韓國 컴퓨터情報學會 2008 제38차 하계학술발표논문집
第16卷 第1號, 2008. 06.

인터넷 여론 정보수집시스템과 관련 국내외 연구 동향 분석

김 시 우*

Internet based opinion collection System with current text filtering techniques survey

Sea Woo Kim*

요 약

웹상에서 자동 데이터 추출과 분석기법은 최근 검색분야의 주요이슈이다. 본 논문은 웹상의 자동 설문조사 시스템에 관한 연구이다. 그리고 기존의 Corpus의 성향을 분석하고 검색 및 분석 시스템의 항목들을 정의하였다. 또한 Corpus를 이용한 웹 검색 및 분석 시스템의 활용 분야를 기술하고 향후 개발 방향을 기술하였다.

▸ Keyword : Corpus, Web Search, BLOGs

• 제1저자 : 김시우
* 숭의 여자대학 인터넷 정보과 조교수

Ⅰ. 개발 기술 개요

1. 개발기술의 개요

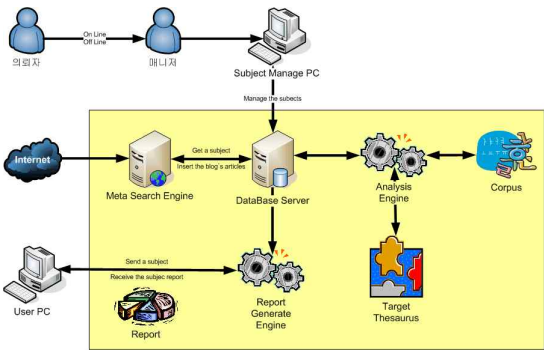
현재 기업들이 세계화된 경쟁구도, 빠른 제품 개발기간, 다양한 고객의 욕구 등의 비즈니스 환경에서 살아남기 위해서 실시간 경영 시스템의 구축에 회사의 사활을 걸고 있다. 특히 기업의 입장에서 고객 즉, 시장이 없이는 어떠한 기업도 생존할 수 없으며 이전의 고객만족경영으로부터 고객가치경영을 실현하고 있는 추세다.

그러나 아직까지 고객의 실질적인 요구 및 반응에 대한 정보를 조사 한 후에 알기 쉽게 정보를 효과적으로 제공해주는 시스템이 없으며 대기업들 정도가 간간이 많은 비용과 시간을 소비하여 여론조사에 대한 정보를 취합하고 있는 실정이다. 이리하여 기업들이 불필요한 비용 지출이 생기게 되는 것이다.

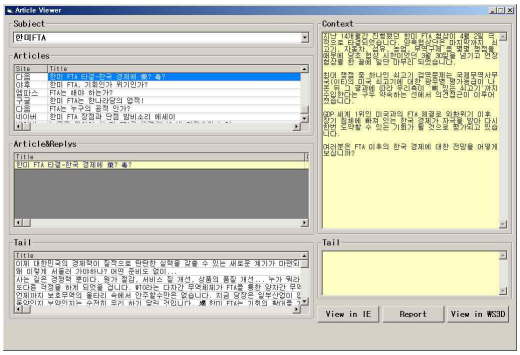
개발하고자 하는 시스템은 실시간으로 특정 대상(회사, 제품, 인물 등)에 대해 특정 웹사이트, 블로그(Blog), 지식검색 웹페이지 등 인터넷 상에 고객(네티즌)들이 올려놓은 정보를 취합, 이를 가공, 분석하여 사용자에게 보고하는 시스템으로 전체 과정이 자동화된 인터넷 모니터링 시스템이다.

1-1. 시스템 구성도

[그림1]은 개발하고자하는 논리적 시스템 구성도로 사용자가 찾고자하는 정보를 입력하면 어휘 분석기, 분석 엔진, 통계평가를 거쳐 그래픽으로 레포트 생성을 하는 구성도이다. 어휘 분석은 기존에 존재하는 코퍼스(Corpus)를 이용할 수 있다.

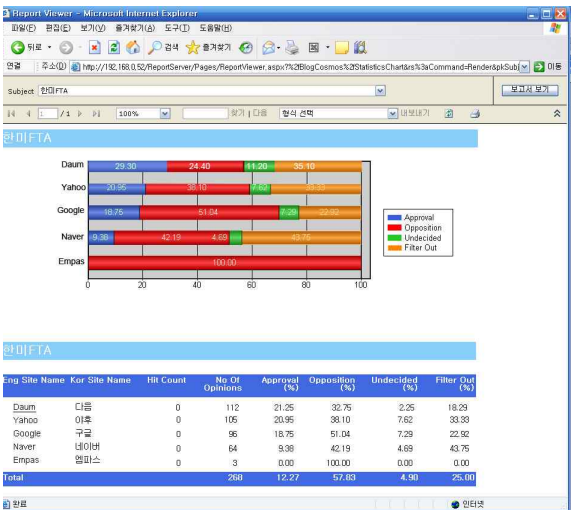


[그림 1] 시스템 구성도

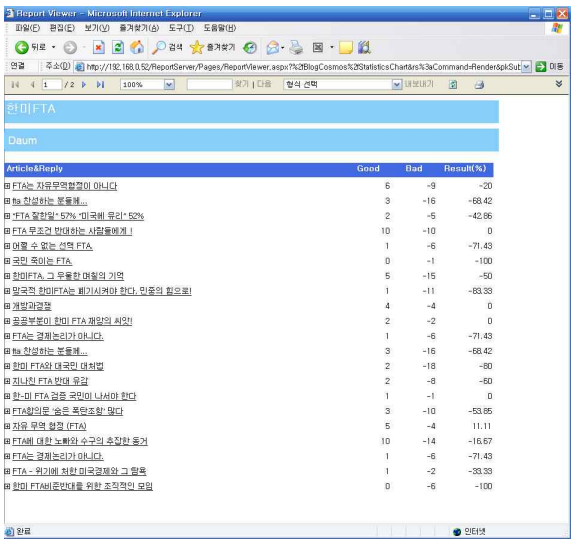


[그림 2] Demo GUI 수집내용

[그림2]는 Demo GUI 수집내용으로 사용자가 찾고자하는 질문을 Subject에 입력하면 각 웹사이트에서 관련된 기사를 Articles에 가져오고 가져온 기사의 답변을 Article&Replys에 나타내며 그 답변의 꼬리말을 Tail에서 나타내준다. Context는 Articles의 목록 중 하나를 선택시 보여주는 기사의 내용이며 [그림3]은 2차원의 그래프 및 수치 형태인 각 웹사이트의 분석 정보를 볼 수 있고 [그림4] 싸이트별 링크를 클릭 하였을 때 각 사이트 별로 수집된 기사의 제목을 볼 수 있다. MSSQL2005 reporting service 프로그램을 이용하여 2차원형태의 표로 만들어 주었다. 자료 수집된 subject를 선택하고 보고서보기 버튼을 클릭하면 [그림3]의 화면이 나타나며 Eng site의 밑줄되어진 Hani를 클릭시 [그림4]의 화면이 보여진다.



[그림 3] 분석 정보



[그림 4] 수집된 기사 제목

따라서, 위의 시스템을 통해서 급변하는 비즈니스 환경하에서 기업의 경쟁력 향상 및 빠르고 올바른 기업의 의사결정이 가능해짐으로 인해 기업과 소비자 모두에게 커다란 가치와 이익을 것으로 기대되며, 특히 중소기업의 경쟁력향상에 크게 기여할 수 있을 것이다.

1-2 웹상 자료 분석 기술

웹상 자료의 자동 분석은 매우 중요한 핵심기술로 미래 Data Mining의 중심과제이다.이에 관련하여 댓글을 통한 자료의 언어처리를 이용한 자동분석은 저비용 고효율 여론 조사 방법이다.

나라와 나라 간의 교류를 원활하게 하는 하기 위해서는 언어 문제가 해결되어야 하며 이 때문에 상대국의 언어를 효과적으로 연구, 교육하는 것은 국가 경쟁력을 높일 수 있는 주요 수단이 된다. 선진 각국에서는 이러한 점을 미리 깨닫고 자국어를 바탕으로 하여 상대국의 언어 정보를 대규모로 수집, 데이베이스화 하여 여러 가지 용도로 활용하고 있는데 이러한 노력의 하나가 대규모 코퍼스 구축이다.

이는 언어간의 효율적인 정보 소통을 위한 통역, 번역 등의 실용적인 측면뿐만 아니라 언어간의 상이한 특징을 비교하는 언어의 대조 분석 연구라는 학문적인 측면에서도 중요한 역할을 한다. 서구선진국가들(특히 유럽의 국가들)에서는 말뭉치의 중요성을 인식하여 자국어와 세계 공용어인 영어를 대상으로 코퍼스를 구축하는 프로젝트를 활발하게 진행하고 있으며 중국에서도 중일/일중 코퍼스를 구축하여 학계에 연구자료를 제공하고 있는 추세이다.

이에 본 시스템을 통한 개발대상기술(제품)은 고객의 제품 사용 후기, 만족/불만 사항, 실제 사용하는 고객에 대한 연령별, 성별, 지역별 등의 정보와, 제품 홍보 상태, 실제 판매가격, 인기순위 등의 정보를 취할 수 있으며, 특히 경쟁제품과의 비교정보를 쉽게 파악할 수 있다.

우선적으로 여론조사 시스템에 있어서 사용자에게 보다 높은 부가가치를 보다 낮은 비용으로 제공해주는 혁신적인 시스템이다. 단순히 고객에 대한 정보를 파악하여 고객 서비스를 잘 하는 것으로 끝나는 것이 아니며, 잘못된 제품개발에 따른 자원의 낭비와 그에 따른 소비자의 불이익을 생각해 본다면 그 파급효과에 대해서는 이론의 여지가 없다 하겠다.

또한, 중소기업에서도 적은 비용으로 가장 취약한 부분인 신제품 기획과 마케팅 능력의 혁신적인 향상을 기대할 수 있다

II. 국내・외 기술동향

가. 국내・외 기술(제품) 동향

현재 기술은 미약하나 국내에 Corpus 관련한 연구가 있는 상황이다. 특히 자연어 처리 연구분야 중에서 Corpus(코퍼스) 연구경향은 다음과 같다.

1) 국내 코퍼스(Corpus)

'코퍼스'란 언어를 연구하는 각 분야에서 필요로 하는 연구 재료로서 언어의 본질적 모습을 총체적으로 드러내 보여 줄 수 있는 자료의 집합을 뜻한다(국어정보학 입문, 서상규, 한영균저). 한국어로는 말뭉치 또는 말모둠으로 번역하여 사용하는데 그 정의는 사람에 따라 다르나 대략 다음과 같은 의미로 쓰이고 있다.

- ○ 대규모 언어 데이터베이스
- ○ 인간의 음성언어(문어, 구어)를 대용량 컴퓨터에 저장하고 이를 필요에 따라 가공하여 언어 연구에 사용하는 것
- ○ 컴퓨터가 판독할 수 있는 형태(Mashine-readable form)로 저장된 자연어의 용례들과 이들 용례에 대한 부속정보(additional information)

컴퓨터에 의하여 대규모 코퍼스가 구축된 것은 1963년에 완성된 브라운코퍼스(100 어절)가 처음이다. 1990년 이후 1억 어절 이상의 대규모 코퍼스(British National Corpus)가 구축되었고, 국내에서도 연세대(1988), 고려대, 카이스트, 국립 국어연구원에서 한국어코퍼스를 구축하고 있다. 1998

년부터 시작된 '21세기 세종계획'이 2007년 완성되어 한국은 세계적 규모의 국가 코퍼스를 가지게 되었다.

① 21세기세종계획
: 21세기 세종계획은 우리 나라의 선진 정보문화를 자주적으로 구현할 수 있는 국어정보화 중장기 발전계획의 수립을 그 목적으로 1998년 국립국어연구원이 중심이 되어 출발

② 고려대 한국어코퍼스
: 민족문화연구원 - 전자텍스트 연구소 연세대한국어코퍼스 언어정보개발 연구원 - 한국어사전편찬실

③ KAIST한국어코퍼스: KAIST 국어정보베이스KAIST 국어정보베이스

2) 국외 코퍼스(Corpus)

가) 아마존 (www.Amazon.com)

아마존에서는 책에 대한 review에 대해 매우 큰 데이터베이스로부터 사전에 알려지지 않은, 유용한 정보를 추출하는 지식 발견 방법인 Data mining을 시도한 사례가 있다. 다시 말해 기업이 보유하고 있는 일일 거래 데이터, 고객 데이터, 상품 데이터 혹은 각종 마케팅 활동에 있어서의 고객 반응 데이터 등과 이외의 기타 외부 데이터를 포함하는 모든 사용 가능한 근원 데이터를 기반으로 감춰진 지식, 기대하지 못했던 경향 또는 새로운 규칙 등을 발견하고 이를 실제 비즈니스 의사결정 등에 유용한 정보로 활용하고자 하는 것이 바로 Data Mining이다.

나) Virtual Language Centre (Grammar zone 연구팀)

어휘와 문법을 기반으로 한 컨텐츠가 다양한 편이다. Web Concordancer에서 ConcApp 프로그램(Monoconc와 비슷한 프로그램)을 무료로 다운받을 수 있고) 이곳에서 사용자 개인의 corpus를 넣고 검색할 수 있다.

다) indexed corpus(색인 코퍼스)

학술서나 연구재료가 될 서적에서는 없어서는 안 되는 중요한 부분이다. 항목선정의 기준이 되는 indexed corpus(색인 코퍼스)는 해당되는 저작물에 대한 항목의 중요도에 있으며 지나치게 상세해도 오히려 색인 본래의 목적을 해칠 우려가 있을 수 있다. "Virtual Language Centre"에서는 전체 코퍼스 중에서 몇 개를 알파벳순으로 정리하여 원하는 단어를 검색할 수 있는 기능이 있어 편리하게 사용할 수 있다.

라) Tom Cobb's Compleat Lexical Tutor at Quebec University

Online Concordancer는 BNC를 비롯한 다른 corpus들 포함하고 있어 앞의 concordancer와 함께 이용하면 더 풍부한 corpus file을 활용할 수 있다.

(http://132.208.224.131/concordancers)

III. 개발 시스템의 목표

국내외 연구 동향을 바탕으로 개발한 시스템은 다음의 효과를 목표로 한다.

가) 기업이미지 혁신

기업이미지는 내부적으로는 새로운 기업문화를 정립, 내적 역량을 결집시켜 성장의 발판을 마련하고 외부적으로는 국민과 소비자에게 비춰지는 신뢰와 믿음에 대한 지표이다. 따라서 각 산업 또는 기업들의 대내외적 경영활동이 국민과 소비자들에게 비춰지는 현 위치를 확인하고 경쟁력 확보를 위한 기초자료로 활용할 수 있도록 하는데 목적이 있다.

나) 제품 평가 및 고객 만족도

기업에서 고객에게 제공하는 상품 및 서비스에 대해 고객이 인지하는 고객만족도 평가를 통하여 경쟁사와의 비교평가는 물론 고객에게 제공하는 제요소에 대한 고객만족 수준을 제고하기 위한 제반 정책을 수립하여 제품에 대한 고객만족도를 조사 할 수 있다. 제품 이용자 만족도, 서비스 이용 만족도, 제품별/매장별 고객 이용자 만족도 평가, Mystery Shopping 조사, pack test를 통한 만족도 조사 등에 활용할 수 있다.

다) 브랜드 시장성 구축

브랜드는 기업의 주요 고객집단에게 해당기업의 가치를 전달하는 핵심도구로, 브랜드 자산의 지속적인 관리전략을 통해 기업가치 증대에 기여하는 브랜드 개발이 요구된다. 이에 따라 마케팅과 기업전략 차원에서의 브랜드 전략 수립을 위한 기초자료 제공을 브랜드 조사의 목적으로 두고 있다.

라) 시장성 분석(위험요소 및 성장성 분석)

신규사업 및 제품 출시 이전 댓글을 통해 예상 고객에 대한 기초 데이터를 수집하고 객관적인 수요를 파악함으로써 예상 고객층 설정 및 제품/서비스 컨셉별 수요변화 파악을 통해 제품/서비스의 구체적인 제공방향을 도출한다. 이러한 종합적인 결과의 분석과 수요예측을 통해 사업성 평가 및 타당성을 분석할 수 있다. 또한 제품 및 서비스 강·약점 분석, 사업모델 비교·분석, 위험요소 및 성장성 분석, 수요 예측 등을 할 수 있다.

마) 비교 검색

최근 소비자들 사이에서는 원하는 상품의 성능 및 가격 등

을 인터넷을 통해 비교해보고 구매하는 '지능형 쇼핑'이 일반화되고 있다. 예전에는 오프라인에서 직접 제품을 탐색하고 인터넷 가격비교사이트에서 가격을 비교한 뒤 전문쇼핑몰에서 구매하였지만, 현재는 쇼핑을 위한 탐색 자체를 인터넷에서 하는 경향이 강해져 네이버, 구글, 엠파스 등 각 포털 업체들의 비교쇼핑 서비스 성장세가 커지고 있다. 포털의 쇼핑 서비스는 저렴하고 품질 좋은 제품을 구매하려는 욕구, 구매 의사결정에 영향을 미치는 풍요로운 정보를 제공함으로써 비교쇼핑 서비스가 큰 호응을 얻고 있다.

Ⅳ. 결론 및 향후 연구

우리 시스템은 웹상의 설문조사 시스템의 기초 개발물이다. 이를 계속하여 보완 개발 하려고 한다. 국내에서도 후기(댓글)를 통해 제품을 평가하는 리뷰사이트가 급증하고 있으며 이 리뷰서비스는 중요한 마케팅 수단으로 부각되고 있다. 이에 본 연구는 네티즌의 댓글에 관한 연구와, 구문 분석, 어휘 분석이 점점 더 필요하며 이를 요약하여 제품 평가 및 소비자 모니터링 기술들을 활용 하려고 한다.

향후 연구는 Corpus 를 우리 시스템에 접목시키고 Thesaurus 를 추가 할 것이다. Ontology를 이용한 각 사용자의 성향을 분석한 맞춤 서비스 구축도 향후에 해야 할 과제이다.

참고문헌

[1] MDinopoulos, Elias, Lewis, Trace R., and David E. M. Sappington 1995. "Optimal Industrial Targeting with Unknown Learning- by-Doing." Journal of International Economics 38: 275-295.

[2] Klimenko, Mikhail, 2004. "Industrial Targeting, Experimentation and Long-run Specialization." Journal of Development Economics 73: 75-105.

[3] Bardhan, Pranab, 1971. "On Optimum Subsidy to a Learning Industry: An Aspect of The Theory of Infant-Industry Protection." International Economic Review 12: 54-70.

[4] Dixit, Avinash, and Gene Grossman. 1986. Targeted Export Promotion With Several Oligopolistic Industries. Journal of International Economics 21: 233-50.

[5] Brin and L.Page, The anatomy of a large-scale hypertextual Web search engine, In Proceedings of the Seventh International World Wide Web Conference; 1998

[6] Evenett Simon. "Study on Issues Related to a Possible Multilateral Framework on Competition Policy," WTO paper WT/WGTCP/W228.

[7] 임신영(ETRI), 박효준(마스시스템), 윤우성,김태윤(고려대). DLC를 이용한 디지털 데이터의 불법복제 방지 시스템 연구, A Study on DLC-based Protection System against Illegal Reproduction, 한국 정보처리학회 추계 학술발표논문집 1999

[8] 임신영,함호상(ETRI), 박효준(마스시스템),윤우성,김태윤(고려대). 디지털 상품의 유통정보 관리기술, Information Management Technology for Online Computer Program, 전자거래학회/한국정보시스템학회 종합학술대회99

[9] 강상승, 임신영, 함호상(ETRI), 박효준(마스시스템), 김태윤(고려대). MP3 미디어데이터의 온라인 유통기술, On-Line Delivery Scheme for MP3 Media Data, 전자거래학회/한국정보시스템학회 종합학술대회99

[10] 박효준(마스시스템),강우준(성균관대). 세계디지털 상품 유통기술 동향과 S.O.Shop, 한국 정보처리학회지 3월호 제7권 2호, 2000년 3월

[11] 강우준, 김응모(성균관대). 디지털 상품의 온라인 유통을 위한 동적 사용권 관리 기술, 한국 정보처리학회 춘계학술대회(4월 15일)

[12] 강우준, 김응모(성균관대). 동적 사용권 관리를 이용한 소프트웨어 상품의 온라인 유통, 정보처리학회 특집호(2000년 6월)