2020
IIBC
CONFERENCE

www.iibc.kr

종합학술대회 논문집

제18권 제1호

일시 | 2020. 11. 12(목)~13(금)

장소 | COEX(삼성동)

주관 및 주최 | (사)한국인터넷방송통신학회(IIBC), (사)국제문화기술진흥원(IPACT), 지식의 숲

후원 | 과학기술정보통신부, 한국연구재단, 한국과학기술단체총연합회

협찬 | ㈜아이티센, ㈜콤텍시스템, 오픈스택㈜, ㈜넥스모어시스템즈

IIBC (사)한국인터넷방송통신학회
The Institute of Internet, Broadcasting and Communication

인덱스 기반의 재귀적 병렬 검색 알고리즘을 활용한 검색 엔진 서비스 모델에 관한 연구

A study on the Search Engine Service Model with Index-Based Recursive Parallel Search Algorithm

이한범*, 조재형, 김영철

Han-Bum Lee* Jae-Hyeoung Cho, R. Young Chul Kim

lkanen16@gmail.com, henrycdoctor@naver.com, bob@hongik.ac.kr

요 약

인터넷 상의 정보가 방대해 짐에 따라 빅 데이터를 다루는 기술이 발전하는 등 데이터 처리 업계는 발전하고 있지만, 검색 엔진 사용자의 입장에서는 원하는 정보의 검색이 점점 더 어려워지고 있다. 일반 사용자 입장 뿐만 아니라, 관련 업계 또한 검색에 소요되는 리소스가 증가하는 것은 큰 손실을 야기할 수 있다. 본 논문에서는 방대한 인터넷 상 정보의 검색에 인덱스 기반의 재귀적 병렬 검색 알고리즘을 도입함으로써 사용자가 원하는 정보를 보다 쉽고 빠르게 얻을 수 있도록 개선하고, 이를 이용한 검색 엔진 프로젝트로 사용자 경험을 향상 시킬 수 있음을 보인다.

키워드 : 인덱싱, 캐싱, 검색엔진, 재귀, 연관어

I. 서 론

본 논문은 2020년 1 · 2학기 홍익대학교 소프트웨어융합학과 종합설계 프로젝트 결과물로써, WWW(World Wide Web) 출시 이래 인터넷 상에 업로드 되는 정보의 양은 기하급수적으로 늘어나고 있다. 이러한 정보의 사용사례는 셀 수 없이 많다. 예를 들어 기업의 경우엔 새로운 아이템을 이용한 신 사업의 검토에 가장 많은 리소스가 소요되는 부분이 바로 정보의 검색이다. 해당 사업 전반의 경쟁 구도, 사업의 운용 전략 구상, 경쟁 기업들의 매출 비중, 연관성 있는 신 기술과 잠재적 이슈는 어떠한가를 확인하는 작업이 모두 자료 검색을 근간으로 하기 때문이다. 그 외에도 특허부터 일반 사용자의 검색까지 정말 다양한 방면에 기하급수적으로 늘어나는 인터넷 상 정보를 직접 혹은 간접적으로 활용한다.

위의 정보 사용 사례들은 사용할 수 있는 정보의 양이 늘어남에 긍정적인 영향을 받는다. 선택의 폭이 넓어지고, 참고할 수 있는 자료가 많아지는건 분명 다행한 일이다. 하지만, 최근 몇 년간의 정보 증가량은 정보를 검색하는 비용을 따져보았을 때, 마냥 긍정적으로 바라보긴 힘들 정도로 기하급수적으로 상승했다. 원하는 정보를 찾기위해 큰 시간과 비용을 들여 반복적으로 검색을 해 보아도, 국가, 언어, 검색 엔진 등 검색 환경에 따라 더 많은 정보가 쏟아지며, 이 뿐 아니라 연관된 단어로도 비슷한 정보가 있을 수 있기 때문에 이때까지의 단순한 정보 검색은 더 이상 가벼히 여길 수 없는 문제로 떠올랐다.

이러한 상황을 타개하고자, 여러가지 검색 엔진을 사용한 자체 알고리즘을 만들고 이를 실제 프로젝트에 도입함으로써, 기업부터 민간 사용자까지 원하는 정보를 보다 쉽고 빠르게 제공하여 최종적으로는 사용자 경험을 크게 향상 시킬 수 있음을 보이고자 한다.

II. 본 문

이 서비스 모델의 핵심 요소는 사용자가 원하는 데이터를 재귀적으로 탐색하여 제공하는 것이므로, 이 탐색에 가장 큰 리소스가 소비된다. 때문에, 이 리소스 부하를 최대한 줄이는 것이 서비스 모델의 네트워크 컨디션과 사용자 경험을 향상 시키는 관건이다.

해당 리소스를 절감하는 핵심 기술로는 인덱싱 및 캐싱을 들 수 있는데, 본 논문에서는 Redis**를 통해 인덱싱과 캐싱

*홍익대학교 컴퓨터정보통신공학과

을 구현한다.

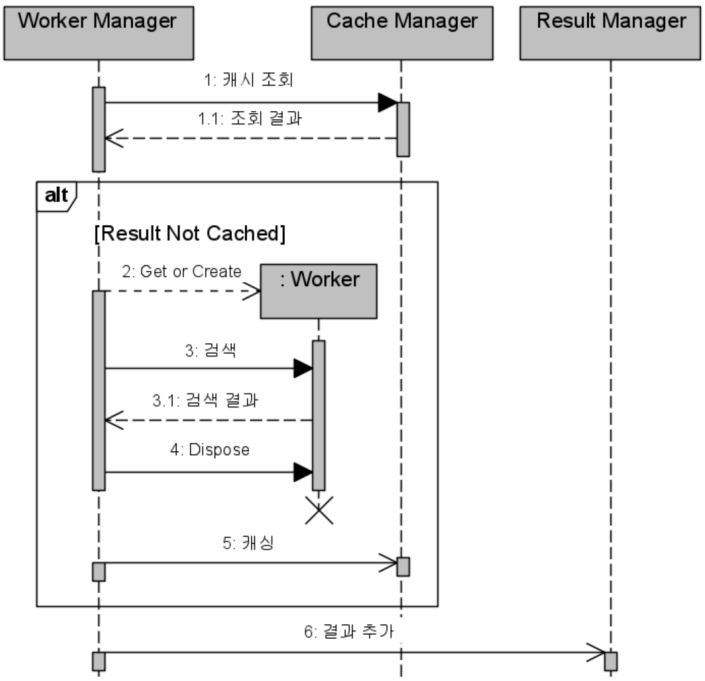

그림 1. Indexed-Cache를 이용한 재귀적 병렬 검색 알고리즘 Sequence Diagram

그림 1은 캐싱이 실제로 알고리즘에서 동작하는 과정을 보여준다. 1절은 실제 쿼리를 실행하기 전, 쿼리 대상이 캐싱되어 있는지를 조회하고, 1.1절은 조회 결과를 리턴해준다. 이때, 조회결과가 true 라면 캐싱된 데이터를 out 키워드를 통해 레퍼런스로 전달해주며, 실제 쿼리를 수행하지 않고 전달받은 데이터를 6절에서 Result Manager에 큐잉하게 된다. 조회 결과가 false라면 2절의 실제 쿼리 수행부분이 진행되며 수행 완료 후, 결과는 쿼리를 수행한 keyword로 인덱싱되어 Redis에 캐싱되게 된다.

그림1의 2, 3, 3.1, 4절은 Worker Thread Pool을 활용한 재귀적 검색 엔진 쿼리 과정을 나타낸다. 조회하려는 데이터가 캐싱되어있지 않을 경우, 2절을 통해 새로운 Worker를 GetOrCreate(Pooling) 하여 풀에 유휴한 Worker가 존재할 경우 해당 Worker를 사용하며, 존재하지 않으면 새로운 Worker를 생성하여 Worker Thread Pool에 등록하게 된다. 생성된 각 Worker는 Worker Manager로부터 정해진 검색 엔진 소스를 통해 주어진 keyword 쿼리를 수행하게 되며, 수행이 완료된 경우 검색 결과를 리턴하고, Dispose를 통해 Pool에 유휴상태로 되돌려지게 된다. WorkerManager는 리턴받은 검색 결과를 Redis에 캐싱하고 Result Manager에 마찬가지로 큐잉하게 된다.

또한, 각 쿼리 수행 결과의 data는 token 별로 분할되어 출현 빈도에 따라 Scoring되고 상위 token부터 연관된 검색어로 판단하여 재귀적으로 위의 단계를 다시 수행하게 된다.

Result Manager는 클라이언트와 SSE 커넥션이 수립될 때, 해당 클라이언트의 Secret(커넥션에 사용될 Key)을 Key로 하는 자체적인 큐를 멤버로 가지고 있으며, 해당 큐가 일정 이상 찬 경우 클라이언트에 결과를 전송하게 되며, 클라이언트가 다른 keyword로 검색을 요청할 경우 기존 커넥션은 terminate 되고, 새로운 Secret을 포함한 SSE 커넥션이 수립되게 된다.

III. 모의 실험

그림 2. Server Side Deploy Diagram

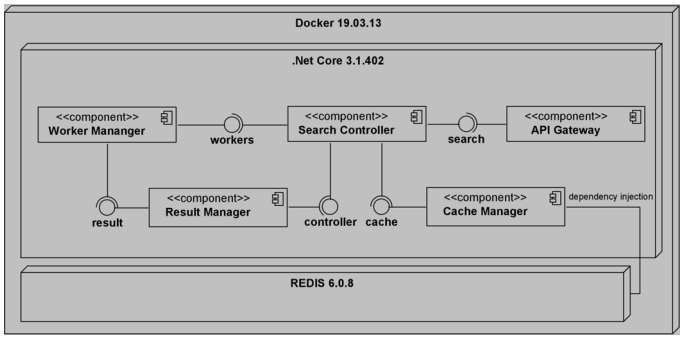

그림 2는 인덱스 기반의 재귀적 병렬 검색 알고리즘이 도입된 서비스 모델의 서버 측 구성도이다. 메인 컨트롤러인 Search Controller가 들어온 검색 요청을 각 서브 시스템들에 전달하고, 각 서브 시스템들은 그림1의 Sequence Diagram에 따라 캐싱, 인덱싱, 재귀 쿼리를 수행한다.


```
"context": { "title": "Google" }, "searchInformation": { "searchTime": 0.628156, "formattedSearchTime":
"0.63", "totalResults": "6810000000", "formattedTotalResults": "6,810,000,000" }, "items": [ { "kind":
"customsearch#result", "title": "Coronavirus Disease 2019 (COVID-19) | CDC", "htmlTitle": "Coronavirus
Disease 2019 (\u003cb\u003eCOVID\u003c/b\u003e-\u003cb\u003e19\u003c/b\u003e) | CDC", "link":
"https://www.cdc.gov/coronavirus/2019-ncov/index.html", "displayLink": "www.cdc.gov", "snippet":
"Coronavirus (COVID-19) Home Page.", "htmlSnippet": "Coronavirus (\u003cb\u003eCOVID\u003c/
b\u003e-\u003cb\u003e19\u003c/b\u003e) Home Page.", "cacheId": "B_FPQc5umtEJ", "formattedUrl":
"https://www.cdc.gov/coronavirus/2019-ncov/index.html", "htmlFormattedUrl":
"https://www.cdc.gov/\u003cb\u003ecoronavirus\u003c/b\u003e/2019-ncov/index.html", "pagemap": {
"cse_thumbnail": [ { "src": "https://encrypted-tbn1.gstatic.com/
images?q=tbn:ANd9GcSOGLQ9sv3zh3YQeIWwxgAQijXvc2GSU6YCrHqimWHVtOgsT_R_u8YidGA", "width": "300", "height":
"168" } ], "organization": [ { "telephone": "Call 800-232-4636", "url": "Email CDC-INFO" }, {
"telephone": "Call 800-232-4636", "url": "Email CDC-INFO" } ], "contactpoint": [ { "url": "U.S.
Department of Health & Human Services" }, { "url": "USA.gov" } ], "metatags": [ { "og:image":
"https://www.cdc.gov/coronavirus/2019-ncov/images/COVID-19-SM-1200px.jpg", "cdc:content_id": "124649",
"article:published_time": "2020-02-11", "twitter:card": "summary_large_image", "og:site_name": "Centers
for Disease Control and Prevention", "cdc:maintained_by": "NATL CENTER FOR IMMUNIZATION AND RESPIRATORY
DISEASES", "og:image:type": "image/jpeg", "og:description": "CDC provides credible COVID-19 health
information to the U.S.", "twitter:creator": "@CDCgov", "cdc:last_published": "2020-10-15T18:32:06Z",
"twitter:site": "@CDCgov", "cdc:last_reviewed": "May 1, 2020", "cdc:content_source": "NATL CENTER FOR
IMMUNIZATION AND RESPIRATORY DISEASES; DIVISION OF VIRAL DISEASES", "og:type": "article",
```


그림 3. 실제 쿼리 결과의 raw 포맷

그림 3은 Worker가 수행하는 실제 쿼리의 raw 결과물이며, 이를 Simplify 후 인덱싱 및 캐싱을 거쳐 아래 그림 4의 포맷으로 클라이언트에게 전송된다.

** Redis란, 일종의 비 관계형 구조 데이터베이스로써 Key-Value 혹은 Document 기반으로 데이터를 저장할 수 있으며, 많은 양의 데이터를 효율적으로 처리할 수 있어서 데이터의 분산 처리 및 빠른 I/O가 필요한 서비스에 주로 채용되는 편이다.

```
"keyWord": "covid-19",
"results": [
    {
        "title": "Coronavirus Disease 2019 (COVID-19) | CDC",
        "snippet": "Coronavirus (COVID-19) Home Page.",
        "link": "https://www.cdc.gov/coronavirus/2019-ncov/index.html",
        "thumbnail": null
    },
    {
        "title": "Coronavirus disease (COVID-19)",
        "snippet": "Information on COVID-19, the infectious disease caused by the most recently
            \ndiscovered coronavirus.",
        "link": "https://www.who.int/emergencies/diseases/novel-coronavirus-2019",
        "thumbnail": null
    },
    {
        "title": "Symptoms of Coronavirus | CDC",
        "snippet": "May 13, 2020 ... COVID-19 is caused by infection with a new coronavirus (called
            SARS-CoV-2) \nand flu is caused by infection with influenza viruses. Because ...",
        "link": "https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html",
        "thumbnail": null
    },
    {
        "title": "COVID-19",
        "snippet": "Everything you need to know to prepare for, and protect yourself from \nCoronavirus
            Disease 2019 (COVID-19)",
        "link": "https://coronavirus.ohio.gov/wps/portal/gov/covid-19/home",
        "thumbnail": null
    },
```


그림 4. 실제 캐싱 및 전송되는 raw 데이터의 simplified 포맷

IV. 결 론

본 논문에서는 인터넷 상의 정보를 탐색하는 비용을 줄여주고자 인덱스 기반의 재귀적 병렬 검색 알고리즘을 활용한 검색 엔진 서비스 모델에 대해서 고찰해 보았다.

사용자는 한 번의 검색만으로도, 원하는 정보를 한 페이지에 모아 전달받을 수 있으며, 정보 검색의 속도는 캐싱 및 인덱싱이 진행됨에 따라 점점 높아지리라 예상된다.

앞으로도 인터넷에 업로드되는 정보의 양은 지속적으로 늘어날 것이며, 이와 비례하여 데이터 기반 사업 또한 다양해질 것이므로, 이 서비스 모델의 잠재적 활용 가치는 점차 높아질 것으로 예상해 볼 수 있다. 앞으로의 연구는 사용되는 검색 엔진의 종류를 보다 다양하게 지원할 예정이며, 이를 통해 사용자가 더욱 풍부한 연관 데이터를 얻을 수 있는 검색 엔진 서비스 모델을 구축할 수 있을 것으로 기대된다.

ACKNOWLEDGEMENT

본 논문은 2019년도 산업통상자원부의 '창의산업융합 특성화 인재양성사업'의 지원을 받아 수행된 기초연구사업임(과제번호 N0000717)

참 고 문 헌

[1] Introduction to Redis, https://redis.io/topics/introduction

[2] Google Custom Search JSON API, https://developers.google.com/custom-search/v1/overview