Proceedings of KIIS Spring Conference 2008 Vol. 18, No. 1.

유전자 온톨로지를 활용한 반지도 클러스터링 기법

Gene ontology based semi-supervised clustering method

고송, 김대원

서울시 동작구 중앙대학교 컴퓨터공학과

E-mail: ssyong20@wm.cau.ac.kr

요 약

본 논문은 유전자의 기능이 비슷한 정도에 따른 사전정보의 값을 부여하며, 클러스터링시 사전정보를 활용할 수 있는 방법을 제시한다. 실세계 문제인 유전자는 각기 다양한 기능을 하는 특징적인 것으로 사전정보의 형태를 1과 0등으로 구분하던 과거의 방식으로는 정의하기가 어렵다. 유전자간의 비슷한 정도에 따라 사전정보의 값이 정해져야 하는 것은 필요하며, 이는 생물학자가 구축해놓은 유전자 온톨로지의 분석을 통하여 산출한다. 유전자 온톨로지는 기능별 카테고리로 분류하며, 세부 기능은 하위의 카테고리로 형성된 거대한 트리 구조의 형태를 띤다. 온톨로지 분석을 통해 형성된 사전정보의 값은 0과 1사이의 연속적인 값으로 형성이 되며, 이 값은 클러스터링 과정 중 거리 계산에 활용함으로써, 그 결과의 성능이 우수함을 보인다.

Key Words : 클러스터링, 반지도 학습, 온톨로지, 바이오정보학

1. 서 론

생명체는 신진대사 등의 다양한 생명 활동을 통하여 생장·번식 등을 할 수 있다[1]. 이러한 생명활동의 근간이 되는 것은 유전자의 발현과 밀접한 연관성을 가진다.

유전자의 분석을 통하여 질병상태의 분석 등 다양한 분야로 응용이 가능하며, 본 논문에서는 클러스터링을 통한 비슷한 기능의 유전자로 그룹을 형성하는 것을 다룬다. 클러스터링 기법을 유전자 분류에 사용하는 것은 수천에서 수만에 이르는 유전자를 생물학자가 실험을 통해 모두 밝히기에는 비용과 시간의 문제에 직면하기 때문이다. 클러스터링 결과를 통해 유전자 그룹의 특징을 추정함으로써, 생물학자는 그 정보를 활용하여 실험이 가능하므로 시간의 절약을 도모할 수 있다.

하지만 클러스터링은 비지도 학습 방법으로써 지도 학습 방법에 비해 일반적으로 낮은 성능을 보이게 된다. 본 논문은 사전정보를 활용함으로써 성능을 향상시킬 수 있는 반지도 클러스터링 기법을 제시하며, 현재 사전정보의 정의 형태가 가지는 단점을 지적하고, 이를 해결할 방법으로 유전자 온톨로지의 분석을 통해 가능함을 보인다.

2. 사전정보의 정의 및 클러스터링 적용

2.1 현재 사전정보의 정의 문제

현재 사용하고 있는 사전정보의 형태는 클래스 레이블의 값을 가진다. 레이블을 활용하여 생성한 사전정보는 같은 레이블을 가지는 Must-link와 다른 레이블을 가지는 Cannot-link의 2가지 방법으로 구분하여 사용한다[2]. 이러한 사용은 현실 세계의 문제에 적합한지에 대한 의문이 생긴다. 같은 레이블을 갖더라도 같은 정도는 각기 다르며, 다른 레이블의 경우도 마찬가지일 것이다. 예를 들어, 암 환자의 경우 말기 환자와 초기 환자의 유전자 발현 상태는 다르며, 정상의 경우 암의 기미가 없는 상태와 발병 직전 상태의 유전자 상태는 또 다를 것이다. 문제는 초기 환자 상태와 발병 직전 상태의 유전자 발현의 정도가 비슷하지 않을까 하는 의문이 든다. 이 문제는 기존의 사전정보 형태로는 극복이 불가능한 문제에 해당할 것이다.

2.2 온톨로지 분석을 통한 사전정보 정의

본 논문에서 제안하는 사전정보의 형태는 0과 1사이의 연속적인 값을 가짐으로써 유전자간의 기능이 비슷한 정도에 따라 산출하며, 사전정보를 가지는 유전자 수가 n개 일때, 사이즈는 n*n의 크기가 된다.

사전정보의 형성은 생물학자가 구축해 놓은 온톨로지의 분석을 통하여 이루어진다[3]. 온톨로지는 생물학적인 과정(biological process)을 중심으로 세부 내용이 하위 노드를 형성한다. 하위 노드에는 세포 주기(cell-cycle)와 신진대사(metabolism)등이 포함되며, 세포 주기의 하위 노드에는 세포간기, 세포분열기 등이 형성되는 거대한 트리 구조를 띤다. 생명체가 일반적으로 보이는 기능들에 대해서 형성한 것이며, 각 유전자는 그 기능이 해당하는 노드에 속하여 나열된다. 유전자는 다양한 기능을 할 시에는 수 개의 노드에 포함된다.

본 논문에서의 사전정보 형성 방법은 온톨로지의 특성을 반영하였다. 같은 노드에 있다는 것은 같은 기능을 한다는 의미이므로 가장 큰 값인 1이 측정되며, 노드 거리가 멀어질수록 그 기능이 같은 정도가 떨어지므로 1보다 작아지는 값으로 측정된다(그림 1)(식 1). 한 가지의 거리를 1로 설정하는 것은 사전정보의 최대 값이 1이 되도록 하기 위함이다.

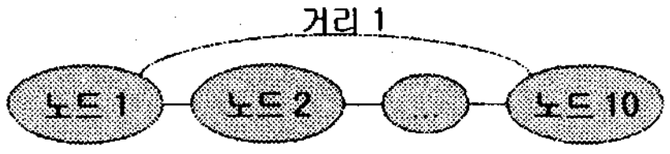

그림 1. 최상위 노드 1과, 최하위 노드10의 거리는 1이며, 노드 1과 노드 2의 사전정보 값은 (1- (2-1)/10) = 0.9 임

사전정보(노드 m, 노드 n)
= 1(상수) - (|m - n|) / 최하위 노드 깊이 (1)

다른 카테고리와의 계산은 현재 단계에서 그 관계를 정의하기가 어려우므로 계산에서 배재하여, 세포주기와 신진대사간의 사전정보는 측정하지 않으며, 개별적으로 측정한다.

2.3 클러스터링의 적용

클러스터링 기법은 k-means의 개념에 사전정보를 적용하였다. k-means는 유전자들로 형성된 그룹의 중심과 유전자의 거리를 측정함으로써 가까운 중심으로 할당하는 방법이다. k-means를 기반으로 하여 사전정보를 통해 (식2)와 같이 유전자간의 거리를 보정해 줄 수 있다.

거리(중심 X, 유전자 a) = 유클리디안(X, a) - α(·) (2)

α(·) = 유클리디안(X, m) × (0.9 + 0.95) / 2 (3)

유전자는 기능이 같다고 하더라도 발현 타이밍은 같지가 않음으로써 실험을 통해 측정된 유전자 발현 측정값은 다른 패턴을 보일 수 가 있어, 거리 측정 시 큰 거리 값을 보일 수가 있다. 이 값을 α(·)를 통해 사전정보를 활용함으로써 같은 그룹에 속하도록 하지만, 큰 차이를 보이는 유전자를 그룹에 속하게 함으로써 발생 가능한 그룹의 평균의 변화가 심하게 되는 것은 방지하도록 하여야 한다. α(·)의 계산은 그림 2를 예로 삼으면서 (식 3)의 과정을 거쳐 계산된다. 그림 3에서 유전자 m은 녹색 그룹의 유전자와 0.9, 095의 사전정보를 가지며, 흰색 그룹의 유전자와 0.1의 사전정보를 가지는 것을 볼 수 있다. 녹색 그룹의 비중이 높으므로 α(·)는 녹색 그룹에 적용이 된다.

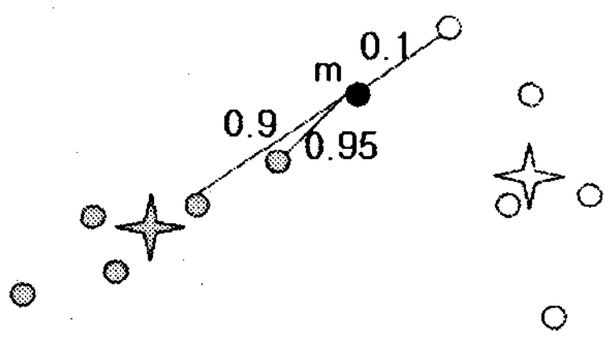

그림 2. 유전자 m은 왼쪽(녹색)의 중심으로 표현되어진 ✦에는 두 개의 유전자와 사전정보를 가지며, 오른쪽(흰색)의 중심으로 표현되어진 ✧ 에는 한 개의 유전자 사전정보를 가짐.

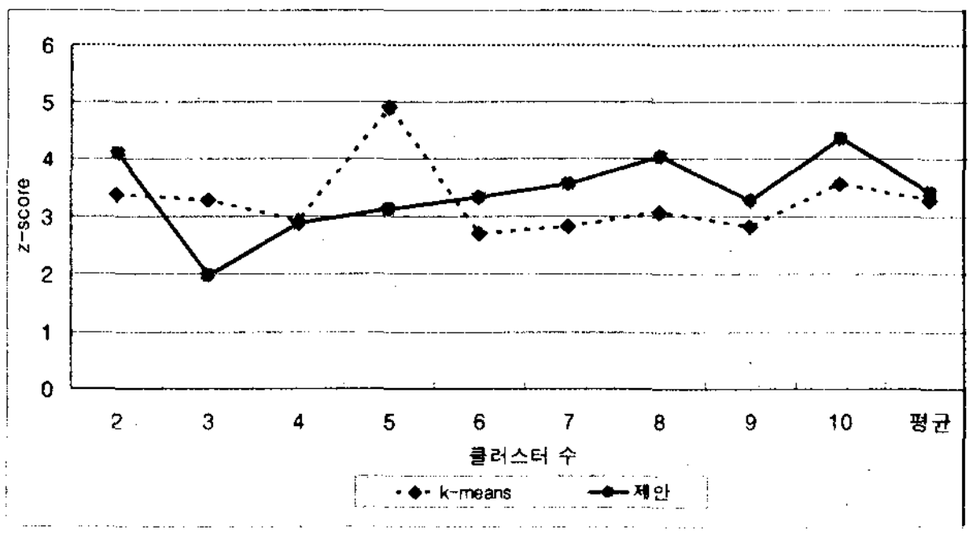

그림 3. Chu의 효모 데이터 중 sporulation 데이터를 활용한 실험 결과임. 비교 대상 방법론 k-means는 점선이며 제안 방법론은 실선으로 표현됨. 가장 오른쪽은 총 평균을 나타냄

3. 실험 및 분석

실험 데이터는 Chu의 sporulation 데이터를 이용하였다[4]. 실험은 그룹의 수를 2~10으로 하며, 10회 적용하여 생성된 결과를 [5]에서 제공하는 평가 툴에 의해 z-score로 판단한다[6].

z-score는 랜덤으로 구분한 결과와 클러스터링의 결과를 비교하는 것으로 차이가 클수록 높은 값이 측정되어 클러스터링 결과가 좋다고 할 수 있다.

제안하는 방법이 비교적 높은 성능을 보이며, 평균적으로 0.14 높은 성능을 보인다[그림 3]. 이것은 일반적인 k-means로는 두 유전자 YAL001C와 YAL002W가 같은 기능을 하지만 발현 타이밍이 달라서 그룹 5개 이상으로 클러스터링 시 같은 그룹에 속하지 못하게 되기 때문이다. 하지만 제안하는 방법으로는 사전정보의 적용으로 이를 극복함으로써 그 외에 같은 기능을 하는 유전자를 같은 그룹에 속하게 할 수 있어 성능을 향상시킬 수 있다.

다만 클러스터 그룹이 3개에서 5개 일 때, 낮은 성능을 보이고 있는 모습을 볼 수 있다. 이것은 α(·)의 적용을 크게 함으로써 발현 패턴의 차이가 큰 유전자가 그룹에 속하게 되어 그 그룹의 평균 값의 변화가 커지기 때문이다. 그룹의 평균은 그룹을 대표하는 값이므로, α(·)의 조절을 통해 평균의 변화가 크지 않도록 해야 한다.

그래서 앞으로의 연구 주제는 사전정보를 활용하는 클러스터링에서 핵심 사항인 α(·)의 추정식의 보완을 통해 안정적인 성능을 보일 수 있도록 할 계획이다.

참 고 문 헌

[1] Berg, Jeremy M et al, "Biochemistry", Freeman, 2002
[2] Kiri Wagstaff et al, "Constrained K-means Clustering with Background Knowledge", Proceeding of the Eighteenth International Conference on Machine Learning, pp 577-584, 2001
[3] http://www.geneontology.org
[4] S. Chu et al, "The Transcriptional Program of Sporulation in Budding Yeast", Science, Vol. 282, pp 699-705, 23 Oct, 1998
[5] http://llama.med.harvard.edu
[6] Francis D. Gibbons and Frederick P. Roth, "Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation", Genome Research, vol.12, No. 10, Jan, 24, 2008

연관관계를 온톨로지 형태로 기술한 문서로, 사회연결망을 표현하는 하나의 수단이다.

본 연구는 사회연결망 서비스에서 발생하는 문제를 해결하고 분석을 돕기 위해 FOAF를 자동적으로 생성하고 활용하는 프레임워크를 제안한다. FOAF의 기본 명세에 기능을 추가하였으며 단점을 극복하는 방안으로 FOAF와 OpenID와 사회연결망을 융합하는 방안을 제안한다. FOAF를 자동으로 성장하도록 하고 관리의 용이성을 부여하였으며 정보 공개의 정도를 조절할 수 있는 방법을 제시하였다.

본 논문의 2장에서는 FOAF에 대한 정의와 장단점을 자세히 살펴보고 3장에서 OpenID에 대하여 소개한다. 4장에서는 앞서 설명한 FOAF와 OpenID와 사회연결망 서비스를 통합하는 자동화된 FOAF 형성 프레임워크를 설계하며, 5장에서는 결론과 향후 연구 방향을 제시하였다.

2. FOAF의 특징과 문제점

2.1 FOAF의 정의

FOAF는 사람과 사람간의 관계와 사람의 주변에 존재하는 다양한 개체들 간의 관계를 기술하는 문서를 말한다. XML의 구문을 이용하고 RDF 규약을 따라 명세를 가지는 개인 온톨로지의 하나이다[4,5]. 기존의 다른 온톨로지들과 차이를 보이는 FOAF의 특징은 FOAF 어휘(Vocabulary)에 있다[4,6]. FOAF 어휘에는 클래스와 속성들을 명시적으로 지정하고 있으며 이해하기 쉬운 표현력을 갖추는 것에 중요성을 두고 있다.

2.2 FOAF의 장점

명세서의 단순함과 쉬움은 일반적인 웹 사용자들에게 시맨틱 웹을 가깝게 느낄 수 있게 하며 정보를 생산하고 소비하는 사람들에게 유용하게 사용될 수 있다. 정보 생산자들에게 다음과 같은 면에서 유용하다[6].

- 커뮤니티 관리에 유용. 커뮤니티의 멤버십을 표현하는 수단으로서, 회원을 관리할 수 있게 함.
- 유일한 ID를 제공. 이메일 주소를 해쉬를 이용해 표현, 타인을 가장하는 것 방지.
- 저자 식별에 도움이 됨. FOAF 툴을 통해 이메일과 문서에 전자 서명을 적용.

또한 FOAF는 다음과 같은 방법으로 정보 소비자들을 돕는다.

- 출처와 책임에 대한 조사. 출처 추적 RDF 툴을 통해 언제 어디서 정보가 획득되었는지 조사할 수 있음.
- 커뮤니티의 신규 회원에게 구성원의 구조와 설명을 제공할 수 있음.
- 공통된 흥미를 지닌 사람들을 찾도록 도움.
- 신뢰할 수 있는 사람으로부터의 이메일인지 판단할 수 있는 근거가 됨.

이외에도 FOAF는 URI대신 rdfs:seeAlso를 통해 다른 FOAF 파일을 링크함으로써 시맨틱 웹 봇들이 크롤링할 수 있게 하고 있다[7].

2.3 FOAF의 문제점

FOAF는 몇 가지 한계점에 의해 비판받고 있으며 이를 해결하기 위한 방안을 모색하고 있다[8]. 제기되고 있는 FOAF의 문제점은 다음과 같다.

- 사생활 침해에 대한 대책이 부족함. 개인의 정보가 웹상에 공개됨에 따라 사적인 관계까지도 노출됨.
- 비전문가는 XML 형식을 다루기 어려움. FOAF-o-Matic과 같은 FOAF문서 제작용 툴을 사용하더라도 효과적인 사용을 위해서는 XML과 FOAF 어휘에 대한 지식이 필요함.
- 정보 갱신의 문제가 있음. FOAF문서를 기술하기 위해서는 사용자가 직접 내용을 입력해야 함. FOAF-o-Matic은 단순한 형식 맞춤 도구에 불과함.
- 명세서 상의 한계로 인한 문제점이 있음. 개인 홈페이지를 기계 가독형으로 전환하고자 했던 초기의 목적에 의해 적용범위가 한정적임. 영어권 사용자들이 주도하여 국제화나 지역화에 대한 고려가 부족함.

3. OpenID와 FOAF

OpenID는 웹 사이트들의 복잡한 가입절차를 간소화 하고, 개인 정보를 효과적이고 효율적으로 관리하기 위한 목적을 가지고 있다. OpenID는 분산형 Digital identity 시스템의 하나로 중앙 집중식 인증센터를 필요로 하지 않는다[7,9]. 이용하는 웹 사이트마다 회원가입을 통해 계정을 갖지 않더라도 자신이 신뢰하여 가입한 Identity 제공자(idP)를 통해 사용자의 해당 ID에 대한 소유권을 OpenID 지원사이트(RPs)에 입증해 줄 수 있다. 최근에는 OpenID 속성 교환 명세서를 발표하여 idP를 통해 사용자의 정보의 일부를 공유하고 통제하는 기능을 가지게 되었다.

FOAF프로젝트는 FOAF의 문제 해결을 위해 OpenID에 주목하고 있다. FOAF 명세서에 openid 속성을 추가하였으며 OpenID를 통해 FOAF를 보강 하는 연구를 진행하고 있다[4].

4. 자동화된 FOAF 형성 프레임워크 설계

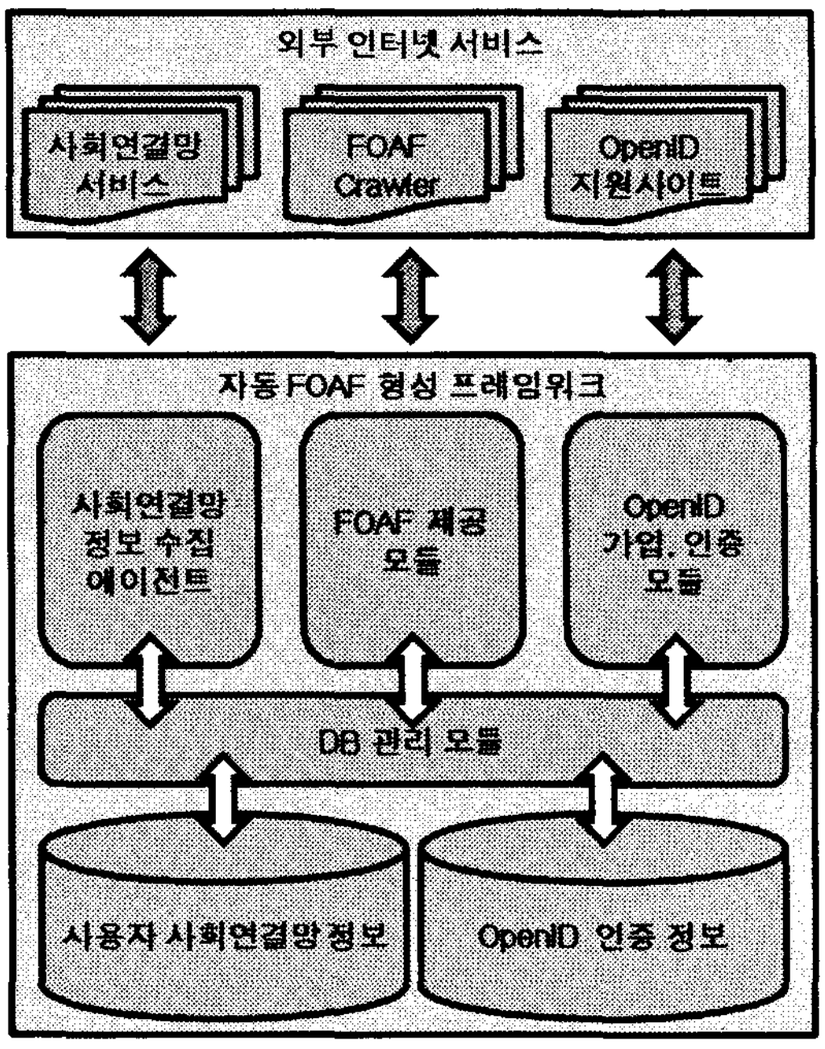

그림 1. 프레임워크 아키텍처

4.1 프레임워크 아키텍처

프레임워크는 네 개의 서브시스템으로 구성되어 있다. 그림1은 프레임워크의 서브시스템들의 관계를 나타낸 것이다.

- FOAF 제공 모듈: 수집된 사용자의 개인정보와 사회연결망 정보를 이용하여 FOAF를 생성한다. 등급에 따라 공개 수준을 결정.
- OpenID 가입 및 인증 모듈: OpenID의 기본적인 인증 기능을 승계하면서, FOAF 생성에 필요한 추가 정보를 사용자에게 요구하고, FOAF를 이용해 향상된 사용자 인증을 수행한다.
- 사회연결망 정보를 수집 에이전트: OpenID 시스템에 등록된 OpenID 지원 사이트나 사용자의 블로그와 같은 사회연결망 서비스로부터 정보를 수집한다.
- DB 와 관리 모듈: 수집 모듈로부터 수집된 데이터를 가공하여 FOAF와 OpenID와 사회연결망 정보를 저장하고 관리하는 모듈

4.2 OpenID 가입 및 인증 모듈의 역할

OpenID가 가져야 하는 기본적인 기능을 모두 수행하면서 FOAF 생성을 위한 추가적인 정보를 사용자에게 요구한다. 추가 정보는 FOAF 생성에 동의한 사람에게만 요청되어야 하며 사용자의 이름과 메일 주소는 필수요소이다. 홈페이지와 전화번호, 직장 정보, 학교와 같은 부분들은 선택적인 요소로 수집한다. 그리고 사회연결망 분석을 위해 OpenID를 사용하지 않았던 기존의 사회연결망 서비스의 계정이나 URL을 입력받는다.

이 모듈은 정보 공개의 등급에 따라 사용자가 원치 않는 정보가 외부 서비스에 노출되는 것을 방지한다. 본 논문에서는 정보 공개의 등급으로 요소별 중요도 분석을 통해 표 1과 같은 정보 등급제를 제안한다. 등급의 숫자가 낮을수록 많은 정보가 공개되며 커질수록 제약사항이 강화된다.

표 1. FOAF 정보 공개 등급

등급	의미
1	모든 정보를 공개
2	Email을 암호화
3	Email, 전화번호와 같은 사생활 침해의 위험이 높은 정보 암호화
4	사회연결망 정보 숨김
5	사용자 지정

본 프레임워크에서는 최초 가입 시 기본으로 3등급으로 설정된다. 이 등급은 FOAF 제작 툴인 FOAF-o-Matic이 제공하는 2등급 수준보다 좀 더 정보가 제한된 상태이다. 기본 등급은 FOAF 정보 요청자가 추가 파라미터를 제공하지 않을 경우 공급하는 FOAF문서의 공개 조건이다. 사용자는 5개 등급중 하나를 골라 기본등급을 변경할 수 있으며, 5번을 선택하여 FOAF의 요소별로 사용자의 의도에 맞게 공개 수위를 조정할 수 있다.

사용자가 기본으로 지정한 것과 다른 등급의 정보를 필요로 하는 특정한 FOAF 수집자들은 프레임워크에 사전 등록 방법을 통해 개별적으로 사용자로부터 인증을 받아야 한다. 프레임워크의 OpenID모듈은 수집자들에게 Open API를 제공하여 FOAF 사용권을 요청받는다. 프레임워크는 이 요청을 각 사용자에게 전달하여 사용자들로부터 정보 제공의 허가를 구한다. 이렇게 인증을 받은 FOAF 수집자만이 다른 등급의 FOAF를 공급받을 수 있게 된다.

4.2 사회연결망 정보 수집 에이전트

사회연결망 정보 수집 에이전트는 다른 모듈과는 독립적으로 동작한다. 사용자 정보 수집을 목적으로 일정 시간마다 실행되어 사용자의 사회연결망 정보를 갱신한다. 4.2 에서 사용자가 입력한 홈페이지, 사회연결망 서비스,

OpenID 인증 사이트들의 정보를 이용해 사용자의 사회관계 정보를 습득하고 데이터베이스에 저장한다. 본 논문에서 제안하는 프레임워크는 수집된 웹 정보들의 분석을 위해 기본 인터페이스를 제공하며, 인터페이스에 맞는 분석 모듈을 구현을 통해 다양한 사회연결망 서비스로부터 정보를 추출하게 된다. 본 연구에서는 싸이월드 분석 모듈개발하고 미니홈피의 주인과 연결된 친구들의 정보를 수집하여 FOAF를 구축하는 실험을 진행하였으며 성공적인 생성이 가능함을 확인하였다.

4.3 FOAF 제공 모듈과 규약

FOAF 제공 모듈은 외부 서비스 또는 시맨틱 크롤러와 같은 시스템에게 FOAF를 제공하는 역할을 수행한다. 수집된 사용자의 개인정보와 사회연결망 정보를 이용해 FOAF 명세서의 규약에 따라 자동으로 문서를 수립하고 이를 웹에 공개한다. 이때 FOAF 정보 요청자에게 상위 등급의 정보를 제공하기 위해서는 다음과 같은 파라미터를 요구한다.

표 2. 상위 등급을 위한 추가 파라미터

파라미터	의미
foaf.sid	FOAF 정보 요청자 ID
foaf.spw	FOAF 정보 요청자의 인증 값
foaf.grade	요구하는 정보 등급

이와 같은 파라미터가 요청에 포함될 경우 정보 인증 작업을 거친 뒤 FOAF를 요청자에게 공급한다. 이를 통해 프레임워크는 자동화된 FOAF 생성, 인증과 등급제를 이용한 사용자의 사생활 보호를 수행하게 된다.

5. 결론 및 향후과제

본 논문에서는 사회연결망서비스에서 발생하는 연관관계의 문제 해결을 위해 FOAF를 생성하고 이것을 이용해 사용자의 사회연결망 수집과 분석을 수행하는 프레임워크를 설계하였다. FOAF의 단점을 극복하는 방안으로서 FOAF와 OpenID와 사회연결망을 융합하는 방식을 설계하였고 이에 필요한 다양한 요소들을 제안하였다. FOAF 정보 공개 등급을 제안하여 노출되는 사용자의 정보를 제한하도록 유도하였으며, 사회연결망 서비스로부터 정보를 추출하여 자동으로 FOAF를 갱신할 수 있도록 하였다. 이를 이용해 사용자의 작업을 최소화하고 FOAF의 구문을 모르더라도 충분히 운영할 수 있도록 하였다. 이 프레임워크를 이용해 다양한 사회연결망 분석을 수행할 수 있을 것으로 기대된다. 사회연결망 분석을 통해 FOAF에 연결 관계 가중치를 추가하는 연구를 진행하고 있으며 생성된 FOAF를 사회연결망에 적용하여 정보를 통제하는 방안도 연구 중이다. 또한 영어권에 한정된 요소들의 국제화를 위한 대안을 마련하는 방안을 연구 중에 있다.

감사의 글

본 연구는 21세기 프론티어 연구개발사업의 일환으로 추진되고 있는 지식경제부의 유비쿼터스컴퓨팅및네트워크원천기반기술개발사업의 08B3-B1-10M 과제로 지원된 것임

참 고 문 헌

[1] Tim O'Reilly, "what is web 2.0," http://www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[2] J. Breslin and S. Decker, "The future of Social Networks on the Internet," IEEE Internet Computing, IEEE Computer Society, vol 6. November December, 2007.

[3] H. Takeda, I. Ohmukai, "Building semantic web applications as information/ knowledge sharing systems, UserSWeb: Workshop on User Aspects of the Semantic Web," 2005.

[4] FOAF Vocabulary Specification 0.91, http ://xmlns.com/foaf/spec/

[5] FOAF project, http://foaf-project.org/

[6] L. Ding, L. Zhou, T. Finin and A. Joshi, "How the Semantic Web is Being Used: An Analysis of FOAF Documents," Proceeding of the 28th Hawaii International Conference on System Science, Track 4, p.113.3 January 03-06, 2005.

[7] 배준현, 김상욱, 개방형 모바일 웹 서비스를 위한 OpenID를 이용한 사용자 인증 메커니즘의 설계 한국컴퓨터 종합학술대회 논문집 vol34, no1 35-39 2007

[8] FOAF Criticism, http://wiki.foaf-project.org/Criticism

[9] OpenID, 위키백과, http://ko.wikipedia.org /wiki/OpenID