

다중 온톨로지를 이용한 시맨틱 웹 포털에서의 의미형 검색

박지형^a, 박상언^b, 이명진^a, 홍준석^b, 김우주^a

^a연세대학교 정보산업공학과
서울 서대문구 신촌동 134
Tel: +82-2-2123-7754, E-mail: {zensents, xml, wkim}@yonsei.ac.kr

^b경기대학교 경영정보학과
경기도 수원시 연통구 이의동 산 94-6
Tel: +82-31-249-9454, E-mail: {supark, Junehong}@kyonggi.ac.kr

Abstract

최근 시맨틱 웹 포털의 등장으로 현재 웹 포털의 많은 문제점들이 해결되고 있다. 하지만, 웹 상의 자원들이 점차 증가하면서 시맨틱 웹 포털에서의 효율적인 탐색에 대한 문제점이 발생하고 있다. 시맨틱 웹 포털은 온톨로지의 사용으로, 기존의 웹 포털과 구조적으로 다르기 때문에 사용자에게 적절한 정보를 제공할 수 있는 새로운 탐색 방법이 필요하다. 이 논문에서는 세 가지 단계로 구성된 의미적 검색 절차를 제안한다. 또한, 전자 분야에서의 다중 온톨로지를 이용한 시맨틱 웹 포털을 구축하여, 본 논문에서 제안하는 의미형 검색 절차 방법을 통해 향상된 검색 성능 결과를 제시한다.

Keywords:

시맨틱 웹 포털, 의미형 검색

1. 서론

차세대 웹으로 주목 받고 있는 시맨틱 웹은 현재 인터넷을 이용하는 다양한 분야에서 응용되고 있으며, 그 분야에 새로운 기회를 제공하고 있고, 다양한 관련 연구를 이끌고 있다. 그 중 시맨틱 웹 포털은 시맨틱 웹 기술들이 통합되어 있는 응용분야로서 주목 받고 있다. 일반적으로 웹 포털은 웹 사용자를 위한 시작점이며, 유용한 정보와 서비스와 연결된 하이퍼링크(hyperlink)를 제공해준다. 게다가, 최근에는 커뮤니케이션 또는 정보 교환을 위한 수단으로 사용되기도 한다. 하지만, 이러한 현재의 웹 포털은 웹의 발달로 인한 정보의 폭발적인 증가로, 부정확하고 이질적인 의미의 정보들이 증가한 상태이다. 또한 현재의 웹 기술의 한계로 인해 의사소통과 정보 교환에 대한 문제가 발생하기도 한다.

시맨틱 웹 포털은 시맨틱 웹 기술을 이용하여 구축된, 보다 진보된 형태의 웹 포털이다. 특정 도메인에 대한 정보와 서비스를 제공하는 웹 어플리케이션으로, 데이터와 정보 자원에 의미를 부여하여 기계가 이해하고 처리할 수 있는 형태의 지식 체계를 기반으로 하고 있다. 하지만, 이러한 시맨틱 웹 포털에서는 또 다른 문제가 존재한다. 웹 자원(resource)의 증가에 따른 포털에서의 정보 탐색의 문제가 바로 그것이다. 시맨틱 웹 포털은 기본적으로 의미적인 정보 처리를 가능하게 하는 지식 체계인 온톨로지를 기반으로 하고 있다. 기존의 포털과는 다른 이런 시맨틱 웹 포털의 특징 때문에, 기존과는 다른 정보 탐색 방법이 필요하다.

본 연구는 다양한 자료들이 함께 있는 상황에서 사용자에게 최대한의 정보를 줄 수 있는 적절한 정보 탐색 방법을 제안하고자 한다. 2장에서는 관련 연구를 소개하며, 3장과 4장에서는 이 논문에서 제안하는 탐색 방법에 대한 자세한 소개가 이어진다. 제안된 탐색 방법을 통한 실험 결과는 5장에서 보여지며, 마지막 6장에서는 결론과 함께 앞으로의 연구 방향에 대해 얘기한다.

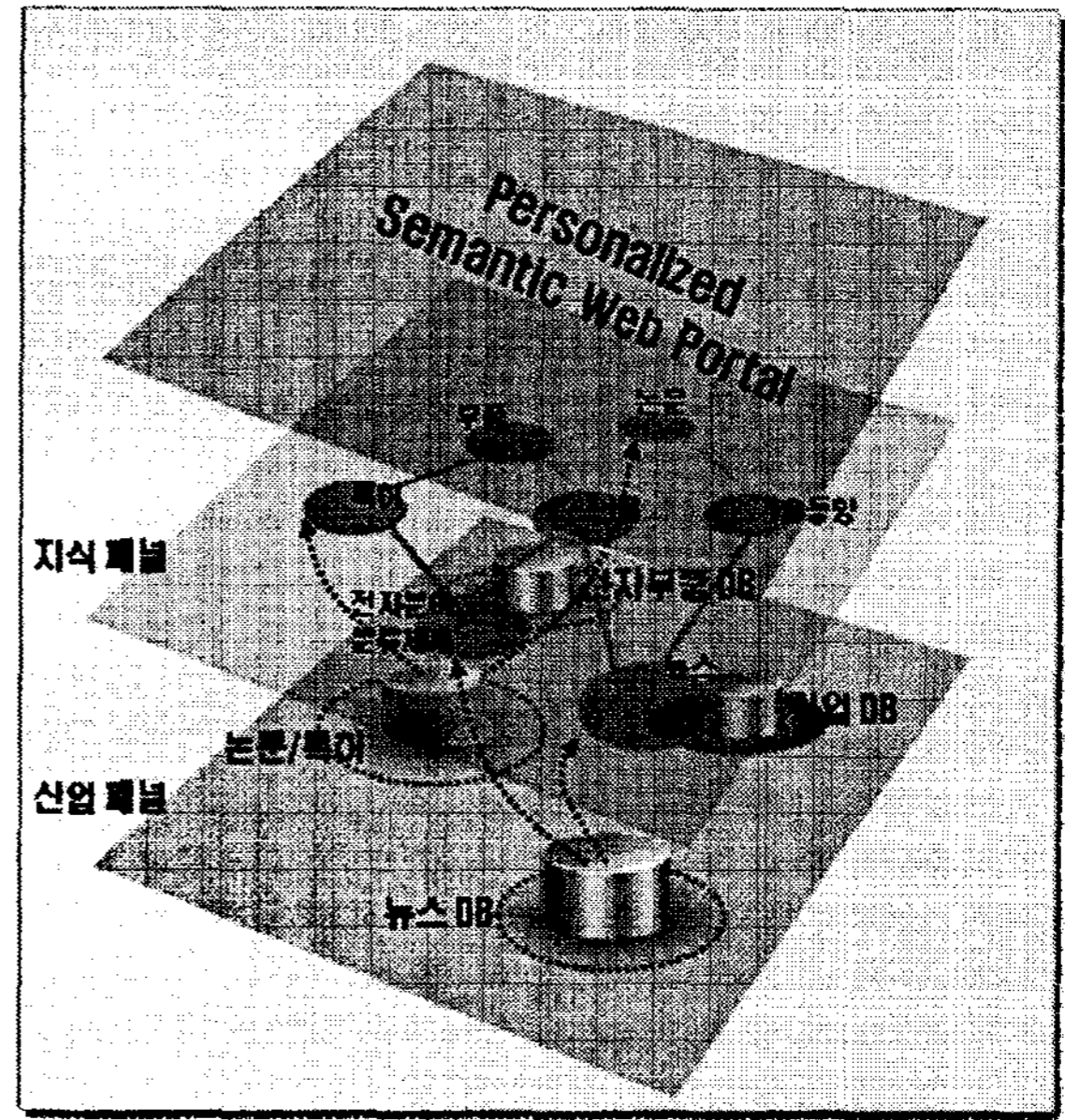
2. 관련 연구

본 논문의 주된 논점은 기존 키워드 검색과는 구별되는 사용자에게 최적의 정보를 제공할 수 있는 시맨틱 검색 절차를 제안하는 것과 그것을 이용한 특정 분야의 시맨틱 포털을 구현하는 것이다. 이러한 관점에서 기존의 잘 구현된 시맨틱 웹 포털과 프로토타입 시스템들을 소개한다. SEAL(SEmantuc web portAL)[9]은 일반적인 시맨틱 포털 구현에 대한 내용을 제안한다. 포털에서의 의미적인 정보 제공과 접근 방법, 또한 정보 유지 방법 등에 대한 내용을 소개하고 있다. 다음으로,

OntoWeb Portal[12]은 시맨틱 웹 관련 학계와 업계의 기관들끼리 정보를 공유할 수 있는 커뮤니티 포털이라 할 수 있다. ODESeW[3]는 각 분야의 온톨로지들로부터의 지식을 자동 생성하는 것에 초점을 두고 있는 시스템이다. 이 시스템은 지식 포털을 위한 다양한 기능들을 제공하고 있다. 마지막으로, OntoWebber[6]는 지식 포털을 구현하고 유지하기 위한 온톨로지 기반의 도구이다. 지식 포털을 모델링하고 지식 포털의 구현과 유지에 관한 다양한 정보를 제공한다.

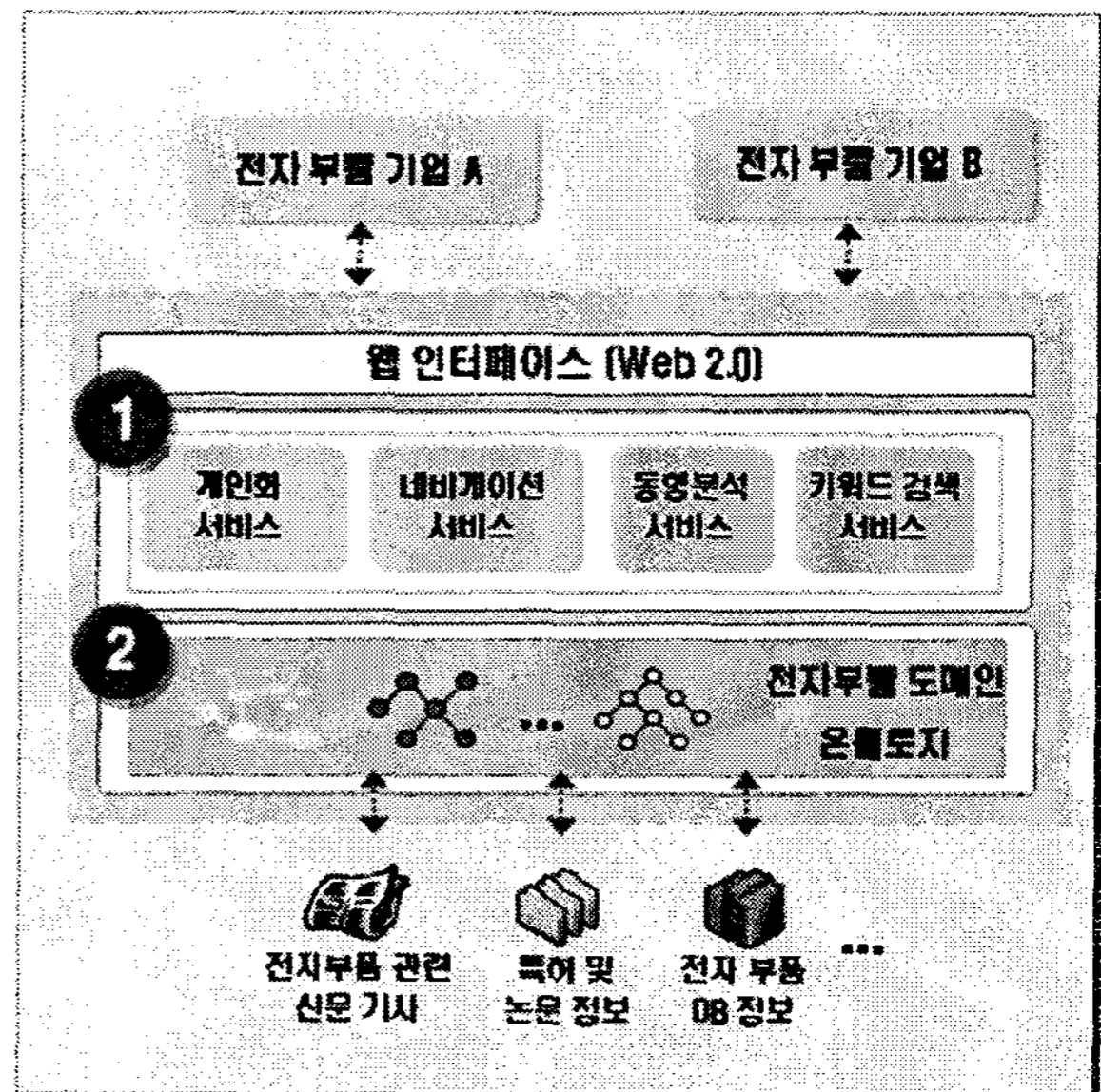
3. 시맨틱 포털 시스템 구조

본 연구의 목적은 전자 분야의 시맨틱 포털을 디자인하고, 기존의 키워드 검색과는 다른, 시맨틱 검색의 절차를 제안한다. 본 논문의 이해를 돕기 위해 앞으로 사용하게 될 시맨틱 포털의 특징에 대해 간략하게 설명하도록 하겠다. 이 시스템의 일반적인 사용자들은 전자 부품을 생산하는 회사의 사람들 또는 전자 부품 산업과 관련된 일을 하는 사람들이다. 이들은 새로운 기술, 동향, 전자 부품 생산에 민감하기 때문에, 항상 전자 관련 신문을 읽으며, 논문, 기술, 특허 관련 정보를 얻기 위해 하루에 최소 5~7개의 사이트를 방문한다. 우리 프레임워크의 주된 목적은 여러 사이트를 방문하며 각각의 정보를 검색해야 하는 사용자들의 번거로움을 없앨 수 있도록 하나의 시맨틱 포털을 통해 각 출처의 정보를 제공해 주는 것이다. 출처의 정보들은 의미적으로 연결된 형태로 제공된다. 예를 들어, 특허 관련 정보가 검색될 경우 이 특허와 관련된 새로운 기술 또는 동향 등의 정보 또한 검색 결과에 포함된다. 또한 이 특허를 소유한 기업에 대한 결과도 마찬가지이다. <그림 1>은 본 연구에서 제안하는 시맨틱 포털 모델을 나타내고 있다. 산업패널에서 정보 출처가 분리되어 있는 것을 볼 수 있다. 보통 각 출처 별로 정보에 대한 형식과 접근 방법이 다르기 때문에 서로 다른 기관들간에 매끄럽고 효율적인 정보 교환과 공유가 불가능하다. 지식 패널은 이러한 문제를 해결하기 위해 정보의 의미적인 통합이 가능하도록 구성되어 있다. 사용자에게 각기 다른 출처로부터의 정보를 검색할 수 있도록 하기 위해, 모든 정보는 지식 패널에서 의미적으로 연관되어 있으며, 각 정보들은 링크(link)를 통해 연결되어 있다. 시맨틱 웹 포털은 지식 패널 상의 이러한 다양한 서비스를 사용자에게 제공해주는 인터페이스이다.



< 그림 1 > 시맨틱 포털 모델

<그림 2>는 시맨틱 포털 시스템의 구조도를 나타낸다. 우리는 먼저 뉴스 DB, 특허 서비스 사이트, 논문 서비스 사이트 등의 서로 다른 정보 출처들로부터 전자 부품 온톨로지를 구성하였으며, 이 정보들은 위에서 언급한대로 의미적인 연결되어 있다. 그 다음으로 우리의 시맨틱 포털의 가장 핵심적인 서비스인 정보 탐색 서비스를 포함한 다양한 서비스를 구현하였다.



< 그림 2 > 시맨틱 포털 시스템 구조도

4. 시맨틱 검색 처리 절차

본 논문에서 제안하는 시맨틱 검색의 처리 과정은 시드(seed) 검색, 시맨틱 탐색(semantic navigation), 순위화의 3가지 단계로 이루어진다. 시맨틱 검색 서비스는 기본적인 검색 기능 외에 두 가지의 추가적인 탐색 기능을 제공한다.

4.1. 온톨로지 시드(Seed) 검색

첫 번째 단계에서는 사용자가 입력한 키워드를 포함하는 시드 후보들을 온톨로지에서 찾는다. 예를 들어, 사용자가 'BLU' 라는 단어(전자 부품의 일종으로 Back Light Unit을 뜻함)를 시맨틱 검색의 키워드로 입력했을 경우, 정확한 사용자의 검색 의도를 모르기 때문에 일단 모든 후보 단어들을 온톨로지로부터 찾게 된다. 후보 단어들은 'BLU' 라는 단어를 포함하는 모든 클래스와 인스턴스를 포함하는 집합이다.

4.2. 시맨틱 탐색 처리

일단 온톨로지로부터 시드들을 찾게 되면, 다음은 시드들과 관련된 개념들을 찾아나가는 단계이다. 일반적인 시맨틱 검색이 기존의 키워드 검색과 가장 차이가 나는 점은 바로 시맨틱 네트워크상의 연결된 모든 클래스들과 인스턴스들을 찾는다라는 것이다. 그래프 구조를 갖는 온톨로지 상의 모든 개념들은 서로 연결이 되어 있다. 그러므로, 이 단계에서는 그래프를 통한 탐색이 이루어진다. 하지만, 이 시맨틱 검색 알고리즘에서는 몇 가지 고려해야 할 사항들이 있다. 첫째, 확장하고자 하는 연결들만 선택해야 한다. 만약 시드와 연결된 온톨로지상의 모든 연결들을 전부 선택하게 되면 그 관계는 너무나 많고 쓸모 없는 검색 결과가 될 것이다. 이 과정에서 우리는 모든 관계에 대한 한계치(threshold)를 계산해야 한다. 연결에 대한 한계치는 연결된 노드(Node)들의 개수에 대한 평균으로 계산된다. 예를 들어, '회사'라는 개념과 '제품'이라는 개념 사이에 '생산하다' 라는 관계가 존재한다고 하자. 전체 온톨로지에서의 인스턴스 수를 비교하였을 때, '삼성전자'라는 회사는 10,000개의 제품을 생산하고, "삼진"이라는 중소기업 회사는 10개의 제품을 생산한다고 한다면, '회사'와 '제품' 사이의 '생산하다' 라는 관계는 5,005라는 한계치를 갖게 된다. 이렇게 계산된 한계치를 탐색 과정에 적용하면, '삼성전자'가 생산한 제품과의 연결은 그 수가 한계치보다 크기 때문에 결과로써 확장되지 않고, 반대로 '삼진'이 생산한 제품과의 연결은 결과로써 확장되게 된다. 또한 여기에서 확장되는 연결의 경우 최대 3개의 제한을 둔다. 그러므로, 검색 결과는 시드들과

연결된 3개의 연결만을 포함하게 되는 것이다. 이러한 확장 개수의 제한은 검색 결과에 많은 영향을 미친다. 이와 관련된 사항은 추후 계속 연구할 계획이다.

4.3. 순위화

위의 두 과정을 거친 후의 결과들은 시드들과 다른 개념들간의 경로를 갖게 된다. 우리는 시드들로부터 연결된 노드들을 확장했기 때문에, 검색된 결과는 여전히 크기가 큰 상태이다. 그러므로, 정확하고 중요한 결과 순으로 정렬하는 과정이 필요하다. 여기에서 사용된 우리의 순위화 알고리즘의 가장 중요한 요소는 시드와 결과 경로의 최종 노드 사이의 연결 개수 즉, 경로의 길이를 말한다. 만약 그 수가 0이라면, 그것은 시드 자체가 그 결과를 의미하는 것이며, 가장 높은 우선순위를 갖게 된다. 그 다음 우선순위는 시드와 직접 연결된 노드들이고, 마지막은 최종 노드와 키워드 사이의 유사도(word similarity)이다. 같은 연결 수를 갖는 노드들은 이 유사도에 의해 정렬된다.

검색 결과는 '회사', '뉴스', '인물', '기사', '제품', '특허', '기술' 등의 주요 7개 그룹으로 구분된다. 따라서, 사용자는 자신의 관심에 따라 쉽게 결과를 확인할 수 있다. 검색 결과에서 주목해야 할 점은 모든 노드가 또 다른 시맨틱 검색의 시작점이 될 수 있는 하이퍼링크를 제공한다는 것이다. 이 하이퍼링크를 통한 검색은 시드를 찾는 과정이 필요 없기 때문에 키워드로부터 검색이 시작되는 것보다 훨씬 간단하게 검색을 진행할 수 있다.

4.4. 추가 탐색 기능

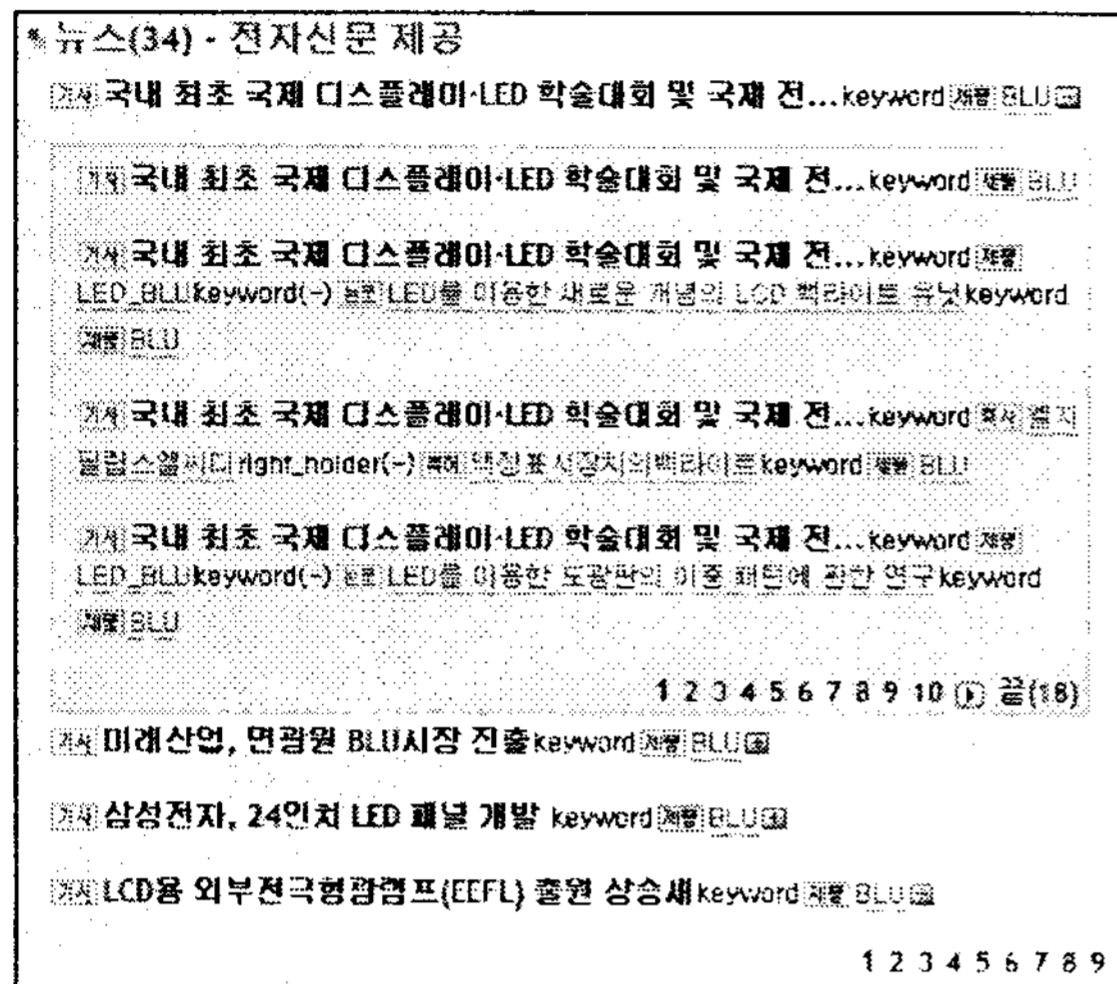
본 절에서는 결과 탐색 과정의 두 가지 추가적인 기능을 소개한다. 첫째는, 'AND' 연산자이다. 만약 사용자가 두 단어와 관계된 개념들을 찾고자 할 경우, 두 개의 키워드 사이에 'AND' 연산자를 사용하여 검색이 가능하다. 'AND' 연산자를 사용한 검색은 그래프 상에서 두 키워드와 연결된 노드들이 결과로 나타나게 된다. 이것은 두 개의 독립적인 검색 결과의 교집합과 같다고 볼 수 있다.

두 번째 기능은 '중심주제어' 검색이다. 만약 사용자가 하나의 주제와 관련된 추가 결과를 얻고 싶다면, 그 주제를 중심주제어로 정하여 검색을 진행할 수 있다. 일단, 사용자가 하나의 키워드를 중심주제어로 등록하면 알고리즘은 검색 결과와 중심주제어 사이의 관계를 보여주는 방식으로 진행된다.

5. 시스템 구현 및 성능 평가

본 논문에서 제안하는 시맨틱 검색이 가능한 프로토타입 시스템을 제한된 온톨로지를 이용하여

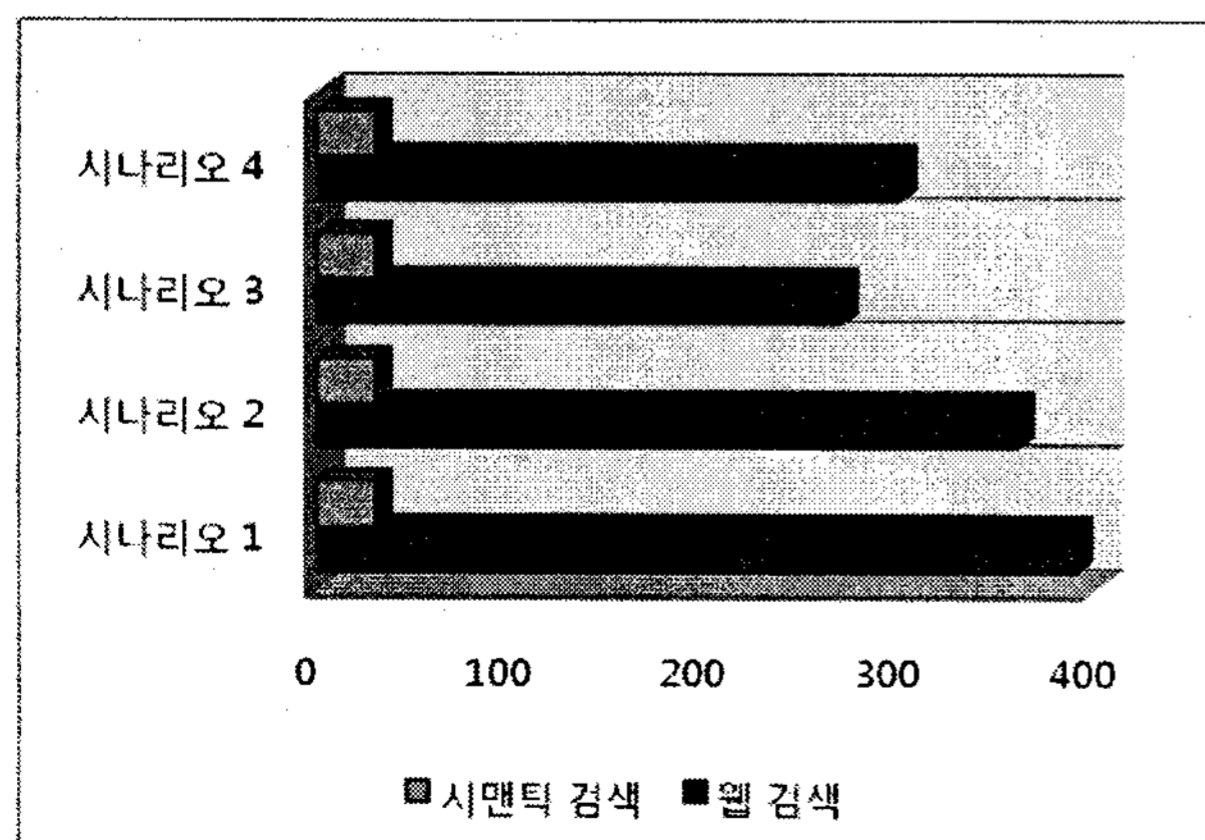
구현하였다. <그림 3>은 개발된 시스템의 검색 결과 중 '뉴스' 그룹 부분을 나타낸다. 시드를 제외한 노드들에는 하이퍼링크가 되어 있고, 시드와 최종 노드사이에 관계된 모든 노드들이 연결되어 있는 것을 볼 수 있다. 노드 앞의 작은 박스에는 그 노드가 속해있는 클래스가 표시되어 있다.



< 그림 3 > 시맨틱 검색 결과

우리는 기존의 웹 검색과 시맨틱 검색의 결과를 비교하기 위한 실험을 진행하였다. 4가지의 시나리오를 통해 실험을 진행하였으며, 진행 시나리오는 아래와 같다.

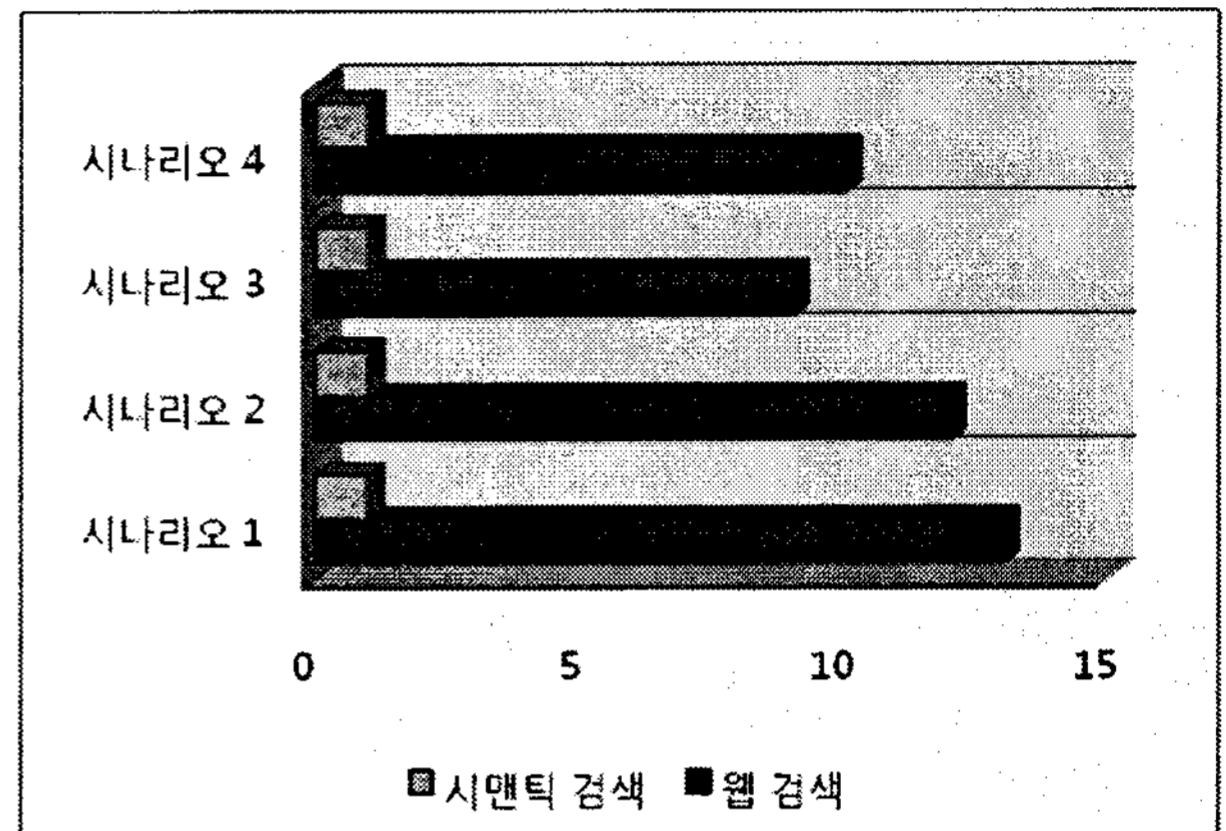
- 시나리오 1- 'BLU'관련 기술 동향 검색
- 시나리오 2- 협력사 현황 파악
- 시나리오 3- 'LED BLU' 현황 파악
- 시나리오 4- 복합기술을 통한 기회



< 그림 4 > 정보 수집 시간 비교

<그림 4>는 기존의 검색과 시스템을 이용한

시맨틱 검색과의 정보 수집 시간 비교를 보여주고 있다. 이 실험은 평균 방문 시간을 계산하기 위해 웹 페이지 접속 통계(<http://inside.daum.net/>)를 통해 이루어졌다. 평균적으로 기존의 웹 검색이 시맨틱 웹 검색보다 11배 정도의 시간을 더 소요하였다. 즉, 우리는 여기에서 10배 이상의 시간을 절약한 것이다.



< 그림 5 > 방문 페이지 수

<그림 5>는 기존의 검색과 시스템을 이용한 시맨틱 검색과의 방문 페이지 수를 비교한 것이다. 결과는 위의 정보 수집 시간 비교와 거의 같다. 기존 웹 검색에서 10페이지 이상을 방문하는 동안 단 한번의 방문으로 동일한 검색 결과를 나타내는 훌륭한 성과를 얻을 수 있었다.

6. 결론

시맨틱 포털은 시맨틱 웹 기술의 중요한 응용 분야이며, 시맨틱 경로 기반의 관련 개념을 찾아내는 시맨틱 검색에 대한 연구는 시맨틱 포털의 폭넓은 활용을 위한 필수 연구 분야이다.

본 논문에서는 전자 부품 분야에서의 시맨틱 검색 절차를 제안하였다. 검색 절차는 시드 검색, 시맨틱 탐색, 순위화의 3가지 단계로 구성되었으며, 여기에 'AND' 연산자 기능과 '중심주제어' 검색 기능을 추가로 제안하였다. 또한 제안된 시맨틱 검색을 제공하는 프로토타입 시스템을 구현하고, 시스템 평가를 위한 실험을 통해 분산되어 있는 다양한 웹 페이지에 대한 기존의 키워드 검색보다 약 10배 이상의 성능 향상을 보일 수 있었다.

현재 이 연구를 확장하여 그래프 상의 두 노드간에 의미적 관계를 측정할 수 있는 측정법을 고안 중이다. 본 연구가 시맨틱 검색과 시맨틱 포털의 실용적인 결과로서 남길 바란다.

7. References

- [1] Berners Lee, T., Hendler, J., and Lassila, O. "The Semantic Web", *Scientific American*, May, 2001..
- [2] Berners Lee, T., and M. Fischetti. "Weaving the web : original design and ultimate destiny of the World Wide Web, by its inventor", Harper, San Francisco, 1999.
- [3] Corcho, Ó., Gómez-Pérez, A., López-Cima, A., López-García, V., and María del Carmen Suárez-Figueroa, "ODESeW: Automatic Generation of Knowledge Portals for Intranets and Extranets", *Proceedings of the 2nd International Semantic Web Conference*, Oct. 20-23, Florida, USA, 2003, pp.802-817.
- [4] Decker, S., Erdmann, M., Fensel, D., and Studer, R., "Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information," *Proceedings of IFIP TC2/WG2.6 8th Working Conference on Database Semantics(DS-8)*, Jan. 4-8, Rotorua, New Zealand, 1999, pp.351-369.
- [5] Goble, C.A., Bechhofer, S., Carr, L., Roure, D.D., and Hall, W., "Conceptual Open Hypermedia = The Semantic Web?," *Proceedings of the 2nd International Workshop on the Semantic Web*, May 1, Hong Kong, China, 2001
- [6] Jin, Y., Decker, S., and Wiederhold, G., "OntoWebber: Model-Driven Ontology-Based Web Site Management," *Proceedings of the 1st Semantic Web Working Symposium*, Jul. 30–Aug. 1, California, USA, 2001, pp.529-547.
- [7] Lausen, H., Stollberg, M., Hernández, R.L., Ding, Y., Han, S.K., and Fensel, D., "Semantic Web Portals–State of the Art Survey," *Technical Report (2004-04-03)of Digital Enterprise Research Institute(DERI)*, 2004.
- [8] Lin, C., Zhang, L., Zhou, J., Yang, Y., and Yu, Y., "SPortS: Semantic + Portal + Service," *Proceedings of the ECAI-2004 Workshop on Application of Semantic Web Technologies to Web Communities*, Aug. 23-27, Valencia, Spain, 2004.
- [9] Maedche, A., Staab, S., Studer, R., Sure, Y., and Volz, R., "SEAL-Tying Up Information Integration and Web Site Management by Ontologies," *IEEE Data Engineering Bulletin* 25(1):10-17, 2002/
- [10] Mäkelä, E., Hyvönen, E., Saarela S., and Viljanen, K., "OntoViews - A Tool for Creating Semantic Web Portals," *Proceedings of the 3rd International Semantic Web Conference*, Nov.7-11, Hiroshima, Japan, 2004, pp.797-811.
- [11] Oscar Corcho, Angel López-Cima, Asunción Gómez-Pérez. "A platform for the development of semantic web portals", *ISWE'06*, July, 11-14, 2006, Palo Alto, California, USA.
- [12] Peter Spyns, Daniel Oberle, Raphael Volz, Jijuan Zheng, Mustafa Jarrar, York Sure, Rudi Studer, Robert Meersman. "OntoWeb – A Semantic Web Community Portal", *Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management*, 189-200, 2002.
- [13] Staab, S., Angele, J., Decker S., Erdmann, M., Hotho, A., Maedche, A., Schnurr, H.P., Studer, R., and Sure, Y., "Semantic community Web portals," *Computer Networks* 33(1-6):473-491, 2000.
- [14] Yufei Li, Yuan Wang, Xiaotao Huang, "A Relation-Based Search Engine in Semantic Web", *IEEE Transactions on Knowledge and Data Engineering*, Volume 19 , Issue 2, 2007.