익명 사용자의 데이터를 포함하는 통합 오피니언 마이닝 시스템

김이준[1] , 윤재열[2] , 임지연[3] , 김응모[4]
성균관대학교 데이터베이스연구실
[1]uk3080789 , [2]vntlffl , [3]01039374479@naver.com
[4]umkim@ece.skku.ac.kr

Integration System of Opinion Mining with Anonymous Data

Iee-Joon Kim , Jae-Yeol Yoon, Ji-Yeon Lim , Ung-Mo Kim
The Laboratory of Database in Sungkyunkwan University

요 약

지금 이 시대를 살아가는 현대인들은 엄청나게 방대한 양의 디지털 정보 속에서 살아가고 있다. 하지만 사람들은 이런 자료들에 노출되어 있다는 것을 망각하고, 많은 유용한 정보들을 자기의 것으로 만들지 못하고 스쳐 지나가고 있다. 가장 큰 문제는 그 정보를 우리는 믿을 수 있는가 이다. 그래서 본 논문에서는 어떠한 정보가 유용하고 필요한 정보인지 고를 수 있게 도울 수 있는 통합 오피니언 마이닝 시스템 흐름도를 제시하고자 한다. 또한 익명의 사용자들이 만들어내는 의견도 포함하여 자료의 풍부함을 도모한다.

1. 서론

우리는 빠르게 변화하는 세상 속에서 살고 있다. 그 결과로, 엄청나게 많은 데이터들이 지금 이 순간에도 생성되고 있고, 혹은 사라지고 있다. 이렇게 많은 데이터 중에서 내가 필요로 하고 찾고자 하는 자료만을 걸러내는 작업만 한다고 해도, 거기에 들어가는 개개인의 시간과 에너지 낭비는 결국 국가적 손실이다. 이러한 상황에 가장 중요한 것은, 얼마나 빠르게 자료들을 다른 사람들보다 먼저 얻느냐 일 것이다.

오피니언 마이닝은 이미 여러 방향으로 연구가 진행되었지만, 이를 활용하여 적용할 분야 또한 여전히 많다. 본 논문에서도 오피니언 마이닝을 활용한 시스템 흐름도를 통해 유용한 정보를 얻을 수 있는 방법을 제안하고자 한다. 일반적으로 사용자가 포털사이트에 찾고자하는 키워드를 입력하면 자료들은 단어의 정확도가 높은 순으로 빠르게 나열되어 결과를 보여주게 된다. 이는 사용자가 원하는 결과에 대한 정확도와는 일치한다고 할 수 없다. 그렇기 때문에 사용자들은 그 속에서 한번더 유용한 정보를 위한 탐색을 해야만 한다.

본 논문에서 우리의 시스템 흐름도를 통한 자료는 이런 추가적인 시간의 사용을 줄여줄 수 있게끔 디자인되었다. 더 나아가 결과물에 익명의 사용자들이 만들어낸 데이터들도 분석을 하여 결과가 더욱 풍성해 지도록 하였다. 예시 결과물은 긍정과 부정의 의견들을 분류하여 순위를 나타내고, 이를 통해 자세한 설명을 사용자가 얻을 수 있게 만들어졌다.

관련 연구인 오피니언 마이닝, LIWC, 필터링 기술들이 2장에서 간략히 설명되고, 제시하는 시스템 흐름도에 대한 설명이 3장에서 이어진다. 마지막으로 4장에서는 결론으로 본 논문이 마무리 된다.

2. 관련 연구

2.1 오피니언 마이닝 (Opinion Mining)

오피니언 마이닝은 요즘 활발하게 텍스트 마이닝 분야에서 연구되고 있는 기술이다. 웹 서버 데이터베이스에 이미 저장되어있는 텍스트를 추출하여 그 중 의견이 들어있는 단어들만을 추려내어 의견을 판단하는 방법이다. 예를 들어 주로 이용되는 분야로는 콘텐츠에 대한 댓글, 영화 감상평, 자유로이 작성된 개인의 의견이나 생각을 적은 글 등이 있다. 이렇게 분석된 자료에서 우리는 찾고자했던 키워드에 대한 의미있는 의견들을 얻을 수 있다[1].

오피니언 마이닝은 크게 감성적 분리(Sentiment Classification)와 특징 기반 오피니언 마이닝(Feature-based Opinion Mining and Summarization) 으로 분리된다. 먼저 감성적 분리는 말 그대로 글속에 포함된 오피니언들을 이용하여 단편적으로 긍정적인지 혹은 부정적인지를 구분하는 이분법적 방식이다. 이 방식은 의견을 빠르게 긍정인지 부정인지를 판단하는 기준을 마련해주기 때문에

빠른 결정과 선택이 필요한 사용자에게 유용한 정보를 제공해 줄 수 있다는 장점을 가지고 있으나, 의도적으로 작성자가 해당 키워드에 대한 감정을 역설적으로 표현한 것에 대한 분석은 정확하지 않을 가능성을 가지고 있다[2][3].

특징 기반 오피니언 마이닝은 텍스트 속에 속해있는 특정 특징들을 통하여 글의 전반적인 의견을 도출해내는 방법이다[4]. 보통 감성적 분리만 가지고 글의 성향을 판단하지는 않는다. 자세하고 정확한 의미를 위해서는 글속에 들어있는 특징들을 분석할 필요가 있다. 마지막으로 이런 자료들을 요약하여 글 전체적인 분위기를 알 수 있게 된다.

2.2 LIWC (Linguistic Inquiry and Word Count)

LIWC는 글쓰기와 심리학에 관한 연구에서 나온 프로그램으로써[5], 전체적인 글에 포함된 단어들을 심리학자들이 만든 테이블을 통하여 카운트하게 되고, 그 결과 글쓴이의 나이, 직업, 성별, 성향, 작성 패턴, 정신상태, 건강상태 등의 정보를 유추해 낼 수 있다[7][8]. 더 나아가 글쓴이가 글에서 거짓된 태도를 보이고 있는지도 확인 할 수 있다. **[그림 1]**은 LIWC 분석 결과의 한 예시이다[9].

[그림 1] LIWC 결과

LIWC Results

Details of Writer: No Details
Date/Time: [redacted]

LIWC dimension	Your data	Personal texts	Formal texts
Self-references (I, me, my)	3.67	11.4	4.2
Social words	5.27	9.5	8.0
Positive emotions	2.38	2.7	2.6
Negative emotions	1.27	2.6	1.6
Overall cognitive words	5.28	7.8	5.4
Articles (a, an, the)	7.71	5.0	7.2
Big words (> 6 letters)	21.40	13.1	19.6

The text you submitted was 15440 words in length.

2.3 FCBF (Handling Redundant features)

우리는 종종 검색엔진을 통한 검색 결과 속에 데이터가 중복되어 나오는 경우를 접할 때가 있다. 또한 스팸(Spam) 정보[10] 들도 그 속에 포함되곤 한다. 이러한 결과는 필요한 정보를 찾는데 있어서 장애물로 다가온다. 이러한 일이 생기는 이유는 기술적으로 검색엔진들은 키워드 중심으로 연관된 정보, 웹 사이트 전부를 사용자에게 출력해주기 때문이다. 이런 데이터 쓰나미 속에서 중복 혹은 스팸을 처리하는 일이 점점 현대사회에서 중요한 요소로 부각되고 있다.

FCBF(Fast Correlation Based Filter)는 방대한 양의 중복 데이터를 처리할 수 있다[11]. 필터는 크게 두 가지 부분으로 나누어지는데, 처음 부분은 각각의 특징의 가치를 계산하고 그 가치에 따라서 정렬을 한다. 그리고 두 번째 부분에서 연관된 특징만 남기고 나머지 중복된 특징들을 모두 제거해 버린다. FCBF 알고리즘은 **[그림 2]**에서 자세히 살펴볼 수 있다.

[그림 2] FCBF 알고리즘

```
input:  S(F_1, F_2, ..., F_N, C) // a training data set
        δ                        // a predefined threshold
output: S_best                   // an optimal subset

1   begin
2     for i = 1 to N do begin
3       calculate SU_{i,c} for F_i;
4       if (SU_{i,c} ≥ δ)
5         append F_i to S'_list;
6     end;
7     order S'_list in descending SU_{i,c} value;
8     F_p = getFirstElement(S'_list);
9     do begin
10      F_q = getNextElement(S'_list, F_p);
11      if (F_q <> NULL)
12        do begin
13          F'_q = F_q;
14          if (SU_{p,q} ≥ SU_{q,c})
15            remove F_q from S'_list;
16            F_q = getNextElement(S'_list, F'_q);
17          else F_q = getNextElement(S'_list, F_q);
18        end until (F_q == NULL);
19      F_p = getNextElement(S'_list, F_p);
20    end until (F_p == NULL);
21  S_best = S'_list;
22  end;
```

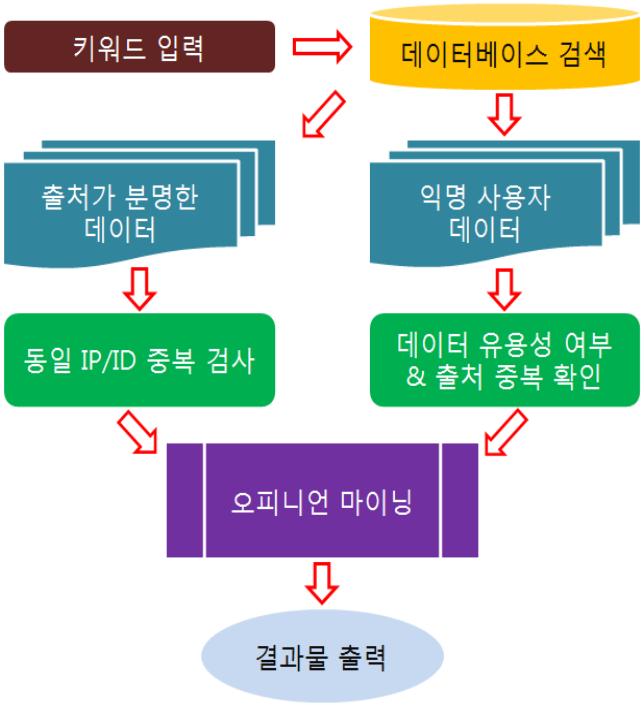
본 논문에서는 이 필터링 기술을 첫 번째로는 중복된 데이터를 제거하는데 사용하고[12], 두 번째로는 익명의 사용자가 작성한 정보가 유용한지 판단하는데 사용한다. 그리고 익명 사용자의 데이터가 같은 사람으로부터 작성된 것인지를 판별하게 된다[13].

3. 제안하는 서비스 흐름도

제안하는 서비스 흐름도 에서는 보통의 검색엔진과 동일하게 사용자가 찾고자하는 키워드를 입력한다. 그러면 데이터베이스에서 어떤 웹 사이트들이 키워드와 연관되어 있는지 검색하게 된다. 검색이 완료되고 나면 우선적으로 중복 검사를 거치게 된다. 첫 번째 필터링 과정에서는 동일한 IP 주소나 작성자ID에서 생성된 데이터들을 제거하게 된다. 이렇게 정리된 자료들은 오피니언 마이닝 프로그램에서 키워드에 관한 오피니언들을 추출되고 요약되어지는 과정을 거치게 된다. 여기서 중요한 부분은 출처의 정보를 알 수 없는 익명의 정보들은 따로 LIWC를 통하여

가용성을 따진 뒤에 오피니언 마이닝 시스템에 적용된다. 따라서 출처가 분명한 정보들과 익명의 정보가 일괄 처리되어 사용자에게 전달된다. 아래 **[그림 3]**에서 서비스 흐름도를 나타낸다.

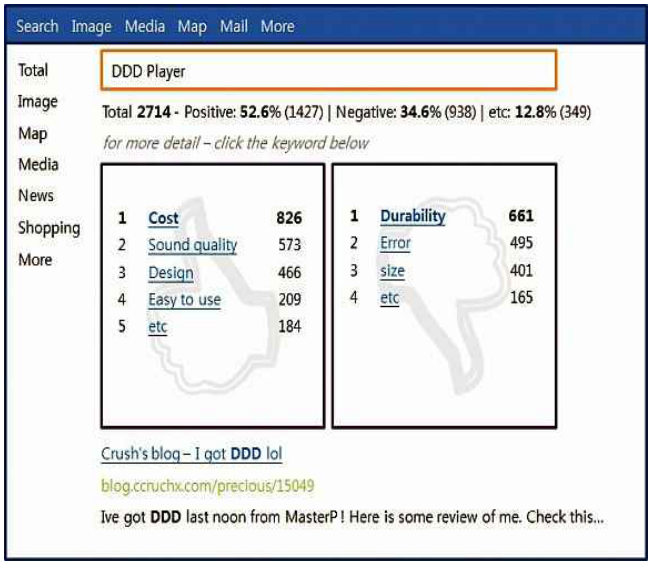
[그림 3] 서비스 흐름도



사용자는 특정 카테고리에 의해 정렬된 자세한 데이터를 결과물로 얻을 수 있다. 키워드는 여러 가지 특징들로 세분화되어질 수 있다. 예를 들어 "DDD" 라고 하는 MP3 플레이어를 검색했다고 가정해보자. 그렇다면 "DDD"의 가격, 품질, 내구성 등과 같이 특징으로 세분화되어 결과물로 출력 된다. "*이 제품은 정말 싸다*", "*시스템이 안정적이다*", "*오작동이 자주 발생한다*"와 같이 이런 특징들에 대한 여론을 종합하여 사용자에게 제공할 수도 있다. 혹은 이런 특정 키워드가 아닌 "취업"과 같이 불분명한 단어를 검색할 경우 취업에 관련된 회사정보, 봉급, 이미지, 환경, 위치 등의 특징에 대하여 세분화 된다.

[그림 4]은 결과물이 어떻게 보여지는지 나타낸다. 우선 시스템 흐름도에 의해 얻어진 결과물이 먼저 출력되어서 한눈에 해당 키워드에 대한 정보와 여론을 알 수 있다. 해당 특징 카테고리 링크 아래에는 여론의 출처가 링크되어있어서 사용자가 일일이 웹 사이트를 찾지 않아도 바로 알 수 있게 디자인 되어있다. 그리고 이어서 일반 사용자들의 웹 사이트가 출력되어 세부 내용을 확인 할 수 있다.

[그림 4] 흐름도의 예상 결과물



4. 결론

본 논문은 단지 시스템 디자인만 그리고 이를 설명하기 위한 리서치에 불과하다. 우선적으로 필요한 것은 이론이 아닌 실제 시스템을 구현하고 거기서 발생할 수 있는 문제점들을 컨트롤하기 위해서는 실험이 필요하다. 가장 먼저 실험에서 필요한 부분은 필터링과 오피니언 분석에서의 속도가 실제로 얼마나 걸리고, 또 그 시간을 최대한 줄이는 것이 목표일 것이다. 그리고 그 다음 부분은 이 데이터들을 얼마나 신뢰할 수 있는지의 여부다. 따라서 앞으로의 연구 방향은 본 시스템의 구현과 사용가능한 속도를 낼 수 있는지의 여부를 알아보는 것이다.

참고문헌

[1] S.M Kim and E hovy, "Extracting opinions, opinion holders, and topics expressed in online news media text", Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text, Sydney, Australi, 2006

[2] Kim, S.M. and Hovy, E, "Determining the sentiment of opinions", Proceedings of the 20th international conference on Computational Linguistics, pp. 1367, 2004

[3] B Pang, and L Lee, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, 2008

[4] Hu, M. and Liu, B., "Mining and summarizing customer reviews", In Proc. of the 10th ACM SIGMOD Conf., p.168-177, 2004

[5] Pennebaker, J.W. and Booth, R.J. and Francis, M.E., “Linguistic inquiry and word count (LIWC2007)”, Austin, TX: LIWC (www. liwc. net), 2007

[6] Pennebaker, J.W. and Graybeal, A. “Patterns of natural language use: Disclosure, personality, and social integration”, Current Directions in Psychological Science, vol. 10, no. 3, pp. 90, 2001

[7] Rude, S. and Gortner, E.M. and Pennebaker, J., “Language use of depressed and depression-vulnerable college students”, Cognition and Emotion, vol. 18, no. 8, pp. 1121-1133, issn 0269-9931, Psychology Press, part of the Taylor \& Francis Group, 2004

[8] Stirman, S.W. and Pennebaker, J.W, “Word use in the poetry of suicidal and nonsuicidal poets”, Psychosomatic Medicine, vol. 63, no. 8, pp. 517, 2001

[9] Pagecar, “Worders”, March 3, 2009, http://themicrowave.wordpress.com/tag/liwc/)

[10] Jindal, N. and Liu, B., “Opinion spam and analysis”, Proceedings of the international conference on Web search and web data mining, pp. 219-230, 2008

[11] Yu, Lei and Liu, Huan, “Efficiently Handling Feature Redundancy in High-Dimensional Data”, 2003

[12] Joanna J´ozefowska, Agnieszka Lawrynowicz, and Tomasz Lukaszewski. “On Reducing Redundancy in Mining Relational Association Rules from the Semantic Web”, 2008

[13] Cho, K.S., Yoon, J.Y., Kim, I.J., Lim, J.Y., and Kim, U.M., “Mining Information of Anonymous user on a Social Network Service” 2011