# User Click Log를 이용한 사용자 검색 의도 분석

지혜성*, 임희석*
*고려대학교 컴퓨터교육학과
hyesung84@korea.ac.kr

# User search intention analysis based User Click Log

Hye-Sung Jee*, Hee-Seok Lim*
*Dept. of Computer Education, *Korea University

요 약

최근 정보검색분야에서는 사용자의 검색 의도를 이해하거나 효과적으로 결과를 전달하고자 하는 시도가 많이 이루어지고 있다. 그러나 현재 제공되고 있는 시스템은 현재 검색 사용자의 의도가 아닌 타인의 의도가 반영된 결과로 실제 사용자의 의도와 상이할 수 있으며, 사용자가 의도하는 바를 유효하게 반영하는 검색 결과를 제시하는 데는 아직 미흡한 실정이다. 따라서 사용자가 원하는 정보를 쉽게 발견할 수 있도록 검색어와 관련된 의도 정보를 제공하거나 검색 결과를 효율적으로 클러스터링 하여 전달하는 기능이 검색의 유용성을 증대시킬 수 있다. 본 논문에서는 검색어에서 사용자의 검색 의도를 자동으로 파악하여 그 의도에 맞는 검색 결과를 제공하기 위하여 사용자 클릭 로그를 사용하여 의도에 맞는 검색결과를 제공하는 방법에 대하여 제안한다.

## 1. 서론

인터넷의 발달에 의하여 정보의 양은 날마다 증가하고 있으며, 그 증가량은 1년마다 2배씩 증가하고 있으며(Google : 98년 약 2,600만 , 00년 약 10억, 08년 약 1조 페이지 색인), 인터넷 호스트의 증가 역시 매년 50% 증가하고 있다(98년 3,000만 -> 06년 4억 3천만). 그러나 늘어나는 정보의 양에 비해, 이를 따르는 정보의 신뢰도는 낮아지고 검색을 통해 원하는 정보를 찾기 위한 노력과 비용은 오히려 증가되고 있다. 따라서 최근 검색 서비스는 검색 알고리즘과 정보의 신뢰도를 강화하려는 전통적인 검색 영역이외에도 사용자의 검색 의도를 파악하여 효과적인 결과 전달을 위한 검색 영역이 확대되고 있다. 즉, 초기 검색 엔들들은 정보공유를 핵심으로 사용자가 원하는 정보를 단순 키워드 매칭이나 개략적인 연관성에 따라 웹 문서에서 대량으로 추출하는 것을 고려하였다면, 최근에는 방대한 정보량 때문에 웹 문서에서 사용자가 원하는 정보를 정확히 파악하여 추출하는 것이 보다 중요해졌다고 볼 수 있다. 그러므로 검색 시스템을 통해서 사용자에게 정확한 정보를 제공하기 위해서는 사용자의 요구를 정확하게 파악하는 문제가 먼저 해결되어야만 사용자의 요구에 따른 정확한 검색이 가능할 것이다. 여기서 사용자의 요구란, 검색 서비스에서 찾고 싶은 것들. 즉, 사용자가 찾고자하는 검색 대상을 가장 잘 표현하는 의도이다. 본 논문에서는 사용자의 요구를 바탕으로 검색 성능 향상을 위해 사용자가 원하는 의도를 자동으로 파악하여 적합한 검색 결과를 도출하는 방법을 제안하고자 한다. 먼저 사용자 질의어에 대한 대표 키워드 집합을 선정한다. 여기서 키워드 집합은 기존 검색 엔진의 사용자 질의어 상위 단어 집합을 말한다. 그리고 이를 기준으로 사용자 Click Log를 기반으로 페이지클러스터링을 수행하고 클러스터링 된 그룹의 대표 키워드를 추출하여 그 군집을 대표하는 키워드로 선택한다. 이 결과를 통해서 나타난 각 군집들의 대표 키워드들이 검색 사용자의 검색 의도를 나타낸 결과로 볼 수 있다.

## 2. 관련 연구 동향

### 2.1. 네이버

네이버의 경우 웹 문서에 대한 클러스터링을 진행하고 있지는 않으며, 대신 이와 유사한 기능을 제공하는 실험을 진행 중에 있다. 네이버의 연관검색어 실험은 사용자가 입력한 검색어의 확장검색어, 상세

검색어, 오타교정/동의어/동음이의어를 제공해준다. 하지만, 사용자가 실제로 탐색한 문서가 아닌 사용자의 키워드에 의존하여 정보를 제공함으로 실제 사용자의 의도와 일치하지 않을 가능성이 높으며, 단어 수준에서 제공하기 때문에 문장의 형태로 존재하는 사용자의 의도를 파악할 수 없다는 단점이 있다.

### 2.2 User Intention based Personalized Search

"User Intention based Personalized Search: HPS(GunWoo Park etc., 2008)"은 웹 기반 정보검색에서 개인화 검색(Personalized search) 측면의 사용자의도(User intention)의 중요성을 강조하면서 서로 다른 의도를 갖는 개인 사용자의 정보 욕구를 충족 시켜줄 수 있는 HPVM(Hierarchical Phrase Vector Mode)을 제안하였다.

구(phrase)기반의 벡터 모델을 통해 사용자의 의도를 계층적으로 확장하고 이를 기준으로 SVM를 이용해 긍정적 혹은 부정적 문서 분류를 통해 사용자의 의도에 적합한 문서를 사용자에게 제공하고자 하였다. 하지만, 이 모델은 사용자 피드백에 의해 질의어를 확장하고, 패턴을 고려하였기 때문에 학습에 대한 내용이 고려되지는 않았다[1].

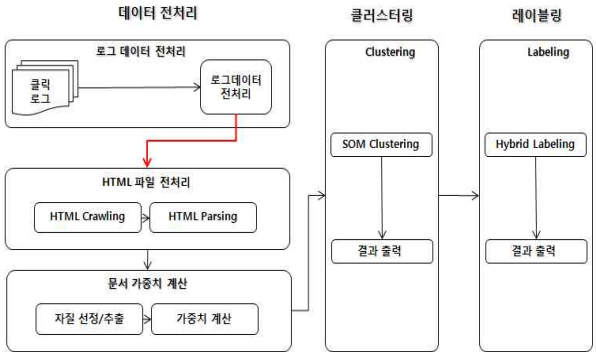
### 2.3 Google wonder wheel

구글 원더 휠은 구글에서 제공하는 부가 검색 도구의 일종으로써, 제시한 검색어에 대한 유사 검색어를 방사형의 형태로 제공하고 있다. 사용자가 입력한 검색어에 맞추어 화면 좌측편에 자동적으로 원더 휠이 그려지며, 단어를 선택할 시에 새로운 방사형 그래프가 나타나는 형태를 띄고 있으며, 뒤로 1단계의 history가 방사형 그래프에 나타나게 된다.

위와 같은 서비스가 모든 검색어에 대해 자동적으로 제시되고 있으나, 사용자의 검색 의도를 다소 놓치는 경향을 보이며 특정 경우에 사이클 현상이 발생하는 등 개선의 여지가 다소 필요하다.

## 3. 클릭로그를 이용한 사용자 의도 분석

본 논문에서는 제안하는 시스템은 [그림 1]과 같이 데이터 전처리, 클러스터링, 레이블링으로 구성되어 있다.




[그림 1] 시스템 구성도

첫째 데이터 전처리 부분에서는 클릭로그 데이터에서 불필요한 부분을 제거한 다음 Log 데이터에 있는 URL에서 데이터를 Crawling, Parsing, 자질 선정 및 추출하여 가중치를 계산하여 벡터로 만든다. 둘째. 클러스터링 부분에서는 클러스터링 알고리즘 선정 실험을 통하여 가장 적절한 알고리즘을 선정하여 문서들이 유사한 주제로 군집화 되도록 클러스터링 한다. 셋째, 레이블링 부분에서는 3가지 방법을 이용하여 클러스터에서 높은 의미를 가지는 단어들을 추출하였다.

### 3.1 데이터 전처리

#### 3.1.1 클릭로그 데이터

사용자가 원하는 의도 정보를 자동으로 파악하기 위한 방법으로, 키워드별로 사용자들이 검색한 페이지에 대한 검색이력을 저장한 클릭로그 데이터를 사용하여 그 정보를 추출하고자 한다. 클릭 로그는 해당 검색어를 입력한 사용자가 실제로 열람한 페이지를 의미하며, 사용자와 검색 시스템 사이에서 기록된 클릭로그 데이터는 이용자의 실제 검색행위를 사실적으로 반영한다[3]. URL에서부터 데이터 Crawling 과정에서 다음 4가지 정보에 관한 부분은 Crawling에서 제외되었다.

첫 번째 일정 빈도 이하의 값을 가지는 URL 데이터를 모두 제거 하였다. 클릭로그 데이터는 사용자들이 검색어를 통하여 찾아본 Page에 대한 로그 데이터이다. 따라서 많은 사람들이 찾아본 Page에 대한 정보도 남아있지만 반대로 많은 사용자들 중에서 단 한사람이 찾아본 Page에 대한 정보도 그대로 저장되어 있다. 이러한 경우는 사용자들의 의도를 반영한 Page라고 볼 수 없다. 따라서 일정 빈도수 이하의 URL는 정보로서의 가치가 없는 것으로 판단하여 Crawling으로부터 제외되었다. 본 논문에서는

2이하의 빈도수를 가지는 URL을 제거하였다.

두 번째 검색 결과 Page를 제거하였다. 검색 결과 Page는 사용자들이 검색어를 입력하고 검색을 하였을 때 나오는 Page를 말한다. 이 Page는 검색어에 대해 자체적으로 정보를 가지고 있지 않고 검색어에 대한 결과 페이지 리스트만을 출력하기 때문에 사용자가 찾고자 하는 의도가 담긴 정보로 보기는 어렵기 때문에 제거 되었다.

세 번째 Error Page를 제거하였다. 잘못된 URL이나 Page가 삭제된 경우에는 가지고 있는 정보가 검색어와 맞기 않기 때문에 데이터로서 가치가 없다. 따라서 본 논문에서는 Error Page 관련 URL은 Crawling 하지 않았다.

마지막으로 연관검색어 Page에 대한 정보를 제거하였다. Click Log 데이터에는 검색어에 대한 URL 정보 뿐만 아니라 검색어에 대한 연관검색어에 대한 URL 정보도 가지고 있다. 그러나 연관검색어의 경우에는 검색어와 유사한 연관검색어도 있지만 전혀 다른 연관검색어도 있다. 예를 들어 '김연아'라는 검색어에는 '김연어 노래'라는 비슷한 검색어가 있는 반면, '아사다 마오'라는 전혀 다른 검색어가 나타나기도 한다. 따라서 연관검색어에 관련된 정보는 모두 제거하였다.

#### 3.1.2 HTML Crawling & Parsing

Click URL Log 데이터에서 앞에서 말한 문제점을 가지고 있는 URL을 제외한 나머지 URL을 가지고 Crawling을 하여 Page를 수집하였다. 그러나 수집된 Page는 HTML 파일이기 때문에 검색어에 대한 정보뿐만 아니라 HTML Tag, Java Script, Link Page 주소등 다른 많은 것들을 포함하고 있기 때문에 검색어에 대한 정보를 제외한 나머지는 제거해야 할 필요성이 있다. 따라서 HTML Parser를 통하여 HTML Tag, Java Script 등을 제거하여 검색어에 대한 정보만이 남도록 하였다.

#### 3.1.3 자질 선정/추출

사용자가 입력한 키워드 및 검색결과에 검색 이력을 기반으로 클러스터링을 하기 위해서 문서별로 자질을 추출하여 가중치를 계산하여 유사한 문서끼리 클러스터링 되도록 문서별 가중치를 계산한다. 자질은 문서 분리도가 높은 단어들로서 주로 개념을 표현하는 명사와 고유명사에 밀집되어 있기에 대상은 주로 명사로 사용되어 왔다[3]. 따라서 본 논문에서는 명사를 자질로 선정하여 추출하도록 하였다.

#### 3.1.4 가중치 계산

본 시스템에서의 가중치 값은 기존 정보검색 방법에서 많이 사용되고 있으며, 정보이론에 따라 정보량이 많은 단어에 가중치를 주고 간단한 계산이 가능하다는 장점이 있는 TF*IDF 값을 사용하였다[4].

$$w_{ij} = tf_{ij} \cdot \log(\frac{N}{df_i})$$

가중치 $w_{ij}$는 $i$ 번째 문서의 j 번째 단어의 가중치를 말하며, $tf_{ij}$는 $i$번째 문서에서 j번째 단어가 나타난 횟수를 말하며, N은 전체에서 존재하는 문서의 개수를, $df_j$는 j번째 단어가 나타난 문서의 수를 말한다. TF값은 문서에서 빈도수가 높은 단어에 가중치를 많이 주며, IDF값은 반대로 여러 문서에서 나타나는 단어의 가중치를 감소시킨다.

계산된 TF*IDF 가중치 값은 클러스터링을 위해서 문서-단어 형태로 만들어지는 벡터 행렬을 만들도록 한다.

### 3.2 Clustering

#### 3.2.1 Cluster 알고리즘 선정

Clustering을 위한 알고리즘은 본 연구에서는 각 구현 방식을 대표하는 알고리즘(SOM, K-means, EM, FarthestFirst,Cobweb)을 대상으로 검색사용 이력 분류 성능 분석을 위한 클러스터링 알고리즘 선정 실험 실시하였다. 알고리즘 선정을 위하여 동일한 Test dataset을 사용하였으며, Test dataset은 기계 학습 분양에서 가장 널리 사용되고 있는 test set 중 하나인 '아이리스'를 사용하여 실험하였다. 실험에 사용된 알고리즘은 SOM(Self Organizing Map), K-means, Cobweb, DBScan 이며, 각 알고리즘 별 분류의 정확도는 [표1]과 같았다.

[표 1] 클러스터링 알고리즘 선정 실험 결과

| Algorithm | SOM | K-means |
|---|---|---|
| Correct clustered instances | 214 (47.56%) | 202 (44.89%) |
| Incorrectly clustered instanced | 236 (52.24%) | 248 (55.11%) |

| EM | FarthestFirst | Cobweb |
|---|---|---|
| 85 (18.89%) | 55 (12.22%) | 50 (11.11%) |
| 365 (81.11%) | 395 (67.78%) | 400 (88.89%) |

실험에서의 결과 같이 SOM 알고리즘이 다른 알고리즘에 비하여 상대적으로 우수한 성능의 결과를 나타내었다. 따라서 본 연구에서는 Clustering 알고리즘은 SOM을 선정하였다.

### 3.2.2 SOM Clustering 실험 결과

Self-Organizing-Map 알고리즘 성능 분석 실험은 클러스터링 알고리즘 성능 분석과 다르게 본 실험에서는 성능향상을 위해 여러 번의 실험 결과를 토대로 각각의 주제어에 맞는 클러스터 수를 도출하기 위해, 문서의 수에 따라 5x5, 6x6의 형태로 클러스터링을 진행하였다.

## 3.3 Labeling

레이블링이란, 클러스터링의 결과로 생성된 각각의 사용자 검색결과 군집을 대표하는 키워드를 선택하는 것을 말한다. 대표 키워드는 크게 두 가지를 의미하는데, 첫 번째는 검색 사용자의 검색의도(intention)을 반영하는 것을 의미하며, 두 번째는 검색 결과의 군집화를 통해 기존 검색엔진의 결과에 비해 효과적으로 검색 결과를 사용자에게 전달 할 수 있음을 말한다.

본 논문에서는 군집 내에서 빈도가 높으면서 클러스터간의 차이를 나타낼 수 있는 Hybrid labeling을 이용해서 레이블링 하였다.

### 3.3.1 Labeling 실험 결과

Hybrid labeling은 internal labeling과 differential labeling의 장점을 가지고 두 방법을 동시에 적용하여 군집 내에서 빈도가 높으면서 클러스터간의 차이를 나타낼 수 있는 단어를 추출하는 기법으로 계산식은 다음과 같다.

$$HL_{i,j} = \alpha freq_{i,j} * \beta idf_{i,j} * \gamma ICF_i$$

$EL$는 Internal Labeling을 말하며, $\alpha$는 단어 발생빈도 가중치, $freq_{i,j}$는 cluster j의 i번째 단어의 발생빈도를 말하며, $\beta$는 idf 가중치, $idf_{i,j}$는 cluster j의 i번째 단어의 internal 문서 가중치이며, $\gamma$는 ICF 가중치, $ICF_i$는 i번째 단어의 Inverse 클러스터 가중치를 말한다. 다음 [표 2]는 검색어 꿈해몽에 대한 클러스터 레이블링 결과이다.

[표 2] 꿈해몽 클러스터 레이블링 결과

| |
|---|
| Cluster 1 : 챙겨가세(0.15) , 구매자(0.15) , strawberry(0.15) , 챙겨가세요(0.15) , 유모차(0.15) |
| Cluster 2 : 실시간게임(0.15) , **로또정보등(0.15)** , **대표운세(0.15)** , 밝은미소(0.15) , **피꿈(0.15)** |
| Cluster 3 : 꿀맛(0.15) , 꿀맛닷컴(0.15) , 채널(0.15) , 규미(0.15) , 규미의(0.15) |
| Cluster 4 : 경숙(0.34) , aaaaaaaa(0.34) , 보석귀걸이(0.34) , 조영옥(0.34) , aaaaaaaaaaaaaa(0.34) |
| Cluster 5 : noises(0.15) , 에스크(0.15) , ask(0.15) , 명예등급(0.15) , 두루미(0.15) |
| Cluster 6 : 무지개(0.22) , 인체(0.22) , 워크넷(0.22) , 워크(0.22) , 해부(0.21) |
| Cluster 7 : 정지기(0.25) , **불나는꿈(0.22)** , 통신사(0.18) , 전화운세상담(0.18) , 정보책임자(0.18) |
| Cluster 8 : 가연(0.17) , 가연결혼정보(0.17) , 안티싱글(0.17) , 시작페이지설정(0.17) , 프로포즈(0.17) |
| Cluster 9 : NAME(0.19) , 진액(0.15) , 옻안타(0.15) , 참옻진액(0.15) , 건영바이오텍(0.15) [ |
| Cluster 10 : 안티싱글(0.17) , 가연결혼정보(0.17) , 가연(0.17), 시작페이지설정(0.17) , 프로포즈(0.17) |
| Cluster 11 : NAME(0.22) , **베스트로또(0.14)** , 스트로(0.14) , 용궁(0.14) , 전대자막제작공간(0.14) |

웹 문서를 이용한 레이블링 결과를 보면, 사용자가 클릭한 웹 페이지가 검색어와 연관된 문장 외에, 필요 없는 문장 및 단어가 많이 포함되어 있기 때문에 뛰어난 성능을 보여주지 못하는 경우가 있다. 하지만, 불용단어 및 주제어와 연관이 없는 단어에 대한 제거가 이루어지지 않았음에도 불구하고 레이블링 결과를 보면 주제어와 밀접하게 연관되는 단어가 추출되었음을 볼 수 있다.

## 5. 결론 및 향후 과제

본 논문에서는 사용자의 검색 의도를 파악하여 사용자의 의도에 맞는 효율적인 검색을 위해서 User Click Log를 이용한 사용자 분석 시스템에 대해 제안하였다. 실험 결과 클릭 로그를 이용하여 클러스터링 및 레이블링 한 결과는 검색하고자 하는 주제어와 관련해서 연관성이 있음을 알 수 있었다.

향후에는 사용자들을 통하여 얼마나 의도를 정확하게 반영하는지에 대한 성능평가와 성능향상을 위한 알고리즘 개선이 필요하며, 실제 검색엔진에 적용하여 어느 정도 효과가 있는지에 대한 평가가 이루어져야 할 것이다.

## 참고 문헌

[1] GunWoo Park, "User Intention based Personalized

Search: HPS", 2008
[2] Zheng Chen, User Intention Modeling in Web Applications Using Data Mining, 2002
[3] 김혜숙, 박상철, "단어/단어쌍 특징과 신경망을 이용한 두 문서간 유사도 측정", 2004
[4] 신동호, "LSA를 이용한 내용기반 검색엔진 시스템", 2000
[5] 박소연, "클릭로그에 근거한 네이버 검색 질의의 형태 및 주제 분석", 2005