구술문서에 기초한 자동 용어 네트워크 구축†

(Automatic term-network construction for Oral Documents)

박 순 철*
(Soon Cheol Park)

요 약 본 연구에서는 문서에 나타나는 용어의 통계값을 이용하여 구술문서자료에 포함되어있는 용어들간의 의미 네트워크를 자동으로 구축하는 시스템을 제안하였다. 본 연구를 위하여 전북 새만금지역에서 채록한 186개의 구술생애사 문서자료를 사용하였으며, 구축된 용어네트워크에서 용어들 사이의 관계는 용어들을 벡터화하여 결정하였다. 새만금 구술문서에서 중요단어로 선택된 단어의 수는 약 1700여 개이다. 단어들 사이의 용어네트워크는 구축 시스템을 통해서 실시간 내에 표현할 수 있었다. 이 용어네트워크는 앞으로 전개될 시멘틱 검색시스템 구축에 새로운 장을 열 것이며, 구술문서 분석에 크게 기여할 것으로 기대한다.

핵심주제어 : 새만금, 구술문서, 용어네트워크, 시멘틱 정보검색, 온톨로지

Abstract An automatic term-network construction system is proposed in this paper. This system uses the statistical values of the terms appeared in a document corpus. The 186 oral history documents collected from the Saemangeum area of Chollapuk-do, Korea, are used for the research. The term relationships presented in the term-network are decided by the cosine similarities of the term vectors. The number of the terms extracted from the documents is about 1700. The system is able to show the term relationships from the term-network as quickly as like a real-time system. The way of this term-network construction is expected as one of the methods to construct the ontology system and to support the semantic retrieval system in the near future.

Key Words : Saemangeum, Oral document, Term network, Semantic IR, Ontology

1. 서 론

최근 들어서 세계 각국은 과거의 문화를 보존하기 위하여 각종 방면으로 노력을 기울이고 있다. 특히 20세기를 증언할 수 있는 사람들의 구술자료를 수집하는데 많은 시간을 쏟고 있으며 다양한 분석 방법을 개발하고 있다[1, 2]. 국내에서도 이러한 경향에 동참해서 일반 대중들을 대상으로 구술자료를 수집하는 일에 힘을 기울이고 있다[3]. 그러나 수집된 구술자료를 과학적으로 분석하는 방법에 대해서는 이렇다 할 논의가 진전되지 못하고 있다. 이러한 시점에서 본 연구는 특정 문서집단(새만금 구술 코퍼스)에서 사용되는 용어들 사이의 관계를 자동적으로 구축하는 시스템을 제안하여 구술문서 분석에 새로운 과학적 방법을 제시하고자 한다.

현재 용어와 용어 사이의 관계(시소러스, 온톨

† 이 연구는 전북대학교 2005년도 연구기반조성기금으로 이루어짐.
* 전북대학교 전자정보공학부 교수

로지 등)를 찾아주는 연구가 여러 곳에서 활발히 진행되고 있다. 특히, 프린스턴대학의 인지과학연구실에서 개발한 WordNet나 Antony Lewis가 개발한 WordWeb 등은 용어들 사이의 관계를 다양한 방법으로 표현하고 있다[4, 5]. 그러나 이 시스템들은, 용어들 사이의 관계를 구하기 위하여, 단어들의 속성을 수동적으로 구축해야 하기 때문에 시스템 구축에 상당한 시간과 노력이 필요하다. 또한 이 시스템은 특정 문서에서 있을 수 있는 용어와 용어 사이의 특별한 관계보다는 일반화된 개념 사이의 관계만을 표현한다. 이 시스템들을 이용하는 한글 시스템의 연구도 진행 중에 있으나 아직은 초기단계에 머물고 있다[6, 7]. 이에 비해 본 연구에서 제안하고 있는 시스템은 특정 문서에서 나타난 용어의 통계정보를 이용함으로써, 사용되는 언어에 관계없이 이용이 가능하다. 또한, 용어 사이의 관계를 자동적으로 구축한다. 따라서 시스템 구축에 상당한 시간과 노력을 절감할 수 있다.

본 논문에서는 새만금지역 거주민의 구술자료로부터 채록한 180여 편의 문서를 토대로 약 1700여 용어를 사용하여 용어네트워크를 자동 구축하는 시스템을 구현하였다. 이 시스템은 특정 문서집단(새만금 구술 코퍼스)에서 사용되는 용어들 사이의 밀접한 관계를 쉽게 구축할 수 있다. 특히 복잡한 구조를 갖는 구술문서의 효과적인 분석에 있어서 본 연구에서 개발된 시스템은 효율적인 기능 및 성능을 보이고 있다. 뿐만 아니라 이 시스템은 앞으로 전개될 온톨로지 구축 및 시멘틱검색시스템[8-15]에도 응용할 수 있을 것으로 기대된다.

본 논문은 서론에 이어서 2장 구술문서의 구조와 특징, 3장 용어네트워크, 4장 시스템구현, 5장 결론과 향후 과제로 구성된다.

2. 구술문서 구조 및 특징

본 연구를 위해서 사용된 구술자료는 2002-2003, 2006-2007년 동안 전라북도의 새만금지역에서 수집한 186개의 구술문서자료[3]이다. 이 자료 가운데 이 지역에 거주하는 '어민 장홍배와 김종길의 구술자료'를 예로 선택하였다. 이 문서는 구술문서의 구조와 형식에 있어서 일반적인 특징을 잘 보여주고 있다.

A. 주제에 따른 용어의 다양성

구술문서의 특성은 구술자의 어휘력에 따라서 차이는 있지만, 일반적으로 하나의 주제를 설명하기 위해서 다양한 용어가 등장한다. 이러한 다양성은 주제에 따른 용어와 용어 사이의 관계를 찾는데 도움을 준다.

아래의 **예문 (1)**에서 '어촌계'와 관련된 어휘들은 크게 두 종류로 나눌 수 있다. 일반적인 의미를 가지는 것(어장, 허가)과 특수한 의미를 가지는 것(단속, 벌금)이 있다. 전자는 어촌계의 일반적인 상황을 설명하며, 후자는 전자와는 달리 새만금지역 어촌계의 특별한 상황을 설명해 주는 용어들이다.

> **예문 (1)** "과거에는 어촌계를 위해서 구획을 해줘 갖고 어촌계에서 이렇게 했는데 그걸 허가낸 사람이 조업을 못하게 해서 우리에게 허가해준다 해가지고 같이 어장을 했다. 근데 그 배들이 경비정들이나 아니면 기관 내 관내 배들한테 단속에 해당 되어가지고 벌금을 문 일이 있다."

B. 불필요한 반복적인 담화

구술문서에는 구술자의 언어습관으로 인해서 의미 없이 반복되는 언설이 많다. **예문(2)**에서는 '인제'가 반복되었고, '말하자믄' 등은 화자가 말을 끌어내기 위해서 사용되는 말들이기 때문에 내용과는 직접 관련은 없다. 이러한 불필요한 용어는 문서를 분석하는 사람에게는 전체 문장을 파악하는데 걸림돌이 된다. 그러나 본 논문에서는 이러한 용어를 제외하여 용어 사이의 관계를 단순화 시켰다.

> **예문 (2)** "인제 그 말하자믄, 새마을사업 부르짖어 가지고 인제 박정희대통령이 새마을사업...." "왜 인제 잘했다. 인제. 예 인제 삼개 마을인데." "말하자믄, 쉽게 말해서"

C. 유사한 내용의 반복

구술언어는 즉흥적이기 때문에 의미를 강조하기 위해서 내용이 반복되는 경우가 많다. 앞에서 말한 내용이 다시 뒤에서 거듭 나오게 된다. 이렇게 반

복되는 내용에 사용되는 용어는 중요어로 구분되어 용어네트워크에 구축된다.

이상과 같이 구술문서는 문자문서에 비해 다양한 내용과 복잡한 구조를 갖고 있다. 그러나 정보기술을 이용함으로써 이러한 구술문서의 복잡성을 단순화하여 손쉽게 용어네트워크를 구축할 수 있다.

3. 용어네트워크

2장에서 언급한 것처럼 구술문서는 문자문서에 비교하여 상당히 복잡한 구조를 갖고 있다. 이러한 문서 구조에서 단어들 사이의 관계를 인지적으로 분석하는 것은 많은 노력과 시간을 필요로 한다. 본 논문에서는 문서집단에서 나타나는 용어들의 통계값을 이용하여 용어들 간의 상호관계(용어네트워크)을 구하고자 한다.

3.1 용어의 가중치

용어의 가중치를 계산하는 방법은 일반적으로 문서집단에서 나타나는 용어의 통계값을 이용한다. 용어의 통계값은 한 문서에 포함되어있는 용어의 수, *Tf*(term frequency)와 용어가 사용된 문서의 수, *idf*(inverse document frequency)로 구분된다. 식(1)은 용어의 가중치를 구하는 계산식이다. *Tf*와 *idf* 값은 Okapi의 계산법을 따랐다[16, 17]. 식 (1)에서 가중치 W_{ij}는 j번째 문서의 i번째 용어가중치이다.

$$W_{ij} = Tf_{ij} \cdot idf_{ij} \qquad (1)$$

식(1)에서 TF_{ij}, idf_{ij} 값은 다음과 같다.

$$Tf_{ij} = \frac{tf_{ij}}{tf_{ij} + 0.5 + 1.5 \times \frac{doclen_j}{avgdoclen}}$$

$$idf_{ij} = \log\left(\frac{N - df_{ij} + 0.5}{df_{ij} + 0.5}\right)$$

3.2 문서벡터 및 문서·용어 행렬

문서 벡터는 일반적으로 각 문서에서 사용한 용어들의 가중치를 기준으로 한다. 다음은 i번째 문서의 벡터이다. 문서집단에서 사용된 용어의 수가 n 인 문서벡터의 크기는 $1 \times n$ 이며 식 (2)와 같이 표현할 수 있다.

$$d_i = (w_{i1} \quad w_{i2} \quad \cdots \quad w_{in}) \qquad (2)$$

모든 문서를 포함하고 있는 문서·용어 행렬 A는 문서벡터의 조합으로 이루어진다. 따라서 문서·용어 행렬은 식 (3)과 같으며 문서의 수가 m 이고 용어의 수가 n 인 행렬의 크기는 $m \times n$ 이다.

$$A = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & & & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \qquad (3)$$

식 (3)으로부터 용어 벡터를 구하기 위해서 가정 1)은 필수적이다. 즉 식 (2)로 표현되는 모든 문서들은 서로 독립적이며 서로 직교화되어 있다. 따라서 용어 벡터는 식 (4)와 같이 구하여진다.

가정 1) A 행렬로부터 문서와 문서는 서로 독립적이고 서로 직교화되어 있다.

가정 1)에 의해서 문서와 문서가 서로 독립적일 때, 용어 벡터 t_i는 다음과 같이 구하여질 수 있다. 즉, t_i는 A 행렬의 i번째 행의 값이다

$$t_i = (w_{1i} \quad w_{2i} \quad \cdots \quad w_{mi}) \qquad (4)$$

정의 1) 용어 t_i와 용어 t_j 사이의 관계값은 두 용어간의 코사인값으로 정의한다.

문서집단에 포함되어있는 i번째 용어와 j번째 용어사이의 값은 정의 1)과 같이 두 용어의 벡터값

을 이용한 코사인 값으로 구한다.

$$\cos(\vec{t_i}, \vec{t_j}) = \frac{t_i^T t_j}{\|t_i\| \|t_j\|} \quad (5)$$

식 (5)와 같이 두 용어 사이의 관계는 코사인값으로 구하여질 수 있다. 그러나 본 논문에서는 수행속도를 향상시키기 위해서 각 벡터를 정규화시킨 후 코사인값을 계산하도록 설계했다. 따라서 식 (5)는 다음과 같이 표현될 수 있다.

$$\cos(\vec{t_i}, \vec{t_j}) = t_i^T t_j \quad (6)$$

3.3 용어네트워크

본 논문에서 구하고자 하는 용어네트워크는 수집된 모든 문서에 포함되어있는 용어와 용어 사이의 관계를 구하는 것이다. n 개의 용어관계를 나타내는 행렬, O의 크기는 $n \times n$ 이며, 용어네트워크로 표현되는 행렬 O는 식 (3)과 식(6)에 의해서 다음과 같이 표현될 수 있다.

$$O = A^T A \quad (7)$$

본 논문에서는 행렬 O가 용어와 용어간의 의미를 나타내는 척도로 사용된다.

식 (6)으로부터 $\cos(\vec{t_i}, \vec{t_j}) = \cos(\vec{t_j}, \vec{t_i})$ 이다. 따라서 $t_i^T t_j = t_j^T t_i$ 이다. 그러므로 행렬 O의 i와 j 번째로 표현되는 $t_i^T t_j$ 와 j와 i번째로 표현되는 $t_j^T t_i$ 값이 같다. 따라서 행렬 O는 대칭적이다. 실제 구현에서는 행렬의 이러한 특징을 이용하여 upper triangle 값이나 lower triangle값만 이용하여 메모리 공간을 절약할 수 있다.

n개의 용어로 이루어진 용어네트워크로부터 용어를 찾기 위해서 검색용어벡터를 구하면 식 (8)과 같이 표현된다

$$Q_i = (q_1\ q_2 \cdots q_n), \quad k = 1, 2, \cdots, n$$
$$\begin{cases} q_k = 1 & when \quad i = k, \\ q_k = 0 & others. \end{cases} \quad (8)$$

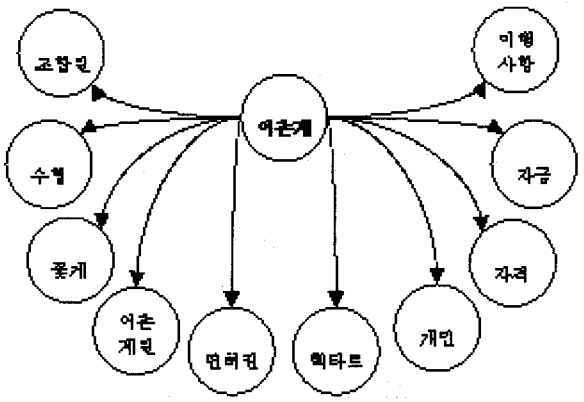
식 (8)의 검색용어벡터는 용어네트워크에서 i번째 존재하는 용어 t_i의 검색용어벡터이다.

식 (8)을 이용하여 i번째 용어 t_i와 관련된 용어들과 관련정도의 값(관계값)을 구하는 관계식은 식 (9)와 같다.

$$Q_i \cdot O \quad or \quad Q_i \cdot A^T A \quad (9)$$

3.4 적용

본 연구에서 제안한 시스템 구축을 위해서 사용한 새만금의 구술자료는 186개 문서이며, 문서에 포함된 용어 중 용어네트워크에 사용된 단어의 수는 1761개이다. 이들 용어를 이용하여 3.3절의 용어네트워크의 내용을 나타내는 행렬 O를 구하였다. 행렬 O에 포함되어있는 용어 중 식 (9)를 이용하여 '어촌계'와 관련 용어를 추출하였다. (그림 1) '어촌계'와 관련된 단어네트워크 결과 중 상위 10개 단어를 선택한 것이다. (그림 1)에는 포함되어있지 않지만 조합원과 어촌계의 관계값은 0.61, 수협 0.54, 꽃게 0.50, 어촌계원 0.47, 면허권 0.46, 헥타르 0.43, 개인 0.43, 자격 0.43, 자금 0.42, 이행사항 0.32 등이다. 용어와 용어사이의 관계값은 높을수록 더 밀접한 관계이며 0과 1사이의 값을 갖는다.

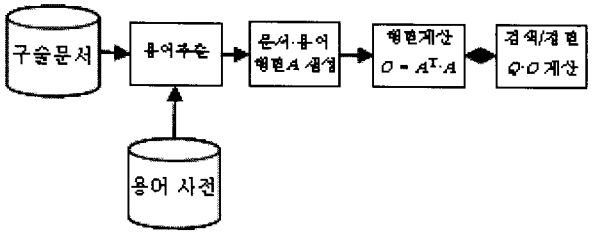


(그림 1) '어촌계'와 관련된 용어네트워크

(그림 1)에서 보이는 것처럼 '어촌계'로 선택된 단어들은 일반적인 개념의 단어 사이의 관계를 포함한 문서집단에서의 특별한 관계를 나타낸다. 예를 들면, '어촌계'와 '조합원', '수협', '어촌계원', '자금', '자격' 등은 일반적인 관계성이 있지만, 일반적인 관계성이 희박한 '이행사항', '헥타르', '꽃게', '면허권' 등은 이들 문서에서만 나타날 수 있는 '어촌계' 용어와 특수한 관계성을 나타낸다.

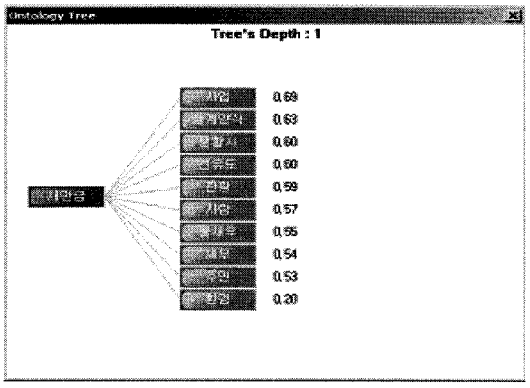
4. 구현 시스템

시스템을 구현하기 위해서 사용된 문서는 186개로 모두 대화체로 구성되어있다. 또한 문서분석과 용어추출을 위해서 사용된 명사사전을 2972의 명사로 이루어졌다. 이 시스템에서 추출된 용어의 수는 1761개이다. 또한 시스템의 총수행시간은 약 13.1초였다. 그러나 시스템이 구축된 후 용어 사이의 관계를 찾기 위해서는 단지 수 *msec*에 불과하다. (그림 2)는 구현된 시스템의 개략구조도이다.



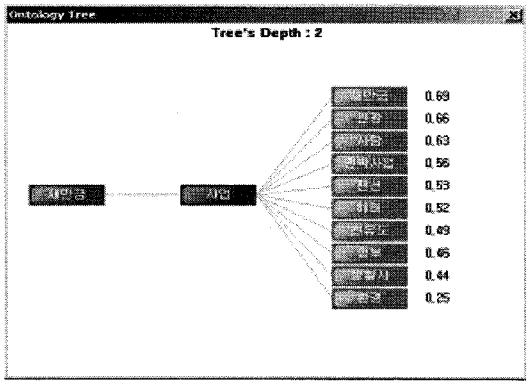
(그림 2) 시스템 구조

(그림 3)은 용어 '새만금'에 대한 용어네트워크의 실제적인 구조를 보인다. 이것은 앞으로 연구될 새만금 주제 온톨로지의 프로토타입 형식의 일부이다. (그림 3)에서 보이는 것처럼 새만금과 관련을 가진 용어가 다음과 같이 검색되었다. 사업/생계양식/생활사/선유도/사람/꽃새우/주민/환경이다. 앞으로 더 많은 실험을 거쳐 보다 정확한 의미 용어네트워크가 구축될 것이라고 생각되지만, 현재로서는 매우 만족할 만한 성과를 이루었다. 또한, (그림 3)은 새만금에 대한 용어사이의 연관관계(용어네트워크)와 함께 통계적자료를 제시함으로써 그 연관성과 정확도을 보여준다. 선택된 용어 중 가장 위에 리스트된 '사업'이라는 용어가 새만금과 가장 밀접한 관계를 갖는다.



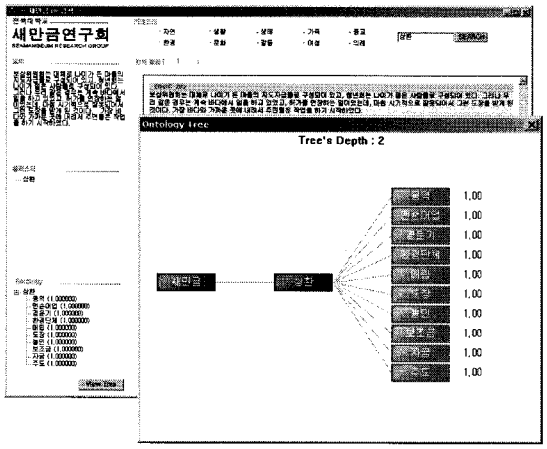
(그림 3) '새만금' 용어네트워크의 트리구조

(그림 4)는 용어 '새만금'에서 '사업'으로 확장된 형태의 용어네트워크의 구조이다. (그림 4)로부터 새만금 사업과 관련된 것은 '관광'사업, '민박사업', '펜션' 등이라는 것은 쉽게 알 수 있다. 이 결과는 주민들의 구술자료로부터 얻어진 것이기 때문에 새만금 사업에 대한 주민들의 관심분야를 예측할 수 있다.



(그림 4) '새만금'과 '사업'의 확장 용어네트워크

(그림 5)은 실제로 구현된 시스템에서 관련용어를 추출하는 프로그램의 사용자인터페이스(GUI)이다. (그림 5)의 바탕창에 나타난 것처럼 구축된 용어네트워크는 정보검색시스템과 같이 동작이 된다. 따라서 이 시스템은 사용자에게 용어네트워크 이외에 다양한 기능을 제공한다.



(그림 5) 정보검색과 용어네트워크 GUI

5. 결론 및 향후 과제

본 연구는 새만금지역 거주민들의 구술을 채록한 문서내용에 포함되어있는 용어들을 중심으로 용어들간의 의미 네트워크를 구축하는 것이다. 본 연구에서 사용한 구술 문서의 수는 186개, 네트워크에 사용된 용어의 수는 1761개이다. 시스템구축을 위한 총수행시간은 약13.1초이다. 그러나 시스템 구축 후 용어와 용어사이의 관계를 표현하는 것은 수 *msec*에 불과하다. 이러한 시스템의 기능과 성능은 다양한 분야에서 사용되고 있는 문서집단의 분석과 용어들 사이의 관계를 분명하고 신속하게 구할 수 있는 가능성을 제공한다.

본 연구에서는 용어들 사이의 관계를 구하기 위하여 문서집단(새만금 구술문서집단)내의 용어 통계자료를 이용하여 벡터화하였다. 이것은 용어네트워크를 수동적으로 구축하던 기존의 방법에서 벗어나 자동화하는 것을 가능하게 한다. 또한 새만금 구술자료와 같이 복잡한 구조를 가지고 있는 다양한 수집문서에 직접 적용할 수 있다. 아울러 용어의 통계값은 언어에 독립적으로 구하여 질 수 있기 때문에 다양한 언어 혹은 다양한 주제를 가진 복잡한 문서집단에 모두 이용될 수 있다. 또한 자동적으로 구축되는 시스템은 우리의 손이 닿기 힘든 미세한 부분 및 많은 양의 자료를 동시에 처리할 수 있다.

본 연구를 통하여 향후 다음과 같은 분야에 더욱 집중적인 연구를 기대할 수 있다. 첫째, 특정 문서집단으로부터, 문서집단이 갖는 주제에 따라 단어들간의 개념을 자동적으로 정립하는 주제 온톨로지시스템으로 발전시킬 수 있을 것이다. 둘째, 구축된 용어네트워크의 관계를 이용한 구술문서 분석으로 지금까지 밝혀지지 않았던 새로운 사실과 지식을 발견하는 문서분석시스템 구축이 가능할 것이다. 셋째로 용어네트워크와 정보검색을 융합시켜 통합적인 시멘틱 정보검색시스템을 구현할 수 있을 것이다. 아울러 추가적이고 집중적인 연구를 통해서 새로운 지식창출 시스템에 많은 공헌을 할 수 있게 될 것으로 기대한다.

참 고 문 헌

[1] http://www.qualitative-research.net/fqs-texte/3-00/3-00orsatti-e.htm

[2] http://www.ccrh.org/comm/slough/oral/oralhis.htm

[3] http://smg21.org/

[4] http://wordnet.princeton.edu/

[5] http://www.wordwebonline.com/

[6] 문유진, 한국어 명사를 위한 WordNet의 설계와 구현, 정보과학회논문지(C) 제2권 제4호, pp. 437 445, 1996. 12

[7] http://dewey.yonsei.ac.kr/memexlee/doc/LeeKim1999.htm

[8] A. Gomez Perez and O. Corcho, "Ontology Language for the Semantic Web", IEEE Intelligent Systems, vol.17, no.1, pp.54 60, 2002.

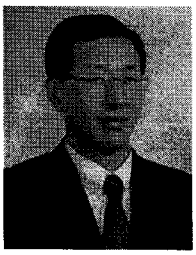
[9] 최호섭, 임지희, 배영준, 최수일, 옥철영, "온토로지 구축 방법과 사례", 정보과학회지, vol.24, no.4, pp.31 44, 2006.

[10] Deborah L. McGuinness and Frank van Harmelen, "OWL Web Ontology Language Overview", 2004, http://www.w3.org/TR/owl features/.

[11] 최중민, "시멘틱 웹의 개요와 연구동향", 정보과학회지, vol.21, no.3, pp.4 10, 2003.

[12] Tim Berners Lee, James Hendler and Ora

Lassila, "The Semantic Web", Scientific American, 2001.
[13] M R Koivunen and E. Miller, "W3C Semantic Web activity", Proceedings of the Semantic Web. Kick off Seminar in Finland, 2001.
[14] 오현목, "시멘틱 웹 발전 방향 및 표준화 개발전략 연구", 한국전산원, 2005
[15] D. Fensel and F. van Harmelen and I. Horrocks and D. McGuinness and P. Patel Schneider, "An Ontology Infrastructure for the Semantic Web". 2001.
[16] Gerard Salton and Christopher Buckley, Term weighting approaches in automatic text retrieval. Information Processing & Management, vol.24(5), pp.513 523. 1988
[17] Rong Jin, Christos Falusos, and Alex G. Hauptmann. "Meta scoring: automatically evaluating term weighting schemes in IR without precision recall," In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001.



박 순 철 (Soon Cheol Park)

• 종신회원
• 1979년 2월 : 인하대학교 공과대학 (공학사)
• 1991년 12월 : (미국)루이지아나 주립대학 (전산학박사)
• 1991년-1993년 : 한국전자통신 연구원
• 1993년-현재 : 전북대학교 전자정보공학부 교수
• 관심분야 : 정보검색, 데이터마이닝