탐색시소러스를 이용한 신문기사 전문데이터베이스의 검색효율에 관한 연구.

A Study on The Retrieval Effectiveness of Newspaper Database using Search Thesaurus.

이 성욱, 사공 철 숙명여자대학교 문헌정보학과

SeongOuk Lee , Chul SaKong Dept. of Library and Information Science, Sook Myung Women's Univ.

본 연구에서는 전문데이터베이스의 자연어 검색에 있어서 탐색시소러스의 검색효율과 퍼지시소러스 관련어 확장검색의 검색효율을 측정하였다. 한국경제신문사 ECONET의 기사 데이터베이스를 대상으로 질문의 기본 탐색어를 계층어와 관련어로 확장검색한 결과 탐색시소러스를 이용한 관련어 확장검색과 종합검색이 정확률은 저하시키지 않고 재현율을 향상시켰다.

1 서론

전문데이터베이스의 자연어 검색에 있어서 검색효율을 향상시키는 수단으로 여러 학자들은 검색과정에서만 사용되는 탐색시소러스의 이용을 제안하였다. 탐색시소러스는 검색에만 사용하는 탐색도구이므로 색인비용에 대한 부담이 없으며 통제어의 장점과 자연어가 가지는 용어의 특정성을 유지할 수 있다.

전문데이터베이스의 하나인 신문기사 데이터베이스는 기사 전문에서 불용어를 제외한 모든 용어로 검색할 수 있기 때문에 주제를 나타내지 않는 용어라도 함께 검색된다. 또한, 신문에 출현하는 용어는 사회가치규범, 풍조, 언어학적인 유행에 따라 동일한 주제를 표현할 때 약어, 유행어, 신조어 등 다양하고 유동적으로 용어를 사용하기 때문에 검색 시에는 질문에서 표현한 기본 탐색어 이외의 관련 용어를 검색에 이용하여야 한다. 그러나 우리나라에서는 전문데이터베이스의 자연어 검색을 보완하는 탐색시소러스의 이용에 대한 평가가 이루어지지 않고 있다.

따라서 본 연구에서는 신문기사 데이터베이스를 대상으로 탐색시소러스의 검색효율과 퍼지시소러스 관련어로 확장한 검색효율을 측정하여 탐색시소러스의 갱신방법으로 제시하고자 한다.

2 실험방법 및 제한

실험대상인 한국경제신문사 ECONET의 기사 데이터베이스는 국내 중앙지 대부분의 기사를 수록하고 있는 데이터베이스로 1988년 9월 27일부터 현재까지 총 32만 8천 여건이 수록되어 있다. 검색기간은 1993년 10월부터 1994년 10월까지 1년으로 수록된 기사 수는 10만 4백 여건이며 검색주제분야는 금융분야로 하였다. 총 질문수는 20개이고, 탐색어 확장을 위한 도구로 한국

경제신문사의 『경제신문 시소러스』와 퍼지관련어를 이용하였다.

퍼지시소러스 관련어 생성을 위한 용어추출은 1994년 5월부터 7월까지 색인한 금융관련기사 총 373건에서 추출하였다. 총 20개의 질문 중 10개의 질문은 기본 탐색어 검색과 퍼지관련어로 확장검색하였고 10개의 질문은 기본 탐색어, 기본 탐색어+계층어, 기본 탐색어+관련어, 기본 탐색어+계층어+관련어로 확장검색하였다.

퍼지관련어 생성은 C언어로 처리하였고 SAS(Statistical Analysis System)로 검색효율을 측정하였다. Wilcoxon 검증으로 기본 탐색어 검색과 확장검색이 유의적인 차이가 있음을 보였다.

본 연구의 제한점은 다음과 같다.

첫째, 한국경제신문사의 기사 데이터베이스는 동의어 사전을 구축하여 통제하고 있기 때문에, 탐색유형에서 동의어 검색은 못하였다.

둘째, 검색된 기사 수가 천 건 이상이면 디스플레이가 되지 않으므로 천 건 이하로 제한하였다.

3 관련어 생성과 탐색식 작성

3.1 용어수집과 색인어 부여

퍼지시소러스의 관련어 생성을 위한 용어는 시소러스 용어수집의 문헌보증(literary warrant)과 이용자보증(user warrant) 원칙에 근거하여(Lancaster, 1986, p23-28) 한국경제신문에서 색인작업을 통하여 수집하였다.

색인어선정과 동의어는 『경제신문 시소러스』의 우선어를 색인어로 선정·통제하였고, 『증권·금융사전』(서울 : 법문사,1992), 『최신 금융·증권 용어사전』(서울 : 한신경제연구소,1991)을 참고로 단일어보다는 복합어 위주로 부여하였으며 실무의 전문가가 검토하였다.

1994년 5월부터 7월까지 색인한 금융관련기사는 총 373건이고, 총 색인어 수는 4034개이다. 기사 한 건당 부여한 평균 색인어 수는 10.8개이며, 중복이 안된 고유한 색인어 총 수는 1430개이다. 이 중에서 『경제신문 시소러스』에 등록된 색인어는 459개이고 미등록된 색인어는 971개이다. 이것은 신문기사의 용어가 신조어가 많고 유행에 따라 표현방식이 다르기 때문이며 이러한 이유로 신문기사 데이터베이스는 학술분야의 시소러스보다 자주 갱신하여야 한다.

3.2 퍼지관련어의 생성

신문기사를 색인하여 추출한 1430개의 색인어로 Miyamoto(Miyamoto, 1990a, p83-123)가 제안한 퍼지시소러스 관련어 생성공식을 이용하여 관련어를 생성하였다.

Miyamoto가 제안한 퍼지시소러스는 용어의 동시출현빈도와 퍼지집합연산을 이용한 것으로 주어진 키워드와 관련있는 용어를 모아놓은 키워드의 집합이며 색인어집합을 W, 데이터베이스 내의 문헌집합을 D라한다.

$W = \{ w_1, w_2, w_3, w_4 \cdots w_n \}$

$D = \{ d_1, d_2, d_3, d_4 \cdots d_n \}$

퍼지시소러스 생성공식에는 관련어($s(w_i, w_j)$)와 계층어($t(w_i, w_j)$) 공식이 있고, 관련어 생성공식은 자카드계수를, 계층어 생성공식은 Salton (Salton, 1971, p133- 141)의 시소러스 자동생성공식을 적용한 것이다(Miyamoto, 1990a, p58 : 1990b, p196) 퍼지관련어 생성공식은 다음과 같다.

$$s(w_i, w_j) = \frac{\sum_k \min[\, h_{ik}, h_{jk} \,]}{\sum_k \max[\, h_{ik}, h_{jk} \,]}$$

$s(w_i, w_j)$ = 용어 w_i 와 w_j 의 관련도

h_{ik} = 문헌 d_k에서 용어 i 의 출현빈도

h_{jk} = 문헌 d_k에서 용어 j 의 출현빈도

3.3 탐색식의 작성

Ilvonen은(Ilvonen, 1989, Kristensen, Jarvelin, 1990, p79 재인용) 신문기자마다 같은 주제에 대하여 서로 다른 동의어나 관련어로 기사를 작성하기 때문에 신문기사 데이터베이스를 검색할 때에는 이를 고려하여 확장검색을 하여야 한다고 하였다. 따라서 경제신문기자와의 상담을 통하여 총 20개의 질문을 만들고, 신문기사검색을 전문으로 하는 실무 전문가와 협의하여 확장탐색식을 작성하였다.

기본 탐색어는 초기 질문에 포함된 용어이고, 확장 탐색어는 퍼지관련어와 『경제신문 시소러스』에서 선정하였다. 퍼지관련어의 선정은 기준치 0.5이상(Miyamoto, 1990a, p91)의 관련도를 가진 용어 중에서 통계정보의 단점을 보완하기 위하여 질문의 기본 탐색어와 의미적으로 타당한 용어를 탐색어로 선정하였다.

질문 10개는 기본 탐색어와 기본 탐색어+퍼지관련어 검색을 하였고, 10개는 기본 탐색어, 기본 탐색어+계층어, 기본 탐색어+관련어, 기본 탐

색어+계층어+관련어 검색을 하였다. 동의어 확장은 한국경제신문사의 시스템 내부에서 자동으로 동의어가 통제되어 검색이 이루어지고 있기 때문에 기본 탐색어 검색에 동의어를 포함하여 검색결과가 출력된다.

4 실험결과의 검색효율성 분석

4.1 검색효율의 측정

본 연구에서는 종합검색(기본 탐색어+퍼지관련어/기본 탐색어+계층어+관련어)에서 검색된 적합기사를 100%로 간주하고 질문지향적 방법인 평균재현율과 평균정확률을 측정하였다.(사공철 등저, 1990, p218-219) 기사의 적합성 판정은 각 탐색유형에서 검색된 기사의 전문을 capture 받아 전문을 보고 수작업으로 판정하였다.

실험결과 검색된 총 기사 수는 1975건, 최소 2건, 최대 539건이었고 적합한 기사는 최소 2건, 최대 157건이었다. 이 중에서 적합한 총 기사 수는 876건, 부적합한 총 기사 수는 1099건이었다.

<표1> 퍼지관련어 확장검색의 검색효율

탐색유형	평균 재현율	최소/최대	평균 정확률	최소/최대
기본탐색어	46.94	14.29/81.25	57.06	32.50/85.00
퍼지관련어	100	100/100	51.05	24.00/90.70

퍼지관련어 확장검색은 기본 탐색어 검색과 비교하여 평균정확률은 약 5% 감소하였으나 평균재현율은 약 53%로 2배 이상 향상되었으므로 퍼지시소러스 관련어 공식을 탐색시소러스의 갱신을 위한 방법으로 적용할 수 있다.

<표2> 4가지 탐색유형의 검색효율

탐색유형	평균 재현율	최소/최대	평균 정확률	최소/최대
기본탐색어	35.46	0/77.78	47.36	0/85.71
계층어확장	47.31	0/88.89	50.23	0/85.71
관련어확장	86.06	52.94/100	49.40	10.00/85.14
종합 탐색	100	100/100	48.18	7.41/88.15

기본 탐색어 검색과 비교하면, 계층어 확장검색은 평균재현율이 약 12%, 평균정확률이 약 3% 향상되었고, 관련어 확장검색은 약 50%, 약 2% 향상되었으며, 종합검색은 약 65%, 약 1% 향상되었다. 계층어 확장검색보다 관련어 확장검색과 종합검색이 재현율을 향상시키는 수단으로 유용하다는 것을 알 수 있다.

<표3> 종합검색의 검색효율(총 20개의 질문)

탐색유형	평균 재현율	최소/최대	평균 정확률	최소/최대
기본탐색어	41.1952	0/81.25	52.21	0/85.72
종합검색	100.00	100/100	49.61	7.41/90.70

총 질문 20개에 대한 확장검색은 퍼지관련어와 『경제신문 시소러스』에서 용어를 선정하여 확장한 것으로 기본 탐색어 검색은 종합검색에 비하여 검색문헌의 약 40%만 검색한다.

4.2 검색결과의 중복도 측정

각 탐색유형은 기본 탐색어의 확장검색으로 적합기사가 중복되는데, 다른 탐색어의 검색은 다른 기사의 검색요인이 되므로 계층어와 관련어 확장검색에서 검색되는 기사가 동일하지 않다.

추가로 검색된 기사 중에서 고유한 기사와 고유한 적합기사의 비율은 고유한 기사의 총합을 추가로 검색된 기사의 총합으로 나누어 구할 수 있다.

<표4> 추가로 검색된 고유한 기사와 적합기사의 비율

탐색유형	추가로 검색된 고유한 기사 비율	추가로 검색된 고유한 적합기사 비율
계층어	68.26	62.71
관련어	94.50	94.69
종합검색	42.92	26.17

각 탐색유형의 중복도는 낮지만, 추가로 검색된 고유한 기사의 비율은 높았으므로 신문기사 데이터베이스를 검색할 때에는 계층어와 관련어 모두를 사용하여야 각각의 탐색어에서 검색되는 고유한 기사들을 검색할 수 있다. 관련어 확장검색에서 검색된 고유한 적합기사의 비율이 가장 높았다.

4.3 검색효율의 검증

확장검색의 검색효율이 기본 탐색어 검색효율과 유의적인 차이가 있는지 알아보기 위하여 비모수통계방법인 Wilcoxon 검증으로 검증하였다.

<표5> 퍼지관련어 확장검색 검색효율의 검증

비교할 탐색 유형 쌍	Wilcoxon 검증	유의수준
평균재현율		
관련어VS기본탐색어	3.9970	0.0001
평균정확률		
관련어VS기본탐색어	-.5669	*

* 는 통계적으로 유의적인 차이가 없다

퍼지관련어 확장검색의 평균재현율은 0.0001 수준에서 통계적으로 유의적인 차이를 보였으나, 평균정확률은 유의적인 차이를 보이지 않았다. 이 검증결과 퍼지관련어 확장검색은 정확률에 영향을 주지않고 재현율을 향상시킨다는 것을 알 수 있으므로 탐색시소러스의 관련어 갱신방법으로 퍼지 관련어 생성공식을 적용할 수 있다.

<표6> 4가지 탐색유형 검색효율의 검증

비교할 탐색 유형 쌍	Wilcoxon 검증	유의수준
평균재현율		
종합탐색 VS 기본탐색어	3.9997	0.0001
관 련 어 VS 기본탐색어	3.3702	0.001
계 층 어 VS 기본탐색어	1.0986	*
평균정확률		
종합탐색 VS 기본탐색어	-.0757	*
관 련 어 VS 기본탐색어	-.0757	*
계 층 어 VS 기본탐색어	0.2652	*

* 는 통계적으로 유의적인 차이가 없다.

관련어 확장검색과 종합검색의 평균재현율은 각각 0.0001과 0.001 수준에서 통계적으로 유의적인 차이를 보였으나 계층어 확장검색은 유의적인 차이를 보이지 않았으며, 각 확장검색의 평균정확률은 통계적으로 유의적인 차이를 보이지 않았다. 이 검증결과 탐색시소러스를 이용한 관련어 확장검색과 종합검색이 정확률을 저하시키지 않고 재현율을 향상시킨다는 것을 알 수 있다.

5 결론 및 제언

5.1 결론

본 연구에서의 결론은 다음과 같다.

첫째, 신문기사에 출현하는 용어는 신조어가 많으므로 시소러스를 자주 갱신하여야 한다.

둘째, 신문기사 데이터베이스 검색의 재현율을 향상시키는 수단으로 퍼지관련어 확장검색을 이용할 수 있고 퍼지시소러스 관련어 공식을 탐색시소러스의 갱신방법으로 적용할 수 있다.

셋째. 신문기사 데이터베이스 검색의 재현율을 향상시키는 수단으로 탐색시소러스를 이용할 수 있고, 계층어보다 관련어 확장검색이 평균재현율을 향상시키는 방법이라고 할 수 있다.

넷째, 다른 탐색어로 검색되는 고유한 기사도 검색하기 위하여 기본 탐색어 이외에 계층어와 관련어로 확장하여야 한다.

다섯째, Wilcoxon 검증결과 관련어 확장검색과 종합검색이 기본 탐색어 검색의 정확률을 저하시키지 않고 재현율을 향상시키는 방법이었다.

5.2 제언

이상의 실험결과를 통하여 다음과 같은 내용을 제언하고자 한다.

첫째, 탐색시소러스를 갱신할 때 통계적인 방법을 이용하여 관련어만이라도 갱신하는 부분갱신 방법을 도입해야 한다.

둘째, 시소러스의 계층어를 자동생성하는 실용적인 연구가 필요하다.

셋째, 본 연구에서는 불논리만을 이용하여 검색하였으므로 전문데이터베이스의 검색에 있어서 근접연산기호방법과 불논리와 근접연산기호와의 비교 연구가 이루어져야 한다.

넷째, 전문데이터베이스의 색인어 부여시 자연어 검색의 특정성을 고려하여 단일어와 복합어에 대한 근본적이고 철저한 연구가 있어야 한다.

참고문헌

사공철 등저, 최신정보검색론, 서울 : 구미무역, 1990

Kristensen, J., and Jarvelin, K., "The Effectiveness of a Searching Thesaurus in Free-Text Searching of a Full-Text Database" *International Classification,* Vol. 17, No. 2 (1990) : pp. 77-84

Lancaster, F. W., *"Vocabulary Control for Information Retrieval"*, 2ed. Arlington, Virginia : Information Resources Press, 1986

Miyamoto, S. *"Fuzzy Sets in Information Retrieval and Cluster Analysis"*, Kluwer Academic Publishers, 1990a.

_____, "Information Retrieval Based on Fuzzy Associations", *Fuzzy Sets and Systems,* Vol. 38 (1990b) : pp. 191-205

Salton, *"The SMART Retrieval System ; Experiments in Automatic Document Processing"*, Englewood Cliffs, NJ., Prentice-Hall, 1971