어휘의미중의성이 인터넷기반 정보검색에 미치는 영향

Effect of Word Sense Ambiguation on Internet-Based Information Retrieval

o

황상규* 오경묵** 변영태* 천윤심**

* 홍익대학교 전자계산학과

**숙명여자대학교 문헌정보학과

Hwang, Sang-Kyu* Yun, Se-Jin**

Oh, Kyung-Mook** Byun, Young-Tae*

* Hong-Ik Univ. **Sookmyung Univ.

요 약

기존의 문헌정보검색에 있어 어휘의미중의성은 검색 효율 저하의 주요 원인 중 하나로 생각되어져 왔다. 어휘의미중의성에 의한 검색 효율 저하란 검색어로 입력한 어휘가 문서에서 서로 다른 의미로 사용됨에 따라 의도하지 않은 다른 문서가 검색될 수 있음을 의미한다. 본 논문에서는 새로운 정보 검색 환경인 인터넷기반 정보검색에 있어 어휘의미중의성이 검색 정확율에 미치는 영향을 살펴보고, 기존에 문헌정보검색에 있어 어휘의미중의성에 관한 연구가 인터넷기반 정보검색에 있어서도 제대로 적용되는지를 확인해 보았다. 또한 실험을 통해 검색어 수와 어휘의미중의성 간의 상관관계를 조사하였으며, 일반 이용자가 인터넷기반 정보검색 수행시 어휘의미중의성에 의한 검색 효율 저하를 최대한 방지할 수는 방법에 대해 모색해 보았다.

1. 서론

정보검색 시스템은 사람과는 달리 어휘의 의미를 구분할 수 없기 때문에, 과일로서의 'apple'과 컴퓨터 상호로 사용된 'apple'을 구분하지 못한다. 이러한 어휘의미중의성 때문에 실제 이용자가 의도한 바와는 전혀 다른 문서가 적합한 문서로 판정지어질 수 있는 것이다. 정보검색에 있어 어휘의미중의성 문제(Word Sense Ambiguation Problem)란 사용자가 입력한 질의어가 실제 웹 문서상에서 여러 가지 다양한 의미로 사용됨에 따라 검색 정확률을 떨어뜨리는 요인으로 작용할 수 있다는 것을 의미한다. 실제 이미 많은 선행 연구자들에 의해 다양한 연구[Weiss73 ; Croft92 ; Voorhees93 ; Sanderson94]가 진행되어 왔으나, 아직까지도 효과적인 해결방법이 제시되지는 못한 상황이다. 여러 가지 다양한 시도에도 불구하고 대부분 기대치에 비해 결과가 뚜렷하게 좋지 못한 편이며, 실제 어휘의미중의성이 발생하는 상황을 살펴보면 무척 다양한 경우의 수가 발생하여 해결책 제시를 어렵게 하고 있기 때문이다. 여러 다양한 주장이 계속되는 가운데 의미 중의성에 관한 연구를 비교 정리한 Sanderson의 연구 결과에서는, "어휘의미중의성 문제는 이용자가 극히 짧은 길이의 질의(very short queries)를 입력한 경우에 한해서만 정보 검색결과에 영향을 끼치게 된다"라고 결론 짓고 있다.

지금까지 어휘의미중의성에 관한 선행 연구들은 문헌정보검색환경 하에서 연구가 진행되어져 왔으며, 새로운 정보검색 환경인 인터넷기반 정보검색환경 하에서는 아직까지 어휘의미중의성이 인터넷기반 정보검색에 미치는 영향에 관해서는 별다른 연구가 진행된바 없다. 인터넷기반 정보검색 역시 문헌정보검색환경과 마찬가지로 텍스트 기반 정보검색을 기본으로 하나, 검색대상이 되는 문서나 써비스 이용자적 측면에서는 기존의 문헌정보검색과는 판이하게 다르다. 일정한 주제를 대상으로 서로 비슷한 성격과 형식을 갖춘 문서들의 모임인 문헌 데이터베이스와는 달리 인터넷은 각기 대상주제가 다른 비 정형화된 문서들의 모임이며, 그 양에 있어서도 상대적으로 훨씬 더 방대하다. 또한 서비스를 이용하는 사용자 역시 소수의 검색 교육을 받은 전문가에서 점점 더 다수의 평범한 일반 이용자들로 확대되어 가고 있다. 매년 인터넷의 이용자수가 2배 이상 증가해 가는 현 상황에서 대부분의 이용자들은 전문적인 웹 정보검색 교육을 받아본 적이 없으며, 앞으로도 새로이 교육을 받을 수 있는 기회 또한 희박한 편이다. 이는 인터넷기반 정보검색을 보다 힘들게 하는 주요 요인으로 작용하게 되는데, 정확한 검색식을 작성한 능력을 갖추지 못한 이용자들은 부정확한 어휘를 검색어로 선정하기 쉬우며, 각기 다른 다양한 주제들을 담고 있는 웹 문서들을 대상으로 한 정보검색이기 때문에 어휘의미중의성에 의한 검색정확율 저하의 가능성은 문헌정보검색에 비해 훨씬 심각한 것으로 확인되었다. 실제 인터넷을 통해 정보검색을 해본 이라면, 인터넷기반 정보검색의 결과가 문헌정보검색에 비해 훨씬 검색 정확율이 낮다는 사실을 쉽게 경험해 보았을 것이다.

어휘의미중의성이 인터넷기반 정보검색에 미치는 영향을 조사하고 기존의 어휘의미중의성에 관한 선행연구결과가 인터넷기반 정보검색환경 하에서도 제대로 적용될 수 있는지를 확인하기 위하여 실제 현실세상에서 그 실태를 조사할 경우에는, 수많은 인터넷 이용자들의 성향 및 그들의 다양한 검색식을 검토해볼 필요가 있다. 따라서 본 연구에서는 설문 조사의 방법 대신 어휘의미중의성 문제가 실제 현실세계에서 발생할 수 있는 상황을 모델링하고, 시뮬레이션을 통해 검증해보기로 한다. 먼저 '극히 짧은 길이의 질의(very short queries)'를 해석하는데 있어서는 Allan[Allan98]의 연구에 excite검색엔진에 이용자 평균 질의어의 개수 2.3개를 기준으로 삼았는데, 웹 정보 검색 이용자의 대부분이 초보자이며 "그들은 대부분 단일질의를 통해 웹 정보검색을 시도한다. [박창호98]"는 현실을 반영해 볼 때, 본 연구에서는 '극히 짧은 길이의 질의'를 검색 질의 길이의 평균치 2.3개보다 작은 1개 혹은 2개인 경우로 가정하였다. 이러한 가정을 전제로 정말 웹 정보검색 이용자가 입력한 질의어의 수가 3개 이상 경우에는 어휘의미중의성에 의한 검색정확율 저하가 더 이상 발생하지 않는지 여부를 웹 정보검색 초보이용자모형을 통한 시뮬레이션을 통해 검증해보았다.

본 실험 결과에서는 인터넷기반 정보검색환경 하에서도 기존의 Sanderson의 가설이 제대로 적용되며, 초보 웹 정보검색 이용자가 모호한 질의를 입력한 경우에는 검색어 수를 3개정도로 유지하는 것이 검색 효율성 측면에서 가장 바람직한 태도임을 확인할 수 있었다.

2.실험 과정

2.1. 실험을 위한 기본전제

본 연구에서는 대상 도메인을 식물로 한정하였으며, DDC의 예시주(example note)에서 일상 생활에 자주 쓰이는 식물명 12개를 선택하였으며, 각각에 대해 검색엔진 알타비스타를 통해 각각 100개씩의 문서를 수집하였다. 검색된 문서의 적합성 여부는 평가시 발생할 수 있는 개인적 성향 차를 감안하기 위하여, 숙대 문헌정보학과 대학원생 6명에 의해 평가되어졌다. 실험을 위해 먼저 2가지 기본 전제를 설정하였으며, 그 내용은 다음과 같다.

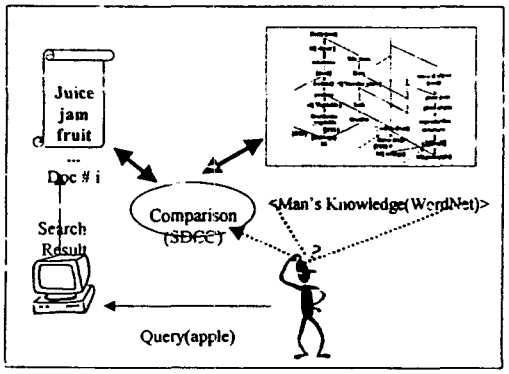
1) 일반적으로 초보 이용자가 입력하는 원 질의어의 수는 1개이며, 본 실험 역시 사용자가 입력한 키워드 1개 외에는 아무런 이용자 추가정보 없이 실험을 수행한다.

2) 일반적으로 초보 이용자가 입력하는 원 질의어는 특수한 식물명이기보다는 일상생활에서 널리 쓰이는 식물명이라는 판단 하에 실험을 위한 키워드를 선

정한다.

2.2. 실험 모형 및 방법

사과를 이용한 음식관련 정보를 찾고자 초보 이용자가 맨 처음 단일 검색 키워드로 'apple'을 입력한 경우, 검색된 문서들에서는 상당히 심한 어휘의미중의성에 의한 검색 정확율 저하를 예상할 수 있다. 이 경우 초보 이용자는 정보검색에 익숙치 못한 관계로 처음부터 적절한 검색식을 작성할 능력을 갖추고 있지 못하다. 또한 초보 이용자가 매 검색단계마다 검색키워드로 선정하는 어휘 역시 전혀 검색의도와는 무관한 단어는 아니지만, 그렇다고 적절하지도 않은 모호한 성격을 지닌 어휘라고 가정한다. 초보 이용자는 다음단계의 재 검색 과정으로 새로운 검색키워드로서 'juice'를 기존의 검색식에 AND연산자로 추가하게 된다. 이때 초보 이용자는 새로운 검색키워드 한 개씩을 추가하는 매 단계마다 검색된 모든 문서의 적합성여부를 확인하며, 전체 검색된 문서 중 상당수가 적합한 문서로 여겨질 경우 정보검색수행을 종료한다고 전제한다.

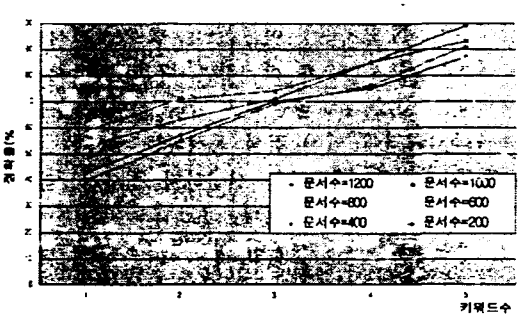


[그림 1] 초보 웹 정보검색 이용자모형

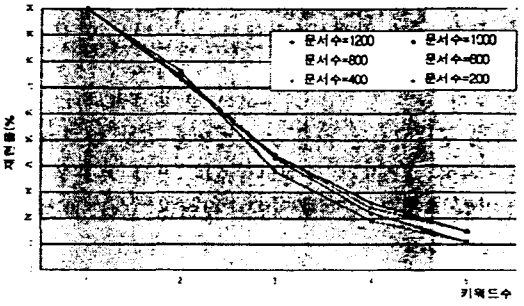
위와 같은 시나리오를 기본으로 하여 본 실험에서는 초보 이용자가 새로운 검색키워드를 생각해내서 재 검색해나가는 과정을 사람을 대신하여 인간의 지식을 대신하게되는 워드넷[WordNet]과 SDCC(Semantic Distance for Common Category)알고리즘[부록 1]이라는 새로운 방법을 이용하여 자동화된 실험을 수행하였다. SDCC알고리즘은 원 질의와 문서에서 추출한 키워드 간에 연관성의 정도를 계산하는 알고리즘이다. 만약 원 질의가 단일키워드 'apple'이고 이를 통해 검색된 문서 D에서 SDCC알고리즘을 통해 문서의 대표키워드로 찾아낸 어휘가 'juice'라면, 이는 앞에 시나리오에서 초보 이용자가 재 검색 과정으로 새로운 검색키워드로서 'juice'를 추가한 경우와 동일한 상태에 해당된다. 이는 초보이용자가 매번 주어진 상황을 고려하여 새로운 검색키워드를 생각해내는 과정을 시스템이 워드넷이라는 지식정보와 SDCC알고리즘을 이용한 일련의 추론 과정을 통해 흉내내게 되는 것이다. 지금까지 언급한 초보 웹 정보검색 이용자 모형을 [그림 1]을 통해 살펴볼 수 있다.

3. 실험 결과 및 향후 계획

문헌정보검색 환경 하에서 사용자가 입력한 질의어의 수가 3개이상인 경우에는 질의어가 문맥상에서 발생하게 되는 의미중의성의 가능성이 극히 희박해진다는 기존의 연구 결과가 웹 정보검색 환경 하에서도 제대로 적용되는지 여부를 검증해보기로 하였다.



[표 1] 키워드 수 증가에 따른 정확률의 변화



[표 2] 키워드 수 증가에 따른 재현율의 변화

초보 웹 정보검색 이용자 모형을 기반으로 한 실험 결과에서는 대체적으로 검색 키워드수가 3개에 이르게 되면, 문서의 적합율이 70%를 넘게 되어 어느 정도 검색결과가 만족할 만 하다는 판단 하에 정보 검색을 종료하게 되었다.([표 1]참조) 이는 기존의 이론이 웹 환경 하에서도 제대로 지켜진다는 것을 의미한다. 여기서 한가지 주목할 점으로 키워드수가 3개 이상 넘어서게 되면, 검색 재현율이 50%이하로 낮아짐을 [표 2]를 통해 살펴볼 수 있다. 특히 검색 키워드수가 5개 정도에 이르게 되면 검색 가능한 문서수가 원래의 10%정도 밖에 되지 않음으로서 무리한 검색이 될 수 있음을 확인할 수 있었다. 이는 문헌데이터베이스에 비해 주제가 명확치 않은 대부분의 일반 웹 문서의 특성을 고려해 볼 때, 검색 키워드 수를 5개 이상 입력하는 방법은 과도한 필터링을 유발하는 비효과적인 검색 태도임을 확인할 수 있었다. 여러 가지 상황을 고려해 볼 때 일반 이용자들이 인터넷기반 정보검색에 있어 적절한 검색어의 수는 3개정도가 타당하리라 여겨진다. 검색 키워드 수를 늘릴수록 검색 정확율은 상당히 높은 수준까지 향상시킬 수 있지만, 이는 초보 웹 정보검색 이용자에게는 큰 부담으로 작용하며, 현실세계에서 초보 이용자는 SDCC알고리즘과는 달리 키워드 수를 늘려 가는 과정에 있어 검색 목적과는 무관한 전혀 엉뚱한 단어를 검색어로 선정하여 쉽사리 '검색된 문서 없음'을 초래하기 쉽다.

본 연구에서는 인터넷기반 정보검색 수행시 발생할 수는 어휘의미중의성과 이에 의한 검색 효율 저하간의 상관관계를 조사하였다. 인터넷기반 정보검색에서는 웹 문서에 특성상 어휘의미중의성의 발생 빈도가 기존의 문헌정보검색에 비해 상대적으로 높으며, 웹 기반 정보 검색시스템에서 보다 나은 검색 효율성을 제공하기 위해서는 어휘의미중의성 해소를 위해 보다 많은 노력을 기울여야 할 필요가 있다.

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

참고문헌

[Weiss73]Weiss SF., "Learning to disambiguate", Information Storage and Retrieval; 9:33-41 .1973

[Croft92]Krovertz R,Croft WB., "Lexical Ambiguity and Information Retrieval",ACM Tansactions on Information Systems, 1992

[Voorhees93]Voorhees EM.,"Using WordNet to disambiguate word sense for text retrieval", Proceeings of ACM SIGIR Conference:16:171-180, 1993

[Sanderson94]Mark Sanderson ,"Word sense disambiguation and information retrieval", Proceedings of SIGIR-94, 17th ACM International Conference on Research and Devlopment in Information Retrieval. Dublin, Ireland. pp 142 - 151. 1994

[Allan98] Ron Papka and James Allan, "Document Classification using Multiword Features", Proceedings of the Conference on Information and Knowledge Management 1998

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

[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 정의를 통해 검색된 문서에서 추출된 DocTerm dti와
QueryTerm qtj가 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,

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