KT Test Set을 이용한 우리말 자연언어검색의 효율성에 관한 비교연구

A Comparative Study on the Effectiveness of Hangul Natural Language Retrieval Using KT Test Set

이 현아, 김 성혁 숙명여자대학교 문헌정보학과

Hyun-A Lee, Sung-Hyuk Kim
Dept. of Library & Information Science, Sookmyoung Women's Univ.

본 연구는 자연언어시스템에서 색인어와 탐색어의 특정성에 기인하는 재현율 감소를 극복하기 위한 방법론으로써 탐색어의 확장을 통한 검색효율을 평가하였다. 이를 위하여 우리말 데이터베이스를 대상으로 주제전문가가 자연언어로 작성한 원 질의문 (Q1), 원 질의문에 사용된 탐색어와 데이터베이스내의 색인어간의 유사도를 이용하여 탐색어를 확장한 질의문 (Q2(0.2), Q2(0.3)), 주제전문가인 이용자가 Q1의 의미적인 관계를 고려해서 자연언어로 탐색어를 확장한 질의문 (Q3)을 검색효율면에서 비교하였다. 실험결과, 평균재현율은 Q2(0.2), Q2(0.3), Q3, Q1의 검색의 순이었다. 평균정확율은 Q3, Q2(0.3), Q1, Q2(0.2)검색의 순으로 나타났다.

1. 서론

최근에 각종 정보기술의 발전으로 색인과 탐색시 자연언어를 사용하는 자연언어시스템의 이용이 증가되었으며, 이러한 자연언어시스템은 미래 정보검색의 중요한 모델로 자리잡을 것이다. 그러나 현재까지의 자연언어시스템의 검색효율에 관한 많은 연구에서 자연언어검색이 여러 가지 장점을 지님에도 불구하고, 자연언어의 특정성으로 인해 재현율이 저조하여, 재현율 향상수단이 필요하다고 결론을 내리고 있다. 따라서 본 연구의 목적은 자연언어시스템에서 색인어와 탐색어의 특정성에 기인하는 재현율 감소를 극복하기 위한 방법론으로 탐색어확장을 통하여 검색효율을 향상시키는 것이다. 이를 위하여 우리말 데이터베이스를 대상으로 주제전문가가 자연언어로 작성한 원 질의문 (original query ; Q1), 원, 질의문에 사용된 탐색어와 데이터베이스내의 색인어간의 유사도를 이용하여 탐색어를 확장한 질의문 (Q2), 주제전문가인 이용자가 원 질의문의 의미적인 관계를 고려해서 자연언어로 탐색어를 확장한 질의문 (Q3)을 검색효율면에서 비교함으로써, 자연언어검색의 재현율을 향상시키는 방안을 제시하였다.

2. 실험

2.1 실험데이터

전술한 3가지 탐색어확장의 검색효율을 비교하기 위하여 실험데이터로 한국통신에서 개발한 KT (Korea Telecommunication)

Test Set (김 성혁, 1994)을 이용하였다. KT Test Set은 정보과학회지, 한국정보과학회지 Proceedings, 정보관리학회지에 수록된 총 1,053건의 문헌레코드로 구성되어 있으며, 각 레코드는 18개의 필드로 구성되어 있다. 본 연구에서는 1,053건의 문헌레코드를 실험데이터로 사용하였고, <keywords>필드에 있는 총 11,305개의 색인어를 대상으로 탐색어와 색인어간의 유사도에 의해 탐색어확장에 사용될 색인어를 추출하였다. 이 색인어는 모두 논문의 표제와 초록에서 사용된 용어를 그대로 색인한 자연언어 색인어로 한 논문당 평균 색인어의 수는 11개이다.

2.2 실험방법

첫째, 주제전문가 4명을 통해 30개의 원 질의문 (이하 Q1)을 작성하였다. 둘째, Q1의 탐색어 확장수단으로 질의문을 구성하는 탐색어와 색인어간의 유사도를 이용하였다. 유사도란 유사한 정도 (likeness)를 수치로 표현한 것으로, 두개의 객체가 하나의 집합일 때 이들의 관계를 측정하는 것이며, 0에서 1의 값을 가진다. 따라서 두 객체가 공유한 속성의 수나 부분 (number or proportion of shared attribute)이 증가하면 유사도 역시 증가하게 된다 (Brian 1974, 50-51). 따라서 용어의 유사도라함은 두 용어간의 유사한 정도를 숫자로 표현한 것으로, 0에서 1의 범위를 지니며, 두 용어가 함께 출현한 문헌이 많을수록 용어의 유사도 역시 증가하게 된다. 본 연구에서는 Q1의 탐색어와 데이터베이스내의 모든 색인어의 동시출현빈도를 근거로, 탐색어와 색인어 사이의 유사도를 산출하여, 주어진 탐색어와 기준치 이상의 유사도를 갖는 색인어들로 확장된 탐색식 (이하 Q2)을 구성 하였다. 탐색어와 색인어간의 유사도를 산출하기 위한 공식은 Tanimtto 유사도공식을 이용하였다. 이는 자연언어로 된 용어간의 유사도를 측정하는 경우 이 공식이 효율적이라는 Dillon과 Calpan (1980, 89)의 주장 에 근거한 것이다. 공식은 다음과 같다.

$$S(x,y) = \frac{n(x,y)}{\{n(x) + n(y)\} - n(x,y)}$$

----------< 공식-1 >

이때, n(x)는 용어 x가, n(y)는 용어 y가 각각 출현한 문헌 수이고, n(x,y)는 용어 x와 y가 같은 문헌에 동시에 출현한 문헌 수이다. 이 공식은 Tanimotto (Roger & Tanimotto, 1962, 1115-1118)가 용어간의 유사도를 집합의 원리를 사용해서 유도한 공식이다. 이 공식으로 N개의 용어를 유사도 행렬로 나타낼 때 N×N의 행렬이 되며, 두 용어 x와 y 사이의 유사도 $S(x, y)$는 0에서 1의 값을 갖는다. 셋째, Q1을 작성한 주제전문가들이 Q1에 대한 의미적인 관계를 고려해서 선정한 자연언어로 확장하여 탐색식 (이하 Q3)을 작성하였다. 넷째, 이상의 3가지 질의문을 불논리검색을 하여 검색효율을 비교하였다. 유사도에 의한 확장용어의 추출과 세가지 질의문의 검색을 위한 프로그램은 Borland C++로 설계하였고, IBM PC 486 DX의 환경에서 실험하였다.

3. 실험의 내용

3.1 Q1의 작성 및 적합성판정

Q1은 정보과학 분야의 전공자 4명을 통해 총 30개를 작성하였다. 이는 1,053건의 문헌을 대상으로 인위적으로 작성한 것이다. Q1의 작성시 'OR' 연산자는 사용하지 않았다. 왜냐하면, 탐색어 확장이 원 질의문의 검색 효율성을 향상하기 위하여 원 질의문에 그와 유사하거나 상위 또는 하위의 개념을 추가하는 것이므로, 이것이 OR개념을 반영하는 것이라고 볼 수 있기 때문이다. 따라서 탐색어 확장방법에 따른 검색효율의 비교라는 본 논문의 목적에 맞도록 'OR' 개념을 사용하지 않고 Q1을 구성하였다. Q1의 주제분야는 대부분 컴퓨터 네트워크, 인공지능, 데이터베이스이며, 사용된 탐색어의 종류는 53종류였다. 한편, 적합성판정은 Q1을 작성한 주제전문가에게 전체 Test Set의 논문 표제와 초

록을 제공해서 적합문헌을 판정하였다.

3.2 유사도에 의한 확장용어추출과 Q2의 구성

Q1에 사용된 53개의 탐색어와 11,305개의 색인어간의 유사도가 전술한 <공식-1>에 의해 53 × 11,305 규모의 행렬로 나타난다. 이 행렬은 다시 기준치에 의해 1 또는 0으로 가공된 53 × 11,305 규모의 이원행렬로 표현되는데, 본 연구에서는 적당한 기준치를 설정하기 위하여 Gotlieb와 Kumar (1968, 499-502)의 Tanimtto 유사도의 변화에 따른 용어수의 변화에 관한 연구를 기초로 하였다. 이들은 $4.0 \leq T^{-1} < 5.0$, $5.0 \leq T^{-1} < 6.0$, $6.0 \leq T^{-1} < 7.0$ 의 영역 안에서의 기준치 변화는 용어군의 규모와 개념에 영향을 주지 않으므로, 기준치를 설정할 때는 이 범위 밖의 값을 선택해야 한다고 하였다. 따라서 본 연구에서는 기준치로 0.2와 0.3을 선택하였다. 즉, Q2는 기준치 (0.2와 0.3)에 따라 두 종류의 질의문 Q2(0.2)와 Q2(0.3)으로 구성되었다. Q2(0.2)는 53 × 11,305 규모의 유사도행렬에서 Q1에 사용된 탐색어와 기준치인 0.2 이상의 유사도를 갖는 색인어를 Q1의 확장용어로 취급하여, 이를 Q1을 구성하는 탐색어와 'OR'로 연결하여 구성하였다. 이와 마찬가지로 Q2(0.3) 역시 <공식 1>에 의해 생산된 유사도 행렬에서 Q1의 탐색어와 유사도 0.3 이상의 관계를 갖는 색인어들로 구성되었다. Q1에 사용된 탐색어 1개당 유사도에 의해 확장된 탐색어의 수는 기준치 0.2에서는 평균 19개, 기준치 0.3에서는 평균 9개였다.

3.3 주제전문가의 탐색어 확장 및 Q3의 작성

주제전문가가 자신이 작성한 Q1과 의미적인 관계에 있는 모든 용어로 탐색식을 작성하였다. 이때, 확장에 사용된 탐색어들은 어떤 어휘집이나 참고도구의 참조없이 주제전문가의 주제지식으로 질의문과 연상되는 용어들로써, 순수한 자연언어로 구성된 것이다. 주제전문가가 탐색어확장에 사용한 용어의 수는 총 217개였고, Q1을 구성하는 탐색어 1개당 평균 4개의 용어로 확장되었다.

4. 실험결과의 분석

4.1 확장된 탐색어(Q2와 Q3)의 분석

Q3의 탐색어들을 분석해 보면, 주제전문가가 탐색어 확장에 사용한 217개의 용어중 73개의 탐색어 즉, 주제전문가가 확장한 용어의 약 33%가 데이터베이스에 색인되지 않은 용어였다. 이것은 한 개념을 자연언어로 표현하는 데 있어서, 다양하고 구체적인 표현이 가능한 자연언어의 특성 곧, 특정성 내지 표현의 다양성을 의미하는 것으로 볼 수 있다. 한편, Q2의 탐색어 수와 Q1을 구성한 탐색어의 데이터베이스내 출현빈도와의 관계를 알아보면, 데이터베이스에 너무 자주 출현하는 탐색어는 확장용어를 적게 갖는다는 것을 알 수 있다. 이는 Q1으로는 검색할 수 없는 적합문헌을 검색하는데 있어서, 데이터베이스에 자주 출현하는 고빈도 용어를 탐색어로 사용하는 것은 확장용어를 많이 갖는 용어를 탐색어로 사용하는 것보다 비 효율적이라고 볼 수 있다. 따라서 유사도에 기초한 탐색어 확장에서 데이터베이스에서 자주 출현하는 용어는 탐색어로 적합하지 않다는 것을 의미한다.

4.2 검색효율성 측정 및 비교분석

검색효율이란 가능한 한 적합문헌은 모두 검색해 내며 동시에 부적합문헌은 검색해 내지 않는 검색시스템의 능력을 평가하는 것으로, 본 연구에서는 전술한 세 가지 탐색어 확장방법들의 검색효율성을 평가하기 위하여 재현율과 정확율을 이용하였다. 주제전문가의 적합성 판정결과를 이용하여 본 실험결과를 평균재현율과 평균정확율로 나타내면 다음과 같다. 평균재현율은 Q2(0.2)의 검색이 44.8%, Q2(0.3)의 검색이 29.9%, Q3의 검색이 26.7%, Q1의 검색이 17.1%의 순이었다. 평균정확율은

Q3의 검색이 69.8%, Q2(0.3)의 검색이 68.9%, Q1의 검색이 62.9%, Q2(0.2)의 검색이 52.8% 순으로 나타났다.

5. 결론

이상의 실험결과를 종합하면 다음과 같은 결론을 내릴 수 있다.

첫째, Q1의 검색결과의 효율성에서, 정확율이 62.9%인 반면 재현율은 17.1%의 낮은 수준을 보임으로써, 자연언어검색에 있어서 탐색어 확장의 필요성이 입증되었다. 둘째, 주제전문가가 어떠한 보조도구도 없이 자연언어로 질의문을 확장하였을 때, 탐색어 확장에 사용된 33%의 용어가 데이터베이스내에 색인되지 않은 용어들이였다. 이것은 자연언어 표현의 특정성 내지 다양성을 의미하는 것으로 자연언어검색에서 탐색어 확장시 보조도구의 필요성을 시사하는 것이다. 셋째, 탐색어와 색인어간의 유사도를 이용한 탐색어 확장은, 개념적인 관계나 의미적인 관계를 고려해서 선정한 탐색어로 검색할 수 없었던 문헌도 검색할 수 있는 좋은 재현율 향상수단이 된다. 또한 정확율의 측면에서도 우수한 성능을 보인다. 따라서 자연언어검색에서 탐색어와 색인어간의 유사도를 탐색어 확장에 이용함은 탐색용 시소러스나 동의어사전의 역할을 보완할 수 있다. 넷째, Q2의 검색효율성을 분석해 보면, Q2(0.3)의 평균정확율이 Q2(0.2)의 평균정확율에 비해 높은 수준을 보인다. 이것은 Q2(0.3)에서 확장된 탐색어들이 Q2(0.2)의 탐색어에 비해 원 질의문의 의미와 개념을 더 잘 반영한다는 것을 시사하는 것이다. 따라서, 같은 문헌에 동시에 자주 출현하는 용어쌍이 같은 주제를 다룬 것이며, 이 용어쌍이 같은 문헌에 덜 출현하는 용어쌍보다 더 주제적으로나 의미적으로 연관되어 있기 때문에 탐색어 확장에 사용할 수 있다는 (Spark Jones & Key 1973, 162-163 ; von Rijsbergen 1979, p.120) 용어의 동시출현빈도의 의의를 나타내는 것이다. 즉, 유사도에 의한 탐색어의 확장에서 유사도의 기준치를 높히는 것은 정확율을 향상시키는 방법이 된다.

따라서 우리말 자연언어시스템에서 탐색어 확장수단으로 탐색어와 색인어간의 유사도를 적용할 수 있을 것이며, 이 유사도에 의한 탐색어 확장에 관한 연구가 전문데이터베이스를 대상으로 체계적으로 진행되어야 할 것이다.

< 참고문헌 >

김 성혁 등. 1994. "자동색인기 성능시험을 위한 Test Set개발." *정보관리학회지* 11(1): 81-102

Dillon, M. & Caplan, D. 1980. "A Technique for Evaluating Automatic Term Clustering." *JASIS* 31 (2): 89-96.

Gilliuno, V. E., et al. 1963. "Linear Association Information Retrieval." In Howerton, P. W. ed. *Vistas in Information Handling* : 30-54.

Gotlieb, C. C. & Kumar, C. 1968. "Semantic Clustering of Index Term." *Journal of the ACM* 15 (4): 493-513.

Roger, D &, Tanimotto, T. 1960. "A Computer Program for Classifying Plants." *Science* 132 (3434): 1115-1118.

Spark Jones, K. 1971. *Automatic Keword Classification for Information Retrieval.* London : Butterworths.

Spark Jone, K. & Key, M. 1973. *Linguistic and Information Science.* New York: Academic Press.

Van Rijsbergen, C. J. 1979. *Information Retrieval.* London : Butterworths.