

리더 가중치를 활용한 오피니언 마이닝

조경수*, 류준석, 김영희, 김응모
 *성균관대학교 정보통신공학부
 e-mail : kisschks@hotmail.com

Usage of Leader Weights for Opinion mining

Kyung Soo Cho*, Joon-suk Ryu, Young Hee Kim, Ung-mo Kim
 *School of Information and Communication Engineering,
 Sungkyunkwan University

요 약

인터넷 사용의 증가와 함께 최근 많은 양의 정보가 인터넷에 쏟아지고 있다. 이들 정보는 다른 사람의 생각을 알고 싶어하는 정보 수집 연구자들에게는 매우 유용한 정보이다. 현재 존재하는 오피니언 마이닝 기법은 매우 다양하다. 그러나 이러한 기법들은 모든 의견들을 동일한 영향력을 지닌 것으로 취급한다. 하지만 현실에서는 모든 의견이 동일한 영향력을 가지고 있지는 않다. 이런 문제점 해결을 위해서 우리는 새로운 오피니언 마이닝 기법을 제안한다.

1. 서론

다른 사람들은 어떻게 생각하는가? 는 이제껏 중요하게 생각되어져 왔다. 오피니언 마이닝은 이러한 내용을 웹 상에서 추출하는 것에 대한 필요로 연구 되었다. 자원들은 기본적으로 고정된 템플릿을 따르는 웹페이지나 DB에서 얻는다. 대단히 가치 있는 정보들을 수집할 수 있기 때문에 이것은 매우 중요한 일이다. 현재 Web 2.0환경에서는 UCC나 블로그 등의 정보가 범람하고 있는데 이러한 환경에서 특히 유용하다.

기존의 오피니언 마이닝은 평가되는 문서가 긍정 혹은 부정 인지 혹은 객체에 대한 호불호나 전체적인 분위기를 구분하기 만 한다. 하지만 실생활에서 보면 모두의 의견이 같은 영향력을 가지는 것은 아니다. 그렇기 때문에 모두 일괄적으로 같은 가치를 주는 것은 현실을 반영하지 못하는 방법이라 할 수 있다. 토론의 집단에서는 언제나 오피니언 리더가 존재하고 그들의 의견이 전체 오피니언 홀더들의 의견에 영향을 주거나 방향을 좌우 한다. 그렇기 때문에 리더의 의견에는 일반 의견에 비해 더욱 큰 가중치를 뒤야한다. 그렇게 하기 위한 방법적 고찰을 본 논문에서 하게 된다.

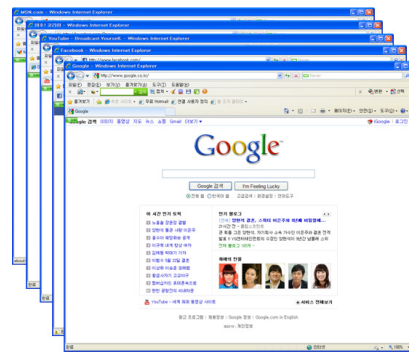
본 논문에서는 어떠한 방식으로 기존의 오피니언 마이닝에서 리더에 속하는 오피니언 홀더들에게 가중치를 부여 하고 그것을 어떻게 활용할 것인지에 대해 논의 한다. 또한 그렇게 함으로써 좀 더 가치 있는 정보를 포함하고 있는 결과물을 추출할 수 있도록 한다.

본 논문은 다음과 같이 구성된다. 2장에서 오피니언 마이닝 기법과 기존의 대표적인 연구에 대해 설명을 한다. 3장에서는 본 논문이 제안하는 리더 가중치를 활용하여 문서를 평가하는 방식을 설명하고 4장에서는 결론과 발전된 연구를 위한 향후 연구 과제를 제시하여 논문을 마친다.

2. 관련연구

2.1 페이지뷰 와 웹사이트 자원

페이지뷰 [1] 는 해당 사이트에 대한 가치 평가의 척도로 사용되고 있으며 페이지 뷰 수가 많다는 것은 관심 사이트거나 인터넷에서 중요한 역할을 하고 있음을 의미한다. 페이지 뷰는 개인 블로그 형식을 위주로 급격히 증가하는 추세다. 페이지뷰는 히트, 방문수, 순방문자로 구분되며 서버나 웹페이지에 심어 둔 코드가 로그에 기록한다.



Site Name	Rank
Google	1
facebook	2
youtube	3
.	.
.	.
.	.
.	.
.	.

<그림 1> 페이지 뷰에 따른 웹사이트 순위

그림 1은 페이지뷰에 따른 전 세계 웹사이트들의 순위의 일부분이다. 페이지뷰가 높은 웹사이트들은 현실에서도 영향력이 큰 사이트들로 평가 되고 있다. 그렇기 때문에 이들 페이지뷰는 e-비즈니스나 광고 마케팅 등 여러 분야에서 활용되고 있는 절대적 수치다. 또한 웹사이트는 오피니언 마이닝의 중요 자원으로 JG Conrad, F Schilder는 웹사이트 자원 중 하나인 합법적인 블로그를 찾아내는 방법을 기술 했다. [2]

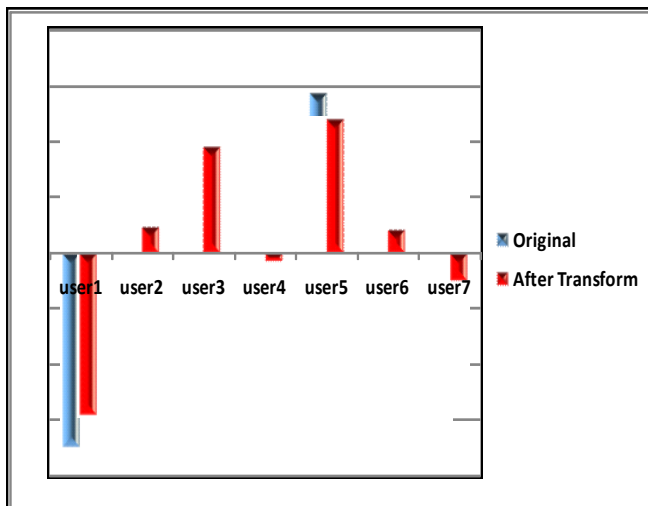
기존에 나온 오피니언 마이닝의 결과물 값에 각각의 페이지뷰 수 값을 전체 페이지뷰 수로 나눈 값을 곱셈 연산을 하게 된다. 이렇게 하면 기존의 결과 값들은 다시 한번 재정리 되게 되고 오피니언 홀더가 가지는 리더 영향력에 의해 수치가 변경되게 된다. 각자가 가지는 페이지뷰수가 적은 것들에 대해서는 그것이 전체 영향에 미치는 힘이 작아질 수밖에 없게 된다. 물론 이러한 방식은 각각의 오피니언 홀더들이 어느 정도의 페이지뷰 수나 그에 상응하는 수치가 어느 정도 확보가 되어야 가능하다. 리더가 존재하지 않는 오피니언 홀더들 사이에서는 적용하기 힘들다는 단점이 있다.

Lw가 적용된 값은 실제로는 어떻게 영향을 받고 있는지를 보여주기 위한 값이기 때문에 원래의 데이터에 대한 유지도 필수이다.

<예시> Car

	Original		PageView		After
User1	Negative : - 0.6453		30		- 0.117327
User2	Positive : 0.3210		10		0.019455
User3	Positive : 0.2134		60		0.077600
User4	Negative : - 0.2124		5		- 0.006436
User5	Positive : 0.5321		30		0.096745
User6	Positive : 0.1421		20		0.017224
User7	Negative : - 0.3323		10		- 0.020139

<그림 5> Lw 적용 데이터 예시



<그림 6> 오피니언 비중 수치 그래프

그림 5는 Car에 대한 유저들의 의견을 추출한 것으로 예시 했다. 또한 실제로 Lw가 적용 되었을 때 어떻게 수치들이 변화하는지를 보여주고 있다. 유저라 표현된 오피니언 홀더들의 반응을 Lw를 적용시켜 각자의 영향력에 맞

게 수치가 변환되어 있다. User1부터 User7까지 나와 있는 그림4의 예시 데이터는 Lw를 적용하기 전에 Car에 대한 Negative 또는 Positive 수치가 Lw 적용 후에는 크게 변할 수 있음을 보여주고 있다.

그림 6에서는 그림5를 기반으로 Lw가 적용되면 영향력이 적용되어 전체 의견에 어느 만큼의 비중을 가지게 되는지에 대한 그래프이다. 첫 번째 항목이 원본 데이터 이며 두 번째가 Lw 적용 데이터이다. 페이지뷰를 기반으로 영향력을 평가 하기 때문에 적은 수치의 페이지뷰를 가졌던 User2, User4 그리고 User7 의 경우에는 비중이 급격히 줄어들었다. 그리고 많은 수의 페이지뷰를 지녔던 User3는 전체 의견에 영향을 미치는 비중이 크게 상승 했다. Lw가 적용된 데이터 결과물은 인터넷 상에서 의견 흐름이 실제로 어떻게 이뤄질 것인지에 대해 예측을 하는데 도움이 된다.

4. 결론 및 향후 연구 방향

본 논문에서는 오피니언 마이닝에서 산출하는 정보가 좀 더 유용하며 현실을 반영할 수 있는 Lw 수식을 제안 하였다. 이는 기존의 방법이 보여줄 수 없는 영향력에 대한 고려를 통해 제안된 것으로 원래의 데이터 값 자체가 쓸모없게 되는 것은 아니다. 다만 Lw가 적용된 후에 나타나는 수치는 전체적으로 의견이 어떻게 제시되고 있는가를 파악 하 는 것에는 매우 유용하게 쓰일 수 있다. 차후에는 논점을 흐리는 오피니언 홀더들에 대한 영향력 (Opinion Spam)을 고려하는 방법에 대한 연구와 페이지뷰와 관련되어 가치가 보증될 수 있는 방법에 대해 좀 더 연구가 필요하다. 심리학적 판단 이론을 적용하여 웹사이트에서 의견을 추출 할 때 신뢰성을 계산하여 좀 더 실제에 가까운 오피니언 마이닝을 하는 방향으로 연구를 지속 하려 한다.

감사의 글

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

참고문헌

- [1] 한은미 “블로그(Blog)의 이용동기 분석 및 서비스 품질이 고객 만족에 미치는 영향”, 숙명여자대학교 정보통신대학원, 2004
- [2] JG Conrad, F Schilder “Opinion mining in legal blogs”, International Conference on Artificial Intelligence and Law, Proceedings of the 11th international conference on Artificial intelligence and law, pp 231-236, 2007
- [3] A. Esuli and F. Sebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining” , In Proceedings of the 5th Conference on

Language Resources and Evaluation (LREC'06),
Citeseer, 2006

[4] A. Esuli and F. Sebastiani, "PageRanking WordNet synsets: An application to opinion mining", Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), Citeseer, 2007

[5] SM Kim and E Hovy "Determining the sentiment of opinions", International Conference On Computational Linguistics, Proceedings of the 20th international conference on Computational Linguistics, 2004

[6] B Pang, L Lee "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, Volume 2 , Issue 1-2 (January 2008), Pp 1-135, 2008

[7] S.M Kim and E Hovy "Extracting opinions, opinion holders, and topics expressed in online news media text", Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text, Sydney, Australia, 2006