개념적 거리와 밀도를 이용한 웹 문서 검색

황희철*, 최 창**, 김판구***
*조선대학교 교육대학원(정보·컴퓨터교육전공)
**조선대학교
***조선대학교 컴퓨터공학과
allicehwang@hanmail.net, enduranceaura@chosun.ac.kr
pkkim@chosun.ac.kr

Web Document Retrieval based on Conceptual Distance and Density

Heechul Hwang*, Chang Choi**, Pankoo Kim***
*Dept. of Information · Computer, Education Graduate School, Chosun University
**Dept of Computer Science, Graduate School, Chosun University
***Dept of Computer Engineering, Chosun University

요 약

최근 인터넷 기술의 비약적인 발전으로 웹상에 많은 양의 정보가 존재하고, 많은 사람들이 이를 검색하고 활용하게 되었다. 그러나 기존의 검색방식은 단순히 텍스트 매칭(Text matching) 방법을 사용하고 있어 많은 자료들 사이에서 자신이 원하는 자료를 찾는데 어려움이 있다. 이에 본 논문에서는 검색할 자료의 정보를 바탕으로 그와 유사한 자료를 검색해주는 웹 문서 검색 시스템을 제안하고자 한다. 이를 위해 울산대학교 어휘 지능망인 U-WIN을 기반으로 개념적 밀도와 단어 간의 유사성 측정을 이용하여 의미적인 검색이 되도록 하였다.

1. 서론

인터넷 사용 증가와 관련 기술의 발달로 무수히 많은 정보들이 웹상에 분포하게 되었다. 이런 정보들을 효율적이고 적합하게 검색하는 방법에 대한 필요성이 대두되면서 각 분야의 관련 연구가 진행되고 있고 관련된 시스템이나 검색 엔진들이 개발되어지고 있다.

그러나 현재의 검색기법은 텍스트 매칭(Text-matching) 기반으로 동음이의어(homonym) 문제, 텍스트 자원이 구조화되지 않은 문제 외에 공통개념이 공유되지 않아 분산자료에서 정보통합의 어려움 등을 가지고 있어 정확한 검색결과에 한계가 있다.[1]

이런 문제들로 인해 차세대 지능형 웹인 시맨틱 웹(Semantic Web)이 등장하게 되었다. 시맨틱 웹 응용의 가장 중심적 개념으로 지식을 서술하는 온톨로지(Ontology)는 웹 기반의 지식 처리나 응용 프로그램 사이의 지식 공유, 재사용 등이 가능한 단어와 관계들로 구성된 일종의 사전으로서 생각할 수 있다.[1][2] 이러한 어휘 사전의 역할 이외에 지식을 효과적으로 표현하기 위해 정보에 의미를 부여하고 정보간의 관계를 설정하며 광범위한 도메인에 적용이 가능하도록 표준을 제시한다. 최근에는 의료나 전자, 기계 등의 특정 도메인에 관련된 단어들을 계층적으로 표현한 Domain Ontology와 WordNet등을 이용해 단어 간의 관계를 정의한 Linguistic Ontology 분야에 대한 연구가 활발히 진행되고 있다.

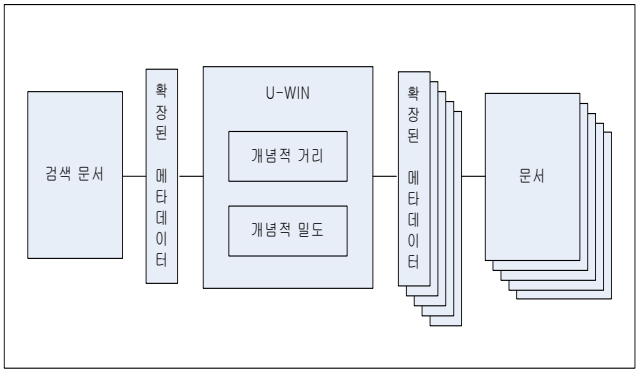
본 논문에서는 Linguistic Ontology의 일종인 울산대학교 어휘 지능망(UOU-Word Intelligent Network : U-WIN)을 기반으로 개념간 의미 유사도 측정방안을 통해 의미적으로 유사한 문서를 검색하였다. 특히 검증을 위하여 분류가 명확한 교육 분야의 웹 문서를 이용해 실험하였다.

2. 관련 연구

지능적인 정보검색을 위해 많이 사용되어지고 있는 기초 기술들은 다음과 같다. 텍스트를 이루고 있는 단어들을 추출하여 통계적인 데이터를 산출하는 문서 색인(indexing)방법과 텍스트를 통계적으로 분석하는 내용 기반 여과(content-based filtering) 방법, 유사한 취향의 다른 사용자가 선호하는 아이템을 추천하는 방식인 협동에 의한 정보 추천(collaborative recommendation), 사용자를 대신해 작업을 수행하는 소프트웨어 에이전트(software agent), 그리고 사용자로부터 피드백을 얻는 방식과 사용자 프로파일을 변경하는 알고리즘에 따라 구분할 수 있는 사용자 기호에 대한 학습 방법 등이 있다.[3] 하지만 이러한 방법들은 여전히 사용자가 원하는 검색 결과를 충족 시켜주지 못하고 있어 의미적 유사도 측정법으로 해결하고자 한다.

3. 전체 시스템의 구성

제안하는 웹 문서 검색 시스템의 전체 구성은 [그림 1]과 같이 사용자 웹 문서의 메타데이터(MetaData) 정보와 대상이 되는 문서의 메타데이터 정보 간의 유사도 측정을 통해 검색한다.



[그림 1] 제안한 웹 문서 검색 시스템의 개념도

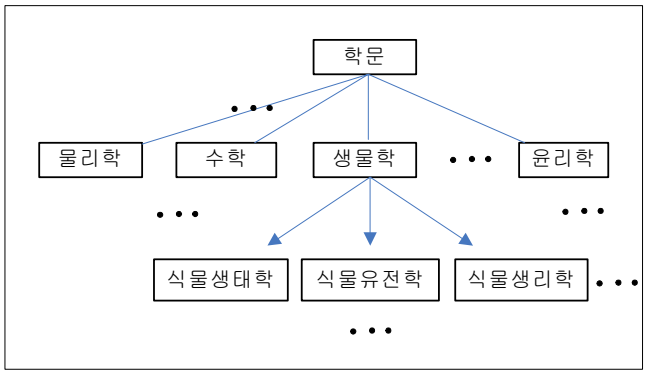
3.1 교육정보 메타데이터

교육정보 메타데이터의 구현방법으로는 LOM 방식과 Dublin Core 방식, ERM 방식, KEM 방식이 있다. LOM 방식은 SCORM (Sharable Contents Object Reference Model)에서 채택해 전 세계적으로 빠르게 확산되어 가고 있는 방식이며, Dublin Core 방식은 교육 분야에서 특정한 한정어를 강화하고 자원의 신속한 검색이 가능하여 가장 범용적인 방식이다. ERM 방식은 e-learning 분야를 포함한 교육 분야, 전자책, 디지털 도서관, 지식 관리 등에서의 교육 자원에 대한 기술 레코드를 생성하는 것이 직접적인 목적으로 하고 있고, KEM (Korea Educational Meta-data)방식은 학습용 컨텐츠를 포함한 교육 정보에 대한 일관성 있는 설명을 도울 수 있는 일반적인 명명법을 제공하는 것을 목적으로 한 방식이다.

교육정보의 일반적인 메타데이터는 자료의 표제(Title), 내용을 기술한 개인 또는 단체명(Creator), 내용에 대한 개요(Description), 발행처 정보(Publisher), 자료내용 유형(Type), 자료처리에 필요한 정보와 크기(Format), 자료접근 정보(Identifier) 등을 이용한다. 특히 본 논문에서는 웹 문서의 제목과 목차의 명사만을 이용하여 메타데이터 정보를 구성한다.

3.2 U-WIN

U-WIN은 울산대학교 한국어처리연구실에서 WordNet을 기반으로 구축한 우리말 어휘 데이터베이스로, 한국어의 명사를 가지고 제작하여 현재 18만개의 어휘로 구성되어 있다.[4] WordNet과 같이 의미를 기준으로 계층적인 구조로 이루어져 있다. [그림 1]은 U-WIN에서 “학문”에 대한 계층적 구조를 보여주고 있다.



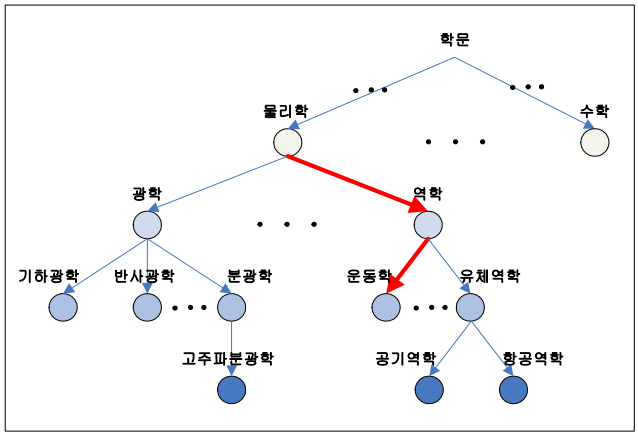
[그림 2] U-WIN에서 “학문”에 대한 구성도

3.3 유사도 측정을 위한 개념적 접근

유사한 문서의 검색을 위해서는 확장된 메타데이터인 단어간 유사도 측정이 필요하다. 단어들 간의 유사도 측정을 위해 고려될 수 있는 특징으로는 깊이(Depth), 정보량(Information Content), 밀도(Density), 거리(Distance)가 있다.[5] 본 논문에서는 단어의 깊이 정보가 상이함에 착안해 단어들 간의 거리와 단어의 중의성 문제를 해결하기 위한 개념적 밀도를 고려하여 유사도를 측정한다. 다음 4장과 5장에서는 각각의 방법에 대해 자세히 살펴보도록 한다.

4. 개념간 유사도 측정

구성된 메타데이터 정보, 즉 단어들의 개념적 거리를 이용해 웹 문서간 유사도를 측정한다. [그림 3]은 "물리학"과 "운동학"의 개념간 구조 및 거리를 예로 나타내고 있다.



[그림 3] "물리학"과 "운동학" 사이의 구조

두 개념간 거리를 이용한 유사도 측정 방법은 아래와 같다.

$$Distance = min[len(C_1, C_2)] \quad \text{----------} \quad (1)$$

$$Sim(C_1, C_2) = \frac{1}{Distance} \quad \text{----------} \quad (2)$$

여기서 $len(C_1, C_2)$는 비교 대상이 되는 C_1과 C_2의 개념간 거리로 최단거리(min 값)를 구한다. 유사도는 거리에 반비례하므로 수식 (2)와 같이 구할 수 있다.

[그림 4]는 위 측정 방법을 이용하여 "운동학"과 "물리학" 간의 최단거리 및 유사도를 측정한 결과이다.

```
HyperTree:*Root*#n#1 인지#n#7 지식#n#2 학문#n#1
과학#n#1 자연과학#n#1 물리학#n#1 역학#n#2 운동학#n#1

HyperTree:*Root*#n#1 인지#n#7 지식#n#2 학문#n#1
과학#n#1 자연과학#n#1 물리학#n#1

최단거리 : 3,          유사도 측정값 = 0.3333333333
```

[그림 4] 거리 및 유사도 측정

위와 같은 방법을 이용하여 <표 1>과 같이 교육과 관련된 웹 문서들의 유사도를 측정해 보았다. A문서는 유사도 측정시 기준이 되는 사용자의 문서로 5개의 메타데이터 정보를 가지고 있고 B문서는 검색 대상이 되는 문서중 하나로 7개의 메타데이터 정보로 구성되어 있다.

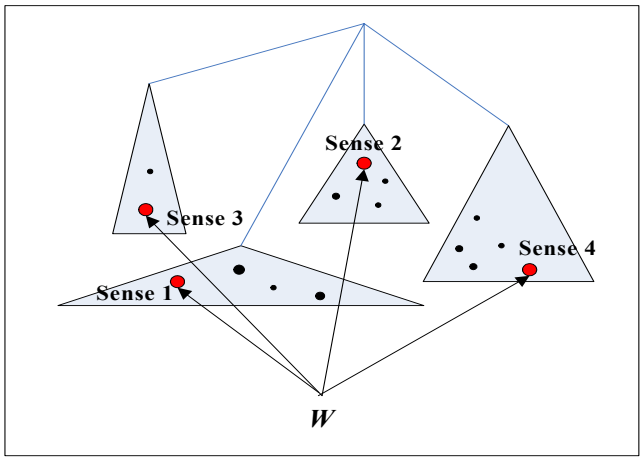
<표 1> 웹 문서간 유사도 측정 결과

B문서 \ A문서		문명	수학	종류	교과	언어
		#n#2	#n#3	#n#2	#n#1	#n#1
정수론	#n#1	0.10000	0.50000	0.14285	0.10000	0.12500
소개	#n#1	0.10000	0.10000	0.14285	0.08333	0.12500
수치	#n#3	0.11111	0.11111	0.16666	0.09090	0.14285
연산	#n#2					
개념	#n#1	0.12500	0.25000	0.20000	0.12500	0.16666
군	#n#16	0.12500	0.12500	0.20000	0.10000	0.16666
범위	#n#2	0.11111	0.11111	0.16666	0.09090	0.14285
평균,계	**0.72460**	0.67222	1.19722	1.01902	0.59013	0.86902

<표 1>에서 평균값 0.72460은 A문서와 B문서 간의 유사도를 측정한 결과이다. 이와 같은 방식으로 n개의 웹 문서들과 비교하여 유사성이 높은 문서 순서로 검색하게 된다.

5. 개념적 밀도 측정

개념적 밀도란 한 단어가 여러 가지 의미를 가지고 있는 경우, 이를 해결하기 위한 것이다. 즉 중의성을 가지고 있는 단어의 각 의미별 영역을 설정한 후 각 영역의 개념적 밀도를 구하기 위해 영역내 노드의 수를 세어 구할 수 있다. 이때 영역내 노드 수가 많을수록 높은 밀도값을 가지고 노드 수가 적을수록 상대적으로 낮은 밀도값을 가진다.[6]



[그림 5] 개념적 밀도

[그림 5]는 개념적 밀도를 통해 단어의 모호성을 해결하는 방법을 보여주고 있다. 예를 들어 단어 W

는 4가지의 의미를 가지고 있고 각각은 U-WIN의 subhierarchy에 포함되며, 이는 각 의미(Sense)별 영역이 된다. 영역내의 각 점은 중의성이 있는 단어 *W*와 *W*가 존재하는 문맥상의 다른 단어들 간의 유사도를 측정하여 그 결과에 따라 4가지 영역중 하나의 영역에 포함된다. 각 영역내 존재하는 점인 노드의 수를 세는 방법을 이용하여 개념적 밀도를 구한다. 개념적 밀도가 클수록, 즉 영역이 클수록 각 문맥에서 정확한 의미를 표현할 가능성이 크다. [그림 5]의 경우 Sense 4의 밀도가 가장 크므로 찾고자 하는 문맥에 가장 알맞은 의미라고 볼 수 있다.

<표 2>는 중의적 의미를 가진 단어를 이용 예로 "수학"이라는 단어는 U-WIN에서 "물의 현상에 관하여 연구하는 학문#n#1"과 "수량 및 공간의 성질에 관하여 연구하는 학문#n#3"으로 중의성이 있어 유사도 측정시 개념적 밀도를 고려해 유사도를 측정한 결과이다.

<표 2> 개념적 밀도에 의한 중의성 해소

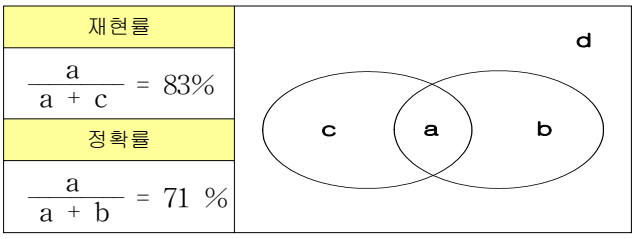
단어		통계학	부호	검정	예제	순위	부여	방식	비교	계
수학	#n#1	0.2500	0.1250	0.1250	0.0769	0.1111	0.1250	0.1428	0.0909	1.0467
	#n#3	**0.5000**	0.1250	0.1250	0.0769	0.1111	0.1250	0.1428	0.0909	**1.2967**

위 결과로 이 웹 문서가 3번째 의미의 "수학"이라는 것을 알 수 있으므로 중의성 문제는 해결 될 수 있다.

6. 실험 및 평가

실험을 위해 본 논문에서는 웹 문서 중 파워포인트 문서를 5개 과목(생물학, 수학, 물리학, 경제학, 윤리학)으로 사전에 분류하였다. 또 각 문서에서의 제목과 목차를 구문 분석하여 명사만을 추출한 후 메타데이터 정보를 구축하였고 U-WIN 1.0 버전을 이용해 실험하였다.

성능 평가의 척도를 위해 재현률과 정확률의 두 가지 방법을 이용하였다. 재현률과 정확률 값을 구하기 위해 각 요소 a, b, c, d의 관계를 [그림 6]과 같이 표현 하였다. a+b+c+d는 전체 데이터베이스이고, a+c는 검색과 관련된 총 웹 문서, a+b는 검색에 의해 검색된 웹 문서로 해석할 수 있다.



[그림 6] 실험 결과 및 요소별 관계

실험 결과에서 측정 결과를 감소시킨 요인으로, 정제된 메타데이터 정보를 사용하지 못했기 때문으로 판단된다. 실험시 사용한 웹 문서는 규격화가 되어있지 않아 메타데이터로 구성할 단어의 수가 적거나 외국어인 경우가 많았고 중요 단어에 비해 비중이 낮은 단어들이 많이 존재해 측정 결과를 떨어뜨렸다.

7. 결론 및 향후연구

본 논문에서는 Linguistic Ontology의 일종인 U-WIN을 기반으로 개념적 거리와 개념적 밀도를 이용하여 유사한 웹 문서 검색을 제안하였다. 위 결과를 토대로 개념 기반 문서 검색 및 문서 분류에 응용할 수 있을 것이다. 또한 좀 더 높은 정확률을 위해서는 메타데이터 관련 연구 및 유사도 측정에 관한 연구가 필요할 것이다.

참고문헌

[1] 정은경 외 3명, "온톨로지 기반의 정보검색" "한국정보과학회 2003년 추계학술대회" VOL. 30, NO. 2-1
[2] 정민선 외 3명, "다양한 도메인의 효율적 표현을 위한 온톨로지 기반의 다계층 컨텍스트 모델링", "한국정보과학회 2005 추계학술대회" VOL.32, NO. 02
[3] 한동일 외 2명, "시맨틱 서비스 에이전트 개발에 관한 연구", "한국정보과학회 2005 추계학술대회" VOL.32, NO. 02
[4] 김준수, 옥철영, "정제된 의미정보와 시소러스를 이용한 동형이의어 분별 시스템", "한국정보처리학회 논문지" VOL. 12, NO. 07, 2005
[5] 최준호 외 2명, "컬러 분포와 WordNet상의 유사도 측정을 이용한 의미적 이미지 검색", "한국정보처리학회 논문지 B" VOL. 11, NO. 04, 2004
[6] 조미영, 김판구, "정보량과 개념적 밀도를 이용한 단어 의미 중의성 해결", "한국정보처리학회" VOL. 12, NO. 02, 2005