지식기반 웹 문서 필터링

o
황상규 김상모 변영태
홍익대학교 전자계산학과

Knowledge-Based Web Document Filtering

Hwang, Sang-Kyu Kim, Sang-Mo Byun, Young-Tae
Dept. of Computer-Science, Hong-Ik Univ.

요 약

인터넷에서 검색 가능한 정보의 양은 폭발적으로 증가하고 있으며, 그에 따라 웹 기반 정보검색시스템은 사용자가 원하는 정보만을 필터링하여 이용자의 정보검색 수행과정에 부담을 덜어줄 필요가 있다. 본 연구에서는 웹 정보검색에 익숙치 못한 초보 이용자들이 실제 웹 정보검색을 수행하는데 있어 발생할 수 있는 문제점을 살펴보고, 초보 이용자들의 보다 편리한 웹 정보검색을 도와줄 수 있도록 하기 위하여 WordNet을 활용한 지식베이스와 SDCC(Semantic Distance for Common Category)를 이용한 효율적인 웹 문서 필터링 알고리즘을 개발하고 그 효율성을 확인하였다.

1. 서론

인터넷은 상호 정보 교환 및 최신 정보를 획득하기 위한 수단으로 이용되고 있으며 인터넷 사용자수의 증가와 함께 그 정보의 양 역시 폭발적으로 증가해왔다. 검색 가능한 정보의 양이 증가할수록 사용자가 원하는 정보를 찾고, 관리하는 시간도 그에 비례하여 증가한다. 시시각각으로 생성, 소멸하는 다량의 웹 정보들 중에서 사용자 스스로 기호에 맞는 정보를 추출 해내는 것은 매우 어렵고 상당한 시간을 소모하는 일이 되어 왔다. 이러한 이용자의 부담을 덜기 위한 정보검색 기법들 가운데 필터링 기법은 이용자가 관심 있어할 만한 정보만을 검색하도록 하는 도구이다.

인터넷상에서 필터링에 관한 연구는 먼저 전자메일이나 뉴스그룹을 대상으로 한 연구가 주로 진행되어져 왔으며, 94년 이후 웹 문서를 대상으로 한 정보검색이 널리 이용됨에 따라 웹 문서 정보검색을 위한 필터링에 관한 연구가 시작되었다. 필터링을 위한 방법 역시 여러 가지 방법들이 사용되어져왔는데, 대표적인 방법으로는 Salton의 TF/IDF방법을 이용한 필터링 기법[Marko94]이 대표적이다. TF/IDF방법은 검색된 문서의 연관성의 정도를 검색된 문서상에서 나타나는 검색어의 빈도수(Term Frequency)와 전체 문서상에서 검색어가 나타나는 비율(Document Frequency)을 통해 계산하게 된다. 위 방법은 필터링 뿐만 아니라, 전통적으로 문서를 대상으로 한 정보검색에 보편적으로 널리 쓰이는 방법이다. 하지만 위 방법은 웹 기반 정보검색 환경 하에서는 근본적인 결함을 가지고 있다. 먼저 WWW상의 존재하는 웹 문서의 수가 기존의 서지 DB에 비해 많아 전체 문서의 개수를 파악하기 어려우며, 또한 인터넷상에 존재하는 웹 문서들은 한 가지 주제를 대상으로 한 문서들의 집합체가 아닌 다양한 주제들을 포함하고 있는 문서들의 집합체이다. 이는 결과적으로 웹 문서상에서 'apple'이란 단어가 과일로서의 'apple'뿐 아니라 'apple computer'를 지칭하는 상호명이나 상품명 혹은 단체명 등 다양한 의미로 사용되어지는 어휘 의미 중의성 문제[황상규99a]를 유발하게 된다. 실제 'apple'이란 검색어를 통해 검색된 문서들을 문서상에서 나타나는 검색어 'apple'의 빈도수로 순위화 하였을 때 대부분의 상위 링크를 차지하는 문서의 대부분이 'apple computer'의 홈페이지와 같은 부적합한 문서들로 채워지는 것을 살펴볼 수 있다. 이러한 문제점을 해결할 수 있는 방법 중에 하나로, 이용자로부터 주어진 질의에 나타나는 단어만으로 검색을 시도하는 것보다는 추론 과정을 통해 질의를 좀 더 구체적으로 확장하여 문서 검색의 정확성을 높일 수 있다[황상규99b].

본 연구는 과학 재단(과제 번호 973010301)의 지원을 받았음

2. 웹 이용자 성향분석

인터넷을 이용하는 대부분의 사용자들은 정보검색에 대한 전문 지식을 갖추고 있지 못한 경우가 대부분이다. 따라서 검색식을 작성하는데 있어 사전에 미리 치밀한 전략을 세우기보다는 필요에 따라 단순히 키워드를 입력하는 형태를 보인다. 이는 초보자의 정보검색 행동 경향분석[박창호98]에서도 잘 나타나는데, 실험 대상자중 과반수에 해당하는 49.4%가 입력한 검색어의 수는 1개이었으며, 검색어 수를 4개 이상 입력한 경우는 거의 찾아볼 수 없었다. 보다 광범위한 조사로 검색엔지 Excite의 log file에 기록된 50만개의 query를 대상으로 사용자 성향을 분석한 결과[Cutting97]에서도, 사용자의 평균 입력 검색어의 수는 2.3개에 불과하였다. 또한 사용자의 95%가 검색엔진이 구절(phrase)검색기능을 지원함에도 불구하고 구절검색기능을 사용하지 않았다고 보고하고 있다. 이는 결과적으로 대부분의 사용자들은 자신이 찾고자하는 대상과 관련된 키워드 한 개를 가지고 웹 정보검색을 시도한다고 생각할 수 있다.

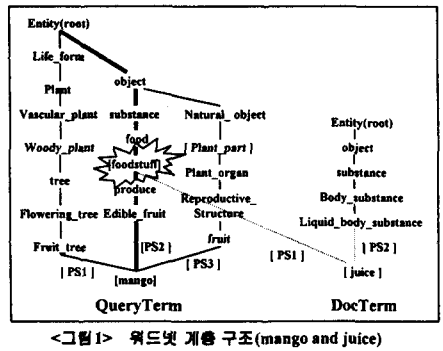
위와 같은 내용을 전제로 동물에 대한 전문 지식베이스를 갖춘 지능형 정보 에이전트 HIIA-1a[이용현99]와 기존 검색엔진들과의 성능 비교데이터들을 보다 광범위한 관점에서 분석하는 작업을 통해 웹 정보검색 이용자들의 성향을 분석해 보았다. 평가 인원은 홍익대 대학원생 19명을 대상으로 실시하였으며, 모든 응답자들은 기존에 웹 정보검색에 대한 교육을 받은 적이 없으며, 웹 정보검색 수행 능력은 대부분 중급이하의 일반 이용자들로 구성하였다. 평가 참여인원은 동물에 관한 질의에 대해서 각각 4개의 웹 정보검색 시스템으로부터 최대 100개까지의 문서를 평가하였다.

실험은 어휘의미중의성의 가능성이 적은 전문단어와 어휘 의미 중의성이 빈번한 일반단어의 경우 각각을 비교하였다. 전문단어 'carnivora'를 분석한 결과는 지식베이스를 기반으로 한 HIIA-1a와 수작업으로 구축된 Lycos가 로봇기반 검색엔진인 Altavista나 Excite에 비해 상대적으로 우수한 성능을 보이고 있다. 하지만 일반단어 'gorilla'를 분석한 결과에서는 의외로 모든 웹 기반 정보검색시스템에서 검색된 문서의 대부분에서 부적합의 비율이 절대적인 모습을 살펴볼 수 있었다. 이러한 현상이 단지 'gorilla'라는 특정 어휘만의 특성인지를 확인해보기 위하여, DDC의 예시주에서 일상생활에 널리 쓰이는 키워드 12개를 선정, 약 1200개의 문서를 수집, 평가해 보았다. 평가 결과 위와 같은 현상은 보편적인 현상이며, 이를 통해 "현재 대부분의 웹 정보검색시스템은 이용하는 초보이용자가 일상생활에 자주 쓰이는 단어로 작성된 단일질의를 query로 입력하였을 때 검색 결과는 어휘 의미 중의성에 의한 심각한 검색 정확율 저하를 동반하게 된다"점을 확인할 수 있었다.

3. 지식 기반 필터링 알고리즘

앞에서 살펴본 'gorilla'의 경우와 같은 심각한 검색 정확율 저하 문제는 대부분의 웹 정보검색시스템에서 *TF*기반 문서순위결정방식이 가지는 근본적인 결함인 어휘의미중의성 문제에서 비롯되며, 이를 해결하기 위한 효과적인 웹 문서 필터링알고리즘이 필요함을 살펴볼 수 있다.

검색어가 일반단어인 'gorilla'의 경우 동물에 대한 전문 지식베이스를 갖춘 지능형 정보 에이전트 HIIA-1a의 경우에 가장 나은 성능을 보이긴 했지만, 여전히 높은 부적합의 비율을 살펴볼 수 있었다. 이 경우 검색 정확율 저하는 알고리즘상에 문제라기보다는 지식베이스의 의한 문제라고 생각되어 지는데, 동물에 대한 전문 지식베이스는 이용자가 일상 생활에 널리 쓰이는 일반단어를 검색어로 입력한 경우에는 그다지 별 도움이 되지 못하는 것으로 확인되었다. 실제 하나의 웹 문서는 다양한 주제들을 포함하고 있는 경우가 대부분이며, 'apple'을 검색어로 입력한 경우 적합 문서로 판정된 문서들의 상당수가 순수한 식물학에 관점에서 'apple'에 관한 문서이기보다는 'apple jam'과 같은 다른 주제에 속한 정보를 포함하고 있는 경우가 대부분이었다. 이는 결국 웹 정보검색에 있어 'apple'이 장미과에 속한다는 전문 지식보다는 잼(jam)이란 어휘가 사과(apple)와 어떠한 연관관계를 가지고 있다는 보통 사람이면 흔히 알고있는 일반 지식정보가 훨씬 더 유용한 정보임을 확인할 수 있었다. 하지만 이러한 보통 사람이면 흔히 알고있는 수많은 일반 지식정보를 지식베이스로 구축하는 작업은 사실상 거의 불가능하다. 따라서 기존의 시소러스나 어휘사전에 정보를 지식베이스로 활용하는 방안을 모색해보았으며, 영어어휘 데이터베이스 워드넷[WordNet]을 일반 영역 지식베이스로 활용할 수 있는 새로운 방식의 지식기반 필터링 알고리즘을 개발하였다.



<그림 1> 워드넷 계층 구조(mango and juice)

지식기반 필터링 알고리즘은 SDCC(Semantic Distance for Common Category)알고리즘[부록 1]을 통해 검색된 문서의 적합, 비적합 여부를 판정 짓게 되는데, 문서에서 추출된 키워드들과 원 질의키워드들과의 연관성 계산을 통해 적합성 여부를 판정하게 된다. 만약 원 질의가 단일키워드 'mango'이고 검색된 문서에서 추출된 j번째 키워드 가 'juice'라면, SDCC알고리즘은 워드넷의 계층 지식정보를 이용하여 둘 사이에 연관성을 찾아내게 된다. <그림 1>을 보면, 두 개의 키워드간에는 공통된 범주 'foodstuff'가 존재함을 살펴볼 수 있으며 계산된 SDCC값은 유효범위 내에 존재하게 된다. 하지만 검색된 문서에서 추출된 키워드들이 대부분 'hotel'같은 단어라면, 원 질의키워드 'mango'와 연관성을 찾기 힘들며, 계산된 SDCC값은 유효범위 내에 존재하지 않는다. 검색된 문서에서 추출된 키워드들의 SDCC값을 계산하였을 때, 그 값이 대부분 유효범위내의 존재하지 않을 때에는 그 문서는 최종적으로 비적합 문서로 판정 짓게 된다.

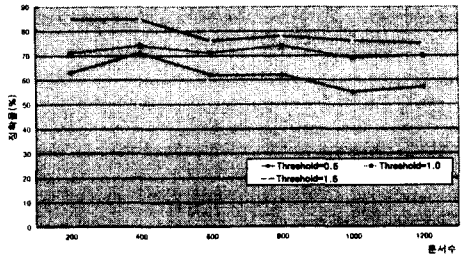
4. 평가 및 결론

성능 평가를 위하여 선정한 키워드 12개는 아래 <표 3>과 같다.

apple cherry kiwi lily mango peanut pepper potato sunflower tomato pear watermelon

<표 3> 키워드 리스트

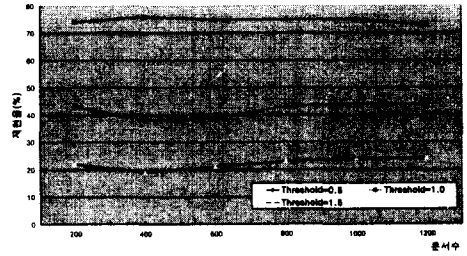
각각에 키워드에 대해 검색엔진 알타비스타를 통해 각각 100개의 문서를 수집하였으며, 알타비스타가 제공하는 Title과 URL 그리고 Description에 나오는 어휘만을 대상으로 하여 필터링 알고리즘을 수행하였다.



<표 4>필터링 알고리즘의 정확율

필터링 알고리즘을 수행한 결과 임계값을 높게 설정할수록 정확율에 있어서 그 결과가 향상되어짐을 볼 수 있었으나, 그에 반비례하여 재현율이 낮아짐을 살펴볼 수 있었다.(<표 4>,<표 5>참조)

<표2>에서 일반단어를 단일검색어로 입력한 경우, 기존의 웹 기반 정보검색시스템들은 모두 정확율이 50%에도 미치지 못하였음에 비하여 본 알고리즘에서는 가장 낮게 설정된 임계값에서도 정확율이 50%이상임을 살펴볼 수 있었다.



<표 5> 필터링 알고리즘의 재현율

본 연구에서는 영어어휘 데이터베이스인 워드넷을 일반 영역 지식베이스로 활용할 수 있는 새로운 방안을 모색하였으며, 이를 통해 지식베이스 구축에 드는 비용절감의 장점을 얻을 수 있었다. 또한 새로운 방식의 지식기반 필터링 알고리즘은 기존의 웹 기반 정보검색시스템의 성능향상에 도움을 줄 수 있으리라 기대되어 진다.

현재 실험에서는 대상 키워드들을 식물 영역만으로 한정하여 테스트를 수행하였으며, 보다 정확한 성능 평가를 위하여 앞으로 대상 영역을 보다 확대하여 검토할 예정이다.

참고문헌

[Marko94] Marko Balabanovic and Yoav Shoham, "Learn ing Information Retrieval Agents: Experiments with Automated Web Browsing", NSF IRI-9411306 with NSF/ ARPA/NASA Digital Library project. 1994.

[황상규99a] 황상규, 변영태, "HIIA-F를 위한 지식베이스와 질의어의 의미적 확장", 한국정보과학회 99년 봄 학술발표논문집, 1999.

[황상규99b] 황상규, 오경묵, 변영태, 천윤심, "어휘의미중의성이 인터넷기반 정보검색에 미치는 영향", 한국정보관리학회 1999년도 학술대회. 1999.

[이용현99] 이용현, "정보통신망에서 지능형 정보 에이전트와 특정 영역에서의 구현", 홍익대학교 박사학위논문 1999.

[박창호98] 박창호, 박민규, 이정모, "가이드라인이 인터넷 정보검색 수행에 미치는 영향", 한국심리학회지: 실험 및 인지, 10권 2호, 1998.

[Cutting97] D. Cutting, "Industry Panel Discussion", SIGIR, 1997(A sample of Excite! queries from September 16, 1997)

[WordNet] Princeton University Cognitive Science Laboratory. WordNet - a Lexical Database for English. http://www.cogsci.princeton.edu/~wn/

[부록 1] SDCC알고리즘

< Semantic Distance for Common Category>

원 질의를 통해 검색된 문서에서 추출된 DocTerm dti와 *QueryTerm qtj*가 존재할 때(dti ≒qtj),

· Set of dti's synsets

=[dti:PS1 , dti:PS2 , ... dti:PSa ... , dti:PSm]

· Set of qtj's synsets

=[qtj:PS1 , qtj:PS2 , ... qtj:PSb ... ,qtj:PSn]

// synset은 원래 term에 의미태그(Possible Sense)정보

// 가 추가 된 것임.

synset dti:PSa와 qtj:PSb의 공통된 범주(Common Category) C:PSm 가 존재하면,

· $\text{SDCC}(\text{C:PSm}) = \frac{1}{p}\sum_{n=1}^{p}\left(\frac{Dn-dn}{Dn}\right)$, n >=2

// 의미거리는 두 노드사이의 링크수의 합으로 계산되어

//진다.

// Dk = 루트로부터 각각의 synset dti:PSa , qtj:PSb

//까지의 의미거리

//dk = C:PSm로부터 각각의 synset dti:PSa , qtj:PSb

//까지의 의미거리

·if (SDCC(K) > θ) then SDCC(K) is valid else invalid.