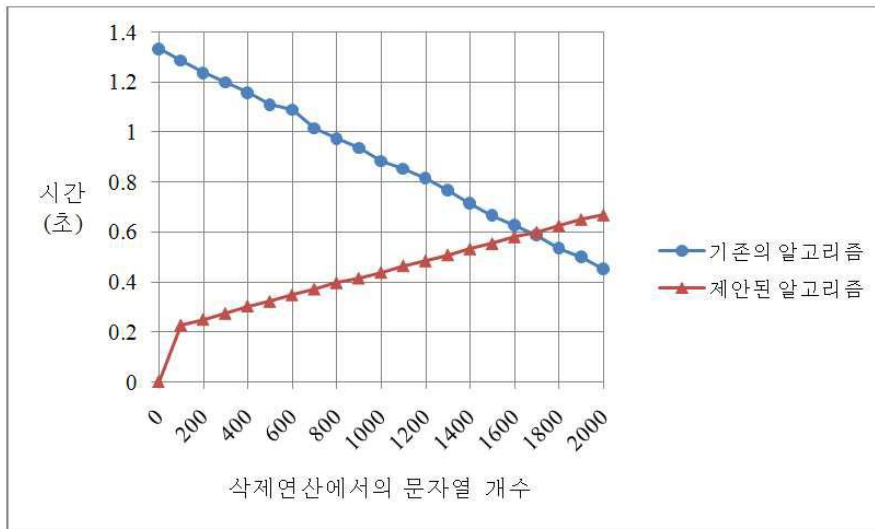
도면10



도면11



도면12



도면13

