# 다 학제 간 캡스톤디자인 답변서 요약본

팀명: 23 조 고리고리

팀원: 맹산하,강길웅,정준권,김사라,이정현

| 심사의견 or 질문 |
| --- |
| 검색에 있어서 단순하게 최종 노드에서 루트까지에 나타나는 모든 키워드를 AND 조건으로 사용한다고 생각되는데, 사용자가 일일이 그 트리를 따라가는 대신에 그냥 바로 그 키워드들을 타이핑하여 검색하는 것이 더 빠를 수 있는데, 굳이 트리를 사용하는 이유가 무엇인지? |
| 답변 |
| 키워드들의 조합을 사용하는 것이 아닌 트리 형식을 사용하는 목적은 크게 두가지 입니다.<br>1. 키워드 조합의 재사용<br>2. 분류를 제공하여 한 눈에 알아보기 쉽게 하고 편의 제공<br><br>먼저 키워드 조합의 재사용입니다. 저희는 검색을 일회성으로 제공하는 것이 아닌, 한 번 작성한 키워드 조합에 대해 지속적인 재사용을 제공하려 합니다. 한번 만들어 놓은 키워드의 조합을 다시 동일하게 이용하려 할 경우 다시 입력하는 것이 아닌, 만들어 놓은 트리 형태의 '도감' 내부에서 키워드를 클릭만 하면 바로 검색을 제공합니다. 초기 '도감'을 제작할 때에는 키워드 조합에 비해 시간이 더 많이 걸리지만 한번 제작한 뒤에는 동일 키워드 검색에 대해 우월한 편의성과 속도를 제공 할 수 있습니다. 단순 일회성의 검색에 초점을 맞춘 것이 아닌, 사용자가 관심있는 키워드에 대한 지속적 검색에 초점을 맞춘 것이기에 트리 형태를 이용하게 되었습니다.<br>그런데, 해당 답변만으로는 트리 구조 사용에 대한 충분한 답변은 아닐 것입니다. 키워드의 조합을 저장해놓고 다시 이용하면 되는 방법이 있기 때문입니다. 두번째 이유가 존재합니다.<br><br>분류를 제공하기 위해 트리 형태를 이용합니다. 만일 키워드들의 조합을 저장해 놓고 이용한다면 지속적인 편리함 제공은 가능할 것입니다. 하지만 키워드의 조합이 많아진다면 시각적으로 알아보기 어려워지고 난잡해지기 때문에 찾기 어려워집니다. 다량의 키워드 조합을 분류하기 위해 결국은 파일 시스템의 구조와 같은 분류 모델이 필요한데 저희는 이것을 트리 구조의 '도감'으로 선택했습니다.<br>트리 구조의 제공으로 키워드 조합에 대한 분류를 통해 시각적으로 이해하기 쉽고 알아보기 편하도록 했습니다. |

| 심사의견 or 질문 |
| --- |
| 도감의 수가 많아질 가능성이 있는데, 도감 자체를 검색하는 기능이 있는지? |
| 답변 |
| 크게 두가지 검색을 제공합니다.<br>1. 내부 도감 검색<br>2. 공유된 도감 검색 |

모바일 내부 사용자가 작성한 도감들에 대하여 제목을 가지고 검색할  수 있는 기능을 가지고 있습니다.
공유된 도감의 경우 도감의 작성자, 도감의 키워드를 가지고 도감을 검색 할 수 있습니다.

심사의견 or 질문

검색시에 상위 계층과 하위계층을 동시에 수행하고 AND 를 수행하는 것보다는 상위계층 검색을 먼저하고, 그 결과 내에서 하위계층을 검색하는 것이 성능이 더 낫지 않을까 생각됨.

답변

질문 해주신 내용이 맞습니다. 현재는 모든 검색 결과에 대해 AND 연산을 수행하고 검색 결과를 얻어냅니다. 상위 계층의 결과를 먼저 걸러 내고 하위 계층에 대해 AND 연산을 수행한다면 성능이 더 나아질 것으로 추측됩니다. 하지만 현재 성능상의 이슈에서 AND 연산에서 사용되는 시간보다 인스타그램에서의  페이지 로드와 HTTP GET 응답 시간에 대한 것이 더 크게 작용하고 있습니다.
또한 AND 연산의 개선으로 나아지는 성능상의 개선이 현재 프로젝트 진행 시간 대비 효율적이라고 말하기 어렵습니다. 인스타그램과의 통신 및 크롤링에 대한 부분을 먼저 개선하고 차후에 개선하도록 하겠습니다.

심사의견 or 질문

사진 검색 시에 실시간으로 인스타그램의 사진들을 크롤링 한다고 하였는데, #유산슬 을 따로 검색한 결과와 #중화요리 를 따로 검색한 결과를 join 하는 과정이 얼마나 잘 되는지가 의문입니다.실제로 인스타그램에 저 두 해시태그를 검색해봐도 공통되는 사진이 몇 장 없습니다. 다만, 실제로 구현해봤을 때 정말 잘 나온다면 크게 문제될 사항은 아니겠으나, 여전히 성능적인 문제가 존재할 것 같다는 생각이 듭니다.

답변

해시태그의 유의어들을 추가로 입력받아 크롤링하게 하여 결과물의 양을 늘렸습니다. 예를 들어, 질문 주신 '#중화요리'의 경우 '#중화요리' 라고 달리지 않고 '#중국집', '#중국음식', '#중식' 과 같은 형태로 달려있을 수도 있습니다. 태그에 대해 같이 달리는 경우가 많은 태그, 유사한 의미의 태그들을 같이 검색하도록 하여 검색 결과의 양을 늘렸습니다.

Join 하는 과정에 대한 AND 연산의 성능적 문제는 연산의 복잡도가 높은 것이 맞습니다.
그래서 join 의 복잡도를 낮추기 위해 병렬 컴퓨팅 성능이 좋은 Go 언어를 선택했고, 자료구조를 변경하여 최적화했습니다. 2 계층에 M 개, 3 계층에 N 개의 게시물이 있다면 join 의 복잡도는 O(MN)입니다.
이 때, 우선 집합 자료구조에 게시물을 저장하도록 하여 복잡도를 O(M)으로 낮추고, M 개의 고루틴을 생성하여 병렬 처리하여 시간 복잡도를 O(1)로 개선하였습니다. 또한 고루틴 간 통신은 채널을 이용하여 별도의 lock 없이 race condition 을 방지했습니다.이는 마치 M 개의 스레드를 생성하여 공간 복잡도가 O(M)이 된 것 같지만, 그렇지 않습니다. 고루틴은 스레드보다 더 저렴합니다.
고루틴을 여러 개 만든다고 해서 스레드가 그만큼 많이 만들어지는 것은 아니며, 스레드는 그것보다 훨씬 적게 만들어지고 여러 개의 고루틴들이 하나의 스레드에 대응됩니다.

한 개의 고루틴이 차지하는 용량은 약 2KB 이며, 이는 OS 에서 제공하는 경량 스레드(약 1MB)보다도 훨씬 저렴합니다.

현재 1 번의 크롤링 시 각 검색어마다 200~250 개씩, 한 계층에 5 개의 검색어가 포함되므로 보통 계층당 1000~1200 여개의 게시물이 존재합니다.

이 점을 고려하면 join 연산의 복잡도를 병렬 처리로 해결한다고 해서 System overhead 가 발생하는 일은 없으며, 이를 테스트를 통해 검증했습니다.

| 심사의견 or 질문 |
| --- |
| 키워드를 tree 구조로 만드는 특별한 이유가 있을까요? 제안 내용이 해결하고자 하는 문제는 '인스타그램에서는 여러 키워드를 한꺼번에 검색하지 못한다' 인 것 같습니다. tree 구조를 따를 필요 없이 곧바로 '여러 해시 태그 검색 가능' 하게 문제를 해결할 수 있지 않을까요? 도감도 비슷하게, 키워드의 묶음으로 표현해도 괜찮지 않을까 싶습니다. 만일 tree 구조를 사용한 특별한 이유가 있다면 조금만 더 이 부분이 드러나게 보고서 등을 작성해주면 좋겠습니다. |
| 답변 |
| 트리 구조의 이용 이유는 첫번째 질문 답변에 자세히 적었습니다. 참고 바랍니다. 보고서를 작성할 때 해당 내용을 반영하여 작성 하겠습니다. |