정보학연구 제12권 제3호 2009년 9월(pp.7-17)
The Journal of Information Technology, Vol.12, No.3

의견정보 검색엔진을 위한 웹 콘텐츠 마이닝 시스템

주 해 종* · 박 영 배** · 최 혜 길***

목 차

Ⅰ. 서 론

최근에 인터넷 사용이 점차 활발해 짐에 따라, 많은 사람들이 인터넷에서 예컨대, 블로그(Blog), 위키(Wiki)와 같은 매체를 통해서 자신의 의견을 표현하고 있는 추세이다. 또한, 특정한 정보의 가치를 평가할 때, 이러한 다른 사람들이 인터넷 상에 올려놓은 의견 정보를 참조하고자 하는 수요도 높아지고 있다.

예를 들면, 인터넷 상에는 상품 리뷰(Review)에서 영화 리뷰까지 다양한 사용자들의 의견이 존재한다. 이러한 각 사용자들의 의견들은 일반 사용자들이 물품을 구매하거나, 영화를 보기 전에 다른 사용자들의 의견을 보고자 하는 경우에도 이용될 수 있으며, 마케팅 담당자나 주식 매매자 등이 각 물품이나 회사에 대한 일반 사용자들의 다양한 의견을 알고자 하는 경우에도 사용될 수 있다. 특히, 일반 사용자들은 특정 물품을 구매

* 명지대학교 컴퓨터공학과 박사과정
** 명지대학교 컴퓨터공학과 교수
*** 경희사이버대학교 정보통신학과 교수

하기 전에 다른 사용자들의 평가를 먼저 보고 나서 이런 물품을 구매하려는 경향이 크다. 하지만, 이러한 인터넷 상에 존재하는 의견들은 개개의 웹사이트들에만 존재하여, 이러한 의견 정보들을 사용하고자 할 경우에는 사용자가 일일이 이러한 개개의 모든 웹사이트를 수동으로 찾아보아야 하는 번거로움이 존재한다. 이러한 모든 웹사이트들을 사용자들이 모두 찾아보기 어려우며 일반 검색으로 다른 사용자들의 의견을 찾고자 하는 경우에는 의견이 있는 웹 문서, 긍정적인 의견이 있는 웹 문서, 부정적인 의견이 있는 웹 문서 등이 혼재하여 효과적으로 다른 사용자들의 의견을 찾아보기 어려운 문제점이 있다.

이러한 문제점을 해결하기 위하여 국/내외 학계를 중심으로 사용자 의견 추출 기술이 활발하게 연구되고 있으며, 정보 검색 분야에서도 2000년도 초반부터 크게 발전하여 다양한 기술이 연구되고 있다. 그러나 기존의 정보 검색 기술은 단순히 키워드가 존재하는 정보에 기반한 검색만 제공해주고 있을 뿐이고, 각 키워드가 등장하는 문서나 문장에서 긍정적/부정적으로 평가된 내용을 기반으로 한 좀더 고차원적인 검색까지 제공해주고 있지 못하고 있다. 최근에 사용자 의견 추출 기술을 정보 검색에 적용하려는 시도가 진행되고 있으나 아직도 단순히 긍정, 부정 문서를 나누는 수준에만 머무르고 있는 실정이다.

본 논문 이러한 문제점을 해결하기 위하여 제안한 것으로서, 본 논문의 목적은 인터넷 상에 존재하는 여러 웹사이트들에 흩어져 있는 웹 콘텐츠에서 사용자 의견 정보들을 자동 추출 및 분석함으로써, 긍정/부정 의견별로 검색 및 통계를 확인할 수 있는 의견 검색 서비스를 간편하게 구현할 수 있으며, 의견 검색 사용자들은 특정 키워드에 대하여 다른 사용자들의 의견을 손쉽게 한눈에 검색 및 모니터링하는 시스템을 용이하게 구현할 수 있도록 한 웹 콘텐츠에서의 의견 추출 및 분석 시스템의 설계를 제공하는데 있다.

이 논문의 구성은 다음과 같다. 2장에서는 본 논문의 이론적 고찰을 위해 기존 웹 마이닝의 특징과 문제점을 살펴본다. 3장에서는 기존 웹 마이닝의 문제점을 해결하기 위한 통계기반 웹 콘텐츠 마이닝 시스템의 설계와 기술요소를 설명한다. 4장에서는 기존 웹 마이닝과의 성능비교를 제시한다. 마지막으로 5장에서는 결론을 맺는다.

II. 관련 연구

그림 1의 기존 웹 마이닝(Web Mining)은 웹 콘텐츠와 서비스로부터 자동으로 정보를 발견하고 추출하기 위해 데이터마이닝 기법을 이용하는 것이다. 즉, 웹 콘텐츠로부터 미리 알려지지 않은 유용한 정보나 지식을 발견하는 과정이라고 정의할 수 있다[3,9].

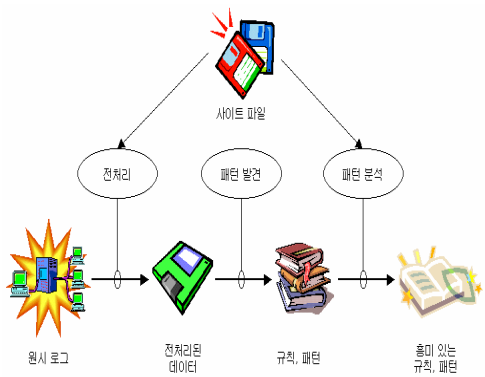

그림 1. 웹 사용 마이닝의 단계

웹 마이닝은 크게 웹 컨텐트 마이닝(Web Content Mining), 웹 구조 마이닝(Web Structure Mining), 웹 사용 마이닝(Web Usage Mining)으로 분류된다[9].

웹 컨텐트 마이닝은 문서, 이미지, 오디오, 비디오 등의 웹 컨텐트, 데이터, 문서로부터 유용한 정보를 발견하는 것이다. 아직까지 웹 컨텐트 마이닝은 연구가 미미한 편이며 텍스트마이닝 분야에

서 주로 연구되고 있다. 웹 컨텐트 마이닝은 비구조화된 문서를 이용하는 정보검색의 관점과 웹 데이터에 복잡한 질의를 수행하기 위해 데이터베이스로 모델화하고 통합하는 데이터베이스 관점으로 분류된다. 웹 컨텐트 마이닝은 웹 문서가 HTML에서 XML로 변환해 감에 따라 그 연구 및 적용 분야가 크게 확대될 것으로 보인다. XML은 태그 셋을 통해 문서 내의 특징들을 스스로 묘사할 수 있기 때문이다. XML 문서는 DTDs (Document Type Definitions)를 질의함으로써 쉽게 문서의 내용을 추출할 수 있다[9,11].

웹 구조 마이닝은 웹의 링크 구조하의 모델을 발견하는 기법으로 웹의 하이퍼링크의 토폴로지(Topology)에 기반한 모델이다. 웹 구조 마이닝은 주로 웹사이트 간의 유사성 및 관계 정보를 생성하는데 적용된다. 웹 구조 마이닝을 이용하여 권위 사이트(Authority Site)와 허브 사이트(Hub Site)를 발견하는 작업에 사용될 수 있는데 웹 권위 사이트는 어느 한 주제에 대해 많이 참조되는 사이트이며 허브 사이트는 많은 권위 사이트들을 가리키고 있는 사이트를 의미한다[9].

웹 사용 마이닝은 웹 서버로그, 브라우져 로그, 사용자 프로파일, 쿠키 등 부차적인 데이터를 이용하여 웹 사용자의 세션(Session)과 행동으로 생성된 데이터로부터 정보를 발견하는 기법으로 현재 가장 많은 연구가 진행되고 있는 분야이다[9,11].

웹 사용 마이닝이 주로 다루는 작업은 다음과 같다. 첫째, 웹 액세스 패턴(Web Access Pattern) 분석과 웹 사용자 네비게이션 행동(Navigational Behavior) 분석한다[9]. 웹 액세스 패턴 분석이 웹 로그를 통해 사용자 프로파일을 학습하거나 각 사용자에 대해 적응 인터페이스(Adaptive Interface)를 모델링하는 개인화된(Personalized) 방법인 반면 사용자 네비게이션 행동 분석은 사용자의 네비게이션 행동을 학습함으로 웹 사이트의 디자인 최적화하는 비개인화된(Impersonalized) 방법이다. 웹 액세스 패턴과 네비게이션 행동 분석의 논점은 얼마나 효율적이고 효과적으로 수행하는 가이다. 이를 위해 웹 로그를 관계형 데이터베이스에 저장한 후 데이터 큐브를 구축하여 OLAP를 통한 다차원 분석과 데이터마이닝을 적용하여 웹 액세스 패턴을 분석하거나, 웹 액세스 패턴을 더욱 효율적으로 발견하기 위해서 조밀한 데이터 저장 구조를 만들어 웹 로그를 저장한 후 데이터마이닝 기법을 적용한다[3,9].

둘째, 웹상에서 사용자에 따른 최적화된 링크를 동적으로 설정한다. 중앙처리장치가 노는 시간에 다음에 요청될 페이지의 URL을 미리 예측하여 미리 동적으로 컨텐트 페이지를 준비함으로써 서버의 용량 낭비를 최소화 하고 서버의 응답시간은 최대화할 수 있다[9].

셋째, 적응 웹사이트(Adaptive Website)를 구축한다. 적응 웹사이트는 사용자 액세스 데이터에 기반하여 구조나 프리젠테이션을 자동으로 개선하는 사이트를 말한다. 적응 사이트가 수행하는 두 가지 작용은 개개 사용자의 필요에 맞게 실시간으로 웹 페이지를 수정하는 개인화(Personalization)와 사용자의 네비게이션이 용이하게 사이트가 스스로 구조를 개선시키는 최적화(Optimization)가 있다[9].

넷째, 웹 페이지와 사용자의 모델링, 웹 페이지와 사용자의 카테고리화, 웹 페이지와 사용자의 매칭(Matching) 등을 수행하는 웹 개인화이다[9,12]. 이를 위해 주로 연관규칙이나 트랜잭션 군집화, 사용 군집화(Usage Clustering) 등의 기법을 이용하여 사용자의 네비게이션 패턴에 기반한 URL 사이의 관계나 사용자와 비슷한 성향을 가지는 군집을 찾아내 개인화에 적용할 수 있다[12,13].웹 사용 마이닝 방법을 이용한 페이지 추천 방법은 협동 필터링(Collaborative Filtering)에 비해 여러 가지 장점을 가진다. 협동 필터링은 개

인화 행동을 결정하는데 사람의 입력 값에 지나치게 의존하기 때문에 입력 값에 나타난 사용자의 주관적 편향을 제거하여야 하며 사용자 프로파일이 정적이어서 시간이 지날수록 수행 성능이 저하되는 단점을 가진다.

III. 통계기반 웹 마이닝 시스템

이 논문에서 제시하는 통계기반의 웹 콘텐츠 의견 추출 및 분석을 위한 웹 마이닝 시스템 플랫폼은 그림 2와 같으며, 인터넷 상에 존재하는 여러 웹사이트들에 흩어져 있는 웹 문서에서 사용자 의견 정보들을 자동 추출 및 분석함으로써, 긍정/부정 의견별로 검색 및 통계를 확인할 수 있는 의견 검색 서비스를 간편하게 구현할 수 있고, 의견 검색 사용자들은 특정 키워드에 대하여 다른 사용자들의 의견을 손쉽게 한눈에 검색 및 모니터링(Monitoring)하기 위한 구조이다.

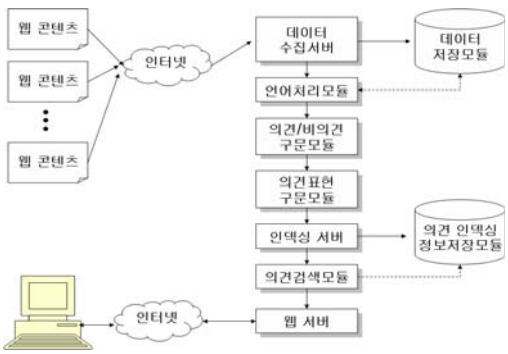

그림 2. 웹 콘텐츠 의견 추출 및 분석을 위한 웹 마이닝 시스템 플랫폼

3.1 시스템 구성요소

그림 2의 제안시스템은 크게 데이터 수집서버, 언어처리모듈, 의견/비의견 구분모듈, 의견표현 구분모듈, 인덱싱 서버, 의견 인덱싱 정보 저장모듈, 의견검색모듈, 웹 서버 및 사용자 단말 등을 포함하여 이루어진다.

3.1.1 데이터 수집서버

데이터 수집서버는 인터넷 상에 존재하는 다양한 웹 콘텐츠들을 수집하는 기능을 수행한다. 즉, 데이터 수집서버는 인터넷 상에 존재하는 각 웹사이트(Web Site)들의 HTML(Hyper Text Markup Language) 정보를 실시간으로 다운로드(Download) 받게 된다. 또한, 데이터 수집서버는 상기와 같이 다운로드(Download) 받은 웹 콘텐츠에서 필요한 정보들 예컨대, 텍스트(Text), 이미지(Image) 또는 비디오(Video) 등의 정보들 중 적어도 어느 하나의 정보 데이터를 추출하여 별도의 데이터 저장모듈에 저장시킬 수 있다.

데이터 수집서버는 표 1과 같이, 의견정보 데이터(즉, 일반 문장/문서 데이터와 이에 대한 긍정/부정 평가가 매겨진 정보 데이터)를 포함하는 웹 콘텐츠들을 선별하여 수집할 수도 있다. 이때, 상기 의견정보 데이터를 포함하는 웹 콘텐츠들만을 선별적으로 수집하는 방법으로는, 의견정보 데이터를 포함하는 특정의 웹 콘텐츠를 선별하고, 후술하는 기계학습 알고리즘(예컨대, SVM, K-NN, Bayseian 등)을 사용하여 웹 콘텐츠 선별 모델을 생성한 후, 상기 생성된 웹 콘텐츠 선별 모델을 사용하여 전체 인터넷 웹 페이지에서 의견정보 데이터가 포함된 웹 콘텐츠들만을 선별적으로 수집할 수 있게 된다.

표 1. 의견 정보 데이터

표현	점수	의견 내용
★★★★★	10	재미있어 신고
★★★★★	10	'똑똑한' 사람들이 살아있는 이야기 신고
★★★★★	8	현명한 사람들의 일상 뜯어고치기! 신고
★★★★★	9	삼촌의 매력에 흠뻑... 신고
★★★★★	8	평범한 사람들의 이야기 신고
★★★★★	10	연기도 좋고 내용도 짬있고 가슴 훈훈해지는 사랑이야기 신고
★★★★★	10	정말 감동할만한 이야기이었어요. 신고
★★★★★	10	보는 내내 가슴 따뜻해지는 영화였습니다. 재미도 있고요 신고
★★★★	6	훈훈하고 코믹하고.. 영화 넘 짧은거 같은데.. 신고
★★★	5	돌고돌고돌아 결국은 뻔한 이야기. 신고

데이터 수집서버를 통해 수집되는 대상 데이터는 상기의 표 1에 나타난 바와 같이, 의견정보 데이터 즉, 일반 문장/문서 데이터와 이에 대한 긍정/부정 평가가 매겨진 정보 데이터들이다. 이때, 상기 긍정/부정 평가는 일정 범위내의 점수로 표현되어지거나, 별표(★)나 기타 기호들을 이용하여 다양하게 평가될 수 있다. 본 논문에서는 이렇게 다양한 방식으로 표현되는 긍정/부정 평가는 모두 동일한 점수 범위로 재계산되어서 사용된다.

이를 구체적으로 설명하면, 본 논문의 실시 예에서 사용하는 점수 범위가 a~b 라고 하였을 때에 수집한 데이터의 점수범위가 c~d 라고 한다면, 해당 수집 점수 x는 수학식 1과 같이 변화된다.

$$PolarityScore(x) = (a-1) + \frac{x-c+1}{d-c+1} \times (b-a+1) \quad \text{... (수학식 1)}$$

예를 들어, 본 논문은 1~10점 사이의 점수를 사용하고(10점에 가까울수록 긍정), 수집한 데이터는 1~5점 사이의 점수를 사용하는 경우에, 수집한 데이터가 2점이라고 한다면, 수학식 2와 같이 계산되어 진다.

$$PolarityScore(2) = (1-1) + \frac{2-1+1}{5-1+1} \times (10-1+1) = 4 \quad \text{... (수학식 2)}$$

3.1.2 언어처리 모듈

언어처리 모듈은 데이터 수집서버로부터 수집되거나 데이터 저장모듈에 저장된 웹 콘텐츠에 대해 문장 단위로 분리하고, 분리된 각 문장에 대해 언어처리를 수행하여 언어적인 자질(Feature)들을 추출하는 기능을 수행한다. 이때, 상기 언어처리는 예컨대, 형태소 분석(Morpheme Analyze) 또는 띄어쓰기(Segmentation) 처리로 수행됨이 바람직하지만, 이외에도 자질(또는 색인어) 추출을 위한 조사 처리, 한국어 굴절 처리, 또는 원형 복귀 처리 등을 수행할 수도 있다.

이후에, 각 도메인별 자질(Feature)들은 예컨대, Naive Baysian, SVM 또는 K-NN, 기타 일반적인 기계학습 알고리즘(Machine Learning Classifier Algorithm)을 이용하여 확률적으로 학습을 하게 된다. 구체적으로 Naive Baysian을 예로 들어서 설명하면, 하기의 수학식3과 같이 표현될 수 있다.

$$p(C|F1,...,Fn) = \frac{p(C)\,P(F1,...,Fn|C)}{p(F1,...Fn)} \quad \text{..... (수학식 3)}$$

여기서, 상기 C는 클래스(Class)를 의미하며 예컨대, 영화, 도서, 상품 등과 같은 도메인이 이에 해당된다. 상기 Fi는 각각의 자질(Feature)을 의미하며 예컨대, 유니그램(Unigram)(저자), 바이그램(Bigram)(저자 도서), 트라이그램(Trigram)(저자 도서 A) 등이 이에 해당된다. 그리고, 상기 P(C)는 클래스 C가 나올 확률이다. 예컨대, 영화 데이터가 5개, 도서 데이터가 12개, 상품 데이터가 8개라고 한다면, P(영화)는 "5/(5+12+8)" 확률이 된다.

그리고, 상기 P(F1,...,Fn)는 각각의 Fi가 동시에 나올 확률인데, 모든 클래스에 대해서 동일하게 적용되기 때문에 생략도 가능하다(모든 클래스에 동일하게 분모로 적용됨). 그리고, 상기 P(F1,...,Fn|C)는 클래스 C가 주어졌을 때, F1,...,Fn가 생성될 확률이다.

3.1.3 의견/비의견 구분 모듈

의견/비의견 구분 모듈은 언어처리 모듈로부터 추출된 각 문장의 언어적인 자질(Feature)들을 이용하여 의견/비의견 문장을 구분하는 기능을 수행한다. 즉, 언어처리 모듈로부터 추출된 문장들은 의견이 있는 문장들도 있고, 의견이 존재하지 않은 일반 문장도 있다. 이러한 문장들은 의견/비의견 구분모을 이용하여 의견이 존재하는 문장과 의견이 존재하지 않은 문장으로 구분할 수 있게 된다. 이러한 의견/비의견 구분 모듈은 상술한 통상의 기계학습 알고리즘을 이용하여 용이하게

구현될 수 있다.

이를 구체적으로 설명하면, 먼저, 의견으로 이루어진 데이터 집합과 사실 정보로만 이루어진 데이터 집합을 수집한다. 이후에, 예컨대, 형태소 분석(Morpheme Analyze)이나 띄어쓰기(Segmentation) 등을 수행하여 적절한 언어적인 자질(Feature)을 추출한다. 여기서, 상기 띄어쓰기(Segmentation)라 함은 입력 문장을 의미를 가지는 단위로 나누는 과정이다. 예를 들면, 입력 문장이 "나는 영화를 재밌게 봤다"라고 한다면, 결과 문장은 "나 는 영화 를 재밌 게 보 았 다"로 변환된다. 그리고, 상기 형태소 분석(Morpheme Analyze)이라 함은 상기 각 나뉘어진 단위에 대하여 어떤 품사(Part Of Speech) 정보를 지니고 있는지 찾아주는 작업이다.

예를 들면, 입력 문장이 "나는 영화를 재밌게 봤다"라고 한다면, 결과 문장은 "나(CTP1 1인칭 대명사) + 는(fjb 보조사) 영화(CMCN 비서술 보통명사) + 를(fjco 목적격조사) 재밌(YBDO 일반동사) + 게(fmoca 보조 연결어미) 보(YBDO 일반동사) +았(fmbtp 과거시제 선어말어미) + 다(fmofd 평서형 종결어미)"로 변환된다.

다음으로, 상기 추출한 언어적인 자질(Feature)을 이용하여 통상의 기계학습 알고리즘인 예컨대, Naive Baysian, SVM, K-NN 이나 기타 모델을 선택하여 학습을 수행한다. 이렇게 학습이 끝나고 나면, 임의의 문장이나 문서가 입력이 되면, 해당 데이터가 의견 데이터인지 사실 데이터인지 구분할 수 있는 의견/비의견 구분 모델 즉, 의견/비의견 구분모듈이 구현될 수 있다.

3.1.4 의견표현 구분 모듈

의견표현 구분 모듈은 의견/비의견 구분 모듈로부터 구분된 의견 문장의 언어적인 자질(Feature)들에 대해 긍정/부정 의견표현으로 구분하는 기능을 수행한다. 이를 구체적으로 설명하면, 인터넷 상에는 영화 리뷰(Review), 상품평, 책서평 등 각종 리뷰가 존재한다. 이러한 리뷰(Review)들은 보통 평가 문장들과 함께 평가 결과도 함께 게시되어 있다.

예를 들어, "이 영화는 최고의 걸작이다."하고 10점을 주거나, "이거는 완전히 쓰레기 영화다."하고 1점을 주는 방식이다. 이러한 의견 데이터를 기반으로 하여 본 논문에서는 각 의미 단위들이 가지는 긍정 점수와 부정 점수를 계산하여 자동적으로 별도의 의견 어휘 저장모듈(미도시)에 저장하게 된다.

만약, 입력 문장이 「"이 영화는 정말 재밌었다" - 10점, "이번에 본거는 꽤 재밌었다" - 9점, "내 생애 최고로 재밌었던 영화" - 9점」이라고 한다면, 상기 언어처리를 수행할 경우, 「"이(SGR 지시 관형사) 영화(CMCN 비서술 보통명사) + 는(fjb 보조사) 정말(SBO 일반 부사) 재밌(YBDO 일반동사) + 었(fmbtp 과거시제 선어말어미) + 다(fmofd 평서형 종결어미)" - 10점, "이번(CMCN 비서술 보통명사) + 에(fjcao 일반부사격조사) 본거(CMCN 비서술 보통명사) + 는(fjb 보조사) 꽤(SBO 일반 부사) 재밌(YBDO 일반동사) + 었(fmbtp 과거시제 선어말어미) + 다(fmofd 평서형 종결어미)" - 9점, "나(CTP1 1인칭 대명사) + 에(fjcao 일반 부사격조사) 생애(CMCN 비서술 보통명사) 최고(CMCN 비서술 보통명사) + 로(fjcao 일반부사격조사) 재밌(YBDO 일반동사) + 었(fmbtp 과거시제 선어말어미) + 던(fmotgp 과거시제 관형형 전성어미) 영화(CMCN 비서술 보통명사)" - 9점」과 같이 언어 단위별로 분리된다. 다음으로, 상기와 같이 분리된 각 언어 단위별로 긍정적/부정적 표현으로 될 확률을 계산한다.

예를 들어, "최고(CMCN 비서술 보통명사)"가

얼마 정도의 긍정/부정을 나타내는지 "최고(CMCN 비서술 보통명사)"라는 단어가 각 점수 대별(1~10)에 어떻게 분포하는지를 확률적으로 하기의 수학식 4와 같은 수식을 거쳐서 계산하게 된다. 아래에서 나타내는 wj는 "최고(CMCN 비서술 보통명사)"이며, 이와 같이 단어와 태그정보(POS – Part Of Speech)의 조합을 나타내거나, "최고" 태그정보를 제외한 하나의 단어를 나타낼 수 있다. 즉, 모든 1~10점의 점수대에 모두 같은 개수의 데이터가 존재한다면 하기의 수학식 4와 같이 구할 수 있게 된다.

$$Score(Wj) = \frac{\sum_{Si \in S} [Score(Si) \times Freq(Wj, Si)]}{\sum_{Si \in S} Freq(Wj, Si)} \quad \text{..... (수학식 4)}$$

여기서, 상기 S는 모든 점수 집합을 의미한다. 예를 들어, 영화 평가문이 1~10점 이 있다면, 1~10점으로 점수가 매겨진 문장 집합을 의미한다. 상기 Score(si)는 해당 점수 집합의 실제 점수를 의미한다. 즉, 10점 점수 집합의 Score(si)는 10이 된다. 그리고, 상기 Score(wi)는 wi의 긍정/부정 점수를 나타낸다. 상기 Freq(wj,si)는 단어 wj가 점수 집합 si에서 나타나는 횟수를 나타낸다.

3.1.5 인덱싱 서버와 의견 인덱싱 저장모듈

인덱싱 서버는 의견표현 구분 모듈로부터 구분된 의견 문장의 언어적인 자질별로 해당 웹 콘텐츠의 의견 정보들이 의견 인덱싱 정보 저장모듈에 저장되도록 인덱싱(Indexing)하는 기능을 수행한다.

여기서, 의견 인덱싱 정보 저장모듈은 인덱싱 서버를 통해 인덱싱된 각 의견 문장의 언어적인 자질별 해당 의견 문장의 요약정보 및 해당 웹 콘텐츠의 기본 및 의견 정보들이 데이터베이스(DB)화하여 저장되는 기능을 수행한다.

3.1.6 의견 검색 모듈

의견 검색 모듈은 웹 서버를 통해 전송된 사용자의 특정 의견 검색 키워드 또는 타입(Type) 정보를 제공받아 인덱싱 서버 또는 인덱싱 정보 저장모듈과 연동하여, 상기 특정 의견검색 키워드 및/또는 타입(Type) 정보와 관련된 웹 문서의 인덱싱 정보들을 검색하여 해당 사용자 단말)로 전송되도록 웹 서버로 전달하는 기능을 수행한다.

즉, 웹 서버에 전달되는 내용은 "키워드(Keyword) : 놈놈놈, 타입(Type) : 긍정/부정/의견"이 될 수 있다. 여기서, 상기 타입 정보 중에서 "의견"이라 함은 긍정 및 부정 의견이 모두 함께 나타나는 검색 결과이며, "긍정"이라 함은 긍정 의견만 나오는 타입이다. "부정"이라 함은 부정 의견만 나오는 타입이다. 이와 같이 특정 의견검색 키워드와 타입을 의견 검색 모듈에 전달하게 되면, 인덱싱 서버 또는 인덱싱 정보 저장모듈에서 해당 특정 의견검색 키워드와 해당 타입에 해당되는 데이터를 읽어 와서 의견의 양이나 날짜 순서 등의 랭킹(Ranking)으로 검색된 결과를 다시 웹서버에 전송해준다. 이때, 상기 검색된 결과 정보는 예컨대, 제목, 링크(Link), 해당 사이트 제목, 긍정 개수, 부정 개수, 긍정 개수, 본문 내용, 본문 요약 내용, 긍정 표현 위치, 부정 표현 위치 등으로 이루어질 수 있다.

3.1.7 웹 서버

웹 서버는 인터넷을 통해 접속된 사용자 단말로부터 전송되는 특정 의견검색 키워드 및/또는 타입(Type) 정보를 제공받아 의견 검색 모듈로 전달하고, 의견 검색 모듈로부터 검색된 의견 검색 결과 즉, 인덱싱 정보 데이터들을 제공받아 해당 사용자 단말의 화면에 디스플레이 되도록 인터페이스(Interface)해주는 기능을 수행한다.

3.2 의견 추출 및 의견 분석 방법

그림 3에 따른 인터넷을 이용한 웹 콘텐츠 의견 추출 및 분석 방법을 설명하면, 데이터 수집서버를 통해 인터넷 상에 존재하는 웹 문서 데이터를 수집(단계 1)한 후, 언어처리 모듈을 통해 상기 단계 1에서 수집된 웹 문서 데이터에 대해 문장 단위로 분리하고, 분리된 각 문장에 대해 언어처리(예컨대, 형태소 분석 또는 띄어쓰기 등)를 수행하여 언어적인 자질(Feature)들을 추출한다(단계 2).

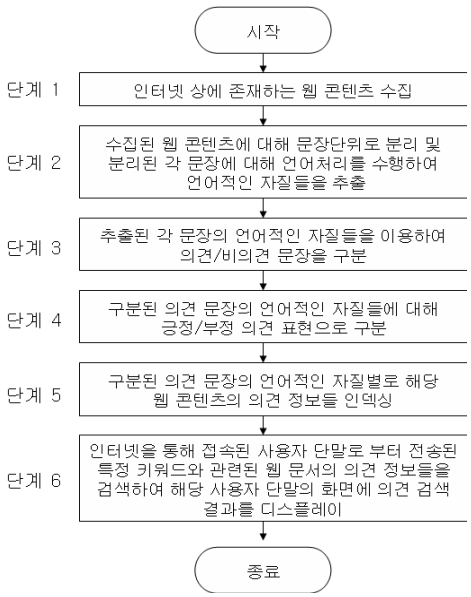

그림 3. 인터넷을 이용한 웹 콘텐츠 의견 추출 및 분석 방법

다음으로, 의견/비의견 구분 모듈을 통해 상기 단계 2에서 추출된 각 문장의 언어적인 자질들을 이용하여 의견/비의견 문장을 구분(단계 3)한 후, 의견표현 구분모듈을 통해 상기 단계 3에서 구분된 의견 문장의 언어적인 자질들에 대해 긍정/부정 의견표현으로 구분한다(단계 4).

이후에, 인덱싱 서버를 통해 상기 단계 4에서 구분된 의견 문장의 언어적인 자질별로 해당 웹 문서의 의견 정보들이 의견 인덱싱 정보 저장모듈에 저장되도록 인덱싱(Indexing)을 수행한다(단계 5). 여기서, 상기 단계 5에서 인덱싱된 각 의견 문장의 언어적인 자질별 해당 의견 문장의 요약정보 및 해당 웹 문서의 기본 및 의견 정보들을 데이터베이스(DB)화하여 별도의 의견 인덱싱 정보 저장모듈에 저장함이 바람직하다.

다음으로, 의견검색을 원하는 사용자는 인터넷 접속이 가능한 사용자 단말을 이용하여 의견검색 서비스를 제공하는 특정의 웹 페이지에 접속하면, 웹 서버는 의견검색을 위한 검색 입력창 및 의견검색 타입(의견/긍정/부정)을 선택하는 타입선택 버튼들을 구비한 메인 검색화면을 제공한다. 이러한 의견검색 서비스 환경에서, 사용자가 원하는 의견검색 키워드를 검색 입력창에 입력한 후, 타입선택버튼들 중 어느 하나의 버튼을 클릭(선택)하면, 웹 서버는 인터넷을 통해 접속된 사용자 단말로부터 전송되는 특정 의견검색 키워드 및/또는 의견검색 타입을 제공받아 의견검색모듈에 전달한 후, 의견검색모듈은 웹 서버를 통해 전달받은 상기 특정 의견검색 키워드와 관련된 웹 문서의 의견 정보들을 인덱싱 서버 또는 의견 인덱싱정보 저장모듈에서 검색하고 그 의견 검색결과를 웹 서버로 다시 전달한다. 이후에, 웹 서버는 의견검색 모듈을 통해 검색된 상기 특정 의견검색 키워드에 대한 의견 검색결과를 해당 사용자 단말의 화면에 디스플레이 해준다(단계 6).

Ⅳ. 성능비교 및 분석

제안한 인터넷을 이용한 통계기반의 웹 콘텐츠 의견 추출 시스템은 국외 'JODANGE' 와 국내 '네이버'가 개발을 진행하고 있으나, 주요 핵심기술인 구체적 의견 어구를 추출하는 기술과 정보 모니터링 구현 기술은 본 논문이 설계 및 구현 중에 있으며 본 논문과 국내외 기술간의 의

견정보 추출 기술 비교를 하면 표 2와 같다.

표 2. 의견정보 추출기술 비교

비교목록	국외 (Jodange)	국내 (네이버)	본 논문
대상 정보	S&P 500 Company 관련 뉴스 의견 정보	10만여개 네이버 문서 (블로그)	블로그, 뉴스 매체, 커뮤니티 등 전체 인터넷
의견 모니터링 지원	0	X	100% 지원
의견 문서 검색 지원	X	0	100% 지원
의견 어구 추출	0	X	100% 지원
긍정/부정 어구 추출	0	X	100% 지원
대상 언어	영어	한국어	한국어/영어
교차언어 의견추출	X	X	100% 지원
의견추출기 학습 방법	Supervised	Supervised	Semi-Supervised
개발 완료 시기	2008.2.23	미정	2009.06.30(예정)

또한, 제안한 인터넷을 이용한 통계기반의 웹 콘텐츠 정보 분석을 통한 모니터링 기술은 구글이 단순히 영어 정보를 한국어로 번역하여 찾아주는 단순 검색 수준인데 반해서, 본 논문에서 제안한 기술은 한국어 정보를 영어 및 기타 다국 언어 정보로 실시간 모니터링을 해주며 사용자들의 다국어 의견정보를 통계 분석하여 제공하는 방식이다. 이를 비교하여 정리하면 표 3과 같다.

표 3. 정보 모니터링 기술 비교

비교목록		국외 (구글)	국내 (네이버)	본 논문
실시간 데이터 수집 기술	수집된 데이터 범위	전 세계	국내 데이터 (자체 데이터 위주)	전세계
	실시간성	X	X	0
다국어 검색 기술	외래어 사전 자동구축	수동	수동	100% 자동
	대상 미디어	텍스트	텍스트	텍스트, 이미지, 동영상
	대상 언어	한 - 영	한 -일	한 - 영
	다국어 번역	통계기반	규칙기반	통계기반
정보 모니터링 기술	다국어 의견 추출 기술	X	X	0
	정보 알림 기술	X (email알림 수준)	X	0

V. 결 론

본 논문에서는 인터넷 상에 존재하는 여러 웹사이트들에 흩어져 있는 웹 문서에서 사용자 의견 정보들을 자동 추출 및 분석함으로써, 긍정/부정 의견별로 검색 및 통계를 확인할 수 있는 의견 검색 서비스를 간편하게 할 수 있으며, 의견 검색 사용자들은 특정 키워드에 대하여 다른 사용자들의 의견을 손쉽게 한눈에 검색 및 모니터링(Monitoring)하는 웹 콘텐츠에서의 통계기반 웹 마이닝을 통한 의견 추출 및 분석 시스템의 설계 방법을 제안하였다.

본 논문의 기대효과는 인터넷 상에 존재하는 여러 웹사이트들에 흩어져 있는 사용자 의견 정보들을 자동 추출 및 분석하여 긍정/부정 의견별로 검색 및 통계를 확인할 수 있도록 의견 검색 서비스를 제공해 줌으로써, 사용자들은 특정 키워드에 대하여 다른 사용자들의 의견을 손쉽게 한눈에 검색 및 모니터링 할 수 있으며, 기존에 다른 사용자들의 의견을 검색하기 위해서 들였던 많은 시간을 크게 단축시킬 수 있는 이점이 있다. 또한, 각 회사의 마케팅 담당자나 주식 투자자, 기업 가치 평가자 등은 방대한 인터넷 상에서 존재하는 해당 기업이나 물품에 대한 여러 사용자들의 의견을 한눈에 확인할 수 있으며, 기존에 사용자들의 의견을 알기 위해서 실시했던 설문조사나 컨설팅 회사에 들였던 비용을 대폭 줄일 수 있으면서 효과적으로 각 사용자들의 의견 추출 및 통계를 내서 활용할 수 있는 이점이 있다.

향후의 과제로는 언어장벽을 해소시켜 모국어로 다른 나라 정보를 모니터링 할 수 있게 하기위한 다국어(한,중.일.영) 검색 및 기계번역 기능을 추가하여 완전한 모니터링 검색엔진을 위한 웹 컨텐츠 의견 검색 시스템을 만드는 것과 의견 모니터링에 대한 신뢰성을 검증하는 것이 필요하다.

참고문헌

[1] 김영만, "통신서비스 시장에서 데이터마이닝을 이용한 이탈고객 분석", 한국과학기술원 석사논문, 1998.

[2] "다변량 데이터의 통계분석", 석정, 2003.

[3] 장남식, 홍성완, 장재호, "데이터마이닝", 대청, 2005.

[4] S. Anand, D. Bell, J. Hughes, "The Role of Domain Knowledge in Data Mining", CIKM 95, 1995.

[5] S. Anand, J. Hughes, "Hybrid Data Mining Systems: The Next Generation", PAKDD '98, Melbourne, Australia, 1998, pp. 13-24.

[6] R. Agrawal, T. Imielinsk, A. Swami, "Mining Association Rules between Sets of Items in Large Databases", In Proceedings of the ACM SIGMOD Conference on Management of Data, Washington D.C., May, 1993, pp. 207-216.

[7] P. Adriaans, D. Zantinge, "Data Mining", Addison Wesley Longman, England, 1996.

[8] J. Berry, G. Linoff, "Data Mining Techniques: For Marketing, Sales, and Customer Support", John Wiley & Sons, 1997.

[9] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", ACM SIGKDD, July, 2000.

[10] B. Lent, R. Agrawal, R. Srikant, "Discovering Trends in Text Databases", In Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August, 1997.

[11] C. H. Lee, H. C. Yang, "A Web Text Mining Approach Base on Self-Organizing Map", In Proceedings of the 2nd International Workshop on Web Information and Data Management, WIDM'99, Kansas City, MO, USA, 1999, 59-62.

[12] M. Mulvenna, S. Anand, A. Büchner, "Personalization on the Net using Web Mining", Communications of the ACM, Vol. 43, No. 8, August, 2000.

[13] B. Mobasher, R. Cooley, J. Srivastava "Automatic Personalization Based on Web Usage Mining", http://maya.cs.depaul.edu/~mobasher/personaliza tion/, 1999.

Web Contents Mining System for Opinion Information Searching Engine

Hae Jong Joo, Young Bae Park, Hae Gil Choi

Abstract

This research is about the design of an opinion drawing and analysis system through statistical based Web Mining of web contents, where data of opinions are automatically drawn and analyzed concerning web documents that are scattered around in various web sites that exist within the internet. Furthermore, provides a search service that can easily classify positive/negative opinions and also provide searching and statistical information. Users, who want to search for opinions, can input a specific keyword to observe opinions of others easily. In addition, there is a merit in materializing the monitoring system.