



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2011-0080667
(43) 공개일자 2011년07월13일

(51) Int. Cl.

G06F 21/20 (2006.01) G06Q 50/00 (2006.01)

(21) 출원번호 10-2010-0000999

(22) 출원일자 2010년01월06일

심사청구일자 2010년01월06일

(71) 출원인

고려대학교 산학협력단

서울 성북구 안암동5가 1

(72) 발명자

정연돈

서울특별시 중구 황학동 롯데캐슬 베네치아
105-812

성민경

서울특별시 강남구 개포동 주공아파트 707동 309
호

(74) 대리인

조영현, 나승택

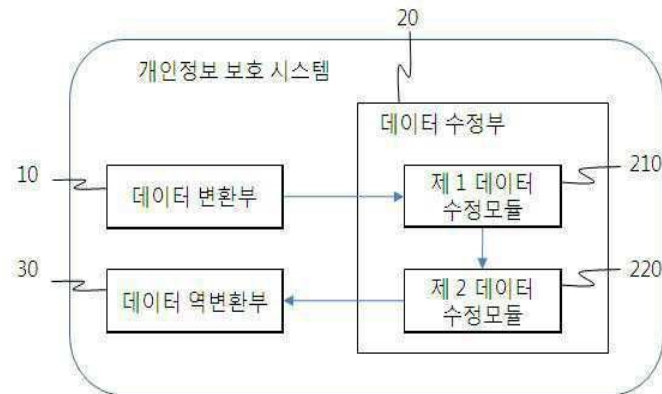
전체 청구항 수 : 총 13 항

(54) 소셜네트워크에서의 개인정보보호 시스템 및 그 방법

(57) 요약

본 발명은 소셜네트워크에서 개인정보보호 시스템 및 방법에 관한 것으로서, 더욱 상세하게는 정점과 간선을 포함하는 소셜네트워크 원시 그래프데이터를 상기 정점이 하나의 레코드로 대응되는 테이블데이터로 변환하는 데이터 변환부와 상기 테이블데이터 내의 상기 정점을 그룹화하고, 실현 가능한 소셜네트워크 그래프데이터 생성을 위해 상기 정점에 연결된 간선의 수인 차수를 수정하여 수정 테이블데이터를 생성하는 데이터 수정부 및 상기 수정된 테이블데이터를 수정된 소셜 네트워크 그래프데이터로 변환하는 데이터 역변환부를 포함하여 정보손실을 최소화하고 개인정보를 보호할수 있는 소셜네트워크에서 개인정보보호 시스템 및 방법에 관한 것이다.

대표도 - 도2



특허청구의 범위

청구항 1

소셜네트워크에서 개인정보보호 시스템에 있어서,

정점과 간선을 포함하는 소셜네트워크 원시 그래프데이터를 상기 정점이 하나의 레코드로 대응되는 테이블데이터로 변환하는 데이터 변환부와;

상기 테이블데이터 내의 상기 정점을 그룹화하고, 실현 가능한 소셜네트워크 그래프데이터 생성을 위해 상기 정점에 연결된 간선의 수인 차수를 수정하여 수정 테이블데이터를 생성하는 데이터 수정부 및;

상기 수정된 테이블데이터를 수정된 소셜 네트워크 그래프데이터로 변환하는 데이터 역변환부를 포함하는 소셜 네트워크에서의 개인정보 보호 시스템.

청구항 2

제 1항에 있어서

상기 테이블 데이터는

특정 개인을 유일하게 식별할 수 있는 데이터인 식별자와;

특정개인과 유일하게 대응되지 않지만 외부정보와 결합되어 특정 개인과 유일하게 대응될 수 있는 정보의 집합인 QI와;

특정 개인과 유일하게 대응될 경우 개인의 프라이버시 문제를 초래하는 데이터인 민감데이터를 항목 데이터로 포함하되,

상기 항목 데이터는 숫자데이터, 계층적 데이터, 차수 데이터 중 선택된 적어도 하나의 데이터를 포함하는 것을 특징으로 하는 소셜네트워크에서의 개인정보보호 시스템.

청구항 3

제 2항에 있어서

상기 데이터 수정부는

상기 테이블데이터 내의 상기 정점의 그룹화를 수행하는 제 1 데이터 수정모듈과;

상기 제 1 데이터 수정모듈에 의해 수정된 테이블데이터에 대해 실현 가능한 소셜네트워크 그래프데이터를 생성하기 위해 상기 정점에 연결된 간선의 수인 차수 수정을 수행하는 제 2 데이터 수정모듈을 포함하는 것을 특징으로 하는 소셜네트워크에서의 개인정보 보호 시스템.

청구항 4

제 3항에 있어서,

상기 제 1 데이터 수정모듈은

상기 테이블데이터에서 차수가 가장 큰 정점을 추출하고 남은 정점 중 상기 추출된 정점과 정보손실 값이 가장 작은 정점을 선택하는 방식으로 k개(여기서, k는 그룹내 포함되는 정점의 수)의 정점을 같은 그룹에 포함시켜 정점의 그룹화를 수행하고,

상기 그룹화된 테이블데이터에서 같은 그룹에 속하는 정점들이 모두 같은 QI와 차수를 갖도록 숫자데이터는 해당그룹에서 가장 작은 값과 큰 값의 범위로 일반화하고, 계층적데이터는 최소 공통조상 값으로 일반화하고, 차

수데이터는 해당 그룹에서 가장 큰 차수값과 모두 같게 수정하는 것을 특징으로 하는 소셜네트워크에서의 개인 정보 보호 시스템.

청구항 5

제 3항에 있어서

상기 제 1 데이터 수정모듈은

상기 차수가 가장 큰 정점과 남은 정점 중 임의적으로 추출된 정점들에 대해 정보손실 값을 산출하는 과정을 반복하여 상기 정보손실 값이 최소가 되는 정점들을 선택하여 그룹화를 수행하되,

상기 정보손실 값은 내용정보손실과 구조정보손실의 합으로 산출되고,

상기 내용정보손실은 하기의 <수학식 1>로 산출하고,

상기 구조정보손실은 하기의 <수학식 2>로 산출하는 것을 특징으로 하는 소셜네트워크에서의 개인정보 보호 시스템.

<수학식 1>

$$AIL(p) = |p| \left(\sum_{i=1}^m \frac{MAX_{N_i} - MIN_{N_i}}{|N_i|} + \sum_{j=1}^n \frac{H(\Lambda(U_{C_j}))}{H(T_{C_j})} \right)$$

여기서, $|p|$ 는 그룹 p 에 속한 정점의 개수, $|N_i|$ 는 연속된 숫자 데이터의 도메인 크기, C_j 는 계층 데이터를 뜻하며, $\Lambda(U_{C_j})$ 는 계층 데이터의 최소공통 조상(least common ancestor)이다. $H(T_{C_j})$ 는 해당 계층 데이터에서 의미 분류트리(Taxonomy tree)의 높이이다.

<수학식 2>

$$SIL(p) = \sum_{i=1}^m \frac{MAX_D - P_{D_i}}{MAX_D - MIN_D}$$

여기서, MAX_D 는 한 그룹 내에서 가장 차수가 큰 정점의 차수이고, MIN_D 는 한 그룹 내에서 차수가 가장 작은 정점의 차수이며, P_{D_i} 는 그룹 p 에서의 각 정점의 차수이다.

청구항 6

제 5항에 있어서,

상기 제 1 데이터 수정모듈은

정점의 그룹화 수행 후 하기의 <수학식 3>에 의해 상기 각 그룹별 내용정보손실의 합(IL_A)과 각 그룹별 구조정보손실의 합(IL_S)을 합산하여 총 정보손실(IL)을 산출하는 것을 특징으로 하는 소셜네트워크에서의 개인정보 보호 시스템.

<수학식 3>

$$IL_A = \frac{1}{N_A} \sum_{p=1}^q AIL(p)$$

$$IL_S = \sum_{p=1}^q SIL(p),$$

$$IL = (\gamma)IL_S + (1-\gamma)IL_A$$

여기서, N_A 는 QI 속성의 개수를 의미하고, γ 는 구조정보와 내용정보의 상대적 중요도를 나타내는 가중치이다.

청구항 7

제 3항에 있어서

상기 제 2 데이터 수정모듈은

상기 제 1 데이터 수정모듈을 통해 수정된 테이블데이터의 각 그룹내 정점차수의 합이 홀수이면 각 정점에 1을 더하여 정점의 합이 짝수가 되도록 차수를 수정하고,

상기 수정을 통해 그룹내 정점차수의 합이 짝수인 경우 실현 가능한 그래프 생성 판별조건인 음수가 아닌 내림차수 정수 $d_1 \leq d_2 \leq \dots \leq d_n$ 에서 모든 상수 $r=1, 2, \dots, n-1$ 에 대하여 하기의 <수학식 4>를 만족하는지 여부를 판단하여, 만족하지 않을 경우 차수의 합이 가장 작은 그룹의 정점차수에 1씩 더하여 상기 실현 가능한 그래프 판별조건을 만족할 때까지 반복적으로 수행하는 것을 특징으로 하는 소셜네트워크에서의 개인정보 보호 시스템.

<수학식 4>

$$\sum_{i=1}^r d_i \leq r(r-1) + \sum_{j=r+1}^n \min(r, d_j)$$

청구항 8

제 7항에 있어서

상기 데이터역변환부는

상기 제 2 데이터 수정모듈을 통해 수정된 테이블데이터의 차수집합(S_r)에서 상기 원시 그래프데이터의 차수 집합(S_o)을 차감한 새롭게 추가된 차수를 의미하는 수정 차수집합(S_s)을 산출하고,

상기 수정 차수집합(S_s) 중 가장 차수가 큰 정점과 두 번째로 차수가 큰 정점 사이에 간선이 없는 경우 간선을 추가하고 해당 정점들의 차수에서 1을 차감하고, 간선이 있는 경우 다음 차수가 큰 정점 사이의 간선이 있는지 여부를 확인하는 방식으로 상기 수정 차수집합(S_s)의 모든 차수가 0이 될 때까지 반복적으로 수행하여 그래프데이터를 생성하는 것을 특징으로 하는 소셜네트워크에서의 개인정보 보호 시스템.

청구항 9

소셜네트워크에서 개인정보를 보호하는 방법에 있어서

- (a) 데이터 변환부가 정점과 간선을 포함하는 소셜네트워크 원시 그래프데이터를 상기 정점이 하나의 레코드로 대응되는 테이블데이터로 변환하는 단계와;
- (b) 데이터 수정부가 상기 테이블데이터 내의 상기 정점을 그룹화하고, 실현 가능한 소셜네트워크 그래프데이터 생성을 위해 상기 정점에 연결된 간선의 수인 차수를 수정하여 수정 테이블데이터를 생성하는 단계 및;
- (c) 데이터 역변환부가 상기 수정 테이블데이터를 수정된 소셜 네트워크 그래프데이터로 변환하는 단계를 포함하는 소셜네트워크에서의 개인정보 보호 방법.

청구항 10

제 9항에 있어서

상기 테이블 데이터는

특정 개인을 유일하게 식별할 수 있는 데이터인 식별자와;

특정 개인과 유일하게 대응되지 않지만 외부정보와 결합되어 특정 개인과 유일하게 대응될 수 있는 정보의 집합인 QI와;

특정 개인과 유일하게 대응될 경우 개인의 프라이버시 문제를 초래하는 데이터인 민감데이터를 항목 데이터로 포함하되,

상기 항목 데이터는 숫자데이터, 계층적 데이터, 차수 데이터 중 선택된 적어도 하나의 데이터를 포함하는 것을 특징으로 하는 소셜네트워크에서의 개인정보 보호 방법.

청구항 11

제 10항에 있어서,

상기 (b) 단계는

상기 테이블데이터에서 차수가 가장 큰 정점을 추출하고 남은 정점 중 상기 추출된 정점과 정보손실 값이 가장 작은 정점을 선택하는 방식으로 k개(여기서, k는 그룹내 포함되는 정점의 수)의 정점을 같은 그룹에 포함시켜 정점의 그룹화를 수행하는 단계와;

상기 그룹화된 테이블데이터에서 같은 그룹에 속하는 정점들이 모두 같은 QI와 차수를 갖도록 숫자데이터는 해당그룹에서 가장 작은 값과 큰 값의 범위로 일반화하고, 계층적데이터는 최소 공통조상 값으로 일반화하고, 차수데이터는 해당 그룹에서 가장 큰 차수값과 모두 같게 수정하는 단계와;

상기 제 1 데이터 수정모듈을 통해 수정된 테이블데이터의 각 그룹내 정점차수의 합이 홀수이면 각 정점에 1을 더하여 정점의 합이 짝수가 되도록 차수를 수정하는 단계 및;

상기 수정을 통해 그룹내 정점차수의 합이 짝수인 경우 실현 가능한 그래프 생성 판별조건인 음수가 아닌 내림차수 정수 $d_1 \leq d_2 \leq \dots \leq d_n$ 에서 모든 상수 $r=1, 2, \dots, n-1$ 에 대하여 하기의 <수학식 5>를 만족하지 않으면 차수의 합이 가장 작은 그룹의 정점차수에 1씩 더하여 상기 실현 가능한 그래프 판별조건을 만족할 때까지 반복적으로 수행하는 단계를 포함하는 것을 특징으로 하는 소셜네트워크에서의 개인정보 보호 방법

<수학식 5>

$$\sum_{i=1}^r d_i \leq r(r-1) + \sum_{j=r+1}^n \min(r, d_j)$$

청구항 12

제 11항에 있어서,

상기 정점의 그룹화를 수행하는 단계는

상기 제 1 데이터 수정모듈이 상기 차수가 가장 큰 정점과 남은 정점 중 임의적으로 추출된 정점들에 대해 정보손실 값을 산출하는 과정을 반복하여 상기 정보손실 값이 최소가 되는 정점들을 선택하여 그룹화를 수행하되,

상기 정보손실 값은 내용정보손실과 구조정보손실의 합으로 산출되고,

상기 내용정보손실은 하기의 <수학식 6>로 산출하고,

상기 구조정보손실은 하기의 <수학식 7>로 산출하는 것을 특징으로 하는 소셜네트워크에서의 개인정보 보호방법.

<수학식 6>

$$AIL(p) = |p| \left(\sum_{i=1}^m \frac{MAX_{N_i} - MIN_{N_i}}{|N_i|} + \sum_{j=1}^n \frac{H(\Lambda(U_{C_j}))}{H(T_{C_j})} \right)$$

여기서, $|p|$ 는 그룹 p 에 속한 정점의 개수, $|N|$ 는 연속된 숫자 데이터의 도메인크기, C_j 는 계층 데이터를 뜻하며, $\Lambda(U_{C_j})$ 는 계층 데이터의 최소공통 조상(least common ancestor)이다. $H(T_{C_j})$ 는 해당 계층 데이터에서 의미 분류트리(Taxonomy tree)의 높이이다.

<수학식 7>

$$SIL(p) = \sum_{i=1}^m \frac{MAX_D - P_{D_i}}{MAX_D - MIN_D}$$

여기서, MAX_D 는 한 그룹 내에서 가장 차수가 큰 정점의 차수이고, MIN_D 는 한 그룹 내에서 차수가 가장 작은 정점의 차수이며, P_{D_i} 는 그룹 p 에서의 각 정점의 차수이다.

청구항 13

제 10항에 있어서

상기 (c) 단계는

상기 데이터역변환부가 상기 제 2 데이터 수정모듈을 통해 수정된 테이블데이터의 차수집합(Sr)에서 상기 원시 그래프데이터의 차수 집합(So)을 차감한 새롭게 추가된 차수를 의미하는 수정 차수집합(Ss)을 산출하는 단계와;

상기 수정 차수집합(Ss) 중 가장 차수가 큰 정점과 두 번째로 차수가 큰 정점 사이에 간선이 없는 경우 간선을 추가하고 해당 정점들의 차수에서 1을 차감하고, 간선이 있는 경우 다음 차수가 큰 정점 사이의 간선이 있는지 여부를 확인하는 방식으로 상기 수정 차수집합(Ss)의 모든 차수가 0이 될 때까지 반복적으로 수행하여 그래프데이터를 생성하는 단계를 포함하는 것을 특징으로 하는 소셜네트워크에서의 개인정보 보호 방법.

명세서

기술분야

[0001] 본 발명은 본 발명은 소셜네트워크에서 개인정보보호 시스템 및 방법에 관한 것으로서, 더욱 상세하게는 정점과 간선을 포함하는 소셜네트워크 원시 그래프데이터를 상기 정점이 하나의 레코드로 대응되는 테이블데이터로 변환하는 데이터 변환부와 상기 테이블데이터 내의 상기 정점을 그룹화하고, 실현 가능한 소셜네트워크 그래프데이터 생성을 위해 상기 정점에 연결된 간선의 수인 차수를 수정하여 수정 테이블데이터를 생성하는 데이터 수정부 및 상기 수정된 테이블데이터를 수정된 소셜 네트워크 그래프데이터로 변환하는 데이터 역변환부를 포함하여 정보손실을 최소화하고 개인정보를 보호할수 있는 소셜네트워크에서 개인정보보호 시스템 및 방법에 관한 것이다.

배경기술

[0002] 최근 들어 LinkedIn, Facebook, Twitter등과 같은 소셜 네트워크 서비스 제공사이트가 늘어나면서 소셜 네트워크상의 개인정보 데이터 양이 늘어나고 개인정보의 유출이 늘어나고 있다.

[0003] 이러한 상황을 방지하기 위해 소셜 네트워크에서 프라이버시를 보호하기 위한 연구들이 최근 진행되고 있다.

[0004] 그러나 나이, 직업, 소속 등을 나타내는 내용정보가 소셜네트워크에서 매우 중요한 정보임에도 불구하고 대부분

의 연구는 소셜 네트워크에서 단지 구조정보만을 고려하고 있다.

- [0005] A Chmpan 등은 구조정보와 정점의 내용정보 모두를 고려한 소셜네트워크에서의 프라이버시 보호기법을 제안하였다. 이 기법은 클러스터링을 기반으로 한 방식으로 정점이나 간선의 추가는 없으나 구조정보의 일부분만을 나타낼 수 있는 단점이 있다.
- [0006] E.Zheleva 등은 구조정보와 간선의 내용정보를 고려한 기법을 제안하였다. 이 기법 또한 클러스터링 기반으로 한 방식을 제안하였으나 정점의 내용정보에 대한 고려는 없다.
- [0007] 데이터 소유자는 데이터를 배포하기 전 데이터에 포함된 개인들의 프라이버시와 관련된 정보 보호를 위해 데이터를 수정해야 한다. 데이터 소유자가 데이터를 수정할 때 두가지 상황에 대한 고려가 필요하다. 첫째로 수정된 데이터가 배포되었을 때 개인의 프라이버시가 보호되어야 하며, 둘째는 원시데이터와 수정된 데이터의 차이를 최소화하여 데이터의 유용성을 최대화하는 것이다. 기존 연구들은 대부분 k-anonymity개념을 기본으로 하여 프라이버시 보호기법을 제시하였으며, 각 데이터 모델의 정의에 따라 데이터 수정 정도를 최소화하도록 노력하였다.
- [0008] 많은 연구들이 테이블로 표현된 관계형 데이터에서 프라이버시 보호기법을 제안하고 있으며 현재도 지속적으로 연구되고 있는 상황이다.
- [0009] K.Sweeney 등은 단순히 이름이나 주민등록번호 등의 식별자를 제거하고 데이터를 제공하는 것만으로는 프라이버시 보호에 충분하지 않다고 지적했다. 그들은 k-anonymity모델을 제시하여 데이터 일반화를 통한 프라이버시 보호개념을 만들었다.
- [0010] A.Machanavajjhala등은 k-anonymity 모델의 단점인 민감한 데이터의 다양성 부족을 지적하고 이를 방지하기 위한 새로운 모델인 1-diversity를 제안하였고
- [0011] N.Liu 등은 t-closeness를 통해 민감한 데이터의 문자적 다양성 뿐만 아니라 의미적 다양성도 함께 고려한 모델을 제시하였다.
- [0012] X.Xiao 등은 일시적 데이터 뿐만 아니라 시간이 지남에 따라 지속적으로 변화하는 데이터에서 프라이버시 보호를 위한 기법인 m-invariance를 제안하였다. 이 기법은 데이터가 추가/삭제되는 상황에서도 프라이버시 보호를 위한 기법으로 데이터 일반화와 모조 데이터(Counterfeit)추가를 통해 프라이버시를 보호하였다.
- [0013] 그러나 이러한 기법들은 그래프로 표현되는 소셜네트워크에 직접적으로 적용될수 없다. 때문에 최근 들어 소셜 네트워크 데이터에서 프라이버시 보호를 위한 기법이 많이 연구되고 있다.
- [0014] L.Backstrom 등은 처음으로 익명화된 소셜네트워크 그래프에서 프라이버시 노출문제를 지적하였다. 그래프는 구조적 특성을 가지기 때문에 구조정보를 고려한 기법이 필요하다고 하였으며 능동적 공격(Active attack)과 수동적 공격(Passive attack)의 새로운 공격모델을 제시하였다.
- [0015] 구조정보 중 부분그래프(Subgraph)를 공격자(Adversary)가 알고 있을 때 프라이버시를 보호하기 위한 기법으로 B.Zou 등 이웃공격(Neighborrhood attack) 모델을 제시하고 이에 대한 보호기법을 제안하였다. 그들은 공격자가 특정 정점과 그와 연결된 이웃들로 이루어진 부분그래프를 알고 있을 때, 전체 그래프에서 해당 부분그래프와 동형 이성(Isomorphism)인 부분 그래프가 최소 k-1개 이상 나타나야 한다는 모델을 제시하였다.
- [0016] 또 하나의 구조정보로 K.Liu 등은 공격자가 정점의 차수를 알고 있을때 프라이버시 보호를 위한 기법을 제안하였다. 그들은 전체 그래프에서 같은 차수를 가지는 정점이 최소 k 개이상 되어야 하는 k-degree anonymity를 제시하고 이를 위한 간선추가 알고리즘을 개발하였다.
- [0017] M.Hay 등은 정점의 이웃그래프와 차수를 동시에 고려한 모델을 제안하였다. 그들은 클러스터링 기반 기법을 이용하여 유사한 정점과 간선을 통합하여 개별 정점과 간선의 프라이버시가 드러나지 않게 하였다. 그들은 k-candidate anonymity을 통해 어떤 질의에도 k 개의 후보 부분그래프가 나타나게 하는 기법을 제시하였다.
- [0018] L.Zou 등도 정점의 이웃그래프와 차수를 동시에 고려한 기법을 제안하였다. 그들은 클러스터링 기반의 기법이 너무 많은 데이터의 구조정보 손실을 초래한다고 주장하며 그래프 수정 기반의 프라이버시 보호기법을 제시하였다. 또한 이 기법은 일시적 데이터 뿐만 아니라 지속적으로 변화하는 데이터까지 처리할 수 있는 특징이 있다.
- [0019] 위에 소개된 관련연구는 소셜네트워크 데이터의 구조정보만 고려하여 공격자가 구조적 정보만 알고 있을 때의 프라이버시 보호만 할 수 있는 단점이 있다. 하지만, 공격자가 소셜네트워크의 구조적 정보만 가지고 있다는 가

정은 현실적이지 않다.

- [0020] 이에 반해 E.Zheleva등은 소셜네트워크에서 각 간선이 가지는 내용정보를 보호하기 위한 기법을 제안하였다. 이를 위해 그들은 클러스터링 기반 기법을 사용하여 다섯 가지 간선 내용정보 보호방법 제시 및 실험을 통해 각 상황에서 프라이버시 노출 정도를 측정하였다.
- [0021] A.Campan 등은 각 정점이 가지는 내용정보를 보호하기 위한 클러스터링 기반기법을 제안하였다. 그들은 유사한 내용정보와 구조정보를 가진 정점들을 하나로 묶어 각 값들을 대표하는 하나의 슈퍼정점을 만들어 슈퍼정점 간의 관계만을 나타내어 원시 데이터의 축약된 구조정보만을 나타내는 방식을 제안하였다.
- [0022] 그러나 종래의 보호 기법은 원시 데이터의 구조정보를 지나치게 많이 왜곡시키고, 개별 정점 간의 관계를 알 수 없으며, 개별 정점과 내용정보 사이의 관계도 알 수 없는 문제가 있었다.
- [0023] 또한, 소셜네트워크 데이터의 구조정보만 고려하여 공격자가 구조적 정보만 알고 있을 때의 프라이버시 보호에만 치우치는 문제가 있었다.

발명의 내용

해결하려는 과제

- [0024] 상기와 같은 문제점을 해결하기 위해 안출된 것으로서 본 발명의 목적은 구조정보뿐만 아니라 내용정보를 고려한 정보손실을 최소화하면서, 정점을 그룹화하고 같은 QI와 같은 차수를 가진 정점을 소셜네트워크 그래프에서 최소 k개(여기서, k는 그룹 내에 포함되는 정점의 수) 이상 가지게 원시데이터를 수정하여 소셜네트워크 개인의 프라이버시 보호가 가능한 소셜네트워크에서 프라이버시 보호 시스템 및 그 방법을 제공하는 데 있다.

과제의 해결 수단

- [0025] 상기와 같은 목적을 달성하기 위해 본 발명에 따른 소셜네트워크에서 개인정보 보호 시스템은 정점과 간선을 포함하는 소셜네트워크 원시 그래프데이터를 상기 정점이 하나의 레코드로 대응되는 테이블데이터로 변환하는 데이터 변환부와 상기 테이블데이터 내의 상기 정점을 그룹화하고, 실현 가능한 소셜네트워크 그래프데이터 생성을 위해 상기 정점에 연결된 간선의 수인 차수를 수정하여 수정 테이블데이터를 생성하는 데이터 수정부 및 상기 수정된 테이블데이터를 수정된 소셜 네트워크 그래프데이터로 변환하는 데이터 역변환부를 포함하는 것을 특징으로 한다.
- [0026] 여기서, 상기 테이블 데이터는 특정 개인을 유일하게 식별할 수 있는 데이터인 식별자와 특정개인과 유일하게 대응되지 않지만 외부정보와 결합되어 특정 개인과 유일하게 대응될 수 있는 정보의 집합인 QI와 특정 개인과 유일하게 대응될 경우 개인의 프라이버시 문제를 초래하는 데이터인 민감데이터를 항목 데이터로 포함하되, 상기 항목 데이터는 숫자데이터, 계층적 데이터, 차수 데이터 중 선택된 적어도 하나의 데이터를 포함하는 것을 특징으로 한다.
- [0027] 상기 데이터 수정부는 상기 테이블데이터 내의 상기 정점의 그룹화를 수행하는 제 1 데이터 수정모듈과 상기 제 1 데이터 수정모듈에 의해 수정된 테이블데이터에 대해 실현 가능한 소셜네트워크 그래프데이터를 생성하기 위해 상기 정점에 연결된 간선의 수인 차수 수정을 수행하는 제 2 데이터 수정모듈을 포함하는 것을 특징으로 한다.
- [0028] 상기 제 1 데이터 수정모듈은 상기 테이블데이터에서 차수가 가장 큰 정점을 추출하고 남은 정점 중 상기 추출된 정점과 정보손실 값이 가장 작은 정점을 선택하는 방식으로 k개(여기서, k는 그룹내 포함되는 정점의 수)의 정점을 같은 그룹에 포함시켜 정점의 그룹화를 수행하고, 상기 그룹화된 테이블데이터에서 같은 그룹에 속하는 정점들이 모두 같은 QI와 차수를 갖도록 숫자데이터는 해당그룹에서 가장 작은 값과 큰 값의 범위로 일반화하고, 계층적데이터는 최소 공통조상 값으로 일반화하고, 차수데이터는 해당 그룹에서 가장 큰 차수값과 모두 같게 수정하는 것을 특징으로 한다.
- [0029] 여기서, 상기 제 1 데이터 수정모듈은 상기 차수가 가장 큰 정점과 남은 정점 중 임의적으로 추출된 정점들에 대해 정보손실 값을 산출하는 과정을 반복하여 상기 정보손실 값이 최소가 되는 정점들을 선택하여 그룹화를 수

행하되, 상기 정보손실 값은 내용정보손실과 구조정보손실의 합으로 산출되고, 상기 내용정보손실은 하기의 <수학식 1>로 산출하고, 상기 구조정보손실은 하기의 <수학식 2>로 산출하는 것을 특징으로 한다.

[0030] <수학식 1>

$$AIL(p) = |p| \left(\sum_{i=1}^m \frac{MAX_{N_i} - MIN_{N_i}}{|N_i|} + \sum_{j=1}^n \frac{H(\Lambda(U_{C_j}))}{H(T_{C_j})} \right)$$

[0032] 여기서, $|p|$ 는 그룹 p에 속한 정점의 개수, $|N|$ 는 연속된 숫자 데이터의 도메인 크기, C_j 는 계층 데이터를 뜻하며, $\Lambda(U_{C_j})$ 는 계층 데이터의 최소공통 조상(least common ancestor)이다. $H(T_{C_j})$ 는 해당 계층 데이터에서 의미 분류트리(Taxonomy tree)의 높이이다.

[0033] <수학식 2>

$$SIL(p) = \sum_{i=1}^m \frac{MAX_{D_i} - P_{D_i}}{MAX_D - MIN_D}$$

[0035] 여기서, MAX_D 는 한 그룹 내에서 가장 차수가 큰 정점의 차수이고, MIN_D 는 한 그룹 내에서 차수가 가장 작은 정점의 차수이며, P_{D_i} 는 그룹 p에서의 각 정점의 차수이다.

[0036] 또한, 상기 제 1 데이터 수정모듈은 정점의 그룹화 수행 후 하기의 <수학식 3>에 의해 상기 각 그룹별 내용정보손실의 합(IL_A)과 각 그룹별 구조정보손실의 합(IL_S)을 합산하여 총 정보손실(IL)을 산출하는 것을 특징으로 한다.

[0037] <수학식 3>

$$IL_A = \frac{1}{N_A} \sum_{p=1}^q AIL(p)$$

$$IL_S = \sum_{p=1}^q SIL(p)$$

$$IL = (\gamma)IL_S + (1-\gamma)IL_A$$

[0041] 여기서, N_A 는 QI 속성의 개수를 의미하고, γ 는 구조정보와 내용정보의 상대적 중요도를 나타내는 가중치이다.

[0042] 그리고 상기 제 2 데이터 수정모듈은 상기 제 1 데이터 수정모듈을 통해 수정된 테이블데이터의 각 그룹내 정점 차수의 합이 홀수이면 각 정점에 1을 더하여 정점의 합이 짝수가 되도록 차수를 수정하고, 상기 수정을 통해 그룹내 정점차수의 합이 짝수인 경우 실현 가능한 그래프 생성 판별조건인 음수가 아닌 내림차수 정수 $d_1 \leq d_2 \leq \dots \leq d_n$ 에서 모든 상수 $r=1, 2, \dots, n-1$ 에 대하여 하기의 <수학식 4>를 만족하는지 여부를 판단하여, 만족하지 않을 경우 차수의 합이 가장 작은 그룹의 정점차수에 1씩 더하여 상기 실현 가능한 그래프 판별조건을 만족할 때까지 반복적으로 수행하는 것을 특징으로 한다.

[0043] <수학식 4>

$$\sum_{i=1}^r d_i \leq r(r-1) + \sum_{j=r+1}^n \min(r, d_j)$$

[0045] 또한, 상기 데이터 역변환부는 상기 제 2 데이터 수정모듈을 통해 수정된 테이블데이터의 차수집합(S_r)에서 상기 원시 그래프데이터의 차수 집합(S_o)을 차감한 새롭게 추가된 차수를 의미하는 수정 차수집합(S_s)을 산출하고, 상기 수정 차수집합(S_s) 중 가장 차수가 큰 정점과 두 번째로 차수가 큰 정점 사이에 간선이 없는 경우 간선을 추가하고 해당 정점들의 차수에서 1을 차감하고, 간선이 있는 경우 다음 차수가 큰 정점 사이의 간선이 있는지 여부를 확인하는 방식으로 상기 수정 차수집합(S_s)의 모든 차수가 0이 될 때까지 반복적으로 수행하

여 그래프데이터를 생성하는 것을 특징으로 한다.

[0046] 한편, 본 발명에 따른 소셜네트워크에서 개인정보 보호 방법은 (a) 데이터 변환부가 정점과 간선을 포함하는 소셜네트워크 원시 그래프데이터를 상기 정점이 하나의 레코드로 대응되는 테이블데이터로 변환하는 단계와 (b) 데이터 수정부가 상기 테이블데이터 내의 상기 정점을 그룹화하고, 실현 가능한 소셜네트워크 그래프데이터 생성을 위해 상기 정점에 연결된 간선의 수인 차수를 수정하여 수정 테이블데이터를 생성하는 단계 및 (c) 데이터 역변환부가 상기 수정 테이블데이터를 수정된 소셜 네트워크 그래프데이터로 변환하는 단계를 포함하는 것을 특징으로 한다.

[0047] 여기서, 상기 테이블 데이터는 특정 개인을 유일하게 식별할 수 있는 데이터인 식별자와 특정개인과 유일하게 대응되지 않지만 외부정보와 결합되어 특정 개인과 유일하게 대응될 수 있는 정보의 집합인 QI와 특정 개인과 유일하게 대응될 경우 개인의 프라이버시 문제를 초래하는 데이터인 민감데이터를 항목 데이터로 포함하되, 상기 항목 데이터는 숫자데이터, 계층적 데이터, 차수 데이터 중 선택된 적어도 하나의 데이터를 포함하는 것을 특징으로 한다.

[0048] 그리고 상기 (b) 단계는 상기 테이블데이터에서 차수가 가장 큰 정점을 추출하고 남은 정점 중 상기 추출된 정점과 정보손실 값이 가장 작은 정점을 선택하는 방식으로 k개(여기서, k는 그룹내 포함되는 정점의 수)의 정점을 같은 그룹에 포함시켜 정점의 그룹화를 수행하는 단계와 상기 그룹화된 테이블데이터에서 같은 그룹에 속하는 정점들이 모두 같은 QI와 차수를 갖도록 숫자데이터는 해당그룹에서 가장 작은 값과 큰 값의 범위로 일반화하고, 계층적데이터는 최소 공통조상 값으로 일반화하고, 차수데이터는 해당 그룹에서 가장 큰 차수값과 모두 같게 수정하는 단계와 상기 제 1 데이터 수정모듈을 통해 수정된 테이블데이터의 각 그룹내 정점차수의 합이 홀수이면 각 정점에 1을 더하여 정점의 합이 짝수가 되도록 차수를 수정하는 단계 및 상기 수정을 통해 그룹내 정점차수의 합이 짝수인 경우 실현 가능한 그래프 생성 판별조건인 음수가 아닌 내림차수 정수 $d_1 \leq d_2 \leq \dots \leq d_n$ 에서 모든 상수 $r=1, 2, \dots, n-1$ 에 대하여 하기의 <수학식 5>를 만족하지 않으면 차수의 합이 가장 작은 그룹의 정점차수에 1씩 더하여 상기 실현 가능한 그래프 판별조건을 만족할 때까지 반복적으로 수행하는 단계를 포함하는 것을 특징으로 한다.

[0049] <수학식 5>

$$\sum_{i=1}^r d_i \leq r(r-1) + \sum_{j=r+1}^n \min(r, d_j)$$

[0051] 여기서, 상기 정점의 그룹화를 수행하는 단계는 상기 제 1 데이터 수정모듈이 상기 차수가 가장 큰 정점과 남은 정점 중 임의적으로 추출된 정점들에 대해 정보손실 값을 산출하는 과정을 반복하여 상기 정보손실 값이 최소가 되는 정점들을 선택하여 그룹화를 수행하되, 상기 정보손실 값은 내용정보손실과 구조정보손실의 합으로 산출되고, 상기 내용정보손실은 하기의 <수학식 6>로 산출하고, 상기 구조정보손실은 하기의 <수학식 7>로 산출하는 것을 특징으로 한다.

[0052] <수학식 6>

$$AIL(p) = |p| \left(\sum_{i=1}^m \frac{MAX_{N_i} - MIN_{N_i}}{|N_i|} + \sum_{j=1}^n \frac{H(\Lambda(U_{C_j}))}{H(T_{C_j})} \right)$$

[0054] 여기서, $|p|$ 는 그룹 p에 속한 정점의 개수, $|N|$ 는 연속된 숫자 데이터의 도메인 크기, C_j 는 계층 데이터를 뜻하며, $\Lambda(U_{C_j})$ 는 계층 데이터의 최소공통 조상(least common ancestor)이다. $H(T_{C_j})$ 는 해당 계층 데이터에서 의미 분류트리(Taxonomy tree)의 높이이다.

[0055] <수학식 7>

$$SIL(p) = \sum_{i=1}^m \frac{MAX_D - MIN_D}{MAX_D - MIN_D}$$

[0057] 여기서, MAX_D 는 한 그룹 내에서 가장 차수가 큰 정점의 차수이고, MIN_D 는 한 그룹 내에서 차수가 가장

작은 정점의 차수이며, P_{D_i} 는 그룹 p에서의 각 정점의 차수이다.

- [0058] 그리고 상기 (c) 단계는 상기 데이터 역변환부가 상기 제 2 데이터 수정모듈을 통해 수정된 테이블데이터의 차수집합(S_r)에서 상기 원시 그래프데이터의 차수 집합(S_o)을 차감한 새롭게 추가된 차수를 의미하는 수정 차수집합(S_s)을 산출하는 단계와 상기 수정 차수집합(S_s) 중 가장 차수가 큰 정점과 두 번째로 차수가 큰 정점 사이에 간선이 없는 경우 간선을 추가하고 해당 정점들의 차수에서 1을 차감하고, 간선이 있는 경우 다음 차수가 큰 정점 사이의 간선이 있는지 여부를 확인하는 방식으로 상기 수정 차수집합(S_s)의 모든 차수가 0이 될 때까지 반복적으로 수행하여 그래프데이터를 생성하는 단계를 포함하는 것을 특징으로 한다.

발명의 효과

- [0059] 본 발명에 따른 소셜네트워크에서 개인정보 보호 시스템 및 그 방법은 공격자가 특정 개인에 대한 정보를 가지고 있다 하더라도 최소 k개의 후보정점이 생겨 민감한 정보에 대한 정확한 예측을 할 수 없게 된다. 따라서, 수정된 소셜네트워크 그래프는 정보의 수정과정에서 발생하는 왜곡을 최소화하고 개인의 내용정보와 구조정보를 공격자로부터 보호할수 있는 탁월한 효과가 발생한다.

도면의 간단한 설명

- [0060] 도 1 은 소셜네트워크에서 사용자의 구조정보와 내용정보를 포함하고 있는 그래프 데이터의 예시이다.
- 도 2는 본 발명의 바람직한 실시예에 따른 소셜네트워크에서 개인정보 보호 시스템을 개략적으로 도시한 시스템 구성도이다.
- 도 3은 본 발명의 바람직한 실시예에 따른 소셜네트워크에서 개인정보 보호방법을 개략적으로 도시한 순서도이다.
- 도 4는 ZIP과 성별에 대한 의미분류 트리의 예시이다.
- 도 5는 본 발명의 바람직한 실시예에 따라 생성된 소셜네트워크 그래프를 나타낸다.

발명을 실시하기 위한 구체적인 내용

- [0061] 이하, 본 발명의 구체적인 실시예에 대하여 도면을 참조하여 상세하게 설명하기로 한다.
- [0062] 도 1은 소셜네트워크 그래프 데이터에 대한 예시이다.
- [0063] 도 1을 참조하면, 소셜네트워크는 그래프로 나타낼 수 있으며 그래프는 $G(V,E)$ 로 나타내어 V 는 정점의 집합을, E 는 간선의 집합을 각각 의미한다.
- [0064] 소셜네트워크에서 각 개인은 하나의 정점에 대응되며, 개인 사이의 관계는 간선으로 나타낸다. 또한, 각 정점은 이름, 나이, 성별, 직업 등의 내용정보를 가지고 있다. 상기 정점에 포함된 내용정보는 식별자(Identifier), QI(Quasi-Identifier), 민감한 데이터(sensitive data)로 분류될 수 있다.
- [0065] 상기 식별자(Identifier)는 개인의 이름이나 주민등록번호 등 개개인을 판별할 수 있는 식별 데이터이고, 상기 QI(Quasi-Identifier)는 나이, 성별, 직업 등 직접적으로 개개인을 판별할 수는 없지만 다른 외부 정보와 합쳐져 개개인을 판별할 수 있는 데이터를 의미하고, 상기 민감한 데이터(sensitive data)는 식별자와 유일하게 대응되어 특정 개인에 대한 정보를 드러내면 안되는 데이터, 즉 질병, 재산 등 타인에게 공개될 경우 개인의 프라이버시 문제를 초래하게 되는 데이터를 의미한다.
- [0066] 이하, 상기와 같이 내용 정보에 대한 공격모델에 대해 설명하기로 한다.
- [0067] 공격자는 소셜네트워크 그래프 데이터에서 특정 개인의 데이터가 있는지 없는지 알고 있고, 특정 개인의 민감한 정보를 제외한 모든 정보(식별자와 QI)를 알고 있으며, 그래프에서 특정 개인의 차수를 알고 있다고 가정할 경우 공격자의 정보에 의하여 특정 개인에 대한 민감한 정보를 추측할 때 $1/k$ 확률보다 큰 확률을 가지고 결정할 수 있을 때 즉, 특정 개인에 대응하는 민감한 정보가 k개 미만일 때 프라이버시가 노출되었다고 정의할 수 있다.

- [0068] 보다 구체적으로, 전체 소셜네트워크 G , 민감한 정보의 집합 $S=\{s_1, \dots, s_n\}$, 공격자의 정보 A_k 에 대해서 $f(G, A_k) = x$ 에서 $x \in S$ 인 x 가 k 개 미만이면 프라이버시가 노출되었다고 한다.
- [0069] 예를 들어, 도 1을 소셜네트워크 G 라고 할 경우 각 정점은 개인을 나타내며 식별자 대신 각 정점을 구분하는 번호로 구별되어 있다. 즉, 이름과 같은 식별자는 데이터에 나타나 있지 않다. 그리고 각 정점은 내용정보를 포함하고 있으며, 각각 나이, 성별, ZIP, 연봉을 의미한다. 여기서, 연봉을 민감한 정보라고 할 경우 공격자가 영희라는 사람이 소셜네트워크 G 에 포함되어 있는 23세의 여성이고 20437의 ZIP을 가졌으며, 소셜네트워크 상에서 차수가 1이라는 것을 알면 그래프 G 에서 해당 정보와 정점1이 유일하게 대응되어, 정점 1에 대응하는 것이 영희라는 것을 알 수 있다.
- [0070] 상기와 같이 영희에 대한 식별자가 그래프 G 에 나타나 있지 않더라도 영희의 연봉이 2000만원이라는 것을 공격자가 알게 되어 영희에 대한 프라이버시가 노출될 수 있다.
- [0071] 상기와 같이 내용정보와 구조정보가 같이 표현된 소셜네트워크 그래프 G 에서 그래프 자체로는 프라이버시가 드러나지 않더라도 공격자가 가진 정보에 의해 프라이버시 노출이 가능하므로 데이터 제공자는 공격자가 자신의 정보를 가지고 특정인의 민감한 정보를 추측했을 때, $1/k$ 확률보다 큰 확률로 추측할 수 없게 되도록 데이터를 수정한 후 데이터를 제공하여야 프라이버시 노출을 방지할 수 있다.
- [0072] 본 발명은 소셜네트워크에서 프라이버시 노출을 방지하기 위해 정보 손실을 최소화하면서 소셜네트워크 원시 그래프 데이터를 같은 QI와 같은 차수를 가진 정점이 최소한 K 개 이상 가지도록 상기 원시 그래프 데이터를 수정하는 것을 특징으로 한다.
- [0073] 도 2는 본 발명의 바람직한 실시예에 따른 소셜네트워크에서 개인정보 보호 시스템을 개략적으로 도시한 시스템 구성도이다.
- [0074] 도 2를 참조하면, 본 발명에 따른 소셜네트워크에서 개인정보 보호 시스템은 데이터변환부(10), 데이터수정부(20), 데이터역변환부(30)를 포함하여 구성될 수 있다.
- [0075] 상기 데이터변환부(10)는 정점과 간선을 포함하는 소셜네트워크 원시 그래프데이터를 상기 정점이 하나의 레코드로 대응되는 테이블데이터로 변환하는 역할을 담당한다.
- [0076] 상기 데이터 수정부(20)는 상기 테이블데이터 내의 상기 정점을 그룹화하고, 실현 가능한 소셜네트워크 그래프 데이터 생성을 위해 상기 정점에 연결된 간선의 수인 차수를 수정하여 수정 테이블데이터를 생성하는 역할을 담당한다.
- [0077] 상기 데이터 수정부(20)는 상기 테이블데이터 내의 상기 정점에 대한 그룹화를 수행하는 제 1 데이터 수정모듈(210)과 실현 가능한 소셜네트워크 그래프데이터 생성을 위해 상기 정점에 연결된 간선의 수인 차수를 수정하는 제 2 데이터 수정모듈(220)로 구분될 수 있다.
- [0078] 상기 데이터 역변환부(30)는 상기 수정된 테이블데이터를 수정된 소셜 네트워크 그래프데이터로 변환하는 역할을 담당한다.
- [0079] 이하, 본 발명의 구체적인 소셜네트워크에서 개인정보 보호방법에 대해 구체적으로 살펴보기로 한다.
- [0080] 도 3은 본 발명의 바람직한 실시예에 따른 소셜네트워크에서 개인정보 보호방법을 개략적으로 도시한 순서도이다.
- [0081] 도 3을 참조하면, 먼저 데이터 변환부(10)가 소셜네트워크 그래프 데이터를 테이블 데이터로 변환한다.(S110)
- [0082] 이를 통해, 소셜네트워크 그래프를 각 정점이 하나의 레코드(Record)에 대응되는 테이블 데이터가 생성되고, 각 정점의 QI, 민감한 데이터, 차수가 각각 속성을 이룬다.
- [0083] 보다 구체적으로 도 1의 소셜네트워크 그래프 데이터를 테이블 데이터로 변환하면 하기의 <표 1>과 같다.

표 1

[0084]

ID	나이	성별	ZIP	연봉	차수
1	23	Female	20437	2000	1
2	51	Female	20372	3500	1

3	26	Male	20372	1000	5
4	45	Male	20412	3000	2
5	31	Female	20412	1500	1
6	39	Male	20343	4000	2

- [0085] 상기 <표 1>을 통해 알 수 있듯이, 도 1의 소셜네트워크 그래프 데이터의 각 정점 번호를 ID로 매핑시키고, 상기 정점의 내용 정보인 나이, 성별, ZIP, 연봉, 차수정보가 테이블로 매핑된다.
- [0086] 이어서, 데이터 수정부의 제 1 데이터 수정모듈이 정보손실을 고려하여 각 정점들의 그룹화를 수행한다.(S120)
- [0087] 상기 정보손실은 내용정보손실과 구조정보손실로 이루어진다. 내용정보손실은 나이, 직업과 같은 QI를 일반화하여 수정하는 과정에서 발생하는 손실이며, 구조정보손실은 소셜네트워크 그래프에서 같은 차수를 가진 정점을 만들기 위해 간선을 수정하는 과정에서 생기는 손실이다.
- [0088] 같은 내용정보와 구조정보를 가진 정점을 k개 이상 만들기 위해서는 k개 이상의 정점을 한 그룹으로 묶어 해당 그룹이 QI와 차수를 각각 같게 만들어 주면 된다.
- [0089] 이때 그래프 상의 모든 정점을 한 그룹으로 묶어 QI와 차수를 각각 같게 만든다면 프라이버시 보호는 되지만 정보 유용성은 매우 떨어진다. 즉 k개 이상의 정점을 한 그룹으로 묶을 때 해당 그룹에 속하는 정점의 수는 k이상의 최소 정수가 되어 수정되는 정보의 양을 최대한 줄이는 것이 좋다.
- [0090] 상기 제 1 데이터 수정모듈은 상기 테이블데이터에서 차수가 가장 큰 정점을 추출하고 남은 정점 중 상기 추출된 정점과 정보손실 값이 가장 작은 정점을 선택하는 방식으로 k개(여기서, k는 그룹내 포함되는 정점의 수)의 정점을 같은 그룹에 포함시켜 정점의 그룹화를 수행한다.
- [0091] 즉, 차수가 가장 큰 정점과 나머지 정점들을 순차적으로 선택하여 정점들 사이의 정보손실을 산출하여 정보손실이 최소가 되는 정점을 추출하여 그룹화를 수행한다.
- [0092] 상기 정보손실은 내용정보손실[AIL(P)]과 구조정보손실[SIL(P)]의 합으로 산출된다.
- [0093] 여기서, 내용정보는 나이와 같은 데이터의 범위가 연속적인 숫자에 한정된 데이터로 숫자데이터와 직업, 국가와 같은 상위에는 추상적인 개념이 들어가고 하위로 갈수록 구체적인 개념이 들어가는 계층적 데이터로 분류될 수 있다. 도 4는 ZIP과 성별에 대한 의미 분류 트리의 예시이다.
- [0094] 상기 내용정보손실은 연속된 숫자데이터 손실과 계층적 데이터 손실의 합으로 나타나며 하기의 <수학식 1>에 의해 산출될 수 있다.

수학식 1

$$AIL(p)=|p|(\sum_{i=1}^m \frac{MAX_{N_i}-MIN_{N_i}}{|N_i|} + \sum_{j=1}^n \frac{H(\Lambda(U_{C_j}))}{H(T_{C_j})})$$

- [0095]
- [0096] 여기서, $|p|$ 는 그룹 p에 속한 정점의 개수, $|N|$ 는 연속된 숫자 데이터의 도메인 크기, C_j 는 계층 데이터를 뜻하며, $\Lambda(U_{C_j})$ 는 계층 데이터의 최소공통 조상(least common ancestor)이다. $H(T_{C_j})$ 는 해당 계층 데이터에서 의미 분류트리(Taxonomy tree)의 높이이다.
- [0097] 그리고 상기 구조정보 손실 측정을 위해서는 추가된 간선의 개수를 측정하고, 간선의 추가가 많을수록 구조정보 손실이 크고 간선의 추가가 적을수록 구조정보 손실이 작다. 상기 구조정보 손실은 하기의 <수학식 2>에 의해 산출될 수 있다.

수학식 2

$$SIL(p) = \sum_{i=1}^m \frac{MAX_D - P_{D_i}}{MAX_D - MIN_D}$$

[0098]

[0099] 여기서, MAX_D 는 한 그룹 내에서 가장 차수가 큰 정점의 차수이고, MIN_D 는 한 그룹 내에서 차수가 가장 작은 정점의 차수이며, P_{D_i} 는 그룹 p에서의 각 정점의 차수이다.

[0100] 상기 정보손실을 고려하여 상기 <표 1>의 각 정점에 대해 그룹화(k=2라고 가정, k는 그룹 내에 포함되는 정점의 수)를 수행하면 정점ID {1,5}, {3,6}, {2,4}로 그룹핑될 수 있다.

[0101] 상기와 같이 정점의 그룹화가 수행되면, 제 1 데이터 수정모듈은 총 정보손실을 산출할 수 있다.

[0102] 총 정보손실은 총 내용정보손실의 합(IL_A)과 총 구조정보손실의 합(IL_S)을 가산하여 산출될 수 있다.

[0103] 상기 총 내용정보손실의 합(IL_A)은 하기의 <수학식 3>에 의해 산출될 수 있다.

수학식 3

$$IL_A = \frac{1}{N_A} \sum_{p=1}^q AIL(p)$$

[0104]

[0105] 여기서, N_A 는 QI 속성의 개수를, p는 각 그룹을 순차적으로 뜻하고, $AIL(p)$ 는 각 그룹별 정보손실을 의미한다.

[0106] 상기 총 구조정보손실의 합(IL_S)은 하기의 <수학식 4>에 의해 산출될 수 있다.

수학식 4

$$IL_S = \sum_{p=1}^q SIL(p)$$

[0107]

[0108] 여기서, SIL(P)는 각 그룹별 구조정보손실의 합을 의미한다.

[0109] 따라서, 총 정보손실 (Information loss) IL 은 하기의 수학식 5로 표현될 수 있다.

수학식 5

$$IL = IL_S + IL_A$$

[0110]

[0111] 상기 <수학식 4>는 내용정보손실과 구조정보 손실에 대한 가중치를 똑같이 고려한 것이다. 그러나 사용자에게 따라 내용정보를 더 중요시 여기거나 구조정보를 더 필요로 하는 요구가 있을 수 있다.

[0112] 때문에 가중치 값(Weight-Parameter) γ 를 정의하여 사용자의 요구에 따라 내용정보와 구조정보 중 더 중요하게 여기는 정보를 유지할 수 있게 한다. γ 이 클수록 구조정보를 더 중요하게 여기는 새로운 총 정보 손실 IL 은 하기의 <수학식 6>과 같다.

수학식 6

$$IL = (\gamma)IL_S + (1-\gamma)IL_A$$

만약 $\gamma=1$ 이라면 내용정보 손실에 대한 고려 없이 구조정보의 손실만 고려한 결과를 얻을 수 있다. 즉 구조정보손실만 고려한 결과를 얻을 수 있다. 즉 구조정보의 손실이 최소화된 결과가 나온다.

본 발명의 바람직한 실시예에 따라 그룹화된 정점ID {1,5}, {3,6}, {2,4}에 대하여, γ 이 0.5인 경우 차례대로 그룹 1, 2, 3이라고 가정하여 총 내용정보손실, 총 구조정보손실 및 총 정보손실은 하기와 같이 산출될 수 있다.

<총 내용정보손실>

$A_{IL}(1)$	$ 2 \cdot \left(\frac{31-23}{51-23} + \frac{0}{1} + \frac{1}{2} \right) = \frac{22}{14}$	$IL_A(G) = \frac{1}{3} \cdot \left(\frac{22}{14} + \frac{55}{14} + \frac{16}{14} \right) = \frac{31}{14}$
$A_{IL}(2)$	$ 2 \cdot \left(\frac{39-26}{51-23} + \frac{1}{1} + \frac{1}{2} \right) = \frac{55}{14}$	
$A_{IL}(3)$	$ 2 \cdot \left(\frac{51-45}{51-23} + \frac{1}{1} + \frac{2}{2} \right) = \frac{62}{14}$	

<총 구조정보손실>

$S_{IL}(1)$	$0+0=0$	$IL_S(G) = \left(0 + \frac{3}{3} + \frac{1}{1} \right) = 2$
$S_{IL}(2)$	$\frac{0}{5-2} + \frac{3}{5-2} = \frac{3}{3}$	
$S_{IL}(3)$	$\frac{0}{2-1} + \frac{1}{2-1} = \frac{1}{1}$	

<총 정보손실>

$IL(G) = IL_A(G) + IL_S(G) = 0.5 \times 2 + (1-0.5) \times (31/14) = 59/28$ 이 된다.

여기서, 그룹 내에 포함되는 정점의 수인 k는 프라이버시 보호를 위해 원시그래프데이터와 프라이버시 강도를 나타내고, 가중치 γ 는 구조정보와 내용정보의 상대적 중요도를 나타낸다.

상기와 같이 정점의 그룹화가 수행되면 테이블 데이터의 내용정보를 수정한다.(S130)

보다 구체적으로, 제 1 데이터 수정모듈은 같은 그룹에 속한 정점들은 모두 같은 QI와 차수를 가지도록 수정하고, 숫자데이터는 해당그룹에서 가장 작은 값과 큰 값의 범위로 일반화하고, 계층적 데이터는 최소 공통조상 값으로 일반화하고, 차수 데이터는 해당 그룹에서 가장 큰 차수 값과 모두 같게 수정한다.

상기와 같이 <표 1>의 테이블 데이터에 대하여 정점을 그룹화하고 데이터를 수정한 결과는 하기의 <표 2>와 같다.

표 2

ID	나이	성별	ZIP	연봉	차수
1	23-31	Female	204**	2000	1
2	45-51	*	20***	3500	2
3	26-39	Male	203**	1000	5
4	45-51	*	20***	3000	2
5	23-31	Female	204**	1500	1
6	26-39	Male	203**	4000	5

- [0127] 이어서, 제 2 데이터 수정모듈이 실현 가능한 그래프 데이터 생성을 위해 차수를 수정한다. (S140)
- [0128] 상기 S130단계에서 수정된 테이블 데이터가 실현 가능한 소셜네트워크 그래프 데이터를 생성하지 못할 수 있으므로 실현 가능한 소셜네트워크 그래프 생성을 위한 차수 수정 작업이 필요하다.
- [0129] 실현 가능한 그래프가 되기 위해서는 우선 모든 정점의 차수의 합이 짝수이어야 한다. 소셜네트워크 그래프에서 각 간선은 두 정점에 연결되어 있으므로 차수를 구할 때 간선은 두 번씩 카운트되기 때문에 모든 정점의 차수의 합은 짝수가 되어야 한다.
- [0130] 만일, 짝수가 아니라면 차수의 합이 홀수인 그룹의 각 정점에 1씩 더해서 전체 합이 짝수가 되게 하여야 한다. 이때, 두 개 이상의 그룹의 차수의 합이 홀수라면 그 합이 가장 작은 그룹의 정점에 1씩 더한다.
- [0131] 차수의 합이 짝수가 되면 주어진 내림차순 정렬된 양수의 차수 집합에서 실현가능한 그래프 생성 판별을 위한 필요충분조건(실현조건)으로 음수가 아닌 내림차순 정수 $d_1 \leq d_2 \leq \dots \leq d_n$ 에서 모든 정수 $r=1, 2, \dots, n-1$ 에 대하여 하기의 <수학식 7> 성립하면 그래프는 실현 가능하다고 판단한다.

수학식 7

$$\sum_{i=1}^r d_i \leq r(r-1) + \sum_{j=r+1}^n \min(r, d_j)$$

[0132]

- [0133] 상기 <수학식 7>을 만족하지 않아서 그래프가 실현 가능하지 않다고 판별되면 차수의 합이 짝수인 그룹 중 그 합이 가장 작은 그룹의 정점에 차수를 1씩 더해서 다시 실현가능한 그래프인지 판별한다. 이 과정을 그래프가 실현 가능하다고 판별될 때까지 반복한다.
- [0134] 예를 들어, 상기 <표 2>에서 차수의 내림차순은 [5, 5, 2, 2, 1, 1]이다. 이것은 차수의 합이 짝수이어야 한다는 조건은 만족하지만 상기 실현가능한 그래프 생성 판별을 위한 필요충분조건을 만족하지 못한다.
- [0135] 그러므로 차수의 합이 짝수이므로 그 합이 가장 작은 그룹인 {1,5}그룹의 정점차수에 1씩 더해서 [5, 5, 2, 2, 2, 2]의 내림차순 차수를 얻을 수 있고, 상기 실현가능한 그래프 생성 판별을 위한 필요충분조건을 만족하여 그래프 생성이 가능한 상태가 된다.
- [0136] 상기 S140단계에 의해 수정된 테이블 데이터는 하기의 <표 3>과 같다.

표 3

[0137]

ID	나이	성별	ZIP	연봉	차수
1	23-31	Female	204**	2000	2
2	45-51	*	20***	3500	2
3	26-39	Male	203**	1000	5
4	45-51	*	20***	3000	2
5	23-31	Female	204**	1500	2
6	26-39	Male	203**	4000	5

- [0138] 이어서, 데이터 역변환부(30)가 상기 수정된 테이블 데이터로부터 소셜네트워크 데이터 그래프를 생성한다. (S150)
- [0139] 보다 구체적으로, 상기 <표 3>과 소셜네트워크 원시 그래프 데이터를 이용하여 프라이버시가 보호되는 수정된 그래프를 생성한다. 원시 그래프의 차수집합(S_o)와 상기 <표 3>의 차수 집합 S_r 을 구할수 있으며, 상기 S_o 와 S_r 은 각각 ID 순서에 의해 정렬이 되어 있다.
- [0140] 여기서, 상기 S_r 에서 S_o 를 뺀 새로운 차수 집합 S_s 를 구할수 있다. 즉 S_s 는 원시 그래프에 비해 새롭게 추가된 차수를 나타내며 이 차수들에 맞게 간선을 추가하면 된다.
- [0141] 우선 S_s 에서 가장 큰 원소를 찾고 그 원소를 제외하고 두 번째로 큰 원소를 찾는다. 원시 그래프를 참조하여 두 원소 사이에 간선이 없으면 간선을 추가하고 원소가 있으면 다음으로 큰 원소와 비교해서 간선 추가 여부를 결정한다. 간선을 추가할 때는 두 정점에 대응되는 S_s 의 원소에서 각각 1씩 빼서 간선이 추가되었음을 확인한다.

이 과정을 모든 원소가 0이 될 때까지 수행한다.

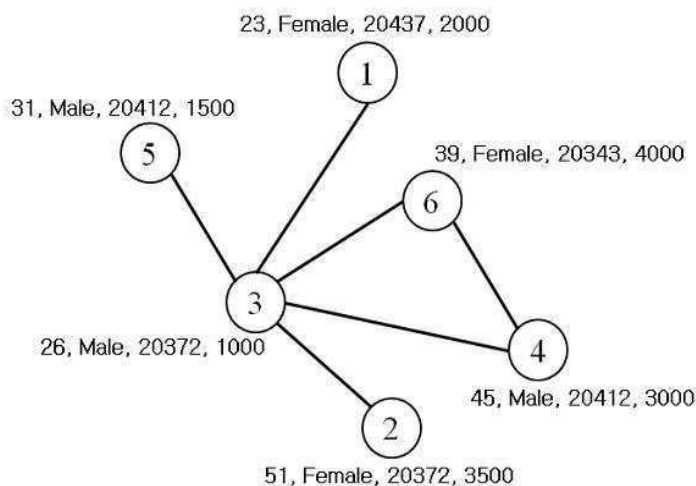
- [0142] 본 실시예에서 So는 [1, 1, 5, 2, 1, 2]이고 Sr은 [2, 2, 5, 2, 2, 5]이다. Sr에서 So를 뺀 Ss는 [1, 1, 0, 0, , 1, 3]이다.
- [0143] Ss의 모든 원소의 합이 6이므로 추가되어야할 간선은 3개임을 알 수 있다.
- [0144] Ss에서 가장 큰 원소인 3과 두 번째 큰 원소인 1을 찾고 그에 대응하는 정점ID6과 ID1사이에 원시 그래프에서 간선이 있는지 확인한다. 간선이 없으므로 두 정점 사이에 간선을 추가하고 ID6과 ID1에서 각각 1씩 뺀 새로운 Ss[0, 1, 0, 0, 1, 2]를 얻는다. 새로운 Ss에서 가장 큰 원소인 2와 1을 찾고 원시 그래프에서 그에 각각 대응하는 정점 ID6과 ID2사이에 간선을 추가하고 Ss의 모든 원소가 0이므로 새로운 Ss[0, 0, 0, 0, 1, 1]을 얻는다 마지막 과정으로 정점ID5, ID6 사이에 간선을 추가하고 Ss의 모든 원소가 0이므로 알고리즘을 마친다.
- [0145] 도 5는 본 발명의 바람직한 실시예에 따라 생성된 소셜네트워크 그래프를 나타낸다.
- [0146] 도 5를 참조하면, 공격자가 특정 개인에 대한 정보를 가지고 있다 하더라도 최소 2개의 후보 정점이 생겨 민감한 정보에 대해 정확한 예측을 할 수 없다.
- [0147] 결국, 공격자가 소셜네트워크 그래프데이터 상에 정점 1에 대응되는 사람의 내용정보(나이, 성별, ZIP)와 소셜네트워크 그래프데이터 상에 차수가 1이라는 것을 알면 그래프에서 해당 정보와 정점 1이 유일하게 대응되어 정점 1에 대응되는 사람의 민감정보(연봉)를 알 수 있게 되는 문제를 도 5의 정점 5와 그룹화하고 일부 내용정보를 일치시키고, 정점에 간선을 추가함으로써 공격자가 알고 있는 정보에 의해서도 대응되는 정점이 그룹내 정점 1, 정점 5가 있게 되어 정점 1에 대응되는 사람의 민감한 정보의 유출을 방지할 수 있게 된다.
- [0148] 이상에서 설명한 본 발명의 상세한 설명에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 본 발명의 보호범위는 상기 실시예에 한정되는 것이 아니며, 해당 기술분야의 통상의 지식을 갖는 자라면 본 발명의 사상 및 기술영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

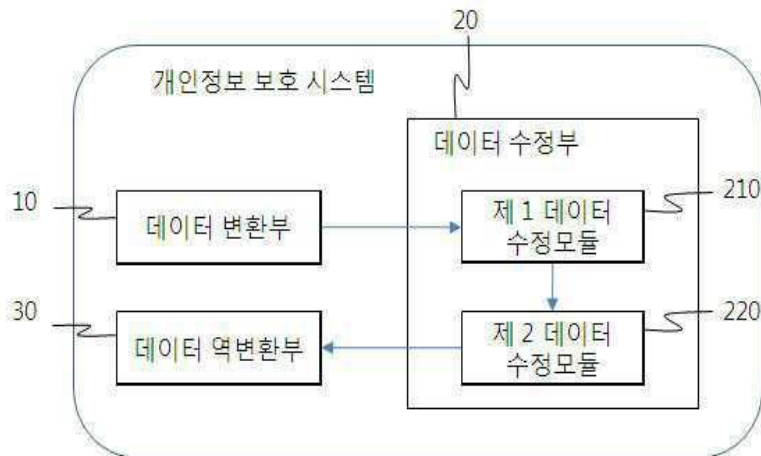
- [0149] 10 : 데이터 변환부 20 : 데이터 수정부
- 210 : 제 1 데이터 수정모듈 220 : 제 2 데이터 수정모듈
- 30 : 데이터 역변환부

도면

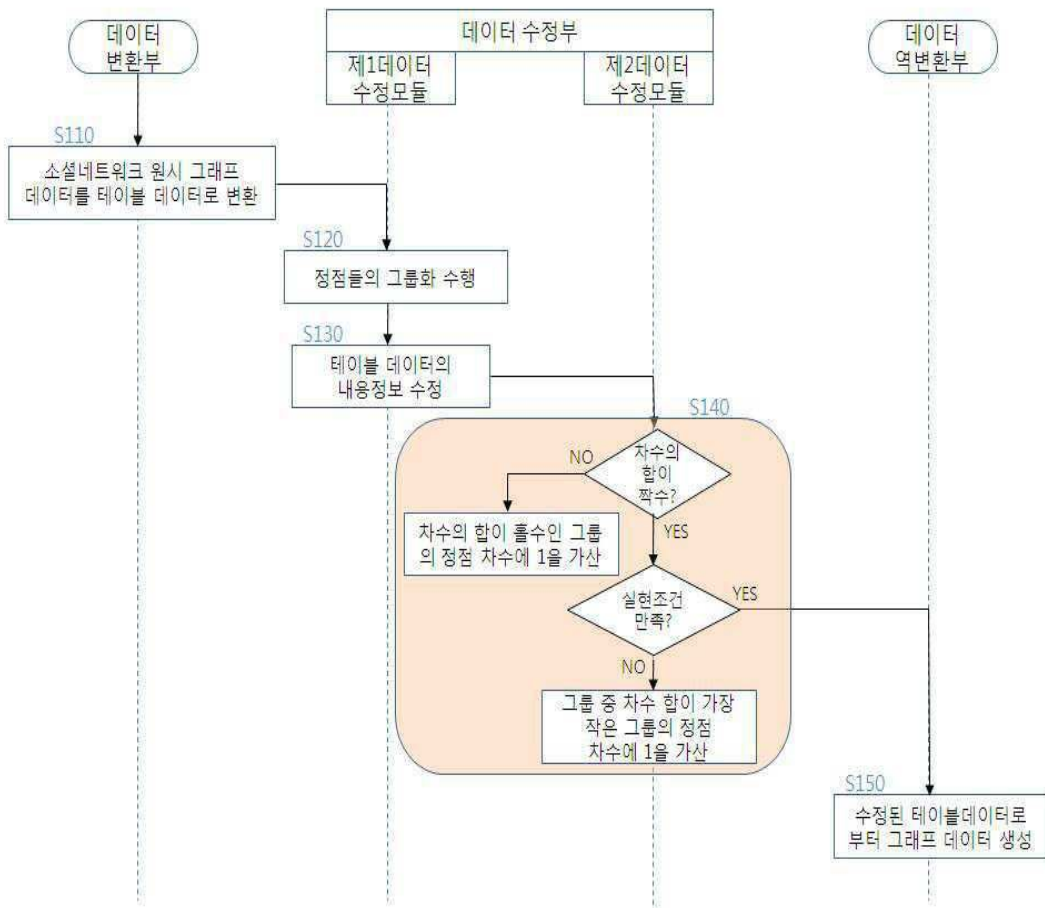
도면1



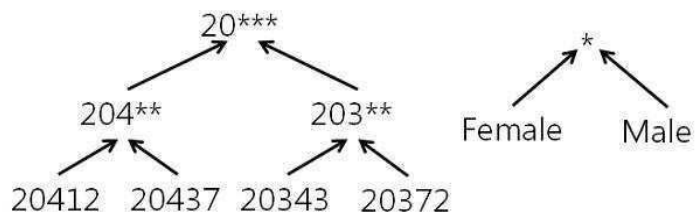
도면2



도면3



도면4



도면5

