http://dx.doi.org/10.3745/KIPSTB.2012.19B.2.149

클릭로그를 이용한 연관키워드 수집

안 광 모[†] · 서 영 훈[††] · 허 정[†††] · 이 충 희[†††] · 장 명 길[††††]

요 약

본 논문은 사용자가 웹 검색을 위해 입력한 키워드와 그 키워드에 의해서 접근한 웹문서의 URL을 이용하여 연관키워드(relevant keyword)를 수집하는데 목적이 있다. 서로 다른 키워드들이라 할지라도 각각의 키워드들이 동일하게 링크된 URL의 수가 많다면, 그 키워드들은 서로 관련성이 높을 것이라는 것이 본 논문의 주된 가정이다. 실제로 이를 검증하기 위해 사용자가 입력한 키워드와 이 키워드를 이용하여 접근한 URL의 정보가 담겨있는 포털사이트의 클릭로그 데이터를 이용하여 URL과 키워드들의 쌍을 추출한 후, 연관키워드 집합을 생성하였다. 그 결과, 실험에서는 최소지지도(minimum support)가 10일 때, 유사어휘 수준에서의 정확도는 89.32%를 보였으며, 유사 어휘는 아니나 관련성이 있는 어휘 수준에서는 99.03%의 정확도를 보였다. 본 논문에서 제안하는 접근 방법은 언어에 독립적이고, 실세계의 데이터로부터 관련성이 있는 단어를 수집할 수 있다는 장점이 있다.

키워드 : 클릭로그, 연관키워드, 키워드 확장

Relevant Keyword Collection using Click-log

Kwang-Mo Ahn[†] · Young-Hoon Seo[††] · Jeong Heo[†††] · Chung-Hee Lee[†††] · Myung-Gil Jang[††††]

ABSTRACT

The aim of this paper is to collect relevant keywords from clicklog data including user's keywords and URLs accessed using them. Our main hyphothesis is that two or more different keywords may be relevant if users access same URLs using them. Also, they should have higher relationship when the more same URLs are accessed using them. To validate our idea, we collect relevant keywords from clicklog data which is offered by a portal site. As a result, our experiment shows 89.32% precision when we define answer set to only semantically same words, and 99.03% when we define answer set to broader sense. Our approach has merits that it is independent on language and collects relevant words from real world data.

Keywords : Clicklog, Relevant Keyword, Keyword Expansion

1. 서 론

컴퓨터로만 인터넷에 접속할 수 있었던 과거의 환경과는 달리 이제는 인터넷에 접속할 수 있는 휴대장치인 스마트폰이 대중화됨에 따라 인터넷의 사용은 일상생활이 되어버렸다. 이러한 환경에서 검색 엔진은 인터넷을 이용하는 사용자들에게 원하는 정보를 간편하고 빠르게 찾을 수 있도록 하는 여러 가지 방법들을 제공해야 한다.

웹 검색을 하는 방법은 여러 가지가 있지만, 현재까지 가장 많이 사용하고 인터넷 사용자에게 가장 익숙한 것은 키워드 기반의 검색 방법이다. 하지만 이러한 키워드 검색의 문제점은 사용자가 요구하는 문서와 관련이 없는 문서들까지 대량으로 제공될 수 있다는 문제점을 가지고 있다. 이러한 문제를 개선하기 위하여 많은 검색 서비스에서 제공하는 방법 중 하나가 사용자가 입력한 키워드와 연관이 있는 키워드들로 확장하는 것이다. 키워드들을 확장하는 기존 연구들은 주로 인터넷 사용자가 접근한 문서의 내용으로부터 키워드들을 확장하는 방법들[2,4,6,7]이다.

본 논문에서 키워드를 확장하기 위한 주된 가정은 한 명 또는 그 이상의 웹 사용자가 입력한 여러 개의 키워드가 있을 때, 그 키워드가 서로 다르다 할지라도 접근한 URL이 같다면 그 키워드들은 서로 관련성이 있을 것이라는 것이다. 그리고 그 키워드들을 이용하여 동일한 URL을 접근한 횟수가 많을수록 해당 키워드 집합들은 서로 관련성이 더 많을 것이라 본다. 본 논문의 가정을 바탕으로 연관키워드들을 수집하는 것이 타당하다는 것을 보이기 위하여 포털

※ 본 연구는 2010년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음.
† 준 회 원 : 충북대학교 컴퓨터공학과 박사과정
†† 종신회원 : 충북대학교 컴퓨터공학과 교수(교신저자)
††† 정 회 원 : 한국전자통신연구원 지식마이닝연구팀 선임연구원
†††† 정 회 원 : 한국전자통신연구원 지식마이닝연구팀 책임연구원
논문접수 : 2011년 8월 2일
수 정 일 : 1차 2011년 10월 10일
심사완료 : 2011년 10월 11일

사이트의 클릭로그 데이터를 이용하였다. 클릭로그는 웹 사용자가 검색어를 입력하고 사용자가 해당 키워드와 그 키워드로부터 나온 결과 페이지에서 사용자가 클릭한 하이퍼텍스트(hyper-text)의 URL 정보를 가지고 있다. 이러한 클릭로그 데이터를 이용하면 키워드와 그 키워드를 이용하여 접근한 URL의 정보를 대량으로 획득할 수 있으며, 키워드와 URL의 쌍으로부터 연관키워드를 추출할 수 있다. 클릭로그를 분석하여 키워드의 주제 분류나 웹 사용자의 성향 등을 분석하는 연구[3,5,8,9]는 있었으나, 연관키워드를 수집하는 연구는 찾아보기가 힘들다. 따라서 본 논문에서 제안하는 방법으로 수집된 키워드들은 언어에 독립적으로 연관키워드들을 수집할 수 있으며, 실제 사용자가 사용하는 비표준 용어까지 처리할 수 있어 실용성이 높아 의미가 있다고 할 수 있다.

본 논문의 2장에서는 클릭로그와 관련된 연구들에 대해서 살펴보고, 3장에서는 본 논문의 핵심 아이디어인 클릭로그를 이용한 연관키워드 확장 방법에 대하여 기술한다. 그리고 4장에서는 본 논문에서 제시한 방법으로 연관키워드를 확장한 실험 결과에 대하여 논하며, 5장에서 결론 및 향후 연구에 대해서 기술하도록 한다.

2. 관련 연구

본 장에서는 연관키워드 수집과 클릭로그를 분석한 기존의 연구들에 대하여 살펴보도록 한다. 먼저 연관키워드를 수집하는 연구들을 살펴보면, [2]의 연구에서는 웹페이지에 동시에 출현하는 단어들의 통계적인 정보를 활용하여 키워드들을 추출하였다. [6]은 연관(association) 클러스터링 기법과 메트릭(metric) 클러스터링 기법, 그리고 스칼라(scalar) 클러스터링 기법을 적절히 조절하여 500개의 경제 관련 기사들로부터 연관키워드를 수집하였다. [4]는 인터넷 광고를 위한 키워드의 확장 방법으로 웹페이지의 성향과 기존 키워드만을 이용하는 적응형 키워드 확장 알고리즘을 제안하였다. [7]에서는 TF-IDF 가중치 모델을 변형하고 이를 뉴스 기사 문서에 적용하여 연관키워드를 수집하였다.

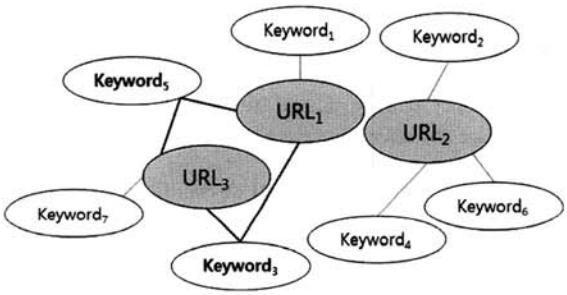
클릭로그를 분석한 연구들도 있는데, [3]의 연구에서는 1년 동안 네이버에 입력된 검색어와 그 검색어를 이용하여 조회한 문서들에 근거하여 국내 웹 검색 질의의 형태 및 주제를 분석하였다. 이 연구는 18,200개의 클릭로그를 분석하여 16개의 주제 범주를 도출하였으며, 비중을 많이 차지하는 범주들은 하위범주로 세분화하였다. [8]은 가격 비교 사이트의 클릭로그 데이터를 이용하여 제품의 가격, 유형, 신뢰정보 등이 사용자의 가격 민감도에 미치는 영향에 대하여 분석하였다. [5]의 연구에서는 협업 필터링과 계층적 k-means 클러스터링 알고리즘을 이용하여 웹 사용자의 웹 사용 습관과 키워드를 분석하였다. [9]는 웹 페이지에 즐겨찾기 추가나 마우스 사용정보와 같은 웹 사용자의 행동 규칙을 만들었다. 그리고 이를 이용하여 의미있는 문서 패턴을 정의하여 의미가 없는 웹 페이지를 제거하였으며, 사용자 질의어 계층구조 모델을 정의하여 사용자의 관심 분야를 분석하고 설계하였다.

위에서 기술한 바와 같이 클릭로그를 분석한 연구나 연관키워드를 수집하는 연구들을 많이 있었지만, 클릭로그를 활용하여 연관키워드를 확장하는 예는 찾아보기가 힘들다. 클릭로그로부터 사용자가 입력한 키워드와 그것이 참조하는 URL의 정보를 이용한 방법은 기존의 방법보다 간단하면서도 효율적으로 연관키워드를 수집할 수 있다.

3. 연관키워드 수집

3.1 URL-키워드 사이의 관계와 클릭로그 데이터

URL은 웹상에서 특정한 웹문서 또는 웹사이트와 1:1로 매핑(mapping)이 되는 인터넷 주소이다. 따라서 URL을 하나의 웹문서 또는 웹사이트로 볼 수 있다. 그런데 하나의 웹문서나 웹사이트들은 비슷한 주제나 또는 그와 관련된 정보들을 담고 있는 경우가 많다. 따라서 같은 URL을 접근한 키워드들 사이에는 서로 관련성이 있을 가능성이 있으며 동일하게 접근한 URL의 수가 많다면 그 키워드들 사이에는 관련성이 있을 가능성이 더욱 높을 것이다. 따라서 본 논문에서는 URL과 그것에 접근한 키워드 정보를 이용하면 연관키워드들을 수집 할 수 있다고 가정한다. 이러한 가정을 실험적으로 증명하기 위하여 본 논문에서는 포털 사이트의 클릭로그 데이터를 이용한다.



(그림 1) 키워드와 URL의 관계

클릭로그란 검색 엔진에 사용자가 실제로 입력한 키워드와 그 키워드로부터 나온 결과들 중 사용자가 실제로 접근한 문서들의 정보를 보관한 로그 파일이다. 검색 엔진을 제공하는 웹 사이트마다 클릭로그의 구조는 다를 수 있으며, 본 논문에서 사용한 클릭로그 데이터의 구조는 다음과 같다.

클릭로그 파일의 구조:
[접속날짜$_1$][접속한IP$_1$][입력키워드$_1$][하이퍼텍스트1][접속한URL$_1$]
[접속날짜$_2$][접속한IP$_2$][입력키워드$_2$][하이퍼텍스트$_2$][접속한URL$_2$]
[접속날짜$_3$][접속한IP$_3$][입력키워드$_3$][하이퍼텍스트$_3$][접속한URL$_3$]
...

본 논문에서는 클릭로그 데이터에서 키워드와 그 키워드로 접근한 URL 정보를 추출하여 연관키워드 수집에 이용한다.

3.2 연관키워드 수집

본 논문에서는 연관키워드를 수집하기 위하여 클릭로그 데이터를 이용한다. 앞 절에서도 기술한 바와 같이 URL과 키워드 정보를 이용하기 위하여 클릭로그 데이터로부터 URL과 키워드 정보를 추출하고 다음과 같이 URL-키워드 집합의 형태로 데이터를 가공한다.

URL-키워드 집합:
[URL1]{[키워드1,1], [키워드1,2], [키워드1,3], …}
[URL2]{[키워드2,1], [키워드2,2], [키워드2,3], …}
[URL3]{[키워드3,1], [키워드3,2], [키워드3,3], …}
…

URL-키워드 집합으로 가공 후 각 데이터의 키워드 집합으로부터 집합의 크기가 2인 키워드의 부분집합을 모두 구한다. URL1에 대한 키워드 집합의 경우라면 {키워드1,1, 키워드1,2}, {키워드1,1, 키워드1,3}, …, {키워드1,2, 키워드1,3}, …과 같은 부분집합들이 생성되게 된다. 이렇게 생성된 집합에서 {키워드i,j, 키워드k,l}과 동일한 집합의 수를 n이라 할 때, n은 키워드i,j와 키워드k,l에 의해서 같이 접속한 URL의 수가 되며, 이것을 본 논문에서는 지지도(support)[1]라 한다. 즉 지지도 n은 다음과 같이 구하게 된다.

$$\text{support } n = \text{Count}(K1 \cup K2)$$

이렇게 추출된 부분집합에 최소지지도(minimum support)를 적용하여 최소지지도 이상의 지지도를 갖는 집합만을 유효 집합(valid set)으로 선별한다. 이렇게 선별된 집합을 다시 각 키워드(대표키워드)와 그 키워드와 같은 집합에 속해 있던 키워드 집합(연관키워드 집합)으로 데이터를 가공하고 정렬하여 연관키워드를 수집한다. <표 1>은 최소지지도 이상의 집합들로부터 수집된 연관키워드를 보여준다.

〈표 1〉 연관키워드 수집

최소지지도 이상의 집합	수집된 연관키워드
{k1, k2} {k2, k3} {k2, k4} {k3, k5} …	k1: {k2} k2: {k1, k3, k4} k3: {k2, k5} k4: {k2} k5: {k3} …

4. 실험 및 분석

4.1 실험 데이터

본 논문에서 제안한 방법이 연관키워드 추출에 효과적임을 분석하기 위해 2009년도에 수집된 포털 사이트의 1년 분량의 클릭로그 데이터를 이용하였다. 클릭로그 데이터의 총 용량은 27.6GB(633개의 파일)이며, 이 중 10개의 파일을 임의로 추출하여 정답셋을 만들었다. 정답셋은 두 가지 유형으로 구축하였는데, 하나는 대표키워드에 대하여 동음이의어나 유사어 수준 또는 직접적으로 관련성이 있는 키워드만을 연관키워드로 한 경우이며, 다른 하나는 대표 어휘에 대하여 뜻은 다르나 그것의 범주가 같아 간접적으로 관계가 있는 키워드까지 연관키워드로 한 경우이다. 정답셋의 구축을 위해서 우선 최소지지도를 3으로 하여 일차적으로 연관키워드를 자동으로 수집하였으며, 이렇게 수집된 결과를 다시 사람이 직접 살펴보며 대표어휘에 대하여 관련성이 없는 연관키워드는 제거하였다. 실험은 정답셋을 구축한 10개의 파일들 중 3개의 파일을 임의로 선택하여 최소지지도를 3에서 10까지 변화시켜가면서 실험을 하였다. 실험에 쓰인 3개의 클릭로그 파일에는 총 701,581건의 클릭로그 데이터가 존재하였으며, 이 중 중복된 클릭로그 데이터(URL과 키워드가 같은 데이터)는 제거하였다. 그리고 이 클릭로그 데이터 중 다른 URL에 비하여 하나의 URL에 비약적으로 많은 키워드가 접근한 경우는 제외하였다. 그 이유는 네이버(naver)나 다음(daum) 등과 같은 대형 포털 사이트의 경우는 서로 관련성이 없는 키워드들이 사용자들로부터 입력이 되므로 이러한 데이터는 연관키워드들을 수집하는데 있어 성능을 감소시키는 원인이 된다. 따라서 본 연구에서는 하나의 URL에 100개 이상의 키워드가 접근한 경우를 제외시켰다. 그리고 하나의 URL에 하나의 키워드만 접근한 경우도 연관키워드를 수집하는데 있어 의미가 없으므로 실험 데이터에서 제외시켰다. 그 결과 총 341,045건의 유효한 클릭로그 데이터가 추출되었다. 다음 <표 2>는 유효한 클릭로그 데이터를 추출한 결과를 정리한 것이다.

〈표 2〉 유효한 클릭로그 데이터 추출 결과
(* 실험에 사용한 데이터)

초기 클릭로그 데이터의 수	701,581
중복된 클릭로그 데이터 수	5,325
유효하지 않은 총 URL의 수	181,649
한 개의 키워드로만 접근한 URL 수	181,474
100 개 이상의 키워드로 접근한 URL 수	175
유효한 총 URL의 수	199,374
유효한 총 클릭로그 데이터*의 수	341,045

4.2 실험 결과 및 분석

최소지지도가 3일 때 클릭로그 데이터로부터 추출된 전체 대표어의 수는 1,781개가 추출되었으며, 그에 대한 전체 연관키워드 수는 2,717개가 추출되었다(하나의 대표어 당 약 1.53개의 연관키워드). 그리고 최소지지도가 10일 때는 92개의 대표어와 103개의 연관키워드가 추출되었다(하나의 대표어 당 약 1.12개의 연관키워드). 실험은 앞 절에서 기술한

1) 본래 본 논문의 지지도(support)란 용어는 Apriori 알고리즘 등의 데이터마이닝[1]에서 쓰이는 용어로 다음 식에 의해서 구하게 된다.

support = (XUY)를 포함하는 트랜잭션수/전체 트랜잭션 수

하지만 본 논문에서는 편의상 전체 트랜잭션의 수를 나누지 않은 값을 지지도(support) 값으로 사용한다.

두 개의 정답셋을 이용하여 다음과 같이 두 개의 그룹으로 나누어 실험을 하였으며, 여기에 사용된 정답셋의 일부를 부록에 제시하였다.

실험1) 대표키워드에 대하여 연관키워드가 동음이의어나 유의어 수준 또는 직접적으로 관련성이 있는 경우

실험2) 대표키워드에 대하여 연관키워드가 대표어휘와 실험1의 경우뿐만 아니라 간접적으로도 관련성이 있는 경우

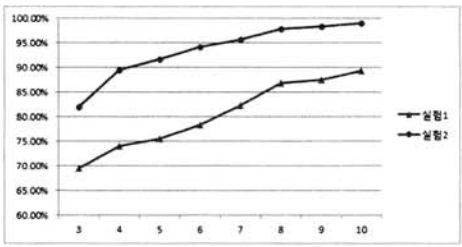
여기서 간접적으로도 관련성이 있다는 것은 대표어와 연관키워드가 그 의미는 다를지라도 두 키워드들 사이의 관계가 비슷한 주제나 범주에 속하거나 어떤 사건에 연관이 있는 경우들이다. <표 3>은 최소지지도가 3일 때의 결과에 대한 평가 예이며, 실험1에서 가수 '강세미'와 직접적으로 관련이 있는 '강세미 남편', '강세미 결혼', 그리고 강세미의 배우자인 '소준'이 추출된 연관키워드 중 정답으로 결정하였다. 그리고 실험2에서 강성연과 같은 가수 카테고리에 속하는 '강성연'이나 강세미가 출현했던 '강심장'이란 예능프로그램 등과 같이 간접적으로 관련성이 있는 연관키워드까지 정답으로 보았다. 실험2의 '박주영'과 같은 경우도 박주영과 같은 축구선수인 '박지성'이 정답에 포함되었다.

〈표 3〉 최소지지도가 3일 때 실험1과 실험2의 평가 예

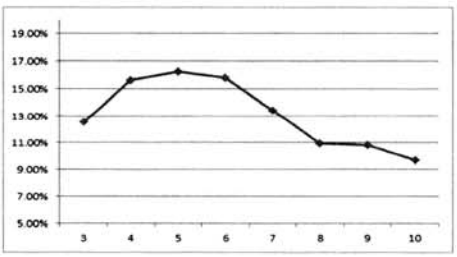
	대표어	정답	오답
실험1	강세미	강세미 남편 강세미 결혼 소준	강병규 아이리스 강심장 강병규 강성연 강세미 결혼
	박주영	박주영 6호골 박주영 5호골	박쥐 박지성
실험2	강세미	강세미 남편 강세미 결혼 강심장 강성연 소준	강병규 강병규 아이리스
	박주영	박주영 6호골 박주영 5호골 박지성	박쥐

그리고 다음의 <표 4>는 실험1, 2에 대한 전체 결과를 정리한 것이다.

최소지지도에 따른 정확도를 분석하여 보면 최소지지도가 높아질수록 정확도가 향상되는 것을 알 수 있으며 이것은 지지도가 높을수록 키워드들 사이에 관련성이 높아진다는 것을 의미한다. 그리고 두 실험 사이의 정확도 차이를 비교해보면 평균 13.16% 정도 실험2의 정확도가 더 높게 나타난다. 이런 정확도의 차이는 간접적으로 관련성이 있는 키워드가 추출되는 비율로 이 차이가 적어질수록 직접적으로 관련성이 있는 데이터들만이 결과로 추출된다는 것이고 차이가 커질수록 간접적으로 관련성이 있는 데이터들이 많이 추출된다는 것을 의미한다. 최소지지도 별로 정확도의 차이를 분석하여 보면 최소지지도가 높아질수록 그 차이가 줄어들 것으로 예상했지만 최소지지도가 3일 때보다 오히려 최소지지도가 5일 때가 그 차이가 더 크게 나타났으며 최소지지도가 5 이상일 때부터는 최소지지도가 커질수록 그 차이가 줄어들고 있음을 볼 수 있다. (그림 2)의 그래프는 최소지지도에 따른 정확도 변화추이를 보여주며, (그림 3)은 최소지지도에 따른 실험1과 실험2의 정확도 차이의 변화를 보여준다.



(그림 2) 최소지지도 변화에 따른 정확도 변화 추이 비교



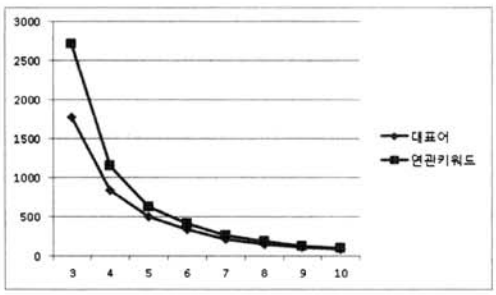
(그림 3) 최소지지도에 따른 실험1과 실험2의 정확도 차이

〈표 4〉 연관키워드 수집 결과(평균 정확도 차이: 13.16%)

최소 지지도	전체 대표어	전체 연관키워드	정답		오답		정확도(%)		정확도 차이(%)
			실험1	실험2	실험1	실험2	실험1	실험2	
3	1,781	2,717	1,887	2,228	830	489	69.45	82.00	12.55
4	840	1,152	852	1,032	300	120	73.96	89.58	15.62
5	498	633	478	581	155	52	75.51	91.79	16.27
6	335	410	321	386	89	24	78.29	94.15	15.85
7	217	260	214	249	46	11	82.31	95.77	13.46
8	155	183	159	179	24	4	86.89	97.81	10.93
9	112	129	113	127	16	2	87.60	98.45	10.85
10	92	103	92	102	11	1	89.32	99.03	9.71

- 실험1: 동음이의어나 유사어 수준 또는 직접적으로 관련성이 있는 키워드를 연관키워드로 본 경우
- 실험2: 간접적으로 관련성이 있는 키워드까지도 연관키워드로 본 경우

다음 (그림 4)의 그래프는 최소지지도에 따라서 추출된 총 대표어의 수와 연관키워드를 나타낸다. 본 논문에서는 재현율을 측정함에 다소 무리가 있어 측정하지 않았지만, 아래 그래프를 보면 재현율의 변화 추이를 예측할 수 있다. 따라서 재현율과 정확도를 고려하여 응용분야에 따라 최소지지도의 적절한 선택을 할 수 있다. 다만, 최소지지도가 낮을수록 그 변화폭이 커서 최소지지도가 작아도 정확도를 높일 수 있는 추가의 연구가 필요하다 할 수 있다.



(그림 4) 최소지지도 변화에 따른 전체 대표어 수 및 연관키워드 수의 변화 추이

5. 결론 및 향후 연구

본 논문에서는 웹 사용자가 입력한 키워드가 달라도 접근한 URL이 같으면 그 키워드들은 서로 관련성이 있다는 간단한 가정을 바탕으로 연관키워드들을 추출하였다. 그리고 연관키워드들을 추출하기 위해 포털 사이트의 클릭로그 데이터를 이용하였다. 관련성이 없는 대량의 키워드가 접근한 URL과 하나의 키워드만이 접근한 유효하지 않은 URL이 포함된 클릭로그 데이터를 제거하여 실험 데이터를 추출하였다. 추출된 실험 데이터로부터 최소지지도를 변화시켜가면서 연관키워드들을 추출하였으며 정확도를 분석하였다. 실험 결과 최소지지도가 10일 때, 유사 키워드 수준에서는 89% 정도의 정확도와 간접적으로 관련성이 있는 키워드 수준에서는 99% 정도의 정확도를 보여 매우 만족할한 성능을 보였다. 본 논문에서 제안한 방법은 클릭로그 데이터로부터 단순한 방법으로, 그리고 언어 독립적으로 연관키워드들을 수집할 수 있다는 장점이 있다. 하지만 클릭로그의 특성 상 데이터의 양이 방대하여 대용량의 클릭로그 데이터를 효과적으로 처리해야 하며, 따라서 대량의 데이터를 효과적으로 처리할 수 있는 방법에 대한 연구가 필요하다.

참 고 문 헌

[1] B. Liu, 'Web Data Mining', Springer, 2006.

[2] Y. Matsuo, M. Ishizuka, "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information," American Association for Artificial Intelligence, Vol.13, No.1, pp.157-169, 2003.

[3] 박소연, 이준호, 김지승, "클릭 로그에 근거한 네이버 검색 질의의 형태 및 주제 분석", 한국문헌정보학회지, 제39권 제1호, pp.265-278, 2005.

[4] 서범준, 이세일, 유승학, 윤성로, "인터넷 광고를 위한 웹 페이지 기반의 키워드 확장 알고리즘", 한국인터넷정보학회 2010년도 학술발표대회, pp.241-246, 2010.

[5] 윤태복, 이승훈, 윤광호, 이지형, "웹 사용 정보에 기반한 다중 성향 키워드 모델의 설계와 응용", 인터넷정보학회논문지, 제10권 제5호, pp.95-105, 2009.

[6] 이상훈, 김기태, "클러스터링 기법을 이용한 키워드 유사도 순위화 알고리즘에 따른 사용자 질의 확장", 한국정보과학회 2003년도 봄 학술발표논문집, 제30권 제1호(B), pp.479-481, 2003.

[7] 이성직, 김한준, "TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법", 한국전자거래학회지, 제14권 제4호, pp.59-73, 2009.

[8] 전종근, 박철, "웹 로그 데이터를 이용한 온라인 소비자의 가격민감도 영향 요인에 관한 연구", Journal of Information Technology Applications & Management, pp.1-16, 2006.

[9] 최동진, 황명권, 김판구, "웹 로그 데이터를 이용한 사용자관심분야 분석 모델 설계", 한국정보기술학회 하계학술대회 논문집, pp.333-336, 2010.

[부 록]

〈표 5〉 연관키워드 정답셋 일부(실험1)

대표어	연관키워드
0.62클라이언트다운	클라이언트0.62다운
070	070인터넷전화
070인터넷전화	070
114전화번호검색	114전화번호안내
114전화번호안내	114전화번호검색
11번가	11번가 쇼핑몰
11번가 쇼핑몰	11번가
…	…
강병규	강병규 CCTV
강병규 CCTV	강병규
강세미	강세미 남편, 강세미 결혼, 소준
강세미 결혼	강세미, 소준
강세미 남편	강세미, 소준
…	…
아사다 마오	아사다마오, 아사다
아사다 일본선수권	아사다마오
아사다마오	아사다, 아사다마오 갈라쇼, 아사다 마오, 아사다마오 눈물, 아사다 일본선수권
아사다마오 갈라쇼	아사다마오, 아사다마오 동영상
아사다마오 눈물	아사다마오
…	…
한국 쓰나미	한국 쓰나미
한국마사회	마사회
한국산업기술대	한국산업기술대
…	…

〈표 6〉 연관키워드 정답셋 일부(실험2)

대표어	연관키워드
0.62클라이언트다운	클라이언트0.62다운
070	070인터넷전화
070인터넷전화	070
114전화번호검색	114전화번호안내
11번가	11번가 쇼핑몰
11번가 쇼핑몰	11번가
…	…
강병규	강병규 CCTV, 서세원, 아이리스
강병규 CCTV	강병규
강성연	강세미
강세미	강세미 남편, 강세미 결혼, 강심장, 강병규, 강성연, 소준
강세미 결혼	강세미, 소준
강세미 남편	강세미, 소준
…	…
아사다 마오	김연아, 아사다, 아사다마오
아사다 일본선수권	김연아, 아사다마오
아사다마오	김연아, 김연아 갈라쇼, 아사다마오 갈라쇼, 아사다 마오, 아사다마오 눈물, 아사다 일본선수권
아사다마오 갈라쇼	김연아 갈라쇼, 아사다마오, 아사다마오 동영상
…	…
한국 쓰나미	한국, 한국쓰나미, 일본 쓰나미
한국관광공사	한국
한국마사회	마사회
한국산업기술대	한국산업기술대
…	…

안 광 모

e-mail : ahnmo@nlp.cbnu.ac.kr
2007년 충북대학교 컴퓨터공학과(공학사)
2009년 충북대학교 컴퓨터공학과 (공학석사)
2009년~현 재 충북대학교 컴퓨터공학과 박사과정
관심분야 : 자연어처리, 기계 학습 등

서 영 훈

e-mail : yhseo@cbnu.ac.kr
1983년 서울대학교 컴퓨터공학과(공학사)
1985년 서울대학교 컴퓨터공학과 (공학석사)
1991년 서울대학교 컴퓨터공학과 (공학박사)
1994년~1995년 Carnegie Mellon 대학 기계번역센터 객원교수
1988년~현 재 충북대학교 컴퓨터공학과 교수
관심분야 : 자연어처리, 한국어 구분분석기, 한영기계번역, 정보검색, 질의응답시스템 등

허 정

e-mail : jeonghur@etri.re.kr
1999년 울산대학교 전자계산학과(공학사)
2001년 울산대학교 전자계산학과 (공학석사)
2001년~현 재 한국전자통신연구원 지식마이닝연구팀 선임연구원
관심분야 : 정보검색, 자연어처리, 정보추출 등

이 충 희

e-mail : forever@etri.re.kr
1996년 한양대학교 전자계산학과(학사)
2001년 연세대학교 컴퓨터과학과(석사)
2001년~현 재 한국전자통신연구원 지식마이닝연구팀 선임연구원
관심분야 : 정보검색, 자연어처리, 질의응답 등

장 명 길

e-mail : mgjang@etri.re.kr
1988년 부산대학교 계산통계학과(이학사)
1990년 부산대학교 계산통계학과 (이학석사)
2002년 충남대학교 컴퓨터과학과 (이학박사)
1990년~1998년 5월 시스템공학연구소 선임연구원
1998년 6월~현 재 한국전자통신연구원 지식마이닝연구팀 책임연구원
관심분야 : 정보검색, 자연어처리, 텍스트마이닝, 시맨틱 웹 및 온톨로지 등