위키피디아를 이용한 검색어의 중의성 해소 및 확장

김성호○, 배상준, 고영중
동아대학교 컴퓨터공학과
{knife16, sjbae0529}@gmail.com, yjko@dau.ac.kr

Ambiguity Resolution and Expansion of Query using Wikipedia

Sung-Ho Kim○, Sang-Joon Bae and YoungJoong Ko
Dept. of computer Engineering. Dong-A University

1. 서 론

정보 검색 (information retrieval)의 궁극적인 목적 중 하나는 사용자의 정보 의도(information need)를 만족시키는 문서를 검색하는 것이다. 하지만, 종종 사용자가 입력하는 초기 검색어(initial query)는 정보 검색을 위한 명확한 정보를 제공하기에는 너무 짧고, 빈약한 경우가 많다[1]. 따라서, 이러한 경우에 초기 검색어에 대한 중의성을 해소 및 개선할 수 있다면, 사용자가 의도하는 정보를 보다 효율적으로 제공해 줄 수 있다. 본 연구에서는 한국어판 위키피디아를 외부 자원으로 이용하여 초기 검색어에 대한 중의성을 해소하고, 검색어 개선 및 확장을 통하여 사용자 의도를 충족 시켜줄 수 있는 방법을 제안한다. 그리고, 실험을 통하여 위키피디아를 이용한 중의성 해소 및 검색어 확장의 유용성을 평가한다.

2. 본 론

본 논문에서 제안하는 방법은 먼저, 위키피디아의 유용한 자원을 바탕으로 다의어, 동의어, 확장 자질을 포함하는 언어 자원 집합을 전처리 과정을 통하여 구축한다. 그리고, 실제 초기 검색어 처리 과정에서 미리 구축한 언어 자원 집합을 활용하여, 검색 성능을 향상시키는 방법을 제안한다.

다의어 집합은 검색어의 중의성을 해소하기 위한 데이터이다. 다의어 집합 구축은 위키피디아의 언어 자원 중에서 “동음이의어 문서”만을 추출하여 구축하였다. 동의어 집합은 하나의 단어를 같은 뜻의 다른 단어로 표현 할 수 있는 단어들의 집합을 의미한다. 동의어 집합 구축은 위키피디아의 자원 중에서 “넘겨주기 문서”를 추출하여 구축하였다. 언어 자원 집합을 구축하기 위한 마지막 과정은 개체에 대한 상세한 설명을 가지는 “일반 문서”로부터 해당 개체를 표현하는 확장 자질을 추출하는 방법이다. 본 논문에서는 [2]에서 제안한 방법을 응용하여, 개체의 확장 자질을 추출하는 방법을 새롭게 제안한다. 일반적으로 개체 x가 개체 y보다 더 일반적인 개념을 설명할 때, 개체 x는 개체 y의 상위 개념, 개체 y는 개체 x의 하위 개념이라고 정의 할 수 있다[2]. 본 논문에서는 이러한 개체와 개체가 포함하고 있는 하이퍼텍스트 단어들을 이용하여 상·하위 관계를 계산하고, 각 개체의 하위 관계를 가지는 하이퍼텍스트 단어들만 해당 개체를 대표할 수 있는 확장 자질로 정의하였다. 두 개체 x, y에 대해서 식(1)의 방법으로 개체 x와 y의 값을 산출하였다면, 개체 x는 개체 y의 하위개념으로 보았다.

$$TF(x \mid y) / N(y) < TF(y \mid x) / N(x) \quad \cdots\cdots \quad (1)$$

여기에서, TF(x | y)는 개체 y가 가지고 있는 하이퍼텍스트 단어 중에서 개체 x의 단어가 출현한 빈도수이고, N(y)는 개체 y가 가지고 있는 하이퍼텍스트 단어의 총 개수이다.

다음은 전처리 작업을 통하여 생성된 언어 자원 집합을 이용하여 사용자의 실제 초기 검색어를 처리

※ 본 논문은 2010년도 정보(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구 사업임.(No. 2010-0016994)

하는 과정이다. 먼저, 사용자의 초기 검색어가 중의성을 가지고 있을 때, 본 시스템에서는 전처리 과정에서 미리 구축한 다의어 집합을 이용하여 사용자의 초기 검색어의 중의적 표현들을 모두 보여주고, 중의적 표현들 중 사용자의 정보 의도를 직접 선택하도록 하였다. 중의성 처리가 끝난 이후에는 전처리 과정에서 미리 구축한 동의어 집합을 이용하여 초기 검색어나 중의성 처리 과정에서 발생한 사용자 피드백 단어의 동의어를 찾아 추가하는 과정을 거친다. 검색어 처리의 마지막 단계는 확장 자질 집합을 기반으로 검색어가 가지는 하위 개념 단어들을 확장 검색어로 추가한다.

본 논문에서 제안한 방법의 실험 평가를 위해서 자체적으로 개발한 한글 정보검색 시스템을 이용하여 2009년 RSS 신문기사를 색인하였다. 그리고, 실험을 평가하기 위해 아래의 [표 1]과 같이 네이버 인기 검색어 중 중의성을 가지거나 빈약한 한 단어로 구성된 20개의 검색어를 사용하였다. 실험은 컴퓨터 공학을 전공하고 있는 총 10명을 대상으로 검색어에 대한 상위 20개의 검색 결과를 보여주고, 사용자 정보 의도에 대한 적합성 여부를 판단하도록 하였다.

[표 1] 실험에 사용된 2008년 ~ 2009년 네이버 인기 검색어

1	정다빈	6	해운대	11	이소연	16	이광재
2	이안	7	우승연	12	오아시스	17	정성훈
3	빅뱅	8	제시카	13	아이리스	18	김태호
4	히어로	9	김민수	14	조인성	19	성인병
5	W	10	KT	15	아오이	20	민주당

[표 2] P-Precision 실험 결과

구 분	P@5	P@10	P@15	P@20
Base 1 … ①	0.1512	0.1562	0.1733	0.1825
Base 2 … ②	0.54	0.4712	0.4287	0.3987
Improvement(①,②)	38.88%	31.5%	25.54%	21.62%
①+M1+M2+M3 … ③	0.6362	0.5625	0.5425	0.5115
Improvement(②,③)	9.62%	9.13%	11.37%	11.28%

위의 [표 2]는 시스템 지향적인 방법을 사용한 P-Precision 결과를 나타낸다. <Base 1>은 하나의 단어로 된 사용자 검색어 이고, <Base 2>는 <Base 1> + 좀 더 세부적인 2~3 단어로 구성된 사용자 검색어이다. <M1>은 다의어 집합, <M2>는 동의어 집합, <M3>는 확장 자질 집합을 나타낸다. Improvement(㉮,㉯)는 ㉮와 ㉯를 비교 했을 때의 성능 향상 차이를 나타낸 것이다. 위의 [표 2]의 실험 결과를 분석하면, 전체 모듈을 사용한 ③에 대한 성능이 P@5에서 63.6%의 성능을 보였다.

3. 결 론

본 논문에서는 온라인 공개 백과사전인 위키피디아를 이용해서 짧고, 빈약한 검색어에 대한 중의성 해소 및 확장하는 방법을 제안하였다. 정보 검색 시스템과 같이 시대상에 민감하게 반응하는 시스템에서 위키피디아는 매우 유용한 정보일 것이다. 본 논문에서는 이러한 중의성 해소 및 확장하기 위한 외부 자원으로써 온라인 공개 백과사전인 위키피디아를 이용하여 높은 성능을 보였다.

향후 연구에서는 좀 더 객관적인 평가 방법으로 실험의 객관성을 확보할 것이다. 아울러, 확장 자질을 추출하는 방법의 다양화하여 실험을 진행해 갈 것이다.

참고문헌

[1] Y. Xu, G. J. Jones, and B. Wang. “Query Dependent Pseudo-Relevance Feedback Based on Wikipedia,” In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp 59-66, 2009.

[2] 이강표, 김현우, 장충수, 김형주, “FolksoViz : Wikipedia 본문을 이용한 상하위 관계 기반 폭소노미 시각화 기법” *정보과학회논문지 : 컴퓨팅의 실제 및 레터*, 제 14권, 제 4호, 2008.