Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology
Vol.7, No.6, June (2017), pp. 759-766
http://dx.doi.org/10.14257/ajmahs.2017.06.41

한국어 비정형 데이터 처리를 위한 효율적인 오피니언 마이닝 기법

남기훈[1)]

An efficient opinion mining technique for processing unstructured data in korean

Kihun Nam[1)]

요 약

본 논문에서는 소셜 네트워킹 서비스(SNS)에서 구조가 복잡한 비정형 데이터를 처리하기 위해 새로운 마이닝 접근법을 제안한다. SNS에 올라온 텍스트 형태의 문장들을 분석하여 의미있는 감성 정보를 보다더 효과적으로 추출하기 위해서이다. 어휘 사이의 관계를 형성하는데 주목적으로 하는 기존 방식의 SentiwordNet은 함축적 표현의 한글 문장 및 초성, 중성, 종성으로 이루어진 독특한 글자 구조인 한글에 적용하기 어려운 문제가 있다. 이를 개선하기위해 레빈쉬타인 거리 알고리즘을 적용하여 한글 문장을 효율적으로 탐색하고 결과에 가중치를 부여하여 감성 사전을 구축했다. 문장 분석을 실시간 처리하여 일반문, 긍정문, 부정문으로 평가 계산식을 사용하여 구별했으며 그 결과를 백분율로 표기하였다. 실험 결과를 통해 본 논문에서 제안하는 방법이 Navie Baysian 방식의 실험보다 문장에서 감정 분류에 효과적이라는 것을 증명했다. 대용량 데이터를 효율적으로 처리하기 위해 하둡 시스템을 이용했다.

핵심어 : 오피니언 마이닝, 비정형 데이터, SentiwordNet, 레벤스테인, 하둡 시스템

Abstract

In this paper, we proposed a new mining approach to process unstructured data with complex structure in Social Networking Service(SNS). The proposed method is to extract meaningful emotional information more effectively by analyzing the texts posted on the SNS. In general, SentiwordNet method, which is the main purpose of forming the relationship between existing vocabularies, has a problem that is difficult to apply to korean sentences of implicit expression and Unique character structure. To improve this, applying the Levenshtein distance algorithm, Korean words are efficiently searched and the sentiment word dictionary is constructed by assigning weights to the result. Sentence analysis was processed in realtime to distinguish general, positive, negative sentences and the results were expressed as a percentage. According to the result of the experiment, we proved that is more effective for classifying emotions in sentences than Navie Baysian. Hadoop system was used for efficient with handle web-scale data.

Keywords : Opinion Mining, Unstructured Data, SentiwordNet, Levnenshtein, Hadoop system

Received (April 3, 2017), Review Result (April 17, 2017)

Accepted (April 24, 2017), Published (June 30, 2017)

[1]02713 Dept. Computer Engineering, SeoKyeong Univ., Jeongneung-dong, Seongbuk-gu, Seoul, Korea
email : namkh@skuniv.ac.kr

ISSN: 2383-5281 AJMAHS

1. 서론

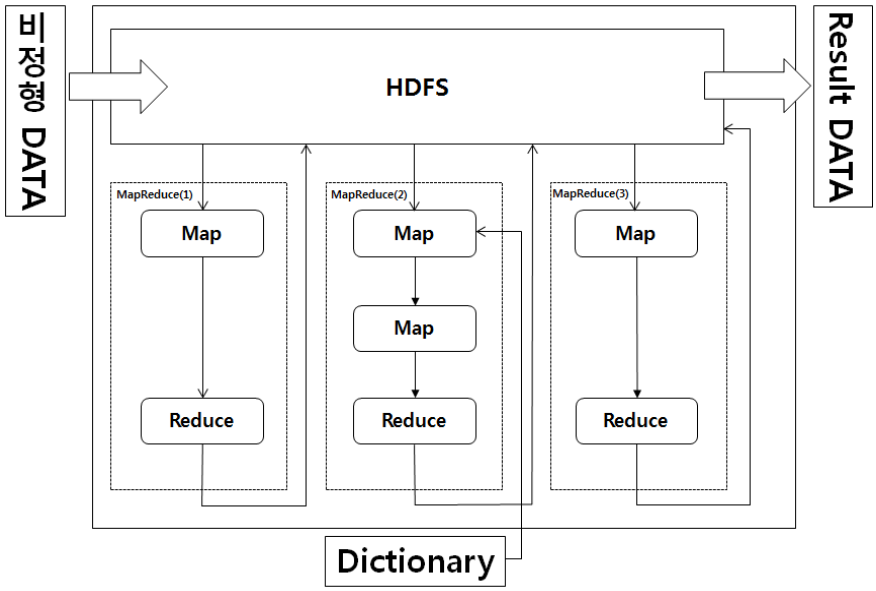
최근 소셜 네트워크 서비스(SNS), 로그 파일, 통신 CDR등에서 형태와 구조가 복잡한 비정형 데이터의 사용 비중이 높아지고 있다. 이를 위해 유용한 정보를 추출, 가공하기 위하여 텍스트 마이닝을 이용한 다양한 자연어처리 기술들이 사용되고 있다. 관련 분야인 오피니언 마이닝은 SNS에서 정형/비정형 텍스트의 긍정, 부정, 일반의 선호도를 판별하는 기술로서 특정 서비스 및 상품에 대한 시장 규모 예측 및 소비자의 반응을 분석한다[1][2].

기존의 사전 구축 방식은 SentiwordNet 접근법을 많이 사용한다[3][4]. 하지만 이는 문장의 접속사나 긍정 또는 부정의 어휘를 우선 구별하게 되어 있어서 영어 문장에는 유용하나 함축적인 의미를 표현하고 받침이 어려운 한글 문장에 적용하기가 쉽지 않다[5]. 이에 영어 문장을 포함하여 초성, 중성, 종성으로 이루어진 한글의 비정형 데이터를 처리하기위해 새로운 마이닝 접근법을 제안한다. 인터넷의 규모가 증가되면서 주관적인 데이터가 증가하여 이를 자동으로 분류할 필요성이 대두되고 있기에 대용량 데이터 처리에 적합한 하둡 시스템에 이를 적용하였다[6].

2. 하둡 시스템을 이용한 비정형 데이터 처리 프로그램

2.1 비정형 데이터 처리 구조

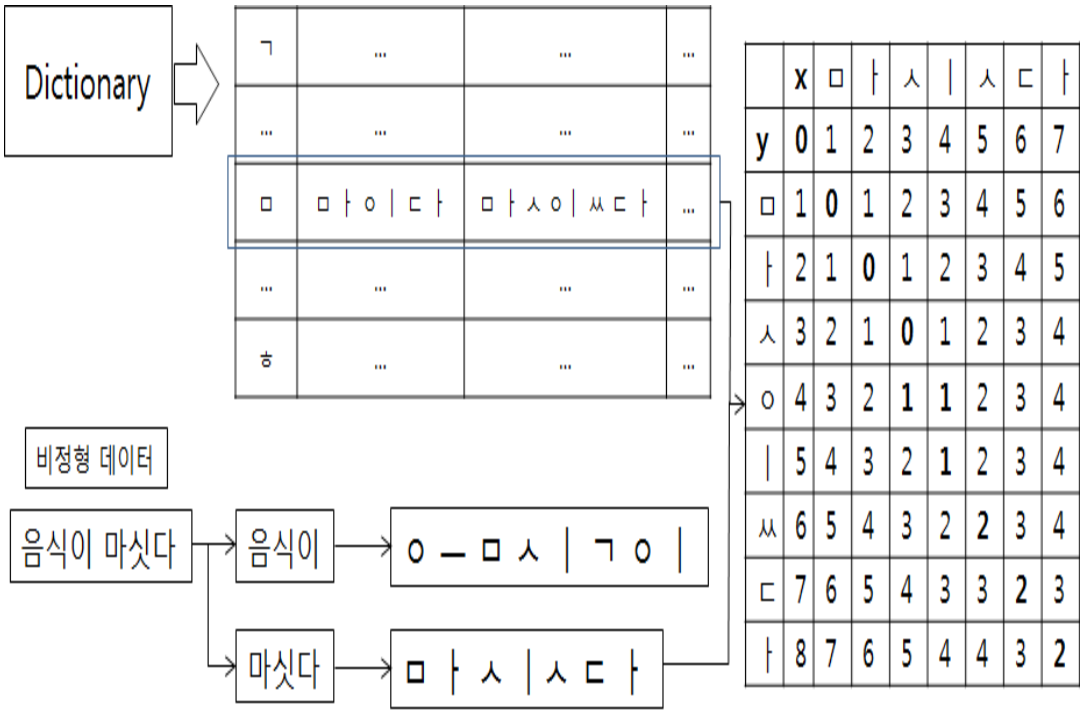
본 논문은 대용량 데이터 처리를 위해 그림 1과 같은 구조를 적용하였으며 사전에 준비된 사전 데이터와 비정형 데이터를 비교 분석한 문장을 이용하여 긍정문인지 부정문인지를 판별했다.



[그림 1] 데이터 처리 구조
[Fig. 1] Data processing structure

2.2 데이터 매칭

본 논문의 Mapreduce(2)의 세부 과정은 그림 2와 같다. 1만개의 단어들로 구성된 사전 DB에서 모든 테이블의 정보를 옮겨 온다. 옮겨온 정보를 자음 순서대로 저장 시 긍정 점수와 부정 점수, 빈도수를 같이 기록한다. ArrayList<'ClassName'> List = new ArrayList<'ClassName'>로 ㄱ~ㅎ을 생성하고 ArrayList<ArrayList>ChosyngList=new ArrayList<ArrayList>()로 자음별 리스트를 Array List 배열에 저장한다.



[그림 2] 사전 DB와 비정형 데이터 매칭 과정
[Fig. 2] Matching process of Dictionary DB and unstructured data

입력된 비정형 데이터를 토큰화한 후, 자음별로 구별되어있는 배열 중 비교할 토큰의 첫 번째 자음과 일치하는 배열을 가져온다. 배열 값과 토큰 사이의 거리 비용을 계산하기 위해 레빈쉬타인 알고리즘을 수행하여 가장 적은 비용을 가지는 인덱스를 선택한다[7]. 오타 어휘인 경우에는 사전에 등록된 정형 데이터로 대체하여 감정 판단을 하게 된다.

2.3 거리 비용 계산

레빈쉬타인 알고리즘은 삽입 삭제 대치를 이용하여 문자 x, y의 변환에 따른 비용을 계산한다. x는 변환할 문자, y는 변환될 문자를 의미한다. 네 칸을 기준으로 (j, i)의 값을 순차적으로 채운다.

	x	ㅁ	ㅏ	ㅅ	ㅣ	ㅅ	ㄷ	ㅏ
y	**0**	1	2	3	4	5	6	7
ㅁ	1	**0**	1	2	3	4	5	6
ㅏ	2	1	**0**	1	2	3	4	5
ㅅ	3	2	1	**0**	1	2	3	4
ㅇ	4	3	2	**1**	**1**	2	3	4
ㅣ	5	4	3	2	**1**	2	3	4
ㅆ	6	5	4	3	2	**2**	3	4
ㄷ	7	6	5	4	3	3	**2**	3
ㅏ	8	7	6	5	4	4	3	**2**

(j-1,i-1)	(j-1, i)
(j, i-1)	(j, i)

[그림 3] 거리 계산 방식

[Fig. 3] Scheme of distance calculation

x[i]와 y[j]가 같은 경우 최단 편집 거리는 (j-1, i-1)이며, 다르면 (j-1)(i-1)에 해당하는 값에(j, i)로 이동하기 위한 이동 거리를 더한 값은 대치가 이루어진다. (j-1)(i)에 해당하는 값에 (j, i)로 이동하기 위한 이동 거리를 더한 값은 삽입이 되고, (j)(i-1)에 해당하는 값에 (j, i)로 이동하기 위한 이동 거리를 더한 값은 삭제가 되는데 그 중에서 가장 작은 값이 최단 거리가 된다. scost x(i) = y(i) 이면 scost = 0이고 그렇지 않으면 scost = 1이 된다.

$$D[j][i] = \min imum(D[j-1][i]+1, D[j][i-1], D[j-1]i[-1]+scost) \quad (1)$$

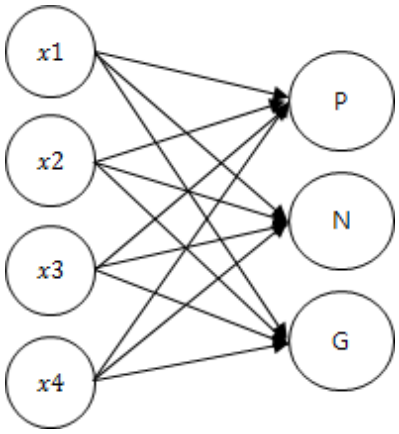
D[1][1]은 'ㅁ'->'ㅁ'의 변환 비용은 수식 (1)에 의해 변화 없음의 연산결과를 얻으며 D[4][4]는 'ㅁ ㅏ ㅅ ㅣ'를 'ㅁ ㅏ ㅅ ㅇ'으로 1의 비용이 기록된다. 자음 list에 들어있는 모든 초, 중, 종성 문자열과 타겟 토큰을 비교하여 자장 적은 변환 비용을 가지는 인덱스를 선택한다. 변환 비용이 같은 인덱스가 존재할 경우, 빈도수의 값이 큰 쪽을 선택하여 처리한다. 연산 결과는 action[][] 배열에 저장한다.

2.4 가중치 부여 및 긍정문과 부정문 판별

action[][] 배열에 저장된 데이터들은 긍정, 부정, 일반 클래스로 구분하여 재배열 한다. 긍정에도 강한 긍정문과 약한 긍정문이 있기에 적절한 가중치를 부여하여한다. 선택된 인덱스의 문자열에 가중치 계산을 수식 (2)을 적용하여 처리한다[8].

$$w_{ji} = \frac{\log(f_{ji}+1)\log(n/\sum_{j} x(f_{ji}))}{\sqrt{\sum_{i}(\log(f_{ji}+1)\log(n/\sum_{j} x(f_{ji}))^2}} \qquad (2)$$

한 문장에서 분류된 어휘들은 그림 4와 같은 긍정(P), 부정(N), 일반(G) 노드들과 결합하여 그 문장이 가지고 있는 긍정의 의미인지 부정의 의미인지를 판독하게 된다.



[그림 4] 두 개의 층으로 구성된 망
[Fig. 4] Two layered network

긍정, 부정, 일반문장 평가 계산은 수식 (3), (4)를 적용하여 처리한다.

$$f_j = \sum_{i=1}^{I} w_{ji} x_i + b_j \qquad (3)$$

$$\tanh(u) = \frac{e^{u} - e^{-u}}{e^{u} + e^{-u}} \qquad (4)$$

3. 실험 및 결과

3.1 실험 데이터

실험에서 사용된 데이터들은 사용자들의 감성이 포함된 문장들로 구성되었으며 트위터에서 스타벅스란 키워드로 광고성 문장들을 수집하였다. 긍정 예시 문으로는 “스타벅스 분위기 좋아요”, “스타벅스 비싸긴 하지만 맛있으니 괜찮음”, “스타벅스 케이크 맛있어요”등 이며, 부정 예시 문으로는 “스타벅스 매장 너무 더러운거 아니예요?”, “스타벅스 이번에 나온 신 메뉴 맛없었음”, 스타벅스 맛은 좋은데 너무 비싸요“등의 문장들이다. 문장 수집은 기간은 2016.11.21부터 26일까지이다.

3.2 실험 환경

실험 조건은 [표 1]과 같은 환경에서 실행하였다.

[표 1] 실험 조건

[Table. 1] Experiment Conditions

구성 요소	세부 내용
Hadoop	2.5.2
HDFS Capacity	3TB
MySQL	5.5
운영체제	CentOS

3.4 실험 비교 및 결과

분류에 가장 대표적인 모델인 나이브 베이지언(Naive Bayesian)과의 실험 결과를 비교하여 본 논문에서 제안하는 판단 방법의 성능 및 타당성을 입증하려 한다[9][10]. 나이브 베이지언의 모델 식은 식 (5)와 같다.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (5)$$

A는 클래스이며 B는 단어로 이루어진 데이터들이다. 수식 (5)를 이용하여 각 단어들이 서로 독립이라고 가정한 후 식을 전개하면 아래와 같은 식들이 나온다.

$$P(A) = \frac{N_A}{N_A + N_{A'}} \qquad (6)$$

$$P(W_i|A) = \frac{T_{A\,W}}{\sum_{W \in V} T_{A\,W'}} \qquad (7)$$

식 (6)의 N_A는 A 클래스의 문서 수이고 식 (7)의 $T_{A\,W}$는 A 클래스에서 단어 W의 출현 빈도 수 이며 $\sum_{W \in V} T_{A\,W'}$는 A 클래스에서 등장하는 모든 단어들의 출현 빈도 수 합이다.

[표 2] Navie Bayesian 방식의 실험 결과
[Table. 2] Result of Navie Bayesian Experiments

Date	Total	Blurb	Positive	Negative	Inability	Hit ratio
16.11.21	118	42	58	3	15	87.2%
16.11.22	882	399	347	40	96	89.1%
16.11.23	769	351	324	56	38	95%
16.11.24	896	410	350	74	62	93.1%
16.11.25	689	299	317	19	54	92.2%
16.11.26	304	107	163	3	31	89.8%

[표 3] 제안된 방식 실험 결과
[Table. 3] Result of proposed method Experiments

Date	Total	Blurb	Positive	Negative	Inability	Hit ratio
16.11.21	118	42	63	5	8	93.2%
16.11.22	882	399	362	43	78	91.1%
16.11.23	769	351	329	59	30	96.1%
16.11.24	896	410	367	66	53	94.1%
16.11.25	689	299	332	16	42	93.9%
16.11.26	302	107	169	6	22	92.8%

4. 결론

본 논문은 SNS에서 비정형 데이터 처리를 위한 비교실험을 했다. 그 결과 판별할 수 없는 문장의 수가 오류 정정을 통해 감소되었으며 영어와 다른 글자의 구조 및 의미가 복잡한 한글 문장들에 있어서 기존 방식보다 긍정과 부정의 감정 평가가 대략 2~6% 정도의 적중률이 개선된 점을 확인 했다. 보완해야 할 점으로는 사용자가 생성한 모든 감성 표현들이 등록되어야 하기 때문에 사전 구축에 필요한 방대한 데이터의 양이 필요하다는 것이다. 추후 연구로 사전 구축을 인공 지능화하여 기본 사전에서 유추하여 감정을 판단하는 시스템 개발이 절실할 것으로 보인다. 그러한 연구를 바탕으로 SNS에서 엄지족 신세대의 자음 표기법(ㅅㄱㅇ, ㄱㅅㅇ)도 감성 분류를 가능하게 되리라 기대한다.

References

[1] Kyeongmee Park, Hogun Park, Hyeonggon Kim, Heedong Ko, Research of opinion mining by SNS, The Journal of Information Science. **(2011),** Vol.18, No.6, pp.68-78.

[2] Soomin Lee, Hye-Jin Rhu, A-Reeum Lee, Jiseon Bang, Editors. A design of Opinion Mining System of Advertisement in SNS Environments. Proceedings of the Korea Computer Congress, **(2014)** June 25-27 Busan, Korea

[3] SeungEun Lee, YongHee Kim, Ung-Mo Kim, An Analysis of the Consumer Behavior using Opinion Mining, The Journal of korea Computer Congress. **(2016)**, Vol.43, No.1, pp.340-342.

[4] Kwang-Mo Ahn, Yun-Suk Kim, Young-Hoon Kim, Young-Hoon Seo, Sentiment Classification of Movie Reviews using Levenshtein Distance, Journal of Digital Contents Society **(2013)**, Vol.14, No.4, pp.581-587

[5] Sukjae Choi, Ohbyung Kwon, The Study of Developing Korean SentiWordNet for Big Data Analytics : Focusing on Anger Emotion, The Journal of Society for e-Business Studies. (**2014**), Vol.19, No.4, pp.1-19.

[6] Mun-Su Kang, Seung-Hee Baek, Young-Sik Choi, Statistical Approach to Sentiment Classification Using MapReduce, The Journal of Science of Emotion & sensibility. **(2012)**, Vol.15, No.4, pp.425-440.

[7] Jong-Sub Lee, Sang-Yob Oh, Vocabulary Retrieve System using Improve Levenshtein Distance algorithm, The Journal of Digital Policy & Management. **(2013),** Vol.11, No.11, pp.367-372.

[8] Hyung-Jin Oh, Ji-Hyun Go, Dong-Un An, Soon-Chul Park, Latent Semantic Indexing Analysis of K-Means Document Clustering for Changing Index Terms Weighting, The Journal B of Information Processing Society, **(2004)**, Vol.10B, No.7, pp.735-742.

[9] http://www.saedsayad.com/naive_bayesian.htm **(2010).**

[10] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, Editor, Introduction to Information Retrieval, Cambridge University Press, New York **(2008)**