오피니언 마이닝을 이용한 스팸 필터링

오진수*, 유준석*, 김응모*
*성균관대학교 컴퓨터공학과
e-mail : glassego@naver.com

Spam Filtering using Opinion Mining

Jin-Soo Oh*, Joon-Suk Ryu*, Ung-Mo Kim*
*School of Information and Communication Engineering, Sungkyunkwan University

요 약

오늘날 사람들의 의견을 제시하는 공간은 폐쇄적인 인쇄물이나 수동적인 답변 수준을 벗어나 무한의 공간을 가지는 웹에서 이루어지고 있다. 불특정 다수를 대상으로 하며 정형화된 틀을 없는, 더욱 유용한 의견을 많이 얻을 수 있는 특징을 가졌기 때문에, 이를 위해 오피니언 마이닝에 대한 연구가 활발히 진행되고 있다. 기본적으로 오피니언 마이닝은 해당 분야에 대한 정확한 정보를 찾는 것을 목적으로 하지만, 그러한 정보를 제외한 나머지 부분에 대해서도 충분히 유용하게 사용할 수 있다. 본 논문에서는 그 나머지 부분을 이용하여 무분별하게 등록되고 있는 스팸성 댓글을 효과적으로 필터링 할 수 있는 방법을 제안한다.

1. 서론

최근 온라인 상의 의견표출이 일상화되고 그 영향력이 커지고 있다. 포털이나 웹사이트, 트위터, 블로그 등을 통해 자신의 의견을 제시하고 다른 이들과 교류하면서 지식의 장을 넓히는 것은 이제 당연한 일이 되었다. 그 양은 방대하고, 실로 많은 사람들의 다양한 의견이 표출된다.

이에 오피니언 마이닝이라는 기법을 통해 그를 효과적으로 이용하는 기술이 부상하고 있다. 필요한 정보가 담겨있는 글 속에서 개개인의 놓치기 쉬운 중요한 의견들을 수집하여 필요한 내용만 구분하여 사용하는 것이 그 기법의 목적이다.

하지만 역시 이를 악용하는 일 또한 활성화 되고 있다. 그 예로 악성 스팸성 댓글이나 욕설 같은, 실제 그 게시물이나 카테고리와는 전혀 관련이 없는 글이 있다. 현재 포털이나 사이트 운영측에서 사용하는 방법은 사용자의 신고를 통한 파악, 모니터링 혹은 자주 쓰이는 문구등을 통한 필터링 등 다소 수동적이거나 비정상적으로 작동할 여지를 가지고 있는 방법을 사용되고 있다.

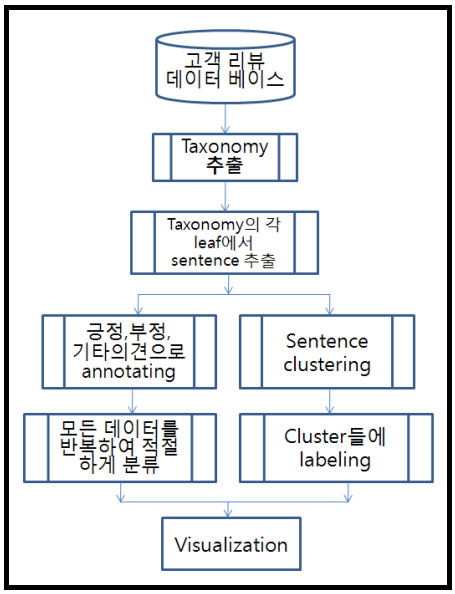
본 논문에서는 활발하게 연구되고 있는 오피니언 마이닝의 부산물이라 할 수 있는, 의견이 아닌 나머지 부분을 이용하여 스팸성 내용에 대한 필터링을 강화할 수 있는 방안을 제시하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2 장은 부정/긍정/그 이외의 내용으로 나누는 방법에 대한 오피니언 마이닝 기술에 대해서 간략하게 설명한다. 3 장에서는 그 이외의 내용에 대한 처리에 대해 기술한 후, 마지막 4 장에서는 결론을 맺고자 한다.

2. 관련 연구

1. Pulse Project

Pulse Project 는 의견, 토픽 내용을 문장 수준에서 탐지할 수 있는 시스템을 구축하는 프로젝트이다[1]. 고객 리뷰 데이터베이스에서 의견을 추출하여 카테고리화 하며 긍정/부정/그 이외의 내용으로 나누어 시각화해주는 일련의 과정을 거쳐 쉽게 필요한 정보를 구분하여 볼 수 있다.



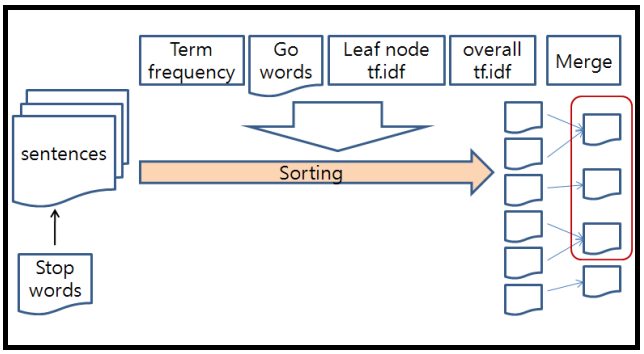
(그림 1) Overview of Pulse System Architecture [1]

2.1 카테고리(분류) /문장 추출

데이터베이스에 간단한 질의를 통해 카테고리와 문장을 추출한다.

2.2 군집 알고리즘

삭제되어야 하는 단어집합인 Stop Words 와 그 도메인에서 잘 알려진 Go Words 로 부터 추출된 군집이 군집 알고리즘의 입력값이 된다. 이 문장 군집은 Porter Stemmer 를 이용하여 어근을 추출한다.[2] 추출된 어근은 정렬과정을 거쳐 문장의 집합에 포함된 내용이 50%이상 중복되는 집합을 병합한다. 이때 화면출력을 위한 top n 개의 군집만을 추출한다.



(그림 2) Diagram of the clustering algorithm [1]

2.3 시각화

Pulse System 시각화의 주요 목적은 토픽과 의견을 화면출력하는 것이다. 창의 좌측에는 클러스터에 포함된 단어를, 우측에는 군집이름이 표시된 Tree Map 시각화[3], 하단에는 좌측에 표시된 단어가 포함된 문장이 전부 보여진다. Tree Map 에는 긍정적인 의견일수록 푸른색으로, 부정적일수록 붉은색으로 구분하였다.

3. 오피니언 마이닝의 부산물을 통한 스팸 내용 필터링 및 관리

본 장은 오피니언 마이닝에서 필요한 주 도메인에 포함되는 의견이 아닌 나머지 부분에 대한 효과적인 사용을 위한 방법을 제안한다. 그 처리 방식은 Pulse Project 에서 제안한 방법과 거의 유사하다[1].

3.1 데이터 마이닝

사용되는 데이터는 군집화의 과정중 최종단계인 병합단계에서 사용했던 top n 개를 제외한 클러스터를 이용한다. 스팸성 데이터는 특성상 주가 되는 자료들의 양보다 훨씬 적을 수 밖에 없고, 따라서 그것이 빈번하게 나타나는 곳은 기존 오피니언 마이닝에서 신경쓰지 않는 기타의견 부분에서 나온다.

일반적으로 웹에 글이 등록이 되면 IP 또한 알 수 있다. IP 가 등록된 글은 문장과 함께 링크시켜 데이터를 처리한다.

3.2 재정렬 단계

Go words 는 기존 웹사이트에서 사용자의 신고 등으로 등록되어 최종적으로 제한된 문장을 2.1 의 단계와 Stop words 를 제거하는 단계를 거쳐 포함시킨다. 3.1 에서 얻어진 데이터는 이미 이전 군집화과정에서 Stop Words 가 제거된 상태이다. 따라서 정렬, 병합만 거치면 된다.

정렬단계에서 유의할 점은, 스팸성 댓글 특성상 길이가 그렇게 길지 않다는 것이다. 따라서 Term Frequency 등 단어빈도수나 가중치가 들어간 계산을 거치는 작업을 경우에 따라 생략해도 무관하다.

3.3 재병합 단계

병합작업은 실제적인 필터링 작업보다는 현재 자주 쓰이는 단어의 추세를 계산하여 그 필터링 수위를 조절하는데 필요하다.

3.4 시각화 단계

기존의 화면출력방식에서 하단의 문장이 뜨는 부분에 IP 정보가 같이 표시된다. IP 정보는 따로 빈도수에 따라 정렬한 것을 다른 팝업 창으로 표시한다.

3.2 와 3.3 에서 정리된 문장과 단어들은 실시간으로 댓글을 체크하는 곳에 사용된다. 문장과 단어의 일치도를 체크하며, 운영하면서 그 수준을 조절하도록 한다. 이는 한 사이트에서만 해서는 효과가 없다. 여러 사이트가 등록되어 있는 상태에서, 툴을 통해 자동으로 등록되는 스팸 댓글을 효과적으로 막을 수 있다.

4. 결론

기존의 스팸 필터링 방식이 한 사이트 관리차원에서 하는 국소적인 방책임을 고려할 때, 오피니언 마이닝 방식을 그대로 사용함으로써 약간의 수고만 더하여 범사이트 적인 효과를 가져올 수 있다는 점을 고려하면 사용할 만한 방법이라 여겨진다.

하지만 차후 실사용을 하기 위해 그 필터링 수준과 정확도를 높이기 위해선 좀더 많은 자료의 수치적인 해석과 연구가 필요하다.

감사의 글

이 논문은 2009 년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0075771).

참고문헌

[1] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger, “Pulse: Mining Customer Opinions from Free Text”, Natural Language Processing, Microsoft Research, Redmond, WA 98052, USA.

[2] Goodman, J, “A bit of progress in language modeling” Technical report, Microsoft Research (2000).

[3] Smith, M.A., Fiore, A.T, “Visualization components for persistent conversations”, In: CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press (2001) 136-143