RDF 지식 베이스의 자원 중요도 계산 알고리즘에 관한 연구

노상규[a], 박현정[b], 박진수[c]

[a]*서울대학교 경영대학 경영전문대학원*
서울특별시 관악구 신림9동 산 56-1번지, 151-742
Tel: +82-2-880-6922, Fax: +82-2-872-0512, E-mail: srho@snu.ac.kr

[b]*서울대학교 경영대학 박사과정*
서울특별시 관악구 신림9동 산 56-1번지, 151-742
Tel: +82-2-880-8794, Fax: +82-2-872-0512, E-mail: sparrow7@snu.ac.kr

[c]*서울대학교 경영대학 경영전문대학원*
서울특별시 관악구 신림9동 산 56-1번지, 151-742
Tel: +82-2-880-9385, Fax: +82-2-876-8774, E-mail: jinsoo@snu.ac.kr

Abstract

The information space of semantic web comprised of various resources, properties, and relationships is more complex than that of WWW comprised of just documents and hyperlinks. Therefore, ranking methods in the semantic web should be modified to reflect the complexity of the information space. In this paper we propose a method of ranking query results from RDF(Resource Description Framework) knowledge bases. The ranking criterion is the importance of a resource computed based on the link structure of the RDF graph. Our method is expected to solve a few problems in the prior research including the Tightly-Knit Community Effect. We illustrate our methods using examples and discuss directions for future research.

Keywords:

Ranking, Semantic Web, RDF Knowledge Base, Resource Importance, Ontology

I. 서론

신속하고 정확한 의사결정 능력은 모든 조직과 개인의 핵심적인 성공 요인이라 할 수 있다. 그러므로 넘쳐나는 데이터 속에서 의사결정에 중요한 정보들을 찾고 랭킹하는 기술에 대한 연구는 계속적인 관심을 기울여야 할 분야일 것이다. 특히 사용자가 원하는 정보 중 가장 관련이 있는 것들을 상위에 보여주는 랭킹 기술에 대한 연구에는 경쟁적인 투자가 지속될 것으로 보인다.

독립된 문서들의 무한한 모임을 대상으로 하는 전통적인 검색 시스템에서는 주로 검색 키워드가 문서 안에서 발견되는 횟수에 의해 문서의 중요도가 결정되었다. 이 후 문서와 문서가 하이퍼링크로 연결된 월드와이드웹(World Wide Web)에서는 거대한 웹 그래프의 문서간 링크 구조를 분석하여 객관적인 중요도 점수를 산출하는 방법이 사용되었다. 대표적인 예로 1998년에 등장하여 빛나는 성공으로 주목을 받아 온 구글 검색 엔진의 페이지랭크[Brin et al., 1998; Page et al., 1998; Haveliwala, 1999] 알고리즘을 들 수 있다. 페이지랭크에서는 임의의 페이지를 가리키는 다른 페이지들이 많을수록, 그리고 이러한 다른 페이지들의 중요도가 높을수록 해당 페이지의 중요도가 올라간다.

월드와이드웹에서 페이지의 중요도를 결정하는 다른 방법으로 Kleinberg의 연구[Kleinberg, 1998]를 들 수 있다. Kleinberg는 특정한 키워드를 포함하는 웹 페이지에 대해 권위 점수(authority score)와 허브 점수(hub score)라는 두 가지 형태의 중요도를 정의하였다. 높은 권위 점수를 가진 페이지는 특정 주제에 대한 우수한 정보를 제공하므로 이 페이지를 가리키는 다른 페이지들이 많다. 그리고 높은 허브 점수를 가진 페이지는 직접적으로 특정 주제에 대한 정보는 포함하고 있지 않지만 많은 추천 사이트를 가리킨다. 많은 좋은 사이트를 가리킬수록 더 좋은 허브이며, 많은 좋은 허브에 의해 추천될수록 더 좋은 권위 페이지가 된다.

시맨틱 웹의 정보공간은 인스턴스 정보 외에 이에 대한 온톨로지를 포함하고 있어 문서간 링크가 모두인 한 가지뿐인 월드와이드웹에 비해 훨씬 복잡하다. 자원과 자원 사이에 다양한 관계가 존재하는 시맨틱 웹에서는 랭킹의 대상이 되는 주제들도 여러

가지이다.

본 논문에서는 RDF 지식베이스에 대한 질의 결과를 자원 중요도에 따라 랭킹하는 문제를 다룬다. 기존 연구에서는 고안된 자원 중요도 계산 알고리즘을 몇 가지 시스템에 적용해본 후 한계점을 보고했다[Mukherjea et al., 2005]. 이것은 주로 링크 구조 분석 방법을 RDF에 적용했을 때 나타나는 근본적인 문제점에서 기인하는 것들이다. 본 논문에서는 이러한 한계점을 온톨로지 스키마상에서 자원 중요도에 영향을 미치는 속성들에 양의 가중치를 설정함으로써 해결할 수 있는 방안을 제시한다. 이러한 가중치는 각 클래스의 특성을 반영하여 설정되며 컨텍스트나 애플리케이션에 따라 유연하게 변화시킬 수 있도록 디자인된다. 그리고 자원 중요도 계산에서 제외되었던 데이터타입 속성도 고려하는 방안과 함께 실험 결과도 제시한다. 본 논문에서 제안하는 자원 중요도 계산 알고리즘은 의미적 관계(semantic association)[Aleman-Meza et al., 2005; Anyanwu et al., 2005; Halaschek et al., 2004; Sheth et al., 2005]를 포함하는 시맨틱 웹의 다양한 주제에 대한 랭킹 알고리즘의 기초가 될 것으로 기대된다.

전체적인 구성은 2장에서 관련연구들에 대해 살펴보고 3장에서 클래스중심 가중치 설정 알고리즘의 기본 아이디어 및 분석을 전개할 것이다. 그리고 4장에서는 실험 방법 및 결과를 제시하고 5장에서 결론 및 향후 연구과제에 대해 언급하겠다.

II. 이론적 배경 및 관련연구

2.1 WWW의 문서 중요도 - 구글(Google)의 페이지랭크(PageRank)

페이지랭크는 월드와이드웹이 단순한 문서 집합과는 달리, 문서와 문서가 하이퍼링크로 연결된 거대한 웹 그래프를 이룬다는 점에 착안하여 이러한 웹 그래프에 내재되어 있는 링크 구조로부터 객관적인 웹 문서의 중요도를 추출해내는 획기적인 방법이다.

한 페이지의 랭크는 이것이 가리키는 다른 페이지의 랭크에 더해질 때 이 페이지에서 나가는 포워드(forward) 링크의 수로 나뉘어 균등하게 배분된다. 최종 랭크는 임의의 초기 해로부터 시작하여 수렴할 때까지 반복적인 계산을 통해 구해진다.

2.2 WWW의 문서 중요도 - Kleinberg의 권위/허브 점수(Authority/Hub Score)

Kleinberg는 하나의 웹 페이지에 대해 권위 점수(authority score)와 허브 점수(hub score)라는 두 가지 유형의 점수를 정의하였다[Kleinberg, 1998]. 많은 좋은 사이트에 의해 링크될수록 더 높은 권위 점수를 가지며, 많은 좋은 사이트를 가리킬수록 더 높은 허브 점수를 갖는다. 페이지랭크에서는 먼저 전체 웹을 대상으로 각 페이지의 랭크를 계산해 놓았다가 텍스트 기반 검색을 했을 때 나오는 결과를 미리 계산된 페이지랭크 점수로 정렬한다. 이에 반해, Kleinberg의 알고리즘은 먼저 텍스트 기반 검색을 하고 이 결과로 얻어진 페이지들로 이루어지는 서브 그래프(sub-graph)에 대해 권위 점수와 허브 점수를 계산한다. 계산 알고리즘은 Figure 1과 같고, 여기에서 x 와 y 는 각각 웹 페이지의 권위 점수와 허브 점수를 나타내는 벡터이다.

```
Iterate(G, k)
    G: a collection of n linked pages
    k: a natural number
    Let z denote the vector (1, 1, 1, ···, 1) ∈ R^n.
    Set x_0 := z.
    Set y_0 := z.
    For i = 1, 2, ···, k
        Apply the I operation to (x_{i-1}, y_{i-1}),
          obtaining new x-weights x_i^•.
        Apply the O operation to (x_i^•, y_{i-1}),
          obtaining new y-weights y_i^•.
        Normalize x_i^•, obtaining x_i.
        Normalize y_i^•, obtaining y_i.
End
Return (x_k, y_k).
```

Figure 1 - Kleinberg의 권위/허브 점수 계산 알고리즘

위 Figure 1의 알고리즘에서 사용한 I(In)와 O(Out) 오퍼레이션(operation)은 각각 다음 식 (1), (2)와 같다. $x^{<p>}$ 는 페이지 p 의 권위 점수, $y^{<p>}$ 는 페이지 p 의 허브 점수이며, E 는 n 개의 웹 페이지 사이에 존재하는 전체 링크 집합이고, (p,q) 는 페이지 p 에서 페이지 q 로 향하는 링크를 의미한다.

$$\text{I operation}: \; x^{<p>} \leftarrow \sum_{q:(q,p)\in E} y^{<q>} \qquad (1)$$

$$\text{O operation}: \; y^{<p>} \leftarrow \sum_{q:(p,q)\in E} x^{<q>} \qquad (2)$$

그리고, 한 번의 반복(iteration)이 끝날 때마다 수행되는 정규화(normalization) 조건은 다음 식 (3)과 같다.

$$\sum_{p}(x^{<p>})^2 = 1, \; \sum_{p}(y^{<p>})^2 = 1 \qquad (3)$$

Kleinberg는 이러한 반복 계산에 의해 구해지는 $x_1, x_2, x_3, \cdots$ 와 $y_1, y_2, y_3, \cdots$ 가 각각의 극한값 x^*, y^* 로 수렴함을 증명하였다. 특히, 페이지들

사이의 링크 연결 구조를 반영하는 $(n \times n)$ 행렬 A (페이지 i 에서 페이지 j 로 향하는 링크가 있으면 $A_{ij} = 1$, 없으면 $A_{ij} = 0$, $1 \le i, j \le n$)에 대해 x^* 는 $A^T A$ 의 제일 고유벡터(principal eigenvector)이고 y^* 는 AA^T 의 제일 고유벡터가 됨을 증명하였다.

2.3 RDF와 RDF Schema

RDF(Resource Description Framework)는 표현하고자 하는 온갖 개념을 자원(resource)으로 보고 이러한 자원을 서로 구별하기 위한 식별자로 URIref(Uniform Resource Identifier reference)를 사용하여 자원의 속성이나 자원과 자원 간의 관계를 기술하는 데이터 모델이다[Klyne et al., 2004; Manola et al., 2004]. RDF의 기본 단위는 '주어부(subject)-서술부(predicate or property)-목적부(object)' 의 세 부분으로 이루어져 흔히 트리플(triple)이라 불리는 서술문(statement)이다. 예를 들어 '연구원1-1(주어부)의 나이(서술부)는 35(목적부)이다' 란 서술문은 '연구원1-1' 의 '나이' 란 속성이 '35' 임을 표현하고 있으며, '연구원1-1(주어부)이 논문2-1(목적부)을 발표하다(서술부)' 란 서술문은 '연구원1-1' 이란 자원과 '논문2-1' 이란 자원이 '발표하다' 란 관계로 연결되어 있음을 나타내고 있다. 이러한 서술문에서 목적부를 나타내는 자원을 다시 주어부로 하여 새로운 서술문을 연결하거나, 하나의 작은 서술문 전체를 목적부로 갖는 큰 서술문 구조(reification)를 이용하면 복잡한 지식도 표현이 가능하다. RDF 서술문은 노드와 링크로 이루어지는 RDF 그래프로도 표현된다. 노드는 서술문의 주어부와 목적부에 자원이 오는 경우에 해당되며 URIref를 포함하는 타원으로 나타낸다. 목적부에 문자열 데이터가 올 때에는 직사각형 안에 그 내용을 기록한다. RDF 그래프의 링크는 서술문의 서술부에 해당되며 주어부에서 목적부로 향하는 화살표와 서술부의 개념에 대한 URIref를 화살표 옆에 명시한다. 다음 Figure 2는 앞에서 언급된 '연구원1-1(주어부)의 나이(서술부)는 35(목적부)이다' 와 '연구원1-1(주어부)이 논문2-1(목적부)을 발표하다(서술부)' 라는 지식을 표현한 RDF 그래프이다.

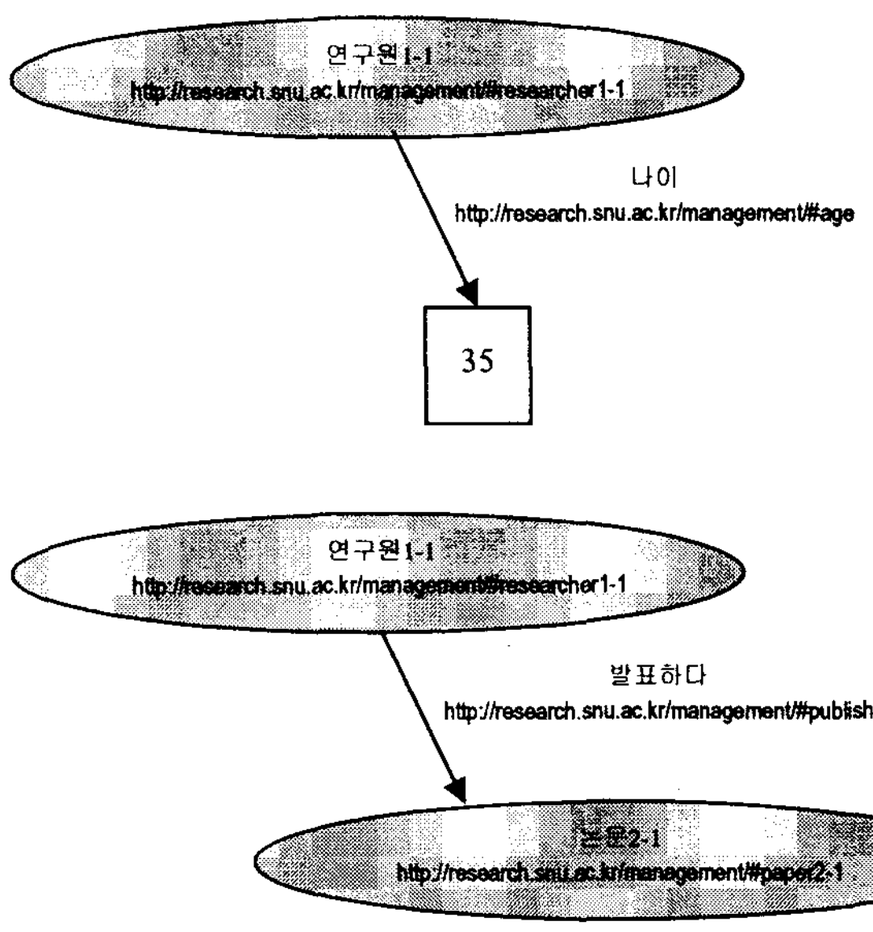

Figure 2 - RDF 트리플 그래프

RDF에는 공통적인 속성을 갖는 자원들을 클래스로 분류하고 이러한 클래스간의 위계구조를 설정하거나 속성의 주어부나 목적부가 될 수 있는 클래스를 제한할 수 있는 기능이 없다. 예를 들어 '김동건은 연구원이다' 와 '연구원은 사람이다' 라는 RDF 서술문이 있을 때 사람은 이것을 보고 '김동건은 사람이구나' 라고 추론할 수 있지만 컴퓨터는 전혀 알 수가 없다. RDF의 이러한 한계점은 2004년 2월에 W3C 권고안(recommendation)이 된 RDF Schema에 의해 어느 정도 해결된다. RDF를 프레임기반으로 확장한 RDF Schema를 이용하면 도메인의 구성 및 상호작용을 묘사하는데 필요한 기본 어휘와 가정들을 정의할 수 있다[Brickley et al., 2004].

2.4 SPARQL 질의

SPARQL은 W3C 데이터 액세스 워킹 그룹(Data Access Working Group)에 의해 연구가 진행되고 있는 RDF 질의 언어로 2007년 3월에 W3C 워킹 드래프트(Working Draft)가 발표된 상태이며[Prud'hommeaux and Seaborne, 2007], 2005년에 팀 버너스리가 발표한 시맨틱 웹 계층에 포함되어 있다. 이전의 RQL, TRIPLE, RDQL 같은 다른 RDF 질의 언어처럼 SPARQL도 SQL과 비슷한 선언적 문법을 사용한다. 이러한 모든 언어들은 질의에 대한 해를 찾기 위한 추론 기능을 제공한다. 그러나 결과 데이터 셋에 대한 텍스트 위주의 정렬기능만 제공할 뿐 자원을 실질적인 중요도에 따라 랭크해서 보여주지는 않는다. Figure 3은 4.1절의 Figure 10과 같은 스키마를 가지고 있는 작은 RDF 지식 베이스에 대해 '분야4-1(fields4-1)로 분류되는 키워드를 가지고 있는 논문(paper)을 발표한 연구원(researcher)과 논문, 논문이 실린 저널(journal)을 보여달라는 SPARQL 질의문이다.

```
PREFIX rs: <http://ontology.snu.ac.kr/research.owl#>
SELECT DISTINCT ?researcher ?paper ?journal
WHERE { ?researcher rs:published ?paper ;
                    rdf:type rs:Researcher .
        ?paper rs:printed_by ?journal ;
               rs:has_keyword ?k .
        ?k rs:classified as rs:field4-1 }
```

Figure 3 - SPARQL 쿼리문 예시

그리고 Figure 4는 이 질의문을 프로티지 3.2의 SPARQL 쿼리 패널(panel)에서 실행시켜 얻은 결과화면을 보여준다.

2.5 시맨틱 웹의 자원 중요도 - 목적부/주어부 점수(Objectivity/Subjectivity Score)

Bamba와 Mukherjea는 중요한 웹 페이지를 찾아내기 위해 효과적으로 사용되어온 월드와이드웹 링크 분석 기술을 수정하여, 시맨틱 웹에 대한 질의 결과를 자원 중요도에 따라 랭킹하는 데에 적용하였다 [Bamba and Mukherjea, 2004; Mukherjea et al., 2005]. 이들이 사용한 자원 중요도 계산 알고리즘은 Figure 1에 소개된 Kleinberg의 알고리즘과 유사하다. 단, 식 (1)의 I 오퍼레이션은 다음 식 (4)에 의해, 식 (2)의 O 오퍼레이션은 다음 식 (5)에 의해 대체된다.

$$O(n) = \sum_{(n_1,n)\in E} S(n_1) \times objWt(e) \qquad (4)$$

$$S(n) = \sum_{(n,n_1)\in E} O(n_1) \times subWt(e) \qquad (5)$$

위 식에서 $O(n)$ 은 자원 n 의 목적부 점수이고 $S(n)$ 은 자원 n 의 주어부 점수이며, $objWt(e)$ 는 링크 e 에 대한 속성의 목적부 가중치이고 $subWt(e)$ 는 링크 e 에 대한 속성의 주어부 가중치이다. 그리고 Kleinberg의 권위 점수는 목적부 점수(objectivity score), 허브 점수는 주어부 점수(subjectivity score)에 해당되며, 두 방법의 차이점은 가중치의 유무이다.

속성에 대한 목적부/주어부 가중치는 속성별로 미리 결정해 놓으며, 각 속성에 대해 서로 다른 목적부/주어부 가중치를 곱하는 것은 시맨틱 웹에서의 속성이 양방향에서 똑같이 중요하지 않을 수 있기 때문이다. 모든 속성의 디폴트 가중치는 1이고 속성에 따라 낮은 목적부 가중치나 주어부 가중치를 가지는 경우는 1보다 작은 양의 가중치로 변경된다.

예를 들어, 이들은 (inventor, invented, patent)라는 트리플이 있을 때 'inventor'의 중요도는 'patent'가 많을수록 올라가지만 'patent'의 중요도는 'inventor'가 많을수록 그대로 증가하는 것은 아니라고 생각하였다. 이것을 반영하기 위해 'invented'라는 속성에는 낮은 목적부 가중치를 주어 'patent'의 목적부 점수가 이 'patent'에 대한 'inventor'의 수에 의해 증가되는 영향력을 감소시켰다.

Mukherjea와 Bamba는 UMLS(unified Medical Language System), Biomedical Patent, TAP 시맨틱 웹에 이들의 자원 중요도 계산 알고리즘을 적용해보고 다음과 같은 한계점을 보고하였다 [Mukherjea et al., 2005].

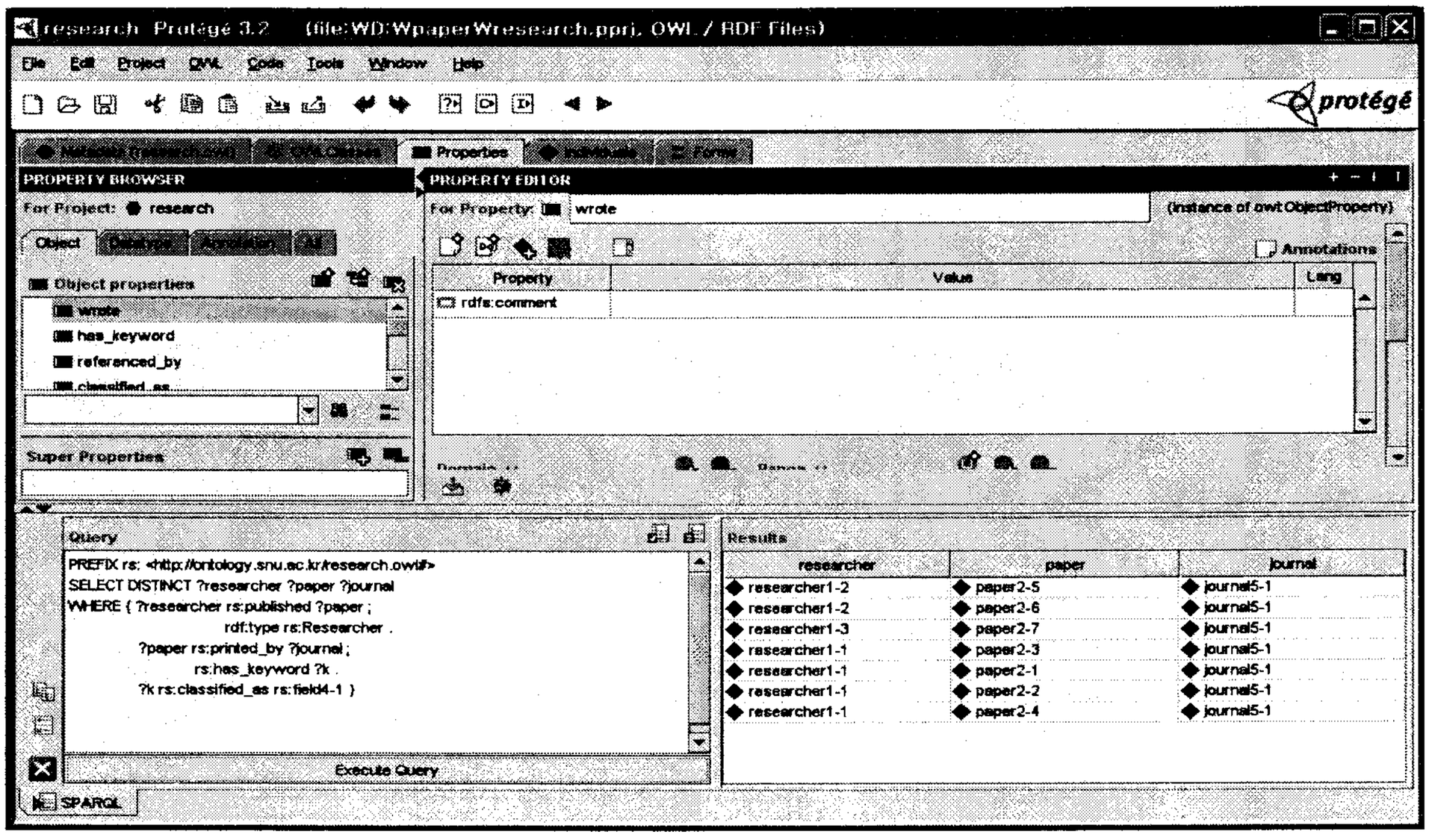

Figure 4 - 프로티지의 SPARQL 실행 및 결과 예

첫째는 도메인에 관련된 대부분의 지식이 표현된 시맨틱 웹에 대해서만 이 알고리즘이 유용하다는 것이며, 둘째는 별로 중요하지 않은 노드들이지만 이들 사이에 링크 연결이 많으면 실제로 중요한 노드들보다 더 높은 점수를 받게 되는 강한 결합 모임 효과(Tightly-Knit Community Effect)이다. 이것은 링크 분석 알고리즘의 근본적인 문제라고 할 수 있다. 셋째는 어떤 자원이 진짜 중요해서가 아니라 매우 흔하기 때문에 높은 점수를 받게 되는 경우가 있다는 것이다.

III. 클래스중심 가중치 설정 알고리즘

3.1 기본 아이디어

3.1.1 속성 중심에서 클래스 중심으로

Mukherjea와 Bamba의 연구에서는 RDF 그래프를 이루고 있는 트리플의 속성이 다양하며, 목적부와 주어부의 중요도를 계산할 때 속성에 따라 양방향 가중치가 모두 1이 아닐 수 있다는 점에 착안하였다. Figure 5에서와 같이 '발명하다'라는 속성을 살펴보면, 발명가의 주어부 점수는 특허의 목적부 점수를 그대로 더하여 계산하지만 특허의 목적부 점수는 특허를 발명한 발명가의 주어부 점수에 1보다 작은 양의 가중치를 곱하여 계산한다. 이것은 하나의 특허를 발명한 발명가들이 여러 명 존재할 경우 단순히 발명가 수가 많아서 해당 특허의 목적부 점수가 높아지는 문제점을 완화하기 위한 것이다.

Figure 5 - '발명하다'의 가중치 설정 예

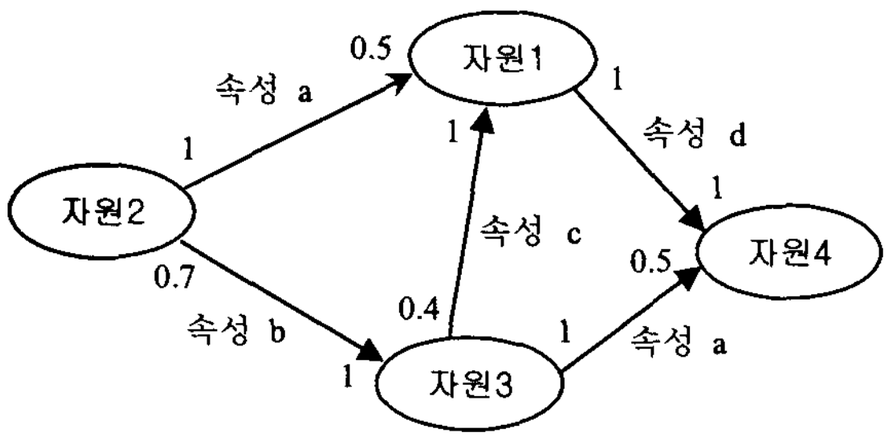

Figure 6 - 속성중심 가중치 설정 예

Figure 6은 이런 방식으로 가중치가 설정된 RDF 그래프의 일부분을 보여준다. 그런데 랭킹을 요하는 거의 모든 질의가 주어부나 목적부에 오는 임의의 클래스에 속하는 자원을 대상으로 이루어진다는 것을 생각해보면 속성이 아닌 클래스 중심으로 속성의 가중치를 설정하는 것이 더욱 합리적이라는 생각을 할 수 있다. 다음 Figure 7은 똑같은 속성 '졸업하다'가 차지하는 비중이 교수 클래스에 속하는 자원과 사업가 클래스에 속하는 자원의 중요도를 계산할 때 서로 다를 수 있음을 보여준다. 대부분의 경우 학력은 사업가 보다는 교수에게 더욱 중요한 영향을 미칠 수 있기 때문이다.

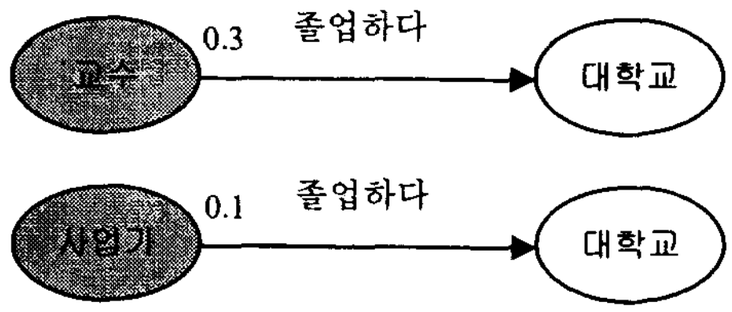

Figure 7 - 클래스에 따라 달라지는 속성 가중치 예

이러한 생각을 확장하면 한 클래스에 속하는 자원의 중요도에 영향을 미치는 속성들의 상대적인 비중을 고려하여 클래스를 중심으로 해당 클래스와 연결된 속성들의 상대적인 가중치를 설정할 수 있을 것이다. 이에 대한 예는 4.1절의 Figure 10과 Figure 11 및 [부록]의 Table 1을 참조하기 바란다.

어떤 클래스에 연결된 속성이 그 클래스에 속하는 자원의 중요도와는 무관하다면 해당 속성에 '0'의 가중치를 설정하여 관계없는 영향력을 제거할 수 있을 것이다. 이렇게 하면 강한 결합 모임효과는 거의 해결될 것으로 예상된다.

3.1.2 데이터타입속성의 반영

기존 연구에서는 데이터타입속성(owl:DatatypeProperty)에 해당되는 링크는 RDF 그래프에서 제거하고 객체속성(owl:ObjectProperty)만으로 중요도를 계산하였다. 이것은 데이터타입속성의 목적부에 오는 값이 자원이 아닌 단순 데이터이고, 이 값에 연결되는 다른 링크가 없기 때문일 것이다. 그런데 어떤 데이터타입속성은 해당 클래스에 속하는 자원의 중요도에 많은 영향력을 미칠 수 있다. Figure 8에서와 같이 책의 중요도에는 판매부수가, 출판사의 중요도에는 매출액이 큰 비중을 차지할 수 있는 바와 같다.

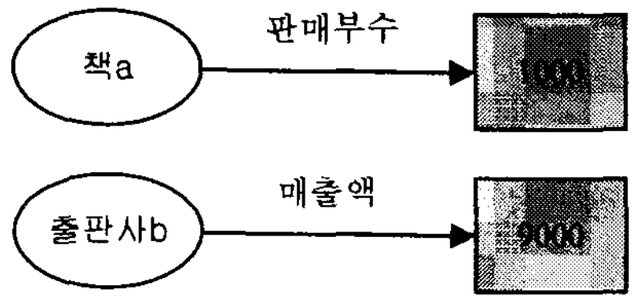

Figure 8 - 자원 중요도에 반영되어야 할 데이터타입 속성 예

이러한 데이터타입속성을 고려할 수 있는 방법에는 크게 두 가지가 있을 것이다. 첫째는 객체속성만을 고려하여 계산된 자원의 중요도에 유의미한 데이터타입속성의 값을 점수화하여 사후적으로 더해주는 방법이다. Figure 8의 여러 책에 대한 판매부수 값을 0에서 1사이의 값을 갖도록 정규화할 수 있고, 링크 분석 방법으로 구해진 자원 중요도도 0에서 1사이의 값으로 정규화하여 책 클래스의 속성 가중치대로 합산할 수 있을 것이다. 이 방법은 데이터타입 속성이 이와 직접적으로 연결된 자원의

중요도에만 영향을 주고, 해당 자원과 연결된 다른 자원들과의 상호작용에서는 제외된다는 단점이 있다. 둘째는 데이터타입속성 값에 해당하는 더미(dummy) 자원을 만들어 링크 분석 계산에 처음부터 포함시키는 것이다. 이것은 데이터타입속성 값을 0에서 1사이의 값으로 정규화하여 이 값에 비례하도록 해당 링크의 가중치를 설정하면 가능해진다.

3.2 중요도 계산 알고리즘과 행렬 연산

3.2.1 중요도 계산 알고리즘 분석

RDF 그래프 $G=(V,E)$ 에서 V 는 n 개의 자원으로 이루어진 자원 집합 $V=\{r_1,r_2,\cdots,r_n\}$ 이고, E 는 V 안에 존재하는 임의의 자원 $r_i\,(1\le i\le n)$ 와 $r_j\,(1\le j\le n)$ 를 연결하는 방향성 있는 링크의 집합이라 하자. 이 때 행렬 A 를 다음과 같이 그래프 G 의 링크 연결구조를 나타내도록 정의한다.

$A_{ij}=1$ (r_i 에서 r_j 로 향하는 링크가 있을 때),

$A_{ij}=0$(r_i에서 r_j로 향하는 링크가 없을 때)

그러면 Figure 1의 i 번째 반복단계는 $x_i^{\bullet}=A^T y_{i-1}$, $y_i^{\bullet}=Ax_i^{\bullet}$ 로 나타낼 수 있고, 정규화 과정을 고려하면 결국 x_i 는 $(A^TA)^{i-1}A^Tz$, y_i 는 $(AA^T)^iz$ 방향으로 향하는 단위 벡터임을 알 수 있다. 여기에서 행렬의 거듭제곱에 관한 성질을 적용하면, $(A^TA)^{i-1}A^Tz$ 는 A^TA 의 제일 고유벡터 방향으로, $(AA^T)^iz$ 는 AA^T 의 제일 고유벡터 방향으로 수렴하게 된다[Burden and Faires, 2001]. 그러므로 권위 점수(x_i)는 A^TA 의 제일 고유벡터에, 허브 점수(y_i)는 AA^T 의 제일 단위 고유벡터에 해당된다. 시맨틱 웹에서 목적부와 주어부 점수를 계산할 때에는 가중치 설정 방식의 변화에 따라 행렬의 원소가 바뀌므로 링크의 목적부 가중치를 나타내는 행렬은 B 로, 주어부 가중치를 나타내는 행렬은 C 로 표기하도록 하자. 중요도를 계산하는 알고리즘의 단계는 Figure 1과 같으므로 목적부 점수 벡터(o_i)는 $(BC)^{i-1}Bz$ 방향, 주어부 점수 벡터(s_i)는 $(CB)^iz$ 방향으로의 단위벡터임을 알 수 있다. 그러므로 목적부 점수 벡터는 BC 의 제일 단위 고유벡터에, 주어부 점수 벡터는 CB 의 제일 단위 고유벡터에 해당된다. 그런데 BC 와 CB 는 A^TA 나 AA^T 와 같이 항상 대칭행렬이 되는 건 아니므로 대각화 가능성을 점검해야 한다. 이것은 매트랩과 같은 수학 프로그램으로 고유벡터 행렬의 랭크(rank)를 구하여 쉽게 확인할 수 있다.

3.2.2 클래스 중심 목적부/주어부 가중치 행렬 예

아주 간단한 예로 클래스 구성과 속성 가중치가 Figure 9와 같은 도메인이 있고, 각 클래스에 속하는 인스턴스가 하나씩만 있다고 가정하자. Figure 1의 i 번째 반복단계에서의 목적부 점수 계산은 식 (6)과 같이, 주어부 점수 계산은 식 (7)과 같이 이루어진다.

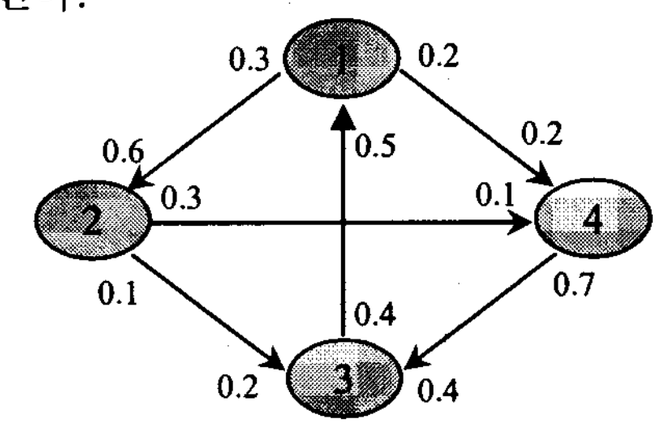

Figure 9 - 클래스 중심 가중치 설정 예

$$o_i^{\bullet}=Bs_{i-1}$$

$$\begin{bmatrix}o_i^{\bullet 1}\\o_i^{\bullet 2}\\o_i^{\bullet 3}\\o_i^{\bullet 4}\end{bmatrix}=\begin{bmatrix}0&0&0.5&0\\0.6&0&0&0\\0&0.2&0&0.4\\0.2&0.1&0&0\end{bmatrix}\begin{bmatrix}s_{i-1}^1\\s_{i-1}^2\\s_{i-1}^3\\s_{i-1}^4\end{bmatrix}\qquad(6)$$

$$s_i^{\bullet}=Co_i^{\bullet}$$

$$\begin{bmatrix}s_i^{\bullet 1}\\s_i^{\bullet 2}\\s_i^{\bullet 3}\\s_i^{\bullet 4}\end{bmatrix}=\begin{bmatrix}0&0.3&0&0.2\\0&0&0.1&0.3\\0.4&0&0&0\\0&0&0.7&0\end{bmatrix}\begin{bmatrix}o_i^{\bullet 1}\\o_i^{\bullet 2}\\o_i^{\bullet 3}\\o_i^{\bullet 4}\end{bmatrix}\qquad(7)$$

$o_i^{\bullet}$ 와 $s_i^{\bullet}$ 는 각각 정규화하기 전의 목적부 점수 벡터와 주어부 점수 벡터이고, B 는 목적부 가중치 행렬, C 는 주어부 가중치 행렬이다. 정규화는 점수 벡터의 크기가 1이 되도록 크기만 조정하는 것이므로 벡터의 방향성에는 영향을 주지 않는다.

3.3 클래스 중심 자원 중요도 계산 알고리즘

본 논문에서는 Figure 1과 같은 반복적인 알고리즘을 사용하지 않고 3.2절의 분석 결과를 바탕으로 다음과 같은 자원 중요도 계산 알고리즘을 구현하였다.

3.3.1 전체 알고리즘

① 온톨로지 스키마 상에서 클래스별로 객체타입 속성에 대한 목적부 가중치와 주어부 가중치, 데이터타입 속성에 대한 가중치를 설정한다.

② 각 클래스의 모든 인스턴스(자원)를 노드로 하는 RDF 그래프에 대해 목적부 가중치 행렬 B 와 주어부 가중치 행렬 C 를 만든다.

③ 행렬 BC의 제일 고유벡터를 계산하여 목적부

점수 벡터를 구하고, 행렬 CB의 제일 고유벡터를 계산하여 주어부 점수 벡터를 구한다.

④ 목적부 점수 벡터와 주어부 점수 벡터를 합하여 최종 랭크 벡터를 구한다.

3.3.2 데이터타입속성의 정규화 및 가중치 변환

클래스 수준에서 설정된 데이터타입 속성 q에 대한 가중치를 $dpwt_q$라 하자. $dpwt_q > 0$인 데이터타입 속성 q에 대한 인스턴스 i의 정규화 점수 g_{qi} ($0 \le g_{qi} \le 1$)는 다음 식 (8)과 같이 계산한다. val_{qi}는 데이터타입 속성 q에 대한 자원 i의 속성값이고, $\min_c$는 데이터타입 속성 q에 대한 클래스 c의 최소 속성값이며, $\max_c$는 데이터타입 속성 q에 대한 클래스 c의 최대 속성값이다.

$$g_{qi} = \frac{val_{qi} - \min_c}{\max_c - \min_c} \quad (8)$$

다음으로 인스턴스 i를 주어부로, i의 데이터타입 속성 q에 대한 속성값인 더미 자원을 목적부로 하는 링크의 목적부 가중치는 1로, 주어부 가중치 dl_sbwt_{qi}는 다음 식 (9)와 같이 설정한다.

$$dl\_sbwt_{qi} = \text{조정계수} \times g_{qi} \times dpwt_q \quad (9)$$

IV. 구현 및 실험 평가

4.1 실험환경 및 데이터 집합

본 논문에서는 객체속성만을 반영하는 시나리오 1과 객체속성과 데이터타입속성을 모두 고려하는 시나리오 2로 나누어 실험 한 결과를 제시할 것이다. 시나리오 1에서는 속성중심으로 가중치를 설정하는 기존 방법(PreRI : Predicate-oriented Resource Importance)과 클래스 중심으로 가중치를 설정하는 방법(ClaRI : Class-oriented Resource Importance)을 비교 분석할 것이다. 그리고 시나리오 2에서는 ClaRI 방식으로 링크 구조를 분석하여 얻은 점수와 데이터타입속성 값을 정규화하여 미리 설정된 가중치대로 합산하는 방법(A)과, 데이터타입속성 값을 인스턴스별 링크 가중치로 변환하여 링크 분석에 처음부터 포함시켜 계산하는 방법(B)을 살펴보기로 하겠다. 시나리오 1은 Figure 10과 같은 스키마를 가진 도메인을 대상으로 하고 있으며, 시나리오 2는 시나리오 1에서 '동호회'와 '홈피' 클래스가 사라지고 데이터타입속성이 추가된 Figure 11과 같은 도메인을 바탕으로 한다. 온톨로지 구성에 있어 RDF Schema 이상에서 제공되는 클래스 간 위계구조와 OWL에서 제공되는 속성 간 위계구조는 단순화하여 모두 한 계층만 있는 것으로 가정하였다. 속성들에 대한 가중치는 각 케이스에 맞는 방식으로 개연성에 따라 설정해 보았으며, 컨텍스트에 따라 다른 가중치를 사용할 수도 있을 것이다. 설정된 가중치 값에 따라 각 방법의 결과가 약간씩 달라질 수는 있겠지만 전반적인 효과성의 비교에는 큰 영향을 주지 않을 것으로 판단된다. 전체적인 실험 내용은 Table 1에 요약되어 있다.

Table 1 – 전체 실험 구성

	케이스 내용	가중치 설정 표	도메인 구성도	고려 속성
시나리오 1	PreRI : 속성중심 가중치 설정 (목적부/주어부 점수 계산)	Table 2 (속성 뷰)	Figure 10	객체속성
	ClaRI : 클래스중심 가중치 설정 (목적부/주어부 점수 계산)	Table 3 (속성 뷰)		
시나리오 2	A: 데이터타입속성값을 따로 구해 가중평균	[부록]의 Table 1 (클래스 뷰)	Figure 11	객체속성 + 데이터타입속성
	B: 데이터타입속성값을 링크 가중치로 변환하여 링크 분석에 포함			

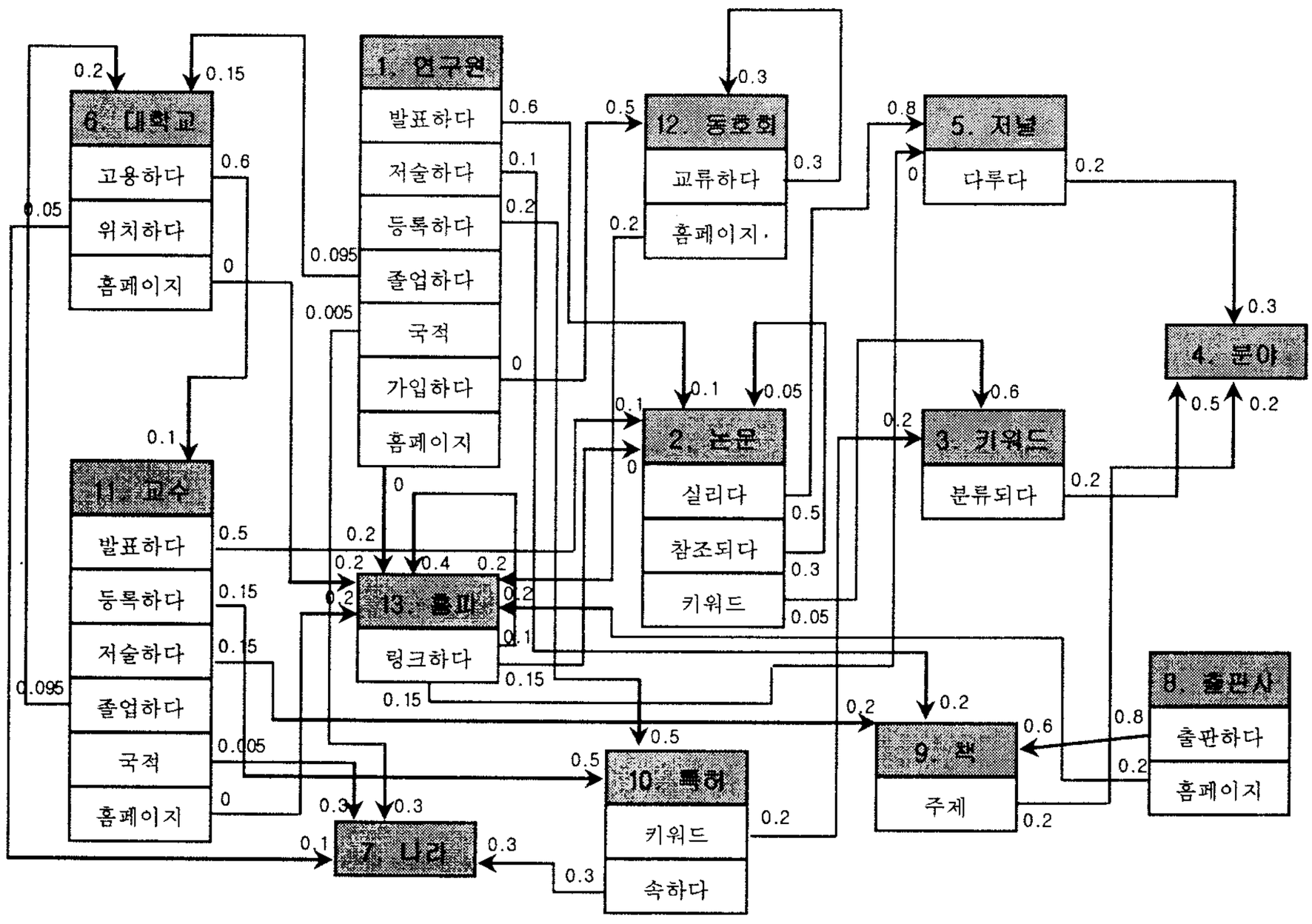

Figure 10 – 시나리오 1의 클래스 구성

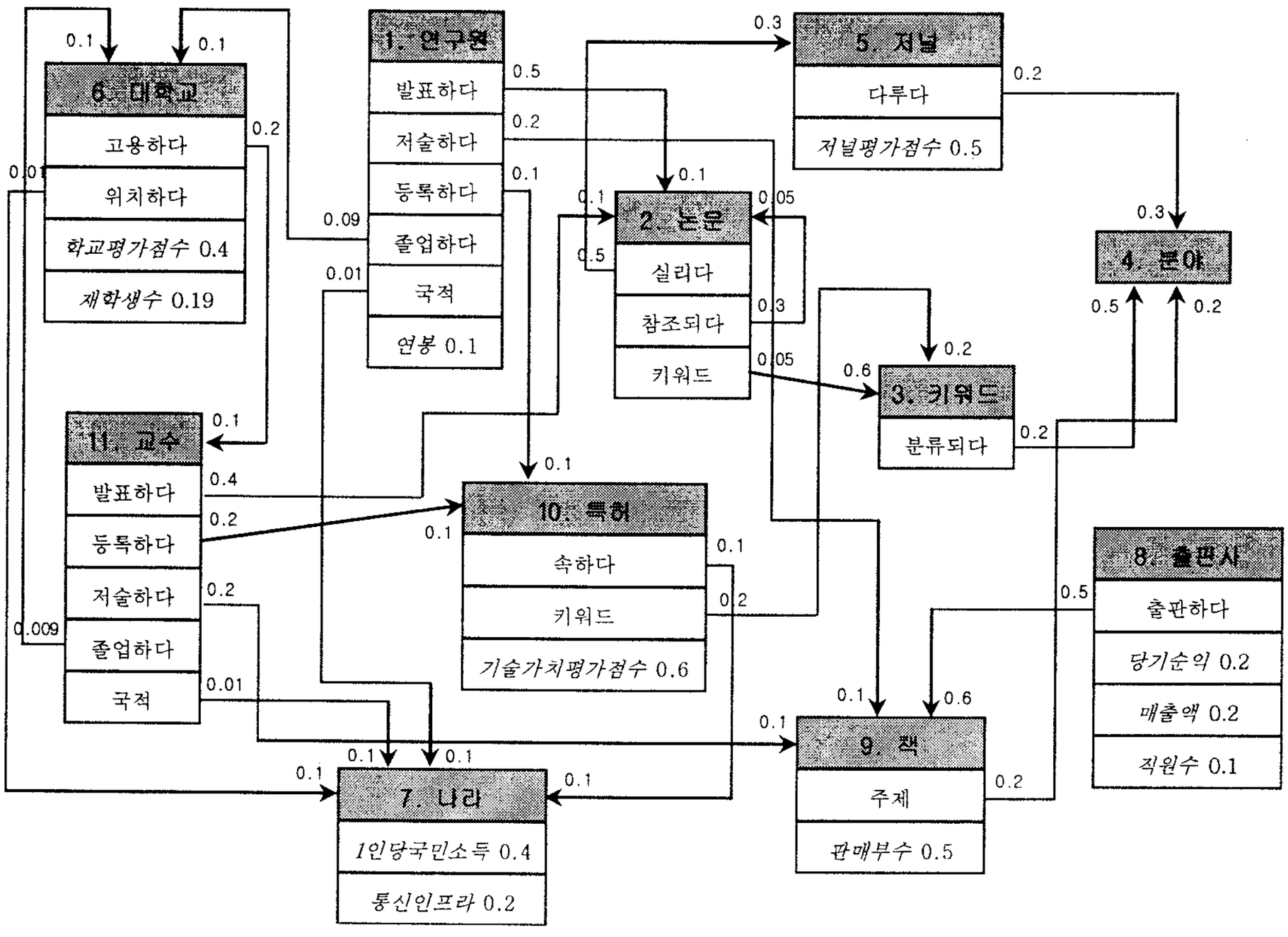

Figure 11 – 시나리오 2의 클래스 구성

Table 2 - 시나리오1 : PreRI 속성 가중치(속성 뷰)

	domain	predicate	range	subwt	objwt
1	교수	국적	나라	0.3	1
2	교수	등록하다	특허	1	0.2
3	교수	발표하다	논문	1	0.3
4	교수	저술하다	책	1	0.3
5	교수	졸업하다	대학교	1	0.8
6	교수	홈페이지	홈피	0.8	1
7	논문	실리다	저널	1	1
8	논문	참조되다	논문	1	0.4
9	논문	키워드	키워드	0.6	1
10	대학교	고용하다	교수	1	0.8
11	대학교	위치하다	나라	0.4	1
12	대학교	홈페이지	홈피	0.8	1
13	동호회	홈페이지	홈피	0.8	1
14	동호회	교류하다	동호회	1	1
15	연구원	가입하다	동호회	0.8	1
16	연구원	국적	나라	0.3	1
17	연구원	등록하다	특허	1	0.2
18	연구원	발표하다	논문	1	0.3
19	연구원	저술하다	책	1	0.3
20	연구원	졸업하다	대학교	1	0.8
21	연구원	홈페이지	홈피	0.8	1
22	저널	다루다	분야	0.4	1
23	책	주제	분야	0.5	1
24	출판사	출판하다	책	1	0.7
25	출판사	홈페이지	홈피	0.8	1
26	키워드	분류되다	분야	0.4	1
27	특허	속하다	나라	0.2	1
28	특허	키워드	키워드	0.6	1
29	홈피	링크하다	논문	1	0.8
30	홈피	링크하다	저널	1	0.8
31	홈피	링크하다	홈피	1	0.8

Table 3 - 시나리오1 : ClaRI 속성 가중치(속성 뷰)

	domain	predicate	range	subwt	objwt
1	교수	국적	나라	0.005	0.3
2	교수	등록하다	특허	0.15	0.5
3	교수	발표하다	논문	0.5	0.1
4	교수	저술하다	책	0.15	0.2
5	교수	졸업하다	대학교	0.095	0.2
6	교수	홈페이지	홈피	0	0.2
7	논문	실리다	저널	0.5	0.8
8	논문	참조되다	논문	0.3	0.05
9	논문	키워드	키워드	0.05	0.6
10	대학교	고용하다	교수	0.6	0.1
11	대학교	위치하다	나라	0.05	0.1
12	대학교	홈페이지	홈피	0	0.2
13	동호회	교류하다	동호회	0.3	0.3
14	동호회	홈페이지	홈피	0.2	0.2
15	연구원	가입하다	동호회	0	0.5
16	연구원	국적	나라	0.005	0.3
17	연구원	등록하다	특허	0.2	0.5
18	연구원	발표하다	논문	0.6	0.1
19	연구원	저술하다	책	0.1	0.2
20	연구원	졸업하다	대학교	0.095	0.15
21	연구원	홈페이지	홈피	0	0.2
22	저널	다루다	분야	0.2	0.3
23	책	주제	분야	0.2	0.2
24	출판사	출판하다	책	0.8	0.6
25	출판사	홈페이지	홈피	0.2	0.2
26	키워드	분류되다	분야	0.2	0.5
27	특허	속하다	나라	0.3	0.3
28	특허	키워드	키워드	0.2	0.2
29	홈피	링크하다	논문	0.15	0
30	홈피	링크하다	저널	0.15	0
31	홈피	링크하다	홈피	0.1	0.4

Table 4는 각 시나리오에서 사용된 클래스들의 인스턴스 수와 이러한 인스턴스들간의 관계 및 데이터타입속성 값을 기술한 트리플의 총 수를 보여준다. 시나리오 2에서 괄호 안에 있는 수는 데이터타입속성에 대한 더미 자원의 수를 나타낸다.

4.2 실험 평가 내용 및 방법

4.2.1 시나리오 1 : 객체속성만을 고려한 비교 분석

시나리오 1에서는 ClaRI에 의한 강한 결합 모임 효과의 해소를 관찰할 수 있는 대상으로 연구원 클래스를 선정하였다. 시나리오 1의 두 가지 방식은 모두 같은 트리플 집합을 사용하였고, 이 안에 포함된 연구원 인스턴스들의 속성 값을 분석해 보면 대략 Table 5와 같다.

모든 트리플 정보를 구성함에 있어 인스턴스와 속성의 이름은 간결성을 위해 URL과 '#'이 없는 단편 식별자(fragment identifier) 형태를 사용하였고, 인스턴스 이름은 '클래스이름 클래스번호-인스턴스번호' 형식으로 부여하였다. 인스턴스번호가 작을수록 Figure 10이나 Table 3 의 기준에 의해 대략 높은 점수를 가지도록 속성값을 설정하여 예상한대로 랭킹 점수 결과가 나오는지를 살펴볼 것이다. 단적인 예로 Table 5에서 살펴볼 수 있는 바와 같이 '연구원 1-1'은 논문을 10편 발표한 반면에 '연구원 1-25'는 발표한 논문이 하나도 없다. 강한 결합을 형성하기 위해 '연구원 21~25'는 동호회에, '연구원 17~25'는 홈피에 연결하였고 동호회와 홈피, 홈피와 홈피, 홈피와 다른 클래스 간에도 링크를 만들어 주었다. '연구원 1-25'는

연구원 중요도 평가에 반영되지 않는 동호회에는 5개나 가입되어 있다.

Table 4 - 시나리오 인스턴스와 트리플 수

	시나리오 1	시나리오 2
1. 연구원	25	20(20×1)
2. 논문	100	100
3. 키워드	15	17
4. 분야	5	5
5. 저널	5	5(5×1)
6. 대학교	3	3(3×2)
7. 나라	3	3(3×2)
8. 출판사	3	3(3×3)
9. 책	15	10(10×1)
10. 특허	10	10(10×1)
11. 교수	9	12
12. 동호회	5	0
13. 홈피	30	0
인스턴스 총 수	228	253
트리플 총 수	1160	873

이 외에 클래스중심 가중치 설정 방법이 다른 클래스에 대해서도 주어진 트리플 정보에 부합하는 랭킹 순위를 보여주는 지와, 특정한 한 자원의 중요도에 영향을 주는 링크 정보를 추가하거나 삭제했을 때 실제로 해당 자원의 랭킹 점수에 영향이 있는지 등의 기본적인 평가도 수행할 것이다.

4.2.2 시나리오 2 : 데이터타입속성도 반영한 경우

시나리오 2에서는 [부록]의 Table 1에서 데이터타입 속성 반영 비율이 높으면서 클래스 인스턴스의 수가 적지 않은 '책' 클래스를 선정하여 A와 B 방법의 적용 결과를 살펴볼 것이다. 데이터타입속성인 '판매부수'에 대한 인스턴스별 속성값은 실험 결과를 보여주는 Table 9와 Table 10에 함께 제시되어 있다.

4.3 구현

일단 RDF 지식베이스의 구축과 쿼리 생성 및 랭킹 점수에 의한 쿼리 결과의 정렬에 이르는 프로세스 예는 Figure 12와 같다.

Table 6 - 시나리오 1 : 연구원 클래스 인스턴스별 속성 값

	논문 수 (논문번호)	책 수 (책번호)	특허 수 (특허번호)	학교	교적	동호회 수 (동호회번호)	홈피유무 (홈피번호)
연구원 1-1	10 편(1~10)	4 권(1~4)	3 개(1~3)	대학교 6-1	나라 7-1	0	no
연구원 1-2	8 편(11~18)	3 권(5~7)	2 개(4~5)	대학교 6-1	나라 7-1	0	no
연구원 1-3	7 편(19~25)	1 권(8)	0	대학교 6-1	나라 7-1	0	no
연구원 1-4	6 편(26~31)	0	0	대학교 6-1	나라 7-1	0	no
연구원 1-5	5 편(32~36)	0	0	대학교 6-1	나라 7-1	0	no
연구원 1-6	5 편(37~41)	0	0	대학교 6-1	나라 7-1	0	no
연구원 1-7	5 편(42~46)	0	0	대학교 6-1	나라 7-1	0	no
연구원 1-8	5 편(47~51)	0	0	대학교 6-1	나라 7-1	0	no
연구원 1-9	5 편(52~56)	0	0	대학교 6-1	나라 7-1	0	no
연구원 1-10	4 편(57~60)	0	0	대학교 6-1	나라 7-2	0	no
연구원 1-11	4 편(61~64)	0	0	대학교 6-2	나라 7-2	0	no
연구원 1-12	4 편(65~68)	0	0	대학교 6-2	나라 7-2	0	no
연구원 1-13	4 편(69~72)	0	0	대학교 6-2	나라 7-2	0	no
연구원 1-14	4 편(73~76)	0	0	대학교 6-2	나라 7-2	0	no
연구원 1-15	4 편(77~80)	0	0	대학교 6-2	나라 7-2	0	no
연구원 1-16	3 편(81~83)	0	0	대학교 6-2	나라 7-2	0	no
연구원 1-17	3 편(84~86)	0	0	대학교 6-2	나라 7-2	0	yes(10)
연구원 1-18	3 편(87~89)	0	0	대학교 6-2	나라 7-3	0	yes(9)
연구원 1-19	2 편(90~91)	0	0	대학교 6-3	나라 7-3	0	yes(8)
연구원 1-20	2 편(92~93)	0	0	대학교 6-3	나라 7-3	0	yes(7)
연구원 1-21	2 편(94~95)	0	0	대학교 6-3	나라 7-3	3(1~3)	yes(6)
연구원 1-22	2 편(96~97)	0	0	대학교 6-3	나라 7-3	3(1~3)	yes(5)
연구원 1-23	1 편(98)	0	0	대학교 6-3	나라 7-3	3(1~3)	yes(4)
연구원 1-24	1 편(99)	0	0	대학교 6-3	나라 7-3	4(1~4)	yes(3)
연구원 1-25	0	0	0	대학교 6-3	나라 7-3	5(1~5)	yes(1,2)

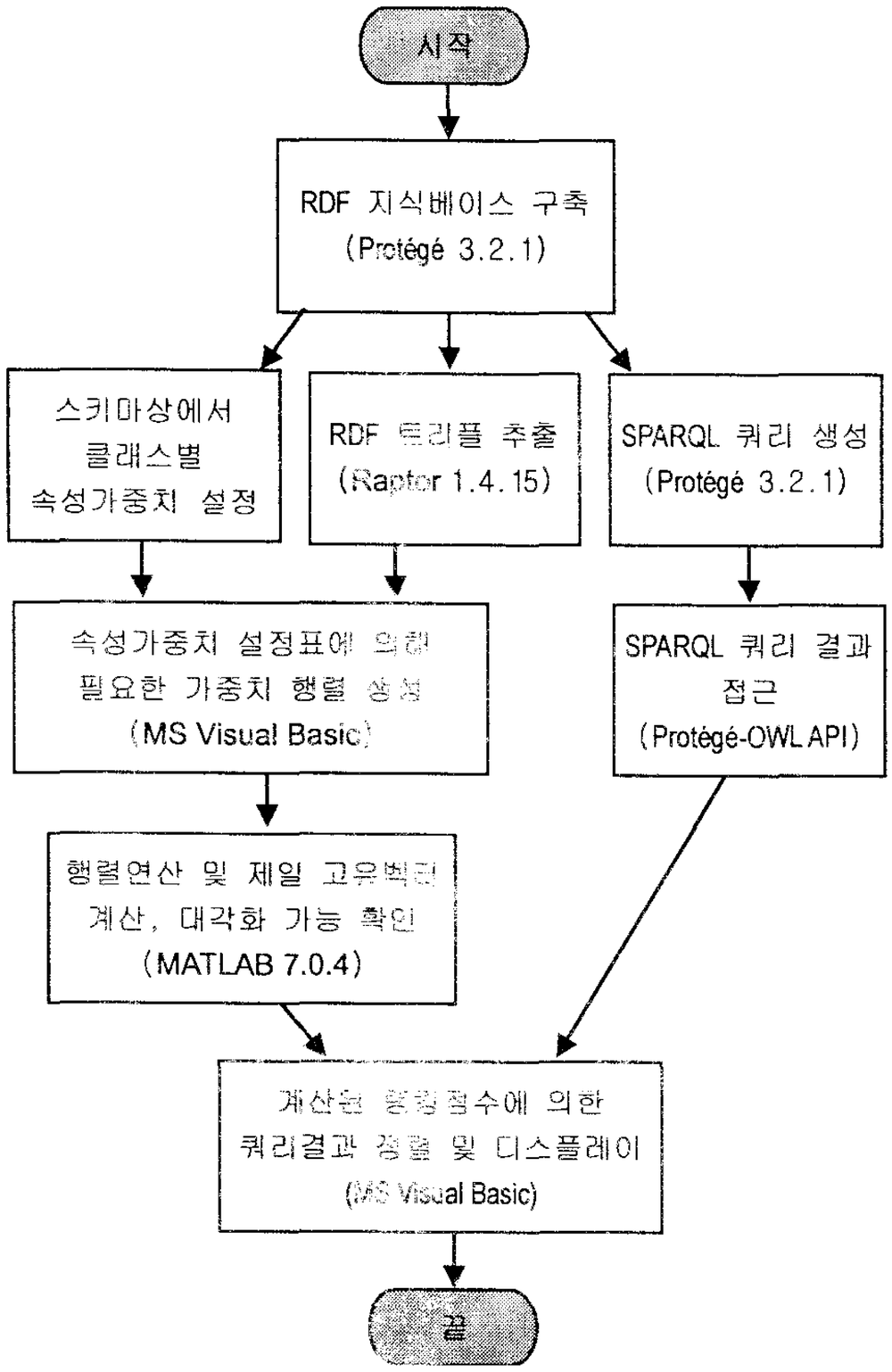

Figure 12 - 전체적인 프로세스 예

본 논문에서는 효율적인 트리플 정보의 구성을 위해 해당 RDF 지식베이스 파일에 대한 RDF 트리플들이 추출되어 있다고 가정하고, MS 엑셀 2003으로 트리플들을 실험 목적에 맞게 구성하여 사용하였다. 가중치 행렬은 MS 비주얼 베이직으로 프로그래밍하여 만들었다. 매트랩(MatLab) 7.0.4로 가중치 행렬의 연산을 수행하고 대각화 가능성을 확인하였으며 매트랩의 결과물(output)을 엑셀 파일로 받아 최종 랭킹 점수를 계산하였다.

4.4 실험 결과 및 평가

4.4.1 시나리오 1의 결과 및 평가

시나리오 1의 PreRI에 의한 연구원 클래스의 랭킹 결과는 Table 6과 같다. 목적부 점수(objectivity)가 모두 0인 것은 Figure 10의 스키마에서 알 수 있는 바와 같이 연구원 클래스에 속하는 인스턴스는 트리플의 목적부가 될 수 없고 주어부에만 올 수 있기 때문이다.

속성 중심으로 가중치를 설정했을 때에는 논문을 7편 발표하고 책을 1권 저술한 '연구원1-3'이나, 논문을 6편 발표한 '연구원1-4'보다 논문을 한 편도 쓰지 않은 '연구원1-25'가 훨씬 높게 랭크 되었음을 살펴볼 수 있다. 그리고 동호회나 홈피로 연결된 다른 연구원들의 중요도도 높이 평가되어 있음을 알 수 있다.

Table 6 - 시나리오 1 : PreRI 연구원 랭킹 결과

	objectivity	subjectivity	ranking score
연구원 1-1	0.0000000000000	0.0280030507814	0.0280030507814
연구원 1-2	0.0000000000000	0.0213430021104	0.0213430021104
연구원 1-25	0.0000000000000	0.0180022646307	0.0180022646307
연구원 1-18	0.0000000000000	0.0127843375684	0.0127843375684
연구원 1-17	0.0000000000000	0.0127130956031	0.0127130956031
연구원 1-20	0.0000000000000	0.0120720263736	0.0120720263736
연구원 1-19	0.0000000000000	0.0117446526760	0.0117446526760
연구원 1-24	0.0000000000000	0.0096891442422	0.0096891442422
연구원 1-21	0.0000000000000	0.0095702531798	0.0095702531798
연구원 1-22	0.0000000000000	0.0095702531798	0.0095702531798
연구원 1-23	0.0000000000000	0.0095574853939	0.0095574853939
연구원 1-15	0.0000000000000	0.0050150163038	0.0050150163038
연구원 1-16	0.0000000000000	0.0038441292781	0.0038441292781
연구원 1-14	0.0000000000000	0.0026835036810	0.0026835036810
연구원 1-3	0.0000000000000	0.0006218765377	0.0006218765377
연구원 1-4	0.0000000000000	0.0005802088374	0.0005802088374
연구원 1-5	0.0000000000000	0.0005623722545	0.0005623722545
연구원 1-6	0.0000000000000	0.0005620909782	0.0005620909782
연구원 1-7	0.0000000000000	0.0005594629999	0.0005594629999
연구원 1-8	0.0000000000000	0.0005506387458	0.0005506387458
연구원 1-9	0.0000000000000	0.0005506387458	0.0005506387458
연구원 1-10	0.0000000000000	0.0004709909670	0.0004709909670
연구원 1-11	0.0000000000000	0.0003827768125	0.0003827768125
연구원 1-12	0.0000000000000	0.0003827768125	0.0003827768125
연구원 1-13	0.0000000000000	0.0003599564785	0.0003599564785

반면에 Table 7에서는 일련번호 순서가 랭크 순위와 거의 일치함을 확인할 수 있다. 여기에서도 Table 6과 같은 이유로 목적부 점수는 모두 0이다. 목적부 점수와 주어부 점수가 다 양의 값을 갖는 클래스 예로 Table 8에 특허 클래스의 랭킹 결과를 제시하였다.

이처럼 클래스 중심으로 가중치를 설정하면 아무리 강한 결합을 보이는 노드들이 있어도 중요도에 영향을 미치지 않는 링크들은 제외시키는 효과가 있으므로 훨씬 안정적이라고 할 수 있다. 강한 결합 모임 말고도 기존 연구의 다른 한계점인 정보 표현의 완전성에 대해서도 효율적인 지침을 제시한다. 온톨로지 스키마 상에서 중요도에 영향을 미치는 속성들에 대해서는 누락된 정보가 없어야 정확한 랭킹 점수를 얻을 수 있다는 것은 당연한 결과인 것이다. 어떤 자원이 흔하기 때문에 높은 점수를 받는 현상도 결국 강한 결합 모임 효과와 맥락을 같이 한다.

Table 7 - 시나리오 1 : ClaRI 연구원 랭킹 결과

	objectivity	subjectivity	ranking score
연구원 1-1	0.0000000000000	0.0155959095082	0.0155959095082
연구원 1-2	0.0000000000000	0.0051286892012	0.0051286892012
연구원 1-3	0.0000000000000	0.0033880976032	0.0033880976032
연구원 1-4	0.0000000000000	0.0028516085955	0.0028516085955
연구원 1-5	0.0000000000000	0.0022749359582	0.0022749359582
연구원 1-6	0.0000000000000	0.0022641151361	0.0022641151361
연구원 1-7	0.0000000000000	0.0018189884972	0.0018189884972
연구원 1-8	0.0000000000000	0.0001034070826	0.0001034070826
연구원 1-9	0.0000000000000	0.0001034070826	0.0001034070826
연구원 1-10	0.0000000000000	0.0000528428733	0.0000528428733
연구원 1-11	0.0000000000000	0.0000094887231	0.0000094887231
연구원 1-12	0.0000000000000	0.0000094887231	0.0000094887231
연구원 1-13	0.0000000000000	0.0000088860387	0.0000088860387
연구원 1-15	0.0000000000000	0.0000088331428	0.0000088331428
연구원 1-14	0.0000000000000	0.0000088331428	0.0000088331428
연구원 1-17	0.0000000000000	0.0000087888710	0.0000087888710
연구원 1-16	0.0000000000000	0.0000087888710	0.0000087888710
연구원 1-18	0.0000000000000	0.0000083393723	0.0000083393723
연구원 1-20	0.0000000000000	0.0000052870769	0.0000052870769
연구원 1-19	0.0000000000000	0.0000052870769	0.0000052870769
연구원 1-21	0.0000000000000	0.0000052870769	0.0000052870769
연구원 1-22	0.0000000000000	0.0000052870769	0.0000052870769
연구원 1-24	0.0000000000000	0.0000052608409	0.0000052608409
연구원 1-23	0.0000000000000	0.0000052608409	0.0000052608409
연구원 1-25	0.0000000000000	0.0000052348641	0.0000052348641

시나리오 1의 ClaRI는 다른 클래스들에 대해서도 트리플 링크 구조를 고려할 때 납득할만한 결과를 보여주었고 특정 자원의 중요 속성에 대한 링크 연결을 추가하거나 삭제했을 경우에도 예상되는 바와 같이 랭킹 점수가 증가하거나 감소하였다. 이에 대한 결과는 따로 제시하지 않는다.

4.4.2 시나리오 2의 결과 및 평가

Table 9는 Figure 11에서 ClaRI에 의해 객체속성만을 반영하여 구한 책 인스턴스들의 링크분석 결과를 정규화한 점수와 데이터타입속성인 '판매부수' 값을 정규화한 점수를 설정된 가중치대로 합산한 내용을 보여준다.

Table 10은 책 인스턴스들의 판매부수 속성값을 정규화하여 인스턴스별 링크가중치로 변환한 후 ClaRI에 의한 링크 분석에 처음부터 포함시켜 계산한 결과를 보여준다. Table 10의 랭킹 점수를 Table 9의 링크분석 점수와 비교할 때 최대값은 더 크고 최소값은 더 작음을 볼 수 있다.

Table 8 - 시나리오 1 : ClaRI – 특허 랭킹 결과

	objectivity	subjectivity	ranking score
특허 10-1	0.0014695855651	0.1185432831844	0.1200128687496
특허 10-2	0.0014695855651	0.1185432831844	0.1200128687496
특허 10-3	0.0014695855651	0.1185432831844	0.1200128687496
특허 10-4	0.0006897810252	0.0030453994039	0.0037351804291
특허 10-5	0.0004257485116	0.0001127339342	0.0005384824457
특허 10-6	0.0002640325136	0.0000407860781	0.0003048185917
특허 10-7	0.0001485768065	0.0000402531125	0.0001888299190
특허 10-8	0.0001450083726	0.0000125036376	0.0001575120102
특허 10-9	0.0001408984515	0.0000125036376	0.0001534020891
특허 10-10	0.0001349454941	0.0000123827349	0.0001473282290

이것은 판매부수 값의 차이가 반영된 결과로 보이며, 판매부수 값도 일련번호가 낮을수록 높게 설정되어 랭킹 순위에는 변동이 없는 것을 볼 수 있다. '책9-10'의 정규화 점수를 0으로 하지 않은 것은 행렬의 대각화 조건을 만족시키기 위해 거의 0과 비슷한 양수를 설정한 것이다. 0으로 했을 때에는 고유벡터 행렬이 풀(full) 랭크를 가지지 않는 경우가 있었는데 이런 경우에도 대각화 가능한 경우와 중요도 점수를 비교해 볼 때 거의 비슷하고 간혹 10^{-13} 이하에서 아주 미미한 차이를 보여주었다. 대각화 가능하지 않은 행렬에 대해서도 거듭제곱했을 때 제일 고유벡터의 실수배로 수렴하는 현상에 대한 연구가 필요한 것 같다.

V. 결론 및 향후 연구과제

시맨틱 웹의 데이터가 축적됨에 따라 자원들을 실질적인 중요도에 의해 랭킹하는 정렬 메커니즘의 구축이 필수적이다. 온톨로지 스키마 상에서 클래스 중심으로 속성의 가중치를 설정하는 방법은 기존 연구의 한계점인 강한 결합 모임 현상을 효과적으로 해결하는 것으로 보인다. 이 방법은 한 클래스에 속하는 자원들을 평가함에 있어 속성의 상대적인 비중을 고려하는 사람들의 평가방식과 유사하다. 컨텍스트에 따라 다양한 조합의 가중치를 설정할 수도 있으므로 여러 상황에 이용될 수 있다. 온톨로지 설계에 있어서도 중요하게 다루어야 할 속성들에 대한 기준을 제시해주며, 양의 가중치를 갖는 속성들에 대한 데이터는 누락하지 않도록 하는 관리지침을 세워주기도 한다. 앞으로 온톨로지 스키마에서 클래스와 속성의 위계구조가 있는 경우에 대한 확장 연구와 대각화 가능한 행렬을 거듭제곱했을 때 제일 고유벡터의 실수배로 수렴하는 성질에 대한 행렬 조건의 완화에 대한 증명 연구도 필요할 것이다.

Table 9 - 시나리오 2 : A - 책 클래스 랭킹 결과(판매부수 속성 가중치=0.5)

	링크분석 점수	링크 정규화 점수	판매부수	판매 정규화점수	최종 랭킹 점수
책 9-1	0.0211158322171767	1.000000000000000	8000	1.000000000000000	1.000000000000000
책 9-2	0.0209980612747133	0.994416209938332	7000	0.833333333333333	0.913874771635833
책 9-3	0.0209718587670499	0.993173889049821	6000	0.666666666666667	0.829920277858244
책 9-4	0.0021112883643479	0.098951058110468	5000	0.500000000000000	0.299475529055234
책 9-5	0.0014356776145794	0.066918805406992	4500	0.416666666666667	0.241792736036829
책 9-6	0.0006028526625789	0.027432667376506	4000	0.333333333333333	0.180383000354920
책 9-7	0.0005946751016884	0.027044950489182	3500	0.250000000000000	0.138522475244591
책 9-8	0.0000572924804120	0.001566409881177	3000	0.166666666666667	0.084116538273922
책 9-9	0.0000454220365117	0.001003604954160	2500	0.083333333333333	0.042168469143747
책 9-10	0.0000242544245480	0.000000000000000	2000	0.000000000000000	0.000000000000000

Table 10 - 시나리오 2 : B - 책 클래스 랭킹 결과(조정계수=10, 판매부수 속성 가중치=0.5)

	판매부수	정규화점수	인스턴스 링크가중치 (조정계수*정규화점수*속성가중치)	ranking score
책 9-1	8000	1.000000000000000	10.00000	0.0286011806011376
책 9-2	7000	0.833333333333333	8.33333	0.0254760658470699
책 9-3	6000	0.666666666666667	6.66667	0.0230398080065576
책 9-4	5000	0.500000000000000	5.00000	0.0027921381886903
책 9-5	4500	0.416666666666667	4.16667	0.0013590619006261
책 9-6	4000	0.333333333333333	3.33333	0.0004335665893297
책 9-7	3500	0.250000000000000	2.50000	0.0004093888546677
책 9-8	3000	0.166666666666667	1.66667	0.0000456496989496
책 9-9	2500	0.083333333333333	0.83333	0.0000298994011345
책 9-10	2000	E-16	5E-16	0.0000163708290858

References

[1] Aleman-Meza, B., Halaschek-Wiener, C., Arpinar, I.B., and Sheth, A. (2003). "Context-Aware Semantic association Ranking," *Semantic Web and Database Workshop Proceedings*, Belin, September 7-8.

[2] Aleman-Meza, B., Halaschek-Wiener, C., Arpinar, I.B., Ramakrishnan, C., and Sheth, A. (2005). "Ranking Complex Relationships on the Semantic Web," *IEEE Internet Computing,* Vol. 9, No. 3, pp. 37-44.

[3] Anyanwu, K., Maduko, A., and Sheth, A. (2005). "SemRank: Ranking Complex Relationship Search Results on the Semantic Web," *International World Wide Web Conference Committee(IW3C2),* Chiba, Japan.

[4] Bamba, B. and Mukherjea, S. (2004). "Utilizing Resource Importance for Ranking Semantic Web Query Results," *Proc. Second Toronto International Work. Semantic Web Databases (SWDB),* pp. 185-198.

[5] Berners-Lee, T., Hendler, J. and Lassila, O. (2001). "The Semantic Web : A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American.*

[6] Brickley, D. and Guha, R.V., eds. (2004). "RDF Vocabulary Description Language 1.0: RDF Schema," W3C Recommendation.

[7] Brin, S. and Page, L. (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Special Issu. 7th International World Wide Web Conf. Computer Networks and ISDN Systems,* Vol. 30, Nos. 1-7, pp. 107-117.

[8] Brin, S., Motwani, R., Page, L., and Winograd, T. (1998). "What can you do with a Web in your Pocket," *Bull. IEEE Computer Society Technical Comm. Data Engineering.*

[9] Burden, R.L. and Faires, J.D. (2001). *"Numerical Analysis,"* seventh edition, BROOKS/COLE.

[10] Gruber, T.R. (1993). "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition,* Vol. 5, No. 2, pp. 199-220.

[11] Halaschek, C., Aleman-Meza, B., Arpinar, I.B., and Sheth, A. (2004). "Discovering and Ranking Semantic Associations over a Large RDF Metabase," *Proceedings of the 30th VLDB Conference,* Toronto, Canada.

[12] Haveliwala, T.H. (1999). "Efficient Computation of PageRank," Unpublished manuscript, Stanford University.

[13] Kleinberg, J. (1999). "Authoritative sources in a hyperlinked environment," *Proc. 9th ACM-SIAM Symp. Discrete Algorithms,* pp. 668-677, 1998. Extended version in J. ACM, Vol. 46, No. 5, pp. 604-632.

[14] Klyne, G. and Carroll, J., eds. (2004). "Resource Description Framework (RDF): Concepts and Abstract Syntax," W3C Recommendation.

[15] Manola, F. and Miller, E., eds. (2004). "RDF Primer," W3C Recommendation.

[16] McGuinness, D.L. and Frank van Harmelen, eds. (2004). "OWL Web Ontology Language Overview," W3C Recommendation.

[17] Mukherjea, S. and Bamba, B. (2004). "BioPatentMiner: An Information Retrieval System for BioMedical Patents," *Proc. 30th Toronto Conf. Very Large Databases (VLDB),* pp. 1066-1077.

[18] Mukherjea, S., Bamba, B., and Kankar, P. (2005). "Information Retrieval and Knowledge Discovery Utilizing a BioMedical Patent Semantic Web," *IEEE Trans. Knowledge and Data Eng.,* Vol. 17, No. 8, pp. 1099-1110.

[19] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report, Stanford University.

[20] Prud'hommeaux, E. and Seaborne, A., eds. (2007). "SPARQL Query Language for RDF," W3C Working Draft.

[21] Schneider, P., Hayes, P., Horrocks, I., eds. (2004). "OWL Web Ontology Language Semantics and Abstract Syntax," W3C Recommendation.

[22] Sheth, A., Aleman-Meza, B., Arpinar, I.B., Halaschek, C., and Ramakrishnan, C. (2005). "Semantic Association Identification and Knowledge Discovery for National Security Applications," *Special Issu. Jour. Database Tech. Enhancing National Security,* Vol. 16, No. 1, pp. 33-53.

[부록]

Table 1 - 시나리오 2 : A, B – 클래스 중심 가중치 설정 예

1. 연구원		
발표하다	0.5	논문
등록하다	0.1	특허
저술하다	0.2	책
졸업하다	0.09	대학교
국적	0.01	나라
연봉	0.1	
이름	0	
생년월일	0	
전화	0	

2. 논문		
실리다	0.5	저널
참조되다	0.3	논문
(obj)발표하다	0.1	(dom)연구원
(obj)발표하다	0.1	(dom)교수
키워드	0.05	키워드
(obj)참조되다	0.05	(dom)논문
:		:

3. 키워드		
(obj)키워드	0.6	(dom)논문
분류되다	0.2	분야
(obj)키워드	0.2	(dom)특허
:		:

4. 분야		
(obj)분류되다	0.5	(dom)키워드
(obj)다루다	0.3	(dom)저널
(obj)주제	0.2	(dom)책
:		:

5. 저널		
(obj)실리다	0.3	(dom)논문
다루다	0.2	분야
저널평가점수	0.5	
저널명	0	
:		:

6. 대학교		
고용하다	0.2	교수
(obj)졸업하다	0.1	(dom)교수
(obj)졸업하다	0.1	(dom)연구원
위치하다	0.01	나라
학교평가점수	0.4	
재학생수	0.19	
:		:

7. 나라		
(obj)국적	0.1	(dom)연구원
(obj)국적	0.1	(dom)교수
(obj)속하다	0.1	(dom)특허
(obj)위치하다	0.1	(dom)대학교
1인당국민소득	0.4	
통신 인프라	0.2	
:		:

8. 출판사		
출판하다	0.5	책
당기순익	0.2	
매출액	0.2	
직원수	0.1	
주소	0	
전화번호	0	
:		:

9. 책		
(obj)출판하다	0.6	(dom)출판사
주제	0.2	분야
(obj)저술하다	0.1	(dom)연구원
(obj)저술하다	0.1	(dom)교수
판매부수	0.5	
:		:

10. 특허		
(obj)등록하다	0.1	(dom)연구원
(obj)등록하다	0.1	(dom)교수
속하다	0.1	나라
키워드	0.2	키워드
기술가치평가점수	0.6	
:		:

11. 교수		
발표하다	0.4	논문
등록하다	0.2	특허
저술하다	0.2	책
(obj)고용하다	0.1	(dom)대학교
졸업하다	0.009	대학교
국적	0.01	나라
전화번호	0	
:		: