어휘의 사용 빈도와 프리시젼을 이용한 키워드 검색 성능의 향상 방안

이상희° 이동주 양종원 이태희 이상구
서울대학교
{louder81°, therocks, zeromus, thlee, sglee}@europa.snu.ac.kr

A Methodology Using Frequency and Precision of Terms for Improving the Searching Performance in Keyword Search

Sanghee Lee°, Dongjoo Lee, Jongwon Yang, Taehee Lee, Sang-goo Lee
Seoul National University

요 약

웹에서의 검색을 수행하기 위해 다양한 연구가 수행되었으나, 일반적으로 키워드 검색 방식이 주를 이루고 있다. 이는 검색 대상에 대한 정보가 충분한 경우에는 원하는 검색 결과를 찾아내기 쉬우나, 그렇지 않은 경우에는 사용자로 하여금 원하는 검색 결과를 추출하기 위해 여러 번의 검색을 추가로 수행하는 수고로움을 요구하곤 한다.

이러한 문제를 해결하기 위하여 어휘 간의 관계에 기반을 둔 확장 검색 방식을 제안한다. 시소러스를 바탕으로 유의어 그룹을 정의하고, 사용자의 검색 키워드 정보를 이용하여 어휘 간의 관계 및 그룹 간의 관계를 정의한다. 정의된 관계를 바탕으로 키워드를 확장하고, 확장된 키워드의 사용 빈도와 프리시젼을 이용하여 사용할 어휘를 선별하여 검색을 수행한다.

1. 서 론

인터넷 등장의 의미 중 하나는 누구나 접근할 수 있는 대량의 정보원이 나타났다는 것이다. 하지만 접근할 수 있는 정보의 양이 늘어난 만큼 원치 않는 정보의 양도 늘어났기에, 결과적으로 원하는 정보를 얼마나 빠르고 정확히 찾아내느냐의 문제가 발생하게 되었다.

이러한 문제에 대해 다양한 접근이 이루어졌으며, 그 결과 현재 웹에서의 대부분의 정보 검색은 키워드 검색을 기반으로 하고 있다. 키워드 검색 방식이란 정보 검색을 위한 데이터에서 해당 키워드가 존재하는 문서만을 추출하고, 이에 랭킹을 적용하여 사용자에게 결과를 보이는 방식이다. 사용자가 검색하고자하는 대상에 대한 정보와 이를 위한 키워드를 숙지하고 있는 경우에는 원하는 결과를 어렵지 않게 추출할 수 있으나, 그렇지 않은 경우에는 원하는 결과를 찾아내기 위하여 여러 번의 검색을 수행해야하는 문제가 발생한다. 또한 이렇게 추출된 검색 결과에 대해 확신할 수 없다는 문제도 발생한다[1].

시소러스를 활용하는 것은 위의 문제를 해결할 수 있는 하나의 대안이 될 수 있다. 키워드에 대한 확실한 정보는 없지만 어느 정도 판단의 근거를 갖고 있는 경우, 시소러스를 통해 원하는 키워드를 얻어낼 수 있다. 또한 웹과 같이 거대한 정보의 집합에서 시소러스를 활용하여 질의 확장을 수행하는 것은 최적화된 결과를 보장하지는 않지만, 직관적으로 더 나은 결과를 보장한다[2].

본 논문에서는 이 상황에서 한 걸음 더 나아가 시소러스를 이용하여 유의어 그룹을 정의하고, 저장된 사용자의 검색 키워드 정보를 바탕으로 어휘 간의 관계를 정의한다. 이러한 어휘 사이의 관계를 유의어 그룹 사이의 관계로 확장하여 키워드 검색에서 활용하는 방안을 제안한다. 결과적으로 사용자로 하여금 더 적은 검색 횟수에 더 많은 정보를 손쉽게 얻는 목표를 만족시킬 수 있도록 한다.

본 연구는 정보통신부의 대학 IT 연구센터 ITRC(Information Technology Research Center)의 지원을 받아 수행되었음

2. 관련 연구

시소러스는 일반적인 정보 검색에서 용어 통제 및 탐색어의 확장이나 축소를 통해 검색 효율을 조절하는데 사용된다. 키워드 기반의 유의어 검색에서는 해당 키워드의 의미를 확장하는데 있어 시소러스가 사용된다. 시소러스의 적절성 여부가 현재 구축하려는 시스템의 성능에 크게 관여하기 때문에 잘 구축된 시소러스가 요구된다.

시소러스는 크게 수동 시소러스와 자동 시소러스로 구분된다[1]. 수동 시소러스는 사람이 직접 수작업으로 구축한 것으로 단어 기반 시소러스와 정보 검색 지향, 구(phrase) 기반 시소러스로 나눠지게 된다. 이러한 시소러스는 대개 시소러스 제작자의 필요에 의해 작성되어 특정 분야에 전문적이다. 구축하는데 많은 비용이 소요되며, 구축 후에도 갱신이 어려운 문제점을 갖는다. 자동 시소러스는 유전 알고리즘 등을 이용하여 문서의 내용을 기반으로 시소러스를 구축하는 방식이다. 대개 문서 등의 자료에 많은 영향을 받게 되므로 자료의 선정이 중요하다. 단순히 상호 출현 빈도 등을 통해 구축되므로 제작자의 의도를 정확히 반영하기 어렵다.

본 논문에서는 [3]의 시소러스와 같이 이미 작성된 시소러스를 이용한다.

3. 키워드 확장과 그 활용 방안

3.1 유의어 그룹을 이용한 키워드 확장

유의어 그룹이란 말 그대로 유사한 의미로 볼 수 있는 어휘의

집합이다. 시소러스를 이용하여 정의한 경우처럼 의미적으로 유사한 단어의 집합부터 다음의 예제에서 나오는 '텔레비전' 그룹처럼 외래어 표기를 나타내는 그룹까지 다양하게 정의하도록 한다. 물론 '텔레비전' 그룹 안의 어휘들 모두 의미는 동일하므로 정의에서 벗어나지 않는다. 동음이의어와 같은 경우에는 의미를 구별할 수 있도록 표기를 서로 다르게 하여 각각의 의미 그룹에 소속시킨다.

키워드는 이렇게 정의된 유의어 그룹에 의해 다른 유사 어휘로 표현될 수 있다. 확장된 어휘 중 몇 개만을 선별하여 검색에 이용하면, 사용자가 원하는 결과에 조금 더 가까워질 수 있다[2].

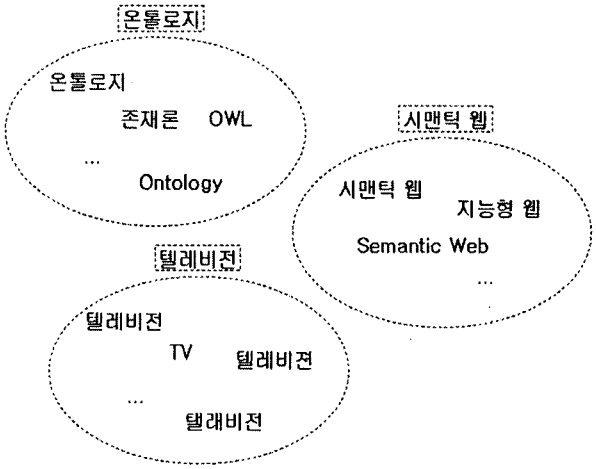

그림 1 유의어 그룹의 예

3.2 어휘 간의 관계 정의를 이용한 키워드 확장

각 어휘 간의 관계를 정의하기 위하여 사용자의 검색어를 이용하도록 한다. 예를 들어 '시맨틱 웹과 온톨로지' 라는 키워드가 검색어로써 여러 번 사용된다면, '시맨틱 웹' 과 '온톨로지' 라는 어휘 간에 밀접한 관련이 있다고 생각할 수 있다. 이러한 관련의 정도를 수학적으로 기술하여, 검색에 활용하도록 한다. 사용자에 의해 사용된 키워드 정보를 모두 저장해두고, 저장된 모든 키워드에 대해 [4]의 AprioriHybrid 알고리즘을 이용하여 키워드 내 어휘 간의 confidence를 계산한다. 이렇게 계산된 confidence를 이용하여 어휘 간의 관계를 정의한다.

한 걸음 더 나아가, 서로 다른 유의어 그룹에 속해있는 어휘 사이의 관계를 바탕으로 유의어 그룹 사이의 관계를 정의할 수 있다. 즉, '시맨틱 웹' 과 '온톨로지' 라는 어휘 사이에 관계가 있다는 것을 바탕으로, '시맨틱 웹' 과 '온톨로지' 라는 유의어 그룹 사이의 관계를 정의할 수 있다. 그룹과 그룹 사이의 관계가 정의되면, 그룹 내의 각 어휘 사이의 관계를 정의할 수 있게 된다. '온톨로지' 와 '지능형 웹' 혹은 'OWL' 과 '시맨틱 웹' 은 모두 그룹 사이의 관계에 따라 각 어휘 사이의 관계로 정의된다. 즉, 그룹 안의 어휘 사이의 confidence가 그룹 전체로 확장되며 이는 다시 각 그룹 내의 어휘 사이의 confidence로 확장된다.

여러 어휘를 하나의 검색 키워드로 사용한다는 것은 두 키워드 모두 사용자가 원하는 검색 결과와 관련이 있다는 것으로 생각할 수 있다. 즉, 어휘 간의 confidence가 높을수록 유사한 결과를 추출할 가능성이 높으며, 이렇게 검색된 유사한 결과는 사용자가 원하는 검색 결과와 일치할 가능성이 높다. 각 어휘 사이의 confidence 값을 정규분포로 변환하여, 해당 confidence가 정규분포의 어느 위치에 존재하느냐에 따라 스코어를 설정한다.

3.3 키워드 검색에서 사용할 어휘의 선정

키워드는 크게 두 방향으로의 확장이 이루어진다. 하나는 유의어 그룹 내 어휘로의 확장이고, 다른 하나는 그룹 사이의 관계를 바탕으로 하는 확장이다. 확장된 키워드 정보를 고려하여 입력된 하나의 키워드에 대해 확장된 여러 키워드로 검색을 수행하게 된다면, 하나의 키워드로 검색하는 기존의 방식에 비해 관련 정보를 더 많이 검색할 수 있다. 하지만 키워드 확장은 키워드의 증가를 의미하며, 검색 루틴이 자동적으로 여러 번 반복되는 효과를 갖는다. 따라서 키워드 확장 결과를 모두 검색에 사용할 수 없다. 질의 확장을 수행할 키워드를 효율적으로 추출하여 소수의 확장 키워드만을 추가 검색에 사용해야한다.

키워드 검색에서 사용할 어휘를 선정하는 첫 번째 기준은 어휘의 precision이다. precision이란 Information Retrieval에서 사용하는 검색 성능에 대한 척도로, 전체 검색 결과와 관련된 문서 사이의 비로 계산된다. 일반적으로 precision이 높을수록 검색 성능이 높다고 판단할 수 있다.

아래의 그림은 여러 어휘를 이용하여 검색할 경우에 precision을 계산하기 위한 다이어그램이다. a_1과 a_2는 각각 검색을 통해 추출된 결과이며, b_1과 b_2는 검색 결과 중 키워드와 관련된 정보를 의미한다. X는 관련된 공통 정보, Y는 검색된 공통 정보를 나타낸다.

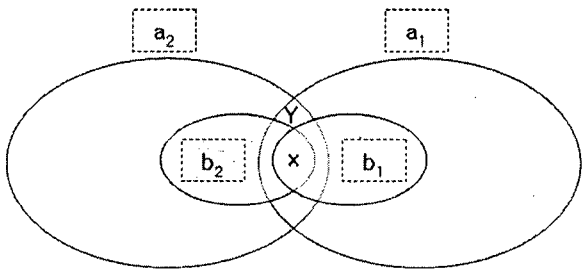

그림 2. precision 계산을 위한 다이어그램

우선 두 검색 결과의 중복이 없다고 가정하자(X, Y = 0). 추가되는 검색을 위한 어휘가 기존 검색 결과의 precision을 증가시키려면, 검색 키워드의 precision보다 큰 어휘를 사용하면 된다.

위의 그림에서 중복이 없다면

한 키워드의 precision = $\frac{b_1}{a_1}$, 추가 검색의 precision = $\frac{b_1+b_2}{a_1+a_2}$

로 정의된다. 추가 검색 결과의 precision이 더 높아야 하므로

$$\frac{b_1}{a_1} < \frac{b_1+b_2}{a_1+a_2} \quad \rightarrow \quad 1$$

1 번식을 정리하면,

$$\therefore \frac{b_1}{a_1} < \frac{b_2}{a_2}$$

검색 결과의 중복이 많은 상태라면 위의 다이어그램에서 Y의 값이 0으로 근사하게 된다. 이러한 조건에서 식을 전개하면 다음과 같다.

위의 그림에서 중복이 존재한다면 $x, y \neq 0$ 이며,

한 키워드의 precision = $\frac{b_1}{a_1} = p_1$,

추가 검색의 precision = $\frac{b_1 + b_2 - x}{a_1 + a_2 - (x+y)}$

로 정의된다. 추가 검색 결과의 precision이 높아야하므로,

$\frac{b_1}{a_1} < \frac{b_1 + b_2 - x}{a_1 + a_2 - (x+y)}$, $y \approx 0 \rightarrow 2$,

2 번식을 정리하면

$x < \frac{a_1 b_2 - a_2 b_1}{a_1 - b_1}$ 가 되고, b_1을 $p_1 a_1$로 치환하면,

$0 < x < \frac{b_2 - p_1 a_2}{1 - p_1}$ 이 되며(중복이 존재하므로 x > 0),

$1 - p_1 > 0$, $b_2 - p_1 a_2 > 0$ 이므로,

$\therefore p_1 < \frac{b_2}{a_2}$ 즉, $\frac{b_1}{a_1} < \frac{b_2}{a_2}$

결국 중복이 존재하는 경우에도 추가 검색의 키워드가 이전 키워드보다 precision이 높으면 더 나은 검색 성능을 보인다. 따라서 입력받은 키워드보다 더 큰 precision 값을 갖는 어휘에 대해서만 추가적인 검색을 수행하도록 한다.

각 키워드의 precision 값은 키워드로 검색된 전체 결과 내에서 통계적으로 사용자에 의해 확인된 결과의 수를 이용하여 정의한다. 확장된 어휘의 precision이 검색 키워드의 precision 보다 크다면 기존 검색 결과에 확장된 어휘에 의한 검색 결과를 추가하고, 반대의 경우라면 확장된 어휘에 의한 검색 결과를 버린다.

키워드 검색에서 사용할 어휘를 선정하는 또 다른 기준은 어휘의 사용 빈도이다. 어휘가 키워드로 많이 사용될수록 키워드로의 가치가 높다고 할 수 있다.

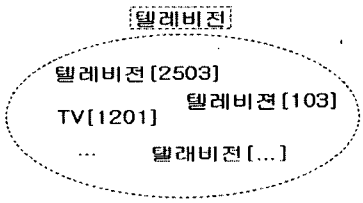

그림 3 어휘의 랭킹을 적용하기 위한 유의어 그룹의 구조

어휘의 랭킹을 계산하기 위해 새로운 파라미터를 정의한다. 위의 그림을 보면 어휘 옆에 빈도가 저장되어 있다. 이 값은 사용자에 의해 학습된 결과로, 키워드 검색에서 얼마나 많이 사용되었는지를 나타낸다. 사용자가 해당 어휘로 검색을 할 때마다 k만큼 증가한다. 여러 개의 키워드로 검색할 경우에는 k/n 만큼 증가한다(n은 검색에 사용한 키워드의 수).

confidence와 마찬가지로 각 어휘의 사용 빈도를 정규분포로 나타내며, 해당 사용 빈도가 어느 정도의 potion에 위치하는지에 따라 어휘의 스코어를 정의한다. 이에 따라 스코어가 높은 어휘만을 선택하여 검색에 사용하도록 한다. 하지만 어휘의 사용 빈다고 높다고 모든 어휘를 검색에 사용할 수 없으므로, 제한을 두어 몇 개의 어휘만을 선택하도록 한다.

실제로 어휘를 확장할 경우에는 두 가지 기준을 모두 이용한다. 어휘의 출현 빈도에 따른 랭킹을 이용하여 많이 사용되는 어휘를 추출한 뒤, 키워드의 precision 보다 높은 precision을 갖는 어휘를 다시 한 번 추출한다. 이렇게 두 번의 과정을 거친 확장된 어휘만을 검색에 활용하도록 한다.

3.4 키워드 검색에서의 활용

유의어 확장 검색이 이뤄진 이후의 결과는 유의어까지 사용하여 검색한 결과까지 모두 포함하므로, 이전의 한 키워드로 검색하는 경우보다 그 수가 늘어나고, 관련된 정보 역시 더욱 많아지게 된다. 따라서 검색 결과에 대한 랭킹이 더욱 중요하다.

유의어를 활용한 검색에서의 랭킹 적용방식은 일반적인 키워드 검색에서의 랭킹 적용방식과 큰 차이는 없지만, 기존의 랭킹 방식에서 term frequency를 사용하는 것과는 달리, 유의어 확장 검색에서는 어휘의 키워드 사용 빈도에 따른 랭킹을 스코어로 변환하여 활용한다. 키워드로 사용된 어휘의 경우에는 이 스코어를 그대로 활용하고, 확장된 어휘의 경우에는 스코어에 confidence에 의한 스코어를 곱하여 이용한다.

4. 결론과 향후 과제

유의어 그룹과 그룹 사이의 관계를 이용함으로써 키워드 검색의 약점을 보완해보았다. 단순 키워드 매칭 방식에서 다양한 유의어의 확장을 통해 검색 결과의 precision을 높일 수 있으며, 더 많은 정보를 한 번의 검색으로 얻어낼 수 있다.

어려운 점 중 하나는 효율적인 시소러스의 구축과 구축된 시소러스를 바탕으로 한 유의어 그룹 사이의 관계 설정이다. 현실적으로 모든 분야에 대한 전문적인 시소러스를 구축하기 어렵기 때문에 자동 시소러스의 구축이 필요하다.

또 다른 문제점으로는 여러 번의 검색에 대한 시간적인 문제이다. 물론 서버의 성능이 점점 좋아지면서 검색에 대한 소요시간은 무시할 만큼 작으나, 여러 번의 검색이 수행되고 그 안에서 이전과는 다르게 몇 가지 계산이 더 필요하게 되므로 한 번의 검색을 위해서는 더 많은 시간이 요구된다. 이를 위한 query 최적화 방안과 효율적인 시스템의 구성에 대한 연구가 앞으로 수행되어야 한다.

5. 참고 문헌

[1] 고삼일, 박사준, 황수철, 김기태, "MeSH를 이용한 개념 기반 검색 엔진 시스템", 정보과학회 춘계학술대회, 2003

[2] Yufeng Jing and W. Bruce Croft, "An Association Thesaurus for Information Retrieval", Proceedings of RIAO, 1994

[3] 한국어 시소러스 사전, http://www.ibookpia.com/index_kthes.html

[4] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", VLDB, 1994

[5] G. Salton, C. Buckley and C. T. Yu, "An Evaluation of Term Dependence Models in Information Retrieval", ACM, 1982

[6] Carolyn J. Crouch and Bokyung Yang, "Experiments in Automatic Statistical Thesaurus Construction", SIGIR, 1992

[7] S. Chaudhuri, G. Das, V. Hristidis, G. Weikum, "Probabilistic Ranking of Database Query Results", VLDB, 2004