웹 상의 제품 리뷰 검색 및 분석을 통한 제품 평가 시스템

강대기
동서대학교 컴퓨터정보공학부

Evaluation System using Automated Search and Analysis of Product Reviews on the Web

Kang, Dae-Ki
Dongseo University
E-mail : dkkang@dongseo.ac.kr

요 약

본 연구에서 우리는 웹 사이트들에서 제품에 대한 사용자들의 리뷰 정보를 수집하고, 수집한 정보들을 분석 및 정련하여 사용자들에게 보이는 서비스에 대해 논하고자 한다. 특정 제품에 대한 리뷰 정보들은 로봇 시스템에 의해 수집되고, 특정 제품에 대한 전체적인 평가 스코어는 두 가지 다른 종류의 스코어들을 고려하여 계산된다. 첫 번째 스코어는 정량적인 스코어(quantitative score)로 각 리뷰들로부터 얻어지는 이른바 별점 값들의 가중 평균값(weighted average)으로 계산된다. 두 번째 스코어는 정성적인 스코어(qualitative score)로, 본 연구에서 제안된 서비스는 각 리뷰들의 텍스트 설명을 자연 언어 처리 기법으로 분석하여 정성적 스코어를 계산한다. 우리는 이러한 스코어 계산 모델에 따라 MP3 플레이어와 Personal Digital Assistant (PDA)에 대해 서비스 시스템 RELLENOS를 설계 및 구현하였다. RELLENOS는 69 개에 달하는 온라인 리뷰 사이트들에서 수집된 정보들을 토대로 정량적인 값과 정성적인 값을 계산하여 서비스를 성공적으로 수행하였다.

1. 서론

웹 상에는 서로 다른 출처에서 나오는 다양한 양의 제품 정보들이 존재한다. 예를 들어 제품을 만드는 제조업자들은, 자신들의 제품에 대한 사람들의 관심을 높이기 위해, 제품들에 대한 설명, 상세한 사양, 그리고 이미지들을 웹을 통해 제공한다. 비슷하게, 소매상들은 고객들을 끌어들이기 위해 자신들이 팔 제품의 정보들을 웹을 통해 제공한다. 그러나, 제조업체나 소매상들 각각은 자신들이 생산하거나 팔려는 제품들의 선전에 치중하므로, 이러한 사이트들에서 얻는 정보는 객관적이 못하다는 문제점이 있다.

반면 제품을 전문적으로 평가하는 독립적인 평론가들 또한 자신들이 사용하는 제품에 대한 정보나 소감을 웹에 올리거나, 잡지나 신문의 기사를 통해 제공되기도 한다. 대부분의 독립적인 평론가들은 특정 제품에 대한 사적인 이득이 없기 때문에, 그들이 제공하는 정보는 제조업체나 소매상들이 제공하는 정보보다 더 객관적인 경우가 많다.

또한 인터넷의 발전으로 많은 개인 사용자들이 자신들이 사용한 제품에 대한 후기나 소감을 웹을 통해 제공할 수 있게 되었다. 이러한 리뷰들은 숙련된 아마추어 사용자들이나 일반 사용자들에 의해 쓰여진다. 이러한 리뷰들은 특정 제품을 구매하려는 사람들에게는 매우 유용할 수 있다. 특히 리뷰어들이 개인적으로 느낀 해당 제품의 장점과 단점을 지적하는 경우, 장래에 구매할 사람들에게 유용한 정보가 된다.

그러나, 이렇게 웹 사이트들의 제품 리뷰들이 유용한 반면, 이러한 웹 사이트들로부터 유용한 데이터를 추출해내는 작업은 시간이 많이 소모되고 어렵다. 이러한 어려움은 주로 다음 세가지로 요약된다.

1. 우선 특정 제품에 리뷰를 찾는 것이 어렵다. 따라서 사용자는 여러 개의 웹 사이트들을 돌아다니면서 원하는 리뷰를 찾아야 한다.
2. 원하는 리뷰를 찾았어도, 그로부터 원하는 정보를 추출하는 것도 어려운 문제이다. 이는 기존의 웹 사이트들이 컴퓨터가 원하는 정보를 쉽게 얻을 수 있는 특징 기반의 구조적 형식으로 되어 있지 않기 때문이다.
3. 특정 제품과 리뷰들이 정해진 상황에서도 사용자의 기호나 요구에 부응하는 정보 추출 및 분석이 용이하지 않다. 예를 들어 여행을 자주 하는 사용자는 노트북 컴퓨터에서 연산 수행 능력보다는 무게와 배터리 수명에 더 관심을 가진다.

그럼에도 불구하고, 현재로서는 이러한 리뷰를 자동으로 찾아서 원하는 정보를 추출하는 방법이 많이 연구되지 않았다. 따라서, 제품 정보를 발견하고 사용자들에게 제공하는 새로운 방법이 요구되게 되었다.

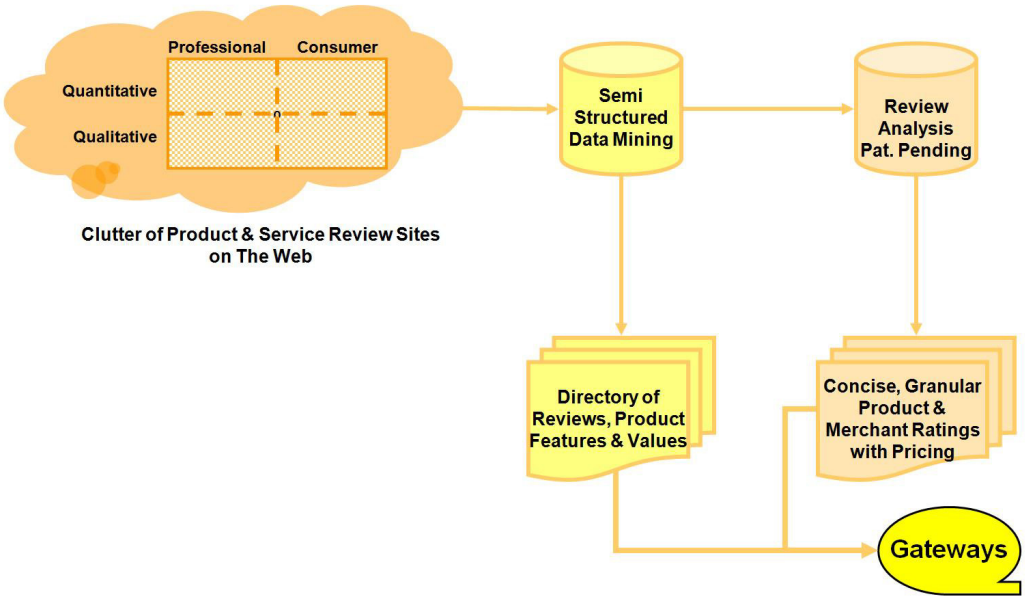

그림 1 리뷰 분석 시스템의 구조도

2. 본론

본 논문에서 제안하는 리뷰 분석 시스템의 구조는 그림 1과 같다. 그림 1과 같이 웹에는 제조업체와 소매상, 그리고 평론가와 소비자들이 각각 만들어낸 정성적인, 그리고 정량적인 리뷰들이 존재한다. 로봇 시스템이 이러한 리뷰 정보(review information)들을 가격(price)이 포함된 제품 정보(product information)와 같이 수집해서 캐쉬 (cache) 데이터베이스에 저장한다. 저장된 정보들 중 가격이 포함된 제품 정보는 가격 색인 (index) 시스템[1,2]으로 색인되고, 리뷰 정보들은 리뷰 분석(review analysis) 시스템을 통해 분석되어 해당 제품에 대한 간결하면서도 통합된 정보로 사용자에게 제공된다. 그림 2는 이러한 가격이 포함된 제품 정보와 리뷰 분석을 통한 통합된 제품 정보 서비스의 한 예이다.

2.1 웹 페이지 정보 수집 로봇

웹 페이지들은 웹 크롤러 (Web crawler) 또는 웹 스파이더 (Web spider) 와 같은 로봇[3]에 의해 시스템으로 수집된다. 이러한 수집 로봇은 내부적으로 주어진 seed URL들로부터 시작하여 온라인 리뷰 사이트, 온라인 상점, 그리고 제조업체의 웹 사이트에서 제품에 대한 정보를 HTML의 형태로 수집한다. 수집을 반복하는 루프에서는 내부의 큐(queue)에서 다음에 수집할 URL을 가져와서 HTML 파일을 수집하고, 수집된 HTML 파일에서 하이퍼링크(hyperlink)들을 추출하여 다시 큐에 추가하는 과정을 반복한다. 또한, 별도의 해쉬 테이블(hash table)에 이미 수집된 URL들을 저장하여, 한 번 수집된 URL들을 다시 수집하지 않도록 한다.

2.2 리뷰 분석 시스템

특정 제품에 대한 전체적인 평가 스코어는 정성적인 스코어와 정량적인 스코어, 두 가지의 다른 종류의 스코어들을 고려하여 계산되었다.

1. 정량적인 스코어(quantitative score)는 각 리뷰들로부터 얻어지는 이른바 별점 값들의 가중 평균값(weighted average)으로 계산된다. 이를 위해 서로 다른 각각의 온라인 리뷰 페이지에서 별점 정보

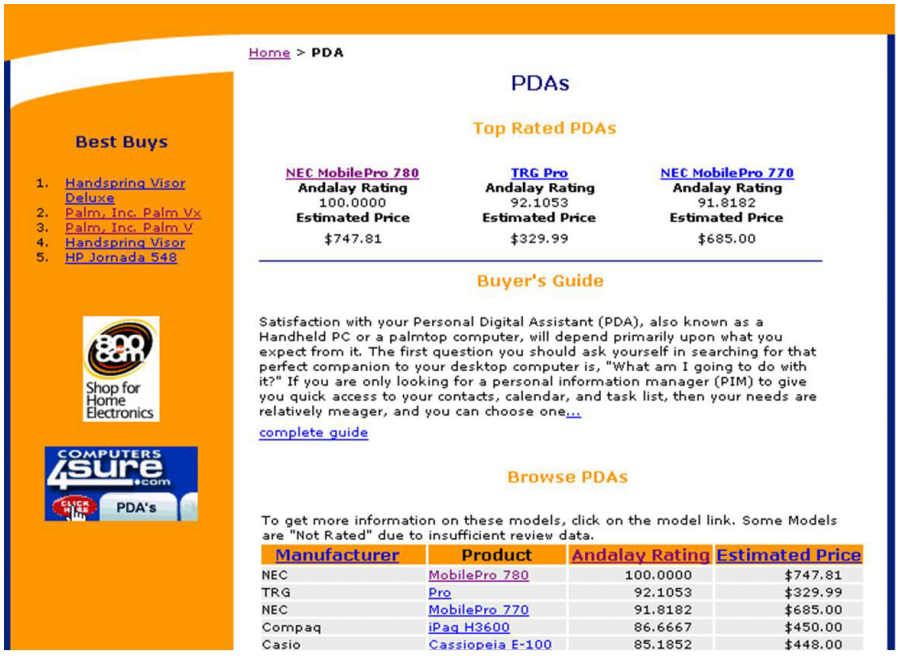

그림 2 가격이 포함된 제품 정보와 리뷰 분석을 통한 통합된 제품 정보 서비스

또는 평가 (rate) 정보를 기존의 패턴 매칭 기법[4]을 통해 분석한다.

2. 정성적인 스코어(qualitative score)는 각 리뷰들의 텍스트 설명을 자연 언어 처리 기법으로 분석하여 계산된다. 이를 위해 사용자의 평가에 관련된 단어들 (“excellent”, “good”, “bad”, “poor” 등등)을 미리 선정하고 각각의 단어들에 대해 가중치를 부여한다. 그리고 주어진 문장들에 대해 형태소 분석 알고리즘 (stemming algorithm), 불용어 (stop word) 제거 등을 통해 전처리를 수행한다. 수행된 결과에서 앞에서 언급한 바와 같이 미리 선정된 단어들의 리스트에 따라 주어진 리뷰의 제품에 대한 정성적 평가를 구한다.

3. 결론

본 연구에서 우리는 웹 사이트들에서 제품에 대한 사용자들의 리뷰 정보를 수집하고, 수집한 정보들을 분석 및 정련하여 사용자들에게 보이는 서비스에 대해 논하였다. 특정 제품에 대한 리뷰 정보들은 로봇 시스템에 의해 수집되고, 특정 제품에 대한 전체적인 평가 스코어는 두 가지 다른 종류의 스코어들을 고려하여 계산된다. 첫 번째 스코어는 정량적인 스코어(quantitative score)로 각 리뷰들로부터 얻어지는 이른바 별점 값들의 가중 평균값(weighted average)으로 계산된다. 두 번째 스코어는 정성적인 스코어(qualitative score)로, 본 연구에서 제안된 서비스는 각 리뷰들의 텍스트 설명을 자연 언어 처리 기법으로 분석하여 정성적 스코어를 계산한다. 우리는 이러한 스코어 계산 모델에 따라 MP3 플레이어와 Personal Digital Assistant (PDA)에 대해 서비스 시스템 RELLENOS를 설계 및 구현하였다. RELLENOS는 69 개에 달하는 온라인 리뷰 사이트들에서 수집된 정보들을 토대로 정량적인 값과 정성적인 값을 계산하여 서비스를 성공적으로 수행하였다.

[참고문헌]

[1] 강대기, 김중배, 손주찬, 함호상, “전자 상거래를 위한 월드 와이드 웹 디렉토리 서비스의 설계,” 한국 정보과학회 춘계 학술 발표 논문집, 제24권 제1호, pp. 449-452, 강원도 춘천 한림대학교, 1997년 4월 25일~26일

[2] Kang, D.-K., and Sohn, K., “Wrapper Induction Based on Minimum Description Length using a Suffix Tree,” Proceedings of the 22nd International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2007), Busan, Korea, July 8-11, 2007.

[3] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, “Chapter 20. Web crawling and indexes,” Introduction to Information Retrieval, Cambridge University Press. 2008. (http://nlp.stanford.edu/IR-book/pdf/20crawl.pdf)

[4] 강대기, 이제선, 함호상, “Web 문서의 효율적인 실시간 검색을 위한 잡음 제거와 패턴 정합 기법,” 한국정보과학회 1998년도 가을 학술발표논문집 제25권 제2호(Ⅱ), 1998. 10. pp. 132~134