

LiveTwitter: 트위터 기반 핫이슈 검색 시스템

성병기⁰, 오진영, 차정원

창원대학교 컴퓨터공학과

fghost@changwon.ac.kr, psyche.oyj@gmail.com, jcha@changwon.ac.kr

LiveTwitter: Hot Issue Search system Based on Twitter

Byung-ki Sung⁰, Jin-Young Oh, Jeong-Won Cha

Dept. of Computer Engineering, Changwon National University

요약

트위터, 페이스북 등의 소셜 네트워크가 이슈가 되는 사건에 의견을 표시하는 수단으로 많이 활용되고 있다. 본 논문에서는 이슈 키워드 추출 및 트위터와 유튜브에 기반한 실시간 검색 시스템을 구현한다. 본 시스템에서는 가장 최근 신문 기사들의 제목과 스니펫을 이용하여 이슈가 되는 키워드를 실시간으로 추출하여 사용자들에게 보여주고 트위터와 유튜브 OpenAPI를 이용하여 추출된 키워드에 대한 컨텐츠들을 실시간으로 사용자들에게 보여준다. 본 시스템을 통해서 이슈가 되는 사건에 대한 실시간 반응을 찾을 수 있다.

주제어 실시간 검색 시스템, 트위터, 유튜브

1. 서론

최근 인터넷과 정보 매체의 발전에 인해 많은 정보와 다양한 컨텐츠가 쏟아지고 있다. 그로 인해 이전에는 적은 정보 속에서 원하는 데이터를 사용자가 직접 찾아 필요한 데이터를 모았으나 현재는 방대한 데이터 속에서 원치 않는 데이터와 필요 데이터를 걸러내는 것이 중요한 요구가 되었다.

이러한 요구중 하나가 현재 이슈가 되는 정보를 제공해 주는 것이다. 최신정보를 통해서 빨 빠르게 여러 사람의 공감대를 이해하고 참여하고자 하는 욕구가 증가한 것이다. 그만큼 실시간 정보의 중요성 또한 증가했다.

현재, 실시간 정보를 이용한 검색 서비스의 성장이 많이 이루어졌다. Crowdeye, Google 실시간 검색, Naver 실시간 검색, Friendfeed 그리고 지금은 서비스화가 중지된 Live-K 등이 실시간 정보를 이용하고 있다.

Live-K[1]는 국내에서는 최초로 시도한 실시간 검색 서비스이다. 트위터, 미투데이, 신문 기사, 블로그 등에서 컨텐츠를 가져와서 실시간으로 자동 업데이트하여 정보를 표시한다. 그리고 시스템 구조에서 데이터 수집 간격이 0~10초 사이이고 데이터베이스를 사용하지 않는 휘발성 서비스라는 특징을 가지고 있다.

Crowdeye[2]는 구글보다 먼저 트위터 실시간 검색 서비스를 내놓았다. 최신 트위터 컨텐츠를 검색 가능하며 어떤 주제에 대해서 가장 인기 있는 링크 목록 또한 제공하고, 타임라인뷰를 통해 시간대별로 결과 검색이 가능하다.

Google 실시간 검색[3]은 트위터, 페이스북, 마이스페이스 등에서 항상 최신 것으로 검색해주고 새로운 데이터가 올라오면 자동으로 새로고침되면서 업데이트가 된다. 하지만 결과의 대부분은 트위터에서 온 것이고 일부는 google buzz의 결과가 나타난다. 그리고 페이스북과

마이스페이스에 대해서는 모든 내용을 검색 결과로 보여주는 것이 아니라 공개페이지로 된 부분만 검색할 수 있도록 되어있다.

Naver 실시간 검색[4]은 키워드를 입력하면 네이버 뉴스, 블로그, 카페, 미투데이, 트위터 지식인과 같은 곳에서 생성된 문서를 5초 단위로 수집하여 스트리밍 방식의 검색 결과로 제공한다.

Friendfeed[5]는 트위터와 인터페이스가 비슷하면서도 훨씬 다양한 웹서비스의 컨텐츠를 다루는 기능을 제공한다. 자신의 트위터와 미투데이, 모든 블로그, 음악스트리밍서비스 등이 프렌드피드에 자동적으로 모아진다. 관심 있는 상대방을 등록하면 그 사람의 컨텐츠를 확인 가능하다. Friendfeed에서 등록된 사람들의 정보를 들고 온다는 점에서 조금의 차이가 있지만 실시간 검색이라는 주제를 놓고 봤을 때는 유사하다.

위에서 설명한 서비스들의 공통적인 특징은 소셜 네트워크 서비스(이하 SNS)를 이용하거나 이슈 키워드 추출 시 사용자의 검색한 질의 횟수가 반영된 것이 대부분이다. 본문 내용을 분석하여 이슈 키워드를 추출한다는 것은 띄어쓰기, 오타와 같은 문제로 어려움이 많다.

본 시스템에서는 이런 문제점을 보안하기 위해서 신문 기사 제목을 이용하였다. 현재 이슈가 되는 것은 신문기사에 반영이 되기 때문이다. 또는 반대로 신문 기사를 통하여 이후에 이슈가 될 수 있는 키워드를 미리 추출 가능하다. 신문 기사 제목은 SNS에 비해 정제된 문서이고 완전한 문장 구조를 가지기 보다는 주요 단어들의 나열로 이루어진 경우가 많으므로 많은 전처리 단계의 비중이 줄어든다.

또한 본 시스템에서는 실시간으로 최신 컨텐츠를 보여주고 있다. 이전까지 대부분의 포털이나 검색시스템을 보면 검색시점 이후에 업데이트된 컨텐츠에 대해서는 고

려치 않는다. 보통의 이슈 키워드 콘텐츠의 양은 검색 시점보다 뒤늦게 업데이트되는 것이 많기 때문에 위와 같은 경우 가장 최신 정보를 알고 싶어 하는 사용자의 욕구를 충족시키지 못하는 문제점이 존재한다. 이를 위해 Live-K와 같이 실시간으로 업데이트된 콘텐츠를 화면에 표시하도록 구현하였다. 사용자는 검색결과 화면을 통해 추가적인 검색 없이도 최신정보를 얻게 되는 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 구현된 시스템에 대한 구조 및 주요 기능을 설명한다. 마지막으로 3장에서 본 개발 도구에 대한 결론을 내린다.

2. LiveTwitter

지금까지의 검색 사이트들은 가장 정확한 정보를 제공해 주기 위해 노력해왔다. 이런 시스템은 변하지 않는 지식을 찾는 곳에서는 유용하다. 그러나 시간에 민감한 정보에 대해서는 대처하지 못한다.

이러한 문제점을 해결하기 위해서 현재 올라오는 기사들을 이용해서 실시간으로 이슈 키워드를 추출하여 현재 시간의 이슈 키워드를 특정 알고리즘에 의해 순위를 계산하여 사용자들에게 이슈 키워드 순위와 콘텐츠를 트위터와 유튜브 API를 이용해서 실시간으로 보여준다. 그럼 1.은 전체적인 시스템 구조도이다.

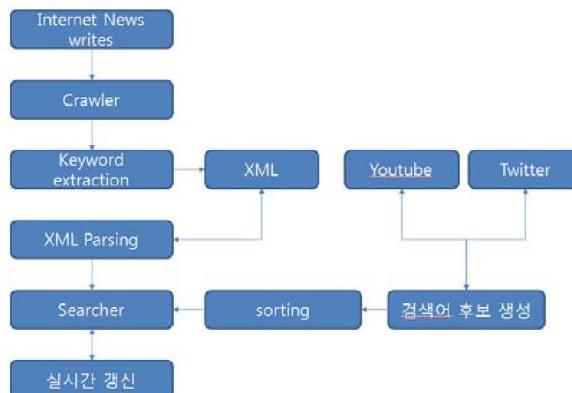


그림 1. 실시간 검색 시스템의 전체적인 구조도

2.1 Crawler

인터넷에서 서비스 중인 네이버, 다음, 네이트 포탈 사이트의 최신 기사들의 시간과 제목 그리고 스니펫을 지정된 시간마다 수집하여 그 시간 순으로 저장한다.

수집한 간격에 따라 이슈키워드에 영향을 주므로, 이것 또한 실험을 통하여 정하였다.

2.2 핫이슈 추출

현재 이슈 키워드는 단어의 빈도수를 이용하여 추출한다. 일정 시간동안 작성된 최신 신문 기사의 제목과 스

니펫을 수집하여 문서의 빈도수를 측정한다. 단, 여러 개의 포털을 적용하기 때문에 같은 신문기사가 여러 포털에 등록될 가능성과 한 포털에 같은 제목이 등록될 가능성이 있다. 현재 이슈 키워드 추출은 같은 포털에서 등장한 같은 제목은 의미가 없다고 생각하여 삭제하였고, 다른 포털에 등록된 경우에는 가중치를 추가한다. 즉, 2개의 포털에 중복된 제목이 등록된 경우 가중치는 2가 된다.

예를 들어 ‘강제구 과장, 유럽안면성형학회 강의’의 제목일 경우 문서의 중복 제목을 모두 삭제 후, ‘강제구’, ‘과장’, ‘유럽안면성형학회’, ‘강의’ 단어 빈도수를 모두 추출한다. 제목의 스코어 값은 각 단어의 빈도수를 모두 더하고 포털 가중치를 곱한 값을 제목을 이루는 단어 수로 나눈 값이다. 이것을 수식으로 표현하면 (1)과 같다.

$$score = \frac{w_s \cdot \sum_{i=1}^n s_i}{n} \quad (1)$$

신문제목 s 는 단어 $\{s_1, s_2, \dots, s_n\}$ 로 이루어 졌고, 제목의 포털 가중치는 w_s 이다.

이렇게 계산되어진 스코어 상위 10개를 이슈 키워드로 추출하여 순위를 정하고, XML 형식으로 저장하면 그 데이터를 계속 갱신하면서 화면에 리스트 형태로 보여준다.

현재 키워드 추출의 문제는 중요 키워드를 포함한 제목을 상위 스코어로 추출하지만 같은 의미의 제목을 그룹화 하는 것이 필요하다. 또한 긴 제목일 경우 이중에서 불필요한 수식어구와 같은 단어는 삭제하는 것이 필요하다.

2.3 검색 시스템

1초마다 트위터, 유튜브 API를 사용해서 JSON 형식으로 받아와서 keyword extraction에 의해 추출된 이슈 키워드 목록에 대한 결과를 보여줄 뿐만 아니라 특정 검색어에 대한 결과 중 필요한 부분만 저장하여 몇 가지 처리과정을 거친 후 실시간으로 자동 업데이트된 정보를 사용자들에게 보여준다.

2.3.1 새로운 검색어 리스트 생성 및 중복 제거

트위터는 전세계 사람들을 대상으로 서비스를 하기 때문에 각 나라마다 그 언어에 따라 검색어에 대해서 다르게 인덱싱을 한다. 영어의 경우에는 어절단위로 검색을 하기 때문에 문제점이 없다. 하지만 한글의 경우 3글자 이상 입력시에는 제대로 검색이 되지 않는다. 그 이유는 트위터에서 한글은 두 글자씩 인덱싱을 하기 때문이다. 이러한 문제를 해결하기 위해서 트위터의 검색어 인덱싱 법을 따라야 한다. 간단한 예를 들면 ‘아이폰’이라고 검색했을 경우 검색 결과로 나오지 않는 데이터가 훨씬 많다. 이 문제를 해결하기 위해서는 ‘아이,이폰’과 같이 두

글자씩 나눠서 콤마를 이용해 새로운 검색어를 만들어줘야만 이 문제를 해결할 수 있다. 영어와 한글이 둘 다 포함된 부분 또한 영어는 어절단위 한글은 위와 같은 방법으로 두 방법을 조합하여 새로운 검색어를 만들어줌으로서 문제점을 해결하였다.

또 다른 문제점은 사용자들이 한어절로 된 단순한 키워드를 통한 검색을 하는 경우도 있고, 여러 어절인 구의 형태로 검색하는 경우도 있다. 구 형태의 검색어를 검색하는 경우 트위터의 한글검색의 경우 모든 단어가 나타난 글을 검색하기 때문에 검색어 즉 키워드에 대한 결과가 추출되지 않을 가능성이 크다. 유튜브 또한 트위터와 마찬가지이기 때문에 이러한 문제점을 해결하기 위해서 구 형태의 검색어를 검색 후 나오는 결과가 없거나 일정 수 이하 일 경우에 어절 단위로 잘라내어서 다시 검색어를 만들어주는 작업을 통해서 위와 같은 문제점을 해결하였다.

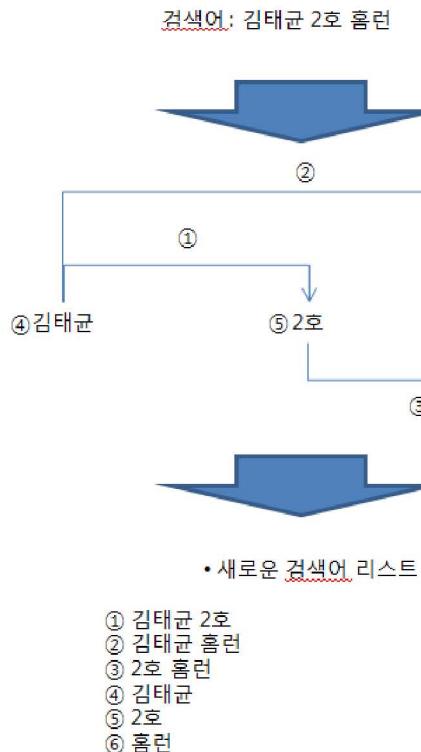


그림 2. ‘김태균 2호 홈런’이라는 키워드를 검색했을 때 추가적인 검색어 목록 생성하는 과정

위의 그림 2. 와 같이 ‘김태균 2호 홈런’으로 검색했을 경우 6 가지의 새로운 검색어를 생성하는데 순서대로 다시 검색을 실시하여 그 결과를 리스트에 저장하는데 이 과정에서 비록 서로 검색어는 다르지만 같은 결과가 나올 수 있기 때문에 결과마다 고유의 아이디를 이용해서 중복된 결과는 제거하고 리스트에 저장한다. 그리고 저장된 결과가 일정수가 넘어가면 나머지 검색어들은 검

색하지 않고 다음 과정으로 넘어간다.

2.3.2 시간순으로 보여주기

검색된 결과를 시간순으로 정렬하는 과정이다. 트위터와 유튜브 API를 이용해서 JSON형태의 결과를 받아왔을 때 단순히 결과를 보여준다면 받아들이는 사용자에게는 데이터의 가치가 떨어진다. 각각의 API를 통해서 최신 순으로 받아왔다고 했을지라도 두 리스트 간의 시간차가 있기 때문에 최종적으로 두 리스트를 합친 후에는 시간 순으로 정렬을 해야 될 필요가 있다. 하지만 여기에서도 문제가 있다. 그것은 트위터와 유튜브 내부에서 서로 시간을 표시하는 형식이 다르기 때문에 무작정 비교가 불가능하다. 결국 서로의 시간을 동일하게 만들어준 다음에 서로의 시간을 비교해서 정렬을 해야만 한다. 그래서 이 과정에서 javascript 의 내부적으로 시간이라고 인식할 수 있는 포맷으로 시간을 바꿔주는데 트위터에서 사용하는 시간은 javascript 내부적으로 인식이 가능한 시간이기 때문에 유튜브의 시간포맷을 트위터의 시간포맷으로 바꿔주는 작업을 거친 뒤 두 리스트를 하나의 리스트로 합쳐서 최신 시간 순으로 정렬 후 아래의 그림 3과 같이 보여준다.

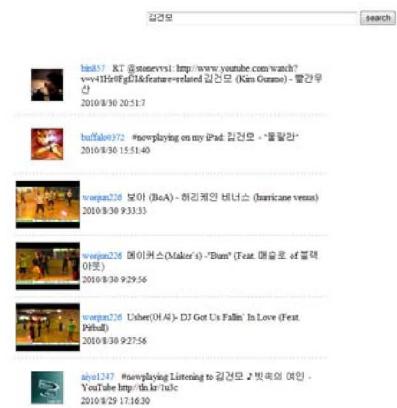


그림 3 검색어에 대해 시간순으로 정렬되어 보여지는 화면

3. 결론

SNS의 발전으로 실시간 검색이 이슈화 되면서 여러 포털에서 SNS를 이용한 실시간 검색을 서비스화 하기 시작했다. 하지만 아직은 단순한 키워드 검색만이 제공될 뿐이고, 본문 내용을 분석하여 이슈 키워드를 추출한다는 것 또한 띄어쓰기, 오타와 같은 문법상의 문제로 어려움이 많기 때문에 현재의 실시간 검색은 기존 포털 시스템에서 부가적인 검색 기능에 불과하다.

본 시스템에서는 이런 문제점을 보안하기 위해서 SNS의 데이터가 아닌 신문 기사 제목을 이용하였다. 현재 이슈가 되는 것은 그 시간의 신문기사에 반영이 되기 때문이다. 반대로 신문 기사를 통하여 이슈가 될 수 있는

키워드를 미리 추출이 가능하다. 신문 기사의 제목은 SNS의 데이터에 비해서 띠어쓰기나 오타와 같은 문법상의 오류가 없고 잘 정제된 문서이다. 그리고 신문 기사의 제목은 신문 기사 내용의 주요 단어들의 나열로 이루어진 경우가 많으므로 전처리 단계의 비중이 줄어든다.

또한 다른 포털들과 마찬가지로 본 시스템에서는 실시간으로 최신 콘텐츠를 보여주고 있다. 이전까지 대부분의 포털이나 검색시스템을 보면 검색시점 이후에 업데이트된 콘텐츠에 대해서는 고려치 않는다. 또한 실시간 검색이라는 웹들도 검색시점과 가까운 시점의 결과를 보여준다는 의미로 사용되는 곳이 많다. 하지만 본 시스템은 Live-K와 같이 실시간으로 자동 업데이트된 콘텐츠를 화면에 표시하도록 구현하였다. 사용자는 검색결과 화면을 통해 추가적인 검색없이도 최신정보를 얻을 수 있기 때문에 현재 서비스 중인 웹들과 차이가 있다고 할 수 있다.

마지막으로 시간이 지날수록 SNS의 이용자들 또한 증가할 것이고 실시간으로 생성되는 정보의 양 또한 증가할 것이다. 그러면 자연적으로 검색에 대해 보여지는 결과물 또한 더 많아질 것이다. 현재 시스템에서 더 많은 곳에서부터 데이터를 가져오고, 올라오는 결과물들에 대해서 조금 전처리를 거쳐 불필요한 정보들을 잡아준다면 검색 엔진의 활용도는 지금 보다 훨씬 커질 것이다.

참고 문헌

- [1] <http://www.google.co.kr/>
- [2] <http://www.friendfeed.com/>
- [3] <http://www.naver.com/>
- [4] <http://www.Live-k.com/>
- [5] <http://www.crowdeye.com/>