사용자 선호도 분석을 통한 검색어 조합 추출

심철우*, 이은주, 김응모
*성균관대학교 정보통신공학부
e-mail:siezka@skku.edu

Finding Correlated Keyword by Analyzing User's Implicit Feedback

Chul-Woo Shim*, Eun Ju Lee, Ung-Mo Kim
*School of Information and Communication Engineering, Sungkyunkwan University

요 약

웹 정보량이 급속히 늘어나면서 원하는 정보를 효율적으로 찾는 검색 기술의 중요성이 커지고 있다. 검색의 정확성을 높이기 위해서는 검색 질의어와 함께 사용자의 환경, 검색 만족도와 같은 다양한 정보가 필요하다. 사용자의 명시적 피드백을 요구하는 것은 거부감을 줄 수 있으므로 사용자의 잠재적 피드백과 연관 검색어 분석을 통해 검색 질의어를 확장하는 연구가 이뤄지고 있다. 그러나 이러한 검색어 확장과 검색 정확성 사이의 상관관계에 대한 분석이 없어 연관 검색어를 정량적으로 평가할 수 없었다. 본 논문에서는 사용자가 검색 질의어를 변경하면서 검색을 반복하는 과정을 사용자의 잠재적 피드백의 하나로 보고 사용자 만족도를 반영하는 페이지 방문 시간과 함께 분석하여 연속적으로 입력된 검색어가 검색 결과의 순위와 사용자 만족도에 미치는 영향을 분석하는 방법을 제안하였다. 마우스 클릭 정보 분석을 통하여 사용자의 검색 만족도를 정량화하였고 특정 주제어에서 관련 검색어가 확장되어 가는 과정은 트리 구조로 표현하였다. 이를 통해 하나의 주제어와 관련해 연속적으로 입력된 검색어 집합으로부터 연관검색어를 추출하고 검색 결과의 정확성을 높일 수 있으며 제안된 트리 구조를 다양한 방향으로 분석하여 검색어, 검색 결과, 사용자 만족도, 배경 지식 등 단순 검색어 분석에서는 나타나지 않는 다양한 정보를 얻을 수 있다.

1. 서론

웹 정보량이 빠른 속도로 늘면서 개인이 찾고자 하는 정보의 불확실성이 높아지고 있다. 방대한 자료 안에서 원하는 정보를 정확하게 찾아주는 것이 검색이다. 검색 서비스에서 중요한 것은 객체화된 지식을 재구성하여 이용자를 중심으로 재배열하고 재구조화 하는 것이다[1]. 검색 결과에는 다양한 정보가 섞여 있으며 검색 서비스는 이용자에게 보다 유용할 것으로 판단되는 정보를 우선적으로 제공한다. 그리고 이용자는 배경 지식과 요약 정보를 바탕으로 검색 결과를 선택한다.

대부분의 검색 서비스에서 다수의 이용자가 선택하는 정보를 보다 정확한 정보로 판단한다. 사용자는 최적의 결과를 찾는 과정에서 상대적으로 정확성이 떨어지는 웹 페이지를 방문해야 한다. 그러나 기존의 알고리즘은 모든 접근을 동등하게 처리하여 가중치를 부여한다[2][3]. 따라서 검색 결과의 상위에 위치하는 검색 결과는 정확도와 무관하게 많은 이용자가 방문하고 이것은 정확성과 신뢰도를 떨어뜨리는 원인이 된다.

또한 사용자는 검색 결과에 원하는 내용이 없으면, 검색어를 변경해 원하는 결과를 얻을 때까지 검색을 반복한다. 이에 따라 사용자가 특정 검색 결과를 선택하는 패턴에 대한 연구가 이루어지고 있다[2][3][4].

최근에는 연속적으로 입력된 검색어의 관계를 분석해 연관 검색어를 추출하는 알고리즘도 연구되고 있다[6][7]. 그러나 검색어가 달라짐에 따라 얼마나 더 정확한 결과를 얻을 수 있는지에 대한 정량적 분석이 없어 연관 검색어의 효용성에 대한 증명이 어렵다.

본 논문에서는 이용자가 검색 결과를 선택하는 과정을 분석하여 사용자 선호도를 측정하는 방법을 제안하고 검색에 사용된 키워드와 검색 결과, 검색 순위, 사용자의 선택에 나타나는 상호 연관성을 파악해 검색 결과의 질을 높일 수 있는 연관 검색어의 패턴을 획득할 수 있는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구 분야를 기술하고 3장에서 사용자의 선호도와 검색 키워드 간의 상호 관계 분석 기법을 제시한다. 4장에서 결론과 응용 분야를 기술한다.

2. 관련 연구

2.1 사용자 선호도

정보의 양이 늘어나면서 전문가 집단에 의한 데이터의 분류보다는 다수의 사용자가 보이는 정보 선택의 경향에 따라 자료의 유용성을 판단하는 방식이 사용되고 있다. 그러나 사용자가 검색 결과로부터 얻은 정보

의 만족도에 대한 측정은 쉽지 않다. 정확한 판단을 위해 명시적인 평가를 요구하는 경우도 있지만 사용자의 거부감을 유발할 수 있어 사용자의 잠재적 피드백을 분석하는 다양한 연구가 이뤄지고 있다[2][3][8].

사용자가 보낸 피드백은 검색 결과의 정확성과 사용자의 개인적 성향에 따라 다른 형태를 보인다. 이러한 피드백을 분석하면 결과의 정확성과 사용자의 성향을 알 수 있다. 따라서 검색 결과의 정확성, 검색 질의어, 사용자 선호도 사이의 연관성을 분석하고 이를 검색 엔진에 학습 시킨 경우, 단순히 선택된 검색 결과만을 분석한 경우보다 높은 정확도를 보인다[5][6][7].

선호도 분석을 위해서는 사용자의 피드백 데이터를 수집해야하는데 사용자의 검색 과정을 방해하지 않고 사용자의 만족도를 측정하는 방법으로 시선 추적(Eye Tracking)과 마우스 클릭 분석(Click Through)이 있다. 시선 추적 방법은 별도의 장비가 필요하여 다수의 이용자를 대상으로 데이터를 수집하기 어렵고, 분석 결과를 일반화시켜 적용하는 것에 한계가 있다. 그러나 비교적 명시적인 사용자의 피드백을 객관화시켜 얻을 수 있기 때문에 정확한 분석이 가능하다[2]. 마우스 클릭 분석은 사용자가 원하는 검색 결과를 얻기 위해 웹 페이지를 클릭하는 행위를 분석하여 검색 결과에 대한 피드백으로 활용한다. 사용자는 자신이 원하는 정보가 있을 것이라고 예상되는 페이지에만 접근하기 때문에 사용자가 방문한 페이지는 상대적으로 질의어와 유사도가 높다고 판단할 수 있다.

사용자의 선택은 검색 결과의 정확도 외에 사용자의 배경 지식, 검색 경향과 같은 다양한 변수에 의해 달라질 수 있지만 단순한 마우스 클릭 정보만으로 이를 분석하는 데는 한계가 있다. 그러나 정보 수집이 용이하고 쉽게 정량화시켜 분석할 수 있기 때문에 많은 검색 엔진에서 사용되며, 효율적인 분석을 위한 모델이 연구되고 있다[3][8]. 표 1에서는 사용자의 마우스 클릭 정보를 분류할 수 있는 기준의 예를 보여주고 있다[3].

<표 1> 마우스 클릭 정보를 이용한 정보 정확도 분석 기준의 예

SA (Skip Above)	최초에 클릭한 검색 결과가 p번째에 위치한다면, p보다 위에 있는 검색 결과는 p번째의 결과보다 덜 적합한 것이라고 예측
SA+N (Skip Above + Skip Next)	화면에 출력된 모든 검색 결과 중에서 사용자가 클릭하지 않은 검색 결과는 클릭한 결과보다 덜 중요하다고 예측
CD (deviation d)	주어진 질의어에 대해 기대한 것보다 더 많이 클릭된(higher-than-expected frequency) 검색 결과가 더 중요한 것이라고 예측
CDiff (margin m)	기대한 것보다 더 많이 클릭된 검색 결과들 중에서 같은 위치(p)에서 상대적으로 더 많이 클릭된 검색 결과가 덜 클릭된 결과보다 좀 더 중요한 것이라고 예측

2.2 연관 검색어 탐색

키워드 기반의 웹 검색에서 질의어의 수가 작을 경우에는 사용자가 원하는 결과를 얻기가 어렵다. 정확한 질의어 해석을 위해 추가적인 정보를 요구할 수 있지만 사용자에게 거부감을 발생시킬 수 있다. 따라서 단어를 의미 범주로 나누어 계층 구조와 연관 관계를 미리 정의해 놓고, 이를 이용하여 사용자의 검색어를 확장하거나[5], 사용자가 사용한 검색어의 히스토리를 분석하여 질의어를 확장하는 방법이 사용된다[6][7].

사용자가 입력한 검색어를 기반으로 연관 검색어를 분석해 사용자에게 제시하면 보다 질 높은 검색 서비스가 가능하다. 연관 검색어를 찾기 위해 데이터 마이닝 기법 중 하나인 연관 법칙을 사용한다. 연관 법칙은 본래 시장 데이터(Market Basket)를 기반으로 연구되는 분야였지만 최근에는 다양한 분야에서 응용되고 있으며, 검색 엔진 데이터에 보다 효율적으로 적용할 수 있도록 변형을 가한 알고리즘이 연구되고 있다[5][6].

기존의 연관 법칙 탐색 기법은 기준 횟수 이상 나타나는 패턴의 집합(large set)을 대상으로 한다. 그러나 연관 검색어 탐색에서는 이렇게 하면 일정 횟수 이상 검색되지 않은 검색어들이 버려진다. 그러나 비교적 적게 검색된 질의어의 연관 법칙도 중요한 의미를 지닌다. 수정된 연관 법칙 알고리즘에서는 이런 집합을 스몰 셋(small set)이라 정의하고 검색 횟수에 대한 함수로 결정되는 낮은 지지도(support)값을 설정한다. 그리고 단순히 많이 검색되었기 때문에 서로 신뢰도가 높은 값들을 제거하기 위해 상호 의존도를 이용한다. 또한 검색어에 대해 만족스러운 결과를 얻었는지에 대한 판단 기준으로 클릭률(click rate)을 사용한다[5][6][8].

3. 선호도 분석을 통한 연관 검색어 추출

3.1 제안 배경

특정 주제와 관련된 연속적인 검색어는 검색 결과에 대한 일종의 피드백이며 사용자가 원하는 검색 결과를 찾는 과정이다. 따라서 검색 이력 데이터를 분석하여 연관 검색어를 추출한다면 제한된 정보로 검색하는 것보다 높은 정확도의 결과를 얻을 수 있다.

그러나 연관 검색어에 대한 적절한 평가가 이루어지지 않고 단순히 함께 검색된 단어의 모임을 연관 검색어로 정의한다면 검색의 정확도를 떨어뜨릴 수 있다. 일반적으로 새로운 검색어를 추가하면 검색의 범위가 줄어드는 대신 더 구체적인 정보를 얻을 수 있다. 이 과정에서 사용자가 기대했던 결과가 누락되거나 낮은 순위에 위치하게 되는 등 검색 결과의 질이 떨어진다면 검색 순위를 조정해야 한다. 다수의 사용자가 특정 검색어를 추가한 후, 검색의 만족도가 낮아졌다면 해당 단어는 연관 검색어 혹은 복합 검색어로서의 가치가 없으며, 검색어 집합과 검색 만족도를 평가하여 이러한 검색어들을 제거해야할 필요가 있다.

3.2 제안 내용

본 논문에서는 마우스 클릭 정보를 기반으로 하는 사용자 선호도 분석을 통하여 웹 페이지 방문 시간을 측정하고, 복합 검색어 집합에 대한 검색 결과와 사용자의 만족도를 평가함으로써 연관 검색어와 복합 검색어의 유용성을 측정할 수 있는 방법을 제안한다.

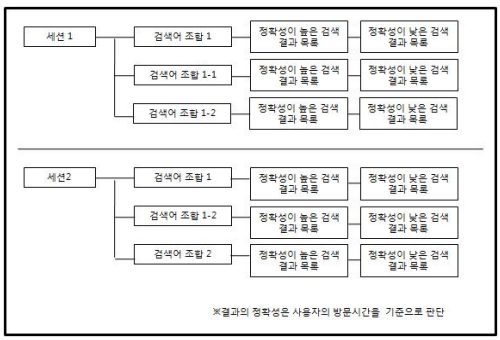
효과적인 분석을 위해서는 데이터를 수집하고 적합한 형태로 나타내는 전처리 과정이 필요하다. 검색어, 선택된 페이지, 피드백과 같은 데이터 수집 후, 의미가 없거나 부정확한 데이터를 제외시킨다.

분석의 기반이 되는 데이터를 수집하기 위해 웹 세션을 이용한다. 사용자가 검색을 시작하면 세션을 시작하고 일정 시간이 지난 후 세션을 종료시키며, 그 사이의 검색 데이터를 쿠키에 기록한다.

특정 검색 결과에 대한 접근이 있을 때 마다 해당 시각을 쿠키에 기록하고, 특정 검색 결과 페이지를 선택한 후에 다른 페이지에 접근할 때까지 걸린 시간을 측정하면 처음 선택했던 페이지에서 보낸 시간을 측정할 수 있다. 이 시간은 해당 페이지의 검색 정확성을 반영한다고 생각할 수 있다. 이렇게 측정된 웹페이지의 주소, 접근 시간, 검색어를 하나의 쌍으로 기록한다.

수집된 데이터 중에서 페이지에 대한 접근 시간이 기준치보다 낮은 기록은 사용자에게 선택된 검색 결과 이지만 정확성이 낮은 정보이다. 기존의 연관성 판단 알고리즘[5][6][7]에서는 검색에 대한 만족도를 사용자의 선택 여부로 측정하였으나 짧은 시간만 방문한 페이지는 사용자가 큰 의미를 두지 않는 페이지로 추가적인 검색으로 이어질 가능성이 높다. 따라서 나중에 검색되었고 사용자가 방문하였더라도 검색의 만족도가 낮은 페이지는 따로 분류하는 전처리 과정이 필요하다.

수집된 데이터는 세션, 검색어, 방문시간으로 구성된다. 분석 단계에서는 효율성을 위해 세션과 검색어 집합을 숫자로 매핑(mapping)시킨다. 그리고 방문시간은 특정 단위로 나누어 사용자 만족도를 나타내는 숫자로 표현하며 방문시간이 길수록 큰 숫자를 할당하고 기준보다 짧은 시간 동안 방문한 경우는 0보다 작은 수를 할당한다.



(그림 1) 전처리 과정 후 데이터의 구조

효율적인 연관 검색어 추출을 위해 (그림 1)과 같이 세션ID, 검색어, 검색 결과를 리스트 형태로 정리한다. 사용자가 선택한 검색 결과는 만족도를 기준으로 둘로 나뉘며 하나의 검색어 집합에 대한 전체 만족도와 만족도가 높은 항목과 낮은 항목의 수와 주소를 저장한다.

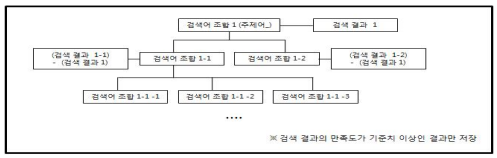
일반적으로 연속된 검색 결과는 상호 연관된 검색일 가능성이 높다[7]. 따라서 한 세션에서 일정 시간동안 수집된 검색 기록은 하나의 주제어를 가지고 있다고 생각할 수 있다. 연관 검색어를 찾기 위해서 각 세션의 주제어를 추출해야 한다. 단독으로 검색되거나 여러 번 검색된 단어가 해당 세션의 주제어라고 할 수 있다.

각 세션에서 주제어가 결정되면 같은 주제어를 가진 세션들을 하나의 집합으로 묶어 연관 검색어를 분석한다. 주제어를 기반으로 검색어 수를 늘리면서 검색어가 확장되는 패턴을 분석한다. 같은 주제어를 가진 세션 전체에서 주제어와 함께 검색된 횟수가 기준치 이상인 검색어는 연관된 검색어의 후보군이 된다. 이런 후보군 중에 검색 결과의 정확성을 높일 수 있는 유효 집합을 찾기 위해 연관 검색어로 검색 후 사용자 만족도의 변화를 쉽게 측정할 수 있어야 한다.

이를 위해 (그림 1)과 같이 정리된 검색어 집합과 검색 결과를 (그림 2)와 같이 트리 구조로 나타낸다. 트리의 루트는 연관된 검색어의 중심이 되는 주제어와 그 검색 결과이며, 특정 노드(node)의 자식 노드는 부모 노드의 검색어에 다른 검색어가 추가된 검색어의 집합이다. 따라서 트리에서 깊이가 k인 노드의 검색어 집합은 주제어로부터 확장된 k개의 검색어로 구성된다. 그리고 각 노드에 검색어 집합 외에 검색 결과, 검색 결과의 만족도를 수치화시켜 저장한다. 사용자 방문시간이 기준치 이하인 만족도가 낮은 결과는 저장하지 않는다. 수치화를 위해서는 다음과 같은 수식을 사용해 간략하게 나타낼 수 있다.

$$\text{사용자 만족도}(s) = \sum[\text{기준치} - \text{방문시간}(t)] \times [\text{가중치}(\text{방문 순서})]$$

또한, 상위 노드의 검색어로부터 검색된 결과는 중복 저장하지 않고 새 검색어를 추가하여 검색한 후에 새로이 추가된 검색 결과만 저장한다. 사용자가 연관된 검색어를 지속적으로 사용할 경우와 그렇지 않는 경우가 발생할 수 있다. 이러한 경우에는 연관성이 없는 키워드를 연관된 검색어로 처리될 수 있으나, 이 알고리즘에서는 사용자의 만족도의 증가, 혹은 감소만을 기준으로 검색어 사이의 연관성을 판단하여 만족도가 늘어나는 검색어는 연관성이 높은 것으로 한다.



(그림 2) 검색어 집합과 검색 결과의 트리 구조

연관 검색어 확장 과정을 트리구조로 나타내고 검색 결과의 변화와 사용자의 만족도간의 상호 관계를 분석하여, 새로운 검색어를 추가했을 때 새로이 검색되는 페이지와 사용자의 만족도 변화를 정량적으로 평가함으로써 효과적으로 연관 검색어를 추출할 수 있다.

또한 수집된 데이터는 다양한 기준에 따라 트리 구조를 형성할 수 있으며, 이를 분석하여 검색 시스템에 대한 의미 있는 결과를 얻을 수 있다. 검색어를 입력된 시간 순으로 정렬해 주제어를 루트로 하는 트리를 만들면 사용자가 하나의 주제어를 바탕으로 검색을 확장하는 형태를 알 수 있다. 동일 주제어를 검색한 다른 세션에서 이러한 트리 구조의 빈발 패턴을 찾으면, 특정 주제어에 대해 검색어를 확장하는 방식을 알 수 있으며, 하나의 검색어 집합을 구성하는 검색어 사이의 의미 관계를 판단하는 근거로 활용할 수 있다.

위와 같이 검색어를 변경하면서 원하는 검색 결과를 찾아가는 과정에서 추가된 검색 결과 페이지의 만족도를 분석하면 특정 주제어에 관해 사용자가 검색을 얼마나 효율적으로 해나가는지 알 수 있으며, 이는 사용자 그룹을 나눌 수 있는 기준이 된다. 이렇게 사용자를 분류하고 특정 분야에 전문적인 사용자의 검색 패턴에 추가적 가중치를 주어 사용자 배경 지식이 검색 과정에 미치는 영향을 검색 결과에 반영할 수 있다.

4. 결론

본 논문에서는 하나의 주제에 대해 연속적으로 입력된 검색어를 트리구조로 나타내고 검색 결과와 사용자 만족도에 미치는 영향을 분석해 연관 검색어를 효과적으로 찾을 수 있는 방법을 제안하였다.

빈번하게 함께 검색된 검색어와 사용자의 클릭률만을 기준으로 연관 검색어를 추출한 후, 검색어 확장에 사용하면 검색 결과의 정확성이 떨어진다. 객관적 분석을 위해 마우스 클릭과 같은 사용자의 잠재적 피드백을 통해 사용자 만족도를 분석하고 주제어를 바탕으로 검색어가 확장되어 가는 방식과 사용자 만족도, 검색 결과의 정확성 사이의 상관관계를 파악할 수 있는 트리 구조를 제안하였다. 연속적으로 입력된 검색어와 마우스 클릭 정보는 검색 과정에서 나타나는 사용자의 검색 흐름과 만족도를 반영하는 잠재적 피드백으로 생각할 수 있으며, 트리 구조를 사용하여 이를 효과적으로 분석하고 검색의 질을 높일 수 있다.

이 논문에서 제안된 트리 구조와 분석 방법을 사용하면 검색어와 검색 결과, 사용자의 만족도 및 배경 지식 등 단순 검색어 분석에서 나타나지 않는 다양한 정보를 얻을 수 있으며 이는 검색 엔진을 구성하는 여러 요소 사이의 의미 관계를 이해하는 기반이 될 것이다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었으며 (IITA-2008-C1090-0801-0028), 21세기 프론티어 연구개발사업의 일환으로 추진되고 있는 지식경제부의 유비쿼터스 컴퓨팅 및 네트워크 원천 기반기술 개발사업의 08B3-B1-10M 과제로 지원된 것임.

참고문헌

[1] 팔란티리2020 “우리는 마이크로 소사이어티로 간다” 웅진윙스

[2] Laura A. Granka, Thorsten Joachims, Geri Gay “Eye-tracking analysis of user behavior in WWW search” Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval pp.478 - 479, 2004 ACM

[3] Thorsten Joachims, Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Geri Gay "Accurately interpreting clickthrough data as implicit feedback" Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval pp. 154 - 161, 2005 ACM

[4] Eugene Agichtein, Eric Brill, Susan Dumais, Robert RagnoLearning "User Interaction Models for Predicting Web Search Result Preferences" Microsoft Research
SIGIR 2006

[5] Gaurav Bhalotia, Arvind Hulgeri, Charuta Nakhe, Soumen Chakrabarti, S. Sudarshan, I.I.T. BombayKeyword “Searching and Browsing in Databases using BANKS” 18th International Conference on Data Engineering (ICDE'02) p. 0431

[6] 문상준, 최재걸 "검색어의 연관법칙" 한국정보과학회, 가을 학술발표논문집 Vol.32, No.2

[7] 김형일(Hyungil Kim), 김준태(Juntae Kim) “질의어 의미별 사용자 선호도를 이용한 웹 검색의 성능 향상” 정보과학회논문지 : 소프트웨어 및 응용 제31권 제8호, 2004. 8 pp. 1101~1112 (12 pages)

[8] Filip Radlinski, Thorsten Joachims “Query Chains: Learning to Rank from Implicit Feedback” Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining pp. 239 - 248 2005 ACM