뉴스 댓글을 통한 인물 인지도 추출

류준석, 김원영, 김응모
성균관대학교 전자전기컴퓨터공학과
e-mail : jsryu@ece.skku.ac.kr

Mining Reputation of People Using Reply of News Article

Joonsuk Ryu, Won young Kim, Ung mo Kim
* Dept of Electrical and Computer Engineering, Sungkyunkwan University

요 약

인터넷의 보편화와 사용자 증가는 사회에 많은 변화를 가지고 왔다. 많은 변화 중 인터넷을 통한 뉴스 제공은 종이 신문과는 다르게 인터넷 사용이 가능한 모든 사람들에게 뉴스를 제공 받을 수 있게 되었으며 언제든 원하는 기사를 다시 제공 받을 수 있게 해주었다. 이러한 이유로 인터넷 뉴스는 다양한 연령대의 사용자들이 뉴스를 접할 수 있게 되었고 인터넷 뉴스를 읽은 많은 사용자중 해당 뉴스에 댓글을 남기게 되었다. 이러한 댓글은 사용자의 의견을 내포하고 있는 것으로 본 논문에서는 사용자들이 남긴 댓글에 오피니언 마이닝을 적용하여 사용자 의견을 추출하여 특정 인물에 대한 인지도를 찾아내는 기법을 제시한다.

1. 서론

인터넷 뉴스 이용자가 증가함으로 예전 종이 신문을 이용할 때 와는 다르게 해당 뉴스에 대한 자신들의 의견을 자유롭게 표현할 수 있게 되었다. 이러한 뉴스 사용자의 의견은 현재 뉴스에서 중점이 되는 부분의 장단점을 알 수 있게 되며 이 장단점을 수치화하여 뉴스의 대상의 되는 인물의 인지도를 알아볼 수 있다. 또한 인터넷 뉴스는 이벤트가 벌어진 후 매우 짧은 시간 내에 뉴스가 작성되고 사용자들이 그 뉴스를 확인 할 수 있으며 같은 이벤트도 서로 다른 내용의 뉴스로 작성될 수 있기 때문에 이벤트 발생 후 사용자들의 다양한 의견을 추출해 낼 수 있다.

특정 인물의 인지도를 측정하기 위해서 지금까지는 직접적인 설문조사로 이루어 졌다. 하지만 많은 사람들을 대상으로 설문조사를 실시하기 위해선 많은 노력과 시간이 필요로 할뿐만 아니라 다양한 사람들의 의견을 모으기란 매우 어려운 일이다. 따라서 본 논문에서는 특정인물을 정하여 그 인물을 대상으로 작성한 뉴스에 사용자들의 의견을 추출하는 기법인 오피니언 마이닝을 적용시켜 사용자들의 의견을 추출한 후 추출된 의견을 수치화하고 인물에 대한 인지도를 알아볼 수 있는 기법을 제시하며 기반이 되는 언어는 영어로 하고 있다.

본 논문의 구성은 다음과 같다. 2 장에서는 배경지식을 간략하게 설명하고 3 장에서는 우리가 제안하는 방식을 설명하게 되며 4 장에서는 결론과 발전 방향에 대해서 제시한다.

2. 배경 지식

본 연구와 연관이 있는 기법은 ParseTree, Association Rule Mining 과 Latent Semantic Analysis 이 있다.

2.1 Association Rule Mining

Association Rule Mining 은 Agrawal et al[1]에 의해서 제안된 알고리즘이며 큰 데이터베이스 내에 존재하는 트랜젝션중 우리가 원하는 패턴을 찾아내는 기술이다.

Association Rule 은 $X \Rightarrow Y$ 로 나타내며 X 는 $Y \subset I$ 와 $X \cap Y = \phi$ 이다. 여기서 rule 을 찾아내기 위해선 두 가지 threshold 를 사용한다. 첫 번째는 Minimum support 로 $(X \cup Y)/N$ 로 표현한다. 즉, 전체 트랜젝션중 X 와 Y 를 모두 가지고 있는 트랜젝션의 개수를 나타난다. 두 번째는 minimum confidence 로 $sup(X \cap Y)/sup(X)$로 표현한다. 즉, X 를 포함하고 있는 트랜젝션중 X 와 Y 를 모두 포함하고 있는 트랜젝션의 개수를 나타낸다. Association Rule Mining 은 기본적으로 두 가지 단계로 이루어져 있다. 첫 번째 단계에서는 트랜젝션내에 존재하는 itemset 중 frequent 한 것들을 추출해내는 단계로 frequent itemset 이란 사용자가 정의해 놓은 minimum support 를 만족시키는 itemset 이다. 다음 단계에서는 이전 단계에서 찾아낸 frequent itemset 을 가지고 rule 을 만들어 낸다. <표 1>은 Association Rule Mining 의 예제이며 minimum support 는 0.4 이다. Minimum support 를 만족하는 룰은 {milk, bread}, {bread, butter}가 된다. 본 논문에서는 frequent itemset 을 찾는 단계까지만 이용된다[2].

<표 1> 트랜젝션 데이터

Transaction ID	milk	Bread	butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

2.2 Part-Of-Speech Tagging

POS tagging 은 자연어 처리 기법을 사용하여 한 문장을 입력 받아 그 문장을 분석하여 단어에 명사, 동사, 형용사 등 품사를 저장하는 기법이다. 이 기법을 통하여 문장의 구성 요서를 파악 할 수 있다. 예를 들면 “When I watched this film, I hoped it ended as soon as possible” 라는 문장이 POS tagging 프로그램에 입력되면, 아래와 같은 결과가 출력된다.

“When/WRB I/PRP watched/VBD this/DT film,/NN I/PRP hoped/VBD it/PRP ended/VBD as/IN soon/NN as/IN possible/JJ.”

POS tagging 를 통한 Opinion mining 적용사례는 [3]에서 찾아 볼 수 있다.

<표 2> POS Tag 의 예

POS tag	Description
JJ	adjective
JJS	adjective, superlative
NN	common noun
CC	coordinating conjunction
DT	determiner

2.3 Latent Semantic Analysis

Latent Semantic Analysis(LSA)는 개념적으로 co-occurrence 정보를 이용한다. co-occurrence 정보를 이용한다는 것은 단어의 '형태(morphology)가 아닌 의미(semantic)'를 이용한다는 뜻으로 '배'라는 단어는 같은 문장에 co-occur 하는 동사가 '타다' 인지 '먹다' 인지에 따라 의미가 갈라지게 된다. 또한, '식당', '맛있게', ' 배부르게' 라는 단어와 같은 문장에 co-occur 하는 처음 보는 단어는 아마 '음식'일 것이다. LSA 는 이론적으로는 선형대수학의 SVD(Singular Value Decomposition)을 이용한다. SVD 는 단어-by-문서 행렬 A 를 3 개의 행렬로 분해하는 것으로 $A_{t\times d} = T_{t\times n} S_{n\times n} (D_{d\times n})^T$ 와 같이 T, S, D 로 분해하게 되며 t 는 단어 개수, d 는 문서 개수 n 은 min(t, d) 이고 T 는 단어, D 는 문서에 대응 되는 행렬이 된다. 그림 1 과 그림 2 는 SVD 의 예제이다.

그림 1 SVD 적용 전

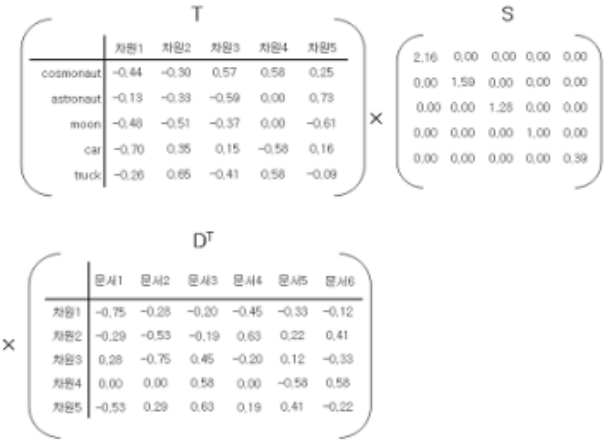

그림 2 SVD 적용 후

위의 행렬을 이용할 경우 단어-단어, 문서-문서, 단어-문서 관계를 알아볼 수 있게 된다. 본 논문에서는 단어와 단어 사이의 유사도 T x S 행렬의 row 간의 유사도로 계산한다.

3. 의견 추출

이번 장에서는 뉴스와 뉴스의 댓글을 이용하여 의견을 추출하고 추출된 의견을 수치화하여 특정 인물에 대한 인지도를 찾아내는 방법을 설명한다.

3.1 뉴스 수집

인터넷 뉴스는 짧은 시간에 많은 뉴스가 제공되기 때문에 사용자의 편의를 위하여 뉴스는 제공과 함께 자신이 속해있는 카테고리로 분류되게 된다. 본 연구에서는 이러한 점을 이용하여 뉴스와 댓글 수집 시 인지도를 찾고 싶은 인물이 속해 있는 이벤트 카테고리를 찾아 뉴스와 그에 해당하는 사용자 댓글을 수집한다.

3.2 ParseTree 생성과 후보 댓글 생성

3.1 장에서 수집된 뉴스와 댓글중 사용자의 의견을 가지고 있는 댓글에 ParseTree 기법을 적용한다. ParseTree 는 POS Tagging 와 같이 한 문장을 입력 받아 문장의 품사를 나타내주는 기법으로 POS Tagging 과는 다르게 문장을 Tree 형식으로 보여주는 기법이다 [4]. 예를 들면 “He studies linguistics at the university”라는 문장이 입력될 경우 그림 1 과 같은 Tree 를 얻을 수 있다..

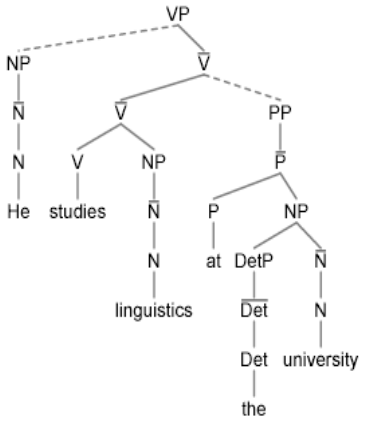

그림 3 ParseTree 예제

위와 같은 Tree 가 생성되는 과정에서 우리는 형용사의 존재 여부를 확인하게 된다. 형용사는 우리가 사물을 표현할 때 사용하는 품사로 사람들의 장/단점 또한 형용사로 표현이 된다. 이러한 형용사가 존재하지 않는 댓글은 인물의 인지도를 찾을 때 사용되지 않는 댓글이기 때문에 제거 대상으로 지정해 놓는다.

3.3 후보 특성과 의견 추출

ParseTree 과정을 거친 Tree 로 구성된 댓글을 이용하며 아래의 알고리즘을 사용하여 추출하게 된다.

```
Input: ParseTree
Output: Nouns, Adjective, Adverb
Steps:
1. Search ParseTree until JJ is detected
2. Let Opinion, Recent  = jj
3. Find first NP ascendant of detected JJ before get to first ascendant S
4. If NP exist as ascendant of JJ
  4.1. Let Recent = NP
  4.2. Let Feature = Recent without jj
  4.3. If RB exist as sibling of Recent or Opinion
    4.3.1. Let RB = RB sibling of Recent or Opinion
5. Else If sibling of JJ is S
  5.1. Let Recent = S
  5.2. Let Feature = VB descendent of Recent
  5.3. If RB exist as sibling of ADJP parent of Recent
    5.3.1 Let RB = RB sibling of ADJP parent of Recent
6. Else
  6.1. Let Recent = VP ascendant of JJ
  6.2. Feature = NP sibling of Recent
7. Store Features(NNs), Opinion(JJ) and RB as [(NN,NN,NN), ( JJ,RB)]
```

3.4 특성 제거

뉴스는 한 이벤트를 설명하기 위해서 많은 사람들의 이름이 언급하게 된다. 즉, 댓글의 의견이 우리가 찾고자 하는 사람에 대한 의견이 아닌 다른 사람에 대한 의견일 가능 성이 있다는 것이다. 또한 자유로운 인터넷 환경에서는 대상의 이름과 특징을 은유로 나타내는 경우가 많다. 이러한 점을 해결하기 위해 LSA 를 사용하게 된다. 3.3 에서 추출된 특성은 LSA 를 적용하기 전 먼저 특성으로 지정된 단어들과 비교하여 같은 단어가 있을 경우 LSA 과정을 통하지 않고 특성과 의견을 저장하게 된다. 만약 단어가 미리 설정해 놓은 특성 리스트에 존재하지 않을 경우 LSA 를 통하여 단어와 단어간의 유사도를 측정하고 유사도가 미니멈 threshold 을 만족시키지 못할 경우 제거 대상이 된다. 만약 유사도가 threshold 를 넘어갈 경우는 가장 유사도가 높은 단어를 매칭대상으로 하여 특성으로 특성리스트에 저장하게 된다.

의견을 저장할 때는 very 나 not 과 같은 부사와 함께 저장하게 된다.

3.5 Apriori Rule Mining 을 통한 특성 추출

LSA 를 통해 생성된 특성은 미리 설정해 놓은 특성만을 추출하게 된다. 그렇기 때문에 우리는 3.4 에서 제거 대상으로 지정된 댓글을 Association Rule Mining 을 이용하여 빈번하게 나타나는 특성을 찾아 낸다. 본 논문에서는 Apriori Rule Mining 의 모든 단계를 거치지 않아 Minimum Confidence 를 설정할 필요가 없이 Minimum Support 만을 이용한 후보만 생성하여 후보를 특성으로 입력하게 된다.

3.6 의견 요약

본 장에서는 앞에서 찾아낸 특성과 그와 관련된 의견을 수치화 하게 된다. 수치화는 WordNET[5]을 이용하게 된다. WordNET 은 프린스턴 대학 조지 밀러(George A. Miller)교수가 주도하는 대규모의 영어 어휘 데이터베이스 프로젝트이며 영어 어휘를 명사(noun), 동사(verb), 형용사(adjective), 부사(adverb)로 크게 나누고 이들 어휘의 동의어(synonym) 집합을 정의한 후 이들 동의어 집합(synset)간의 의미적 상관관계를 컴퓨터로 처리가능 하도록 체계적으로 정리하고 계속해서 유지 발전시키고 있다. WordNET 을 이용하여 같은 의미의 단어들의 상관관계를 알 수 있으며 이 상관관계를 이용하여 사용자의 의견의 긍정과 부정의 깊이를 알 수 있다. <표 3>은 좋다라는 의미인 단어에 대한 상관 관계를 숫자로 나타낸 테이블이다.

<표 3> 의견 값

의견	Value
Fine	0.6
Good	0.8
Excellent	1

의견을 나타내는 형용사와 Very 를 함께 사용하여 좀더 자세한 의견에 대한 값을 표현 할 수 있게 이용하게 되며 아래의 예제가 있다.

또한 특성에도 상관 관계를 두어 인물에 대한 의견일 경우 그 값을 1 로 하여 단어가 전체 특성 단어의 개수를 기반으로 현재 특성 단어가 차지한 %를 그 값으로 정하게 된다. 한 문서에 김연아라는 단어가 50 번 외모 20 번, 성격 10 번, 국민여동생 40 번, 프리스케이팅 25 번, CF 5 번 나왔다고 가정할 경우 <표 4>

와 같은 결과가 나온다.

<표 4> 특성에 따른 값

특성	Value
김연아	1
국민여동생	0.26
외모	0.13
성격	0.07
프리스케이팅	0.16
CF	0.03

위의 값을 WordNET 을 통해 구한 값에 곱하여 주어 그 값을 모두 합쳐주면 인물에 대한 인지도를 구할 수 있다. 위에 의견에 대한 값과 특성에 대한 값은 개발자와 사용자에 의해서 임의적으로 바뀔 수 있다.

4. 결과 및 향후 연구

본 논문에서는 인터넷 상에서 인물에 대한 인지도를 찾을 수 있는 자동 추출할 수 있는 방법을 제시하였다. 향후 연구로는 특성과 의미를 나타내는 값을 구하기 위해 더욱 정확한 방법을 연구하여 정확성을 높이는 방법을 연구하게 될 것이다. 또한 현재 본 논문에서 제시한 방법은 한글에 적용하기는 부적합하다. 그 이유는 아직 한글을 대상으로 한 WordNET[5]같은 단어의 semantic 을 중심으로 구축된 단어 사전이 존재하지 않을뿐더러 아직 한글의 POS tagging 을 사용한 문장 분석에 어려움을 많이 겪고 있기 때문이다. 이처럼 앞으로는 자연어 기법에 대한 연구가 더욱 많이 필요하고 WordNET 같은 단어 사전을 구축해 나갈 필요가 있다.

감사의 글

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

참고문헌

[1] Agrawal, R. and Srikant, R. 1994. “Fast algorithm for mining association rules.” VLDB’94, 1994

[2] Minqing Hu and Bing Liu, “Mining and Summarizing Customer Reviews”, KDD’04, Washington, 2004

[3] Lipika Dey and SK Mirajul Haque. “Opinion mining from noisy data,” Proceedings of 2nd workshop on Analytics for noisy unstructured text data, Singapole. pp 83-90, 2008

[4] Stanford Parser Version 1.6.
2008. http://nlp.stanford.edu/software/lex-parser.shtmlWordNet – About [5] WordNet
http://wordnet.princeton.edu