Regular paper http://dx.doi.org/10.13067/JKIECS.2013.8.11.1719

데이터 변형성 기반 유사성 연결을 위한 단어 추천 알고리즘

김분희*

Words Recommendation Algorithm for Similarity Connection based on Data Transmutability

Boon-Hee Kim*

요 약

기존의 데이터 처리 방식과는 다른 접근이 필요한 빅데이터는 데이터의 분량, 데이터의 변화 속도, 데이터의 다양성의 특징을 가진 비정형 데이터이다. 트위터의 트윗(tweet)이 국내만 보더라도 하루 500만건이 넘는 상황이다. 이렇게 많은 데이터는 저렴해진 저장시스템과 분석정보에 대한 수요 증대로 인해 연구가치가 높아지고 있다. 본 논문에서는 이러한 기술에서 요구되는 요소 기술로써 데이터 변형성의 특징을 기반으로 우선순위기반 단어 추천 알고리즘을 제안한다.

ABSTRACT

Big data which requires a different approach from existing data processing methods, is unstructured data with a variety of features. The features mean the volume of data, the rate of change of the data, the data with a variety of features. Tweets of twitter in only Korea are more than 5 millions per day. So much cheaper data storage and analysis system due to the increasing demand for information, the value of research is increasing. In this paper, the technology required by the deformation characteristics of the data elements as a technology priority-based word-based recommendation algorithm is proposed.

키워드

Big Data, Visual Information, Transmutability, Recommendation
빅 데이터, 시각정보, 변형성, 추천

Ⅰ. 서 론

비정형 데이터의 집합인 빅데이터(BigData)는 단순히 데이터의 양이 크다는 특징만으로 설명할 수 없다. 트윗(tweet)메시지처럼 현재 발생되고 있는 데이터에 대한 빠른 처리를 전재로 하고 데이터의 형태가 매우 다양하다. 기존에는 버려졌던 데이터 또한 분석의 대상이 되는 빅데이터 처리 시스템은 저장시스템 또한 그 특성에 맞게 확장되어야 한다[1-4].

빅데이터는 데이터를 효과적으로 추출하고 분석하는 작업이 중요한데, 일반 데이터가 아닌 그래프와 같은 형식으로 시각화하는 작업 또한 활발한 연구가 진

* 교신저자(corresponding author) : 동명대학교 미디어공학과(bhkim@tu.ac.kr)
접수일자 : 2013. 08. 30 심사(수정)일자 : 2013. 10. 21 게재확정일자 : 2013. 11. 15

행되고 있다[5-6].

기존 데이터베이스 기술로는 처리하기 힘든 빅데이터 기술은 비정형 데이터, 즉 기존의 정형화된 데이터가 아니어서 어떠한 틀에 맞게 정형화하기 힘들어 버려졌던 데이터들을 대상으로 한다. 빅 데이터는 이러한 비정형 데이터에 대해 새로운 관심의 흐름에서 발생되었으므로 정해진 목적에 맞게 데이터를 효과적으로 저장하고 분석하는 기존의 데이터베이스 기술과는 다른 형태로 분석되어야 한다. “이러한 빅데이터는 그림 1과 같이 3V(Volume, Variety, Velocity)의 특징을 가지고 있다. 이러한 빅데이터는 사람에 의해서 만들어지는 수많은 데이터를 기반으로 유용한 정보를 얻어내기 위해 이용되는데 컴퓨터 프로그램의 특성에 인간의 기억의 변형성의 특징을 가미한 기계 학습 기술이 적용되어야 좀더 유용한 정보를 얻을 수 있다. 물론 빅데이터는 통계를 바탕으로 결론을 도출하고, 이를 통해 예측을 할 수 있다.”[7] 그러나 인간이 가지고 있는 기억의 메커니즘을 감안해 보면 원본 데이터가 그대로 유지되기 보다는 유사 데이터로 변형되어 기억되는 변형성으로 인해 이를 반영한 기술이 요구된다.

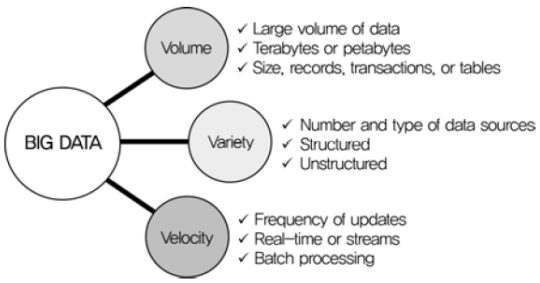
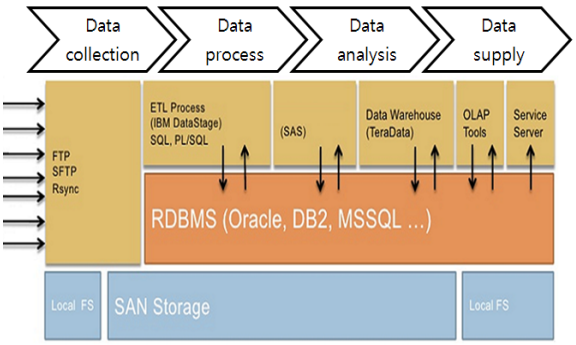


그림 1. 빅데이터의 특징 [1]
Fig. 1 Characteristics of BigData [1]

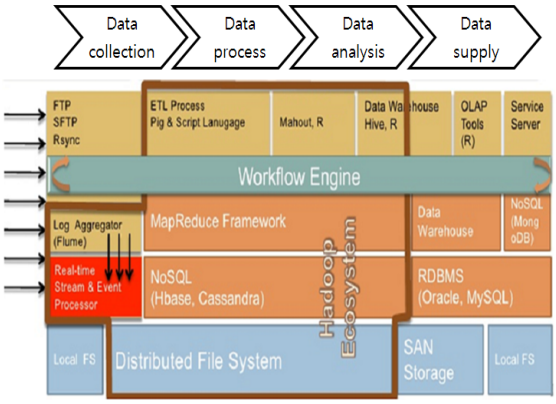
본 연구에서는 빅데이터 기술에서 요구되는 요소기술로써 데이터 변형성의 특징을 기반으로 우선순위 기반 단어 추천 시스템을 제안한다. 데이터 변형성이란 비정형 데이터를 만들어내는 인간의 학습 과정과 기억 메커니즘에서 나타나는 원본 데이터의 유사 변형의 특성을 의미한다.

II. 관련 연구

기존의 데이터 분석처리 시스템과 달리 빅데이터 분석 시스템의 경우 다양한 데이터와 이를 처리하는 프레임워크를 감당할 수 있는 형태의 워크플로우 통합 관리시스템이 필요하다.



(a) analysis system



(b) big data analysis system

그림 2. 데이터 분석 시스템 [2]
Fig. 2 Data analysis system [2]

그림 2는 기존의 데이터 분석 시스템의 구성과 빅데이터를 대상으로 한 데이터 분석 시스템의 구성적 차이를 보여준다. 데이터 수집, 데이터 처리, 데이터 분석의 과정을 통해 발생된 결과 정보를 제공하는 과정에서 대용량 데이터를 저장하고 분석하는데 오픈소스 기반의 하둡(Hadoop), NoSQL, R등을 활용할 수 있다. 특히 하둡과 오픈소스 기반의 시스템들은 저가의 범용 서버와 네트워크를 클러스터 형태로 구성해

서 사용할 수 있어 기존 시스템에 비해 확장성의 부담이 적다. 비정형 데이터들은 상호 의미관계를 정립하기 힘들어서 가치 있는 데이터를 분석해내는데 매우 어려운 특징이 있다.

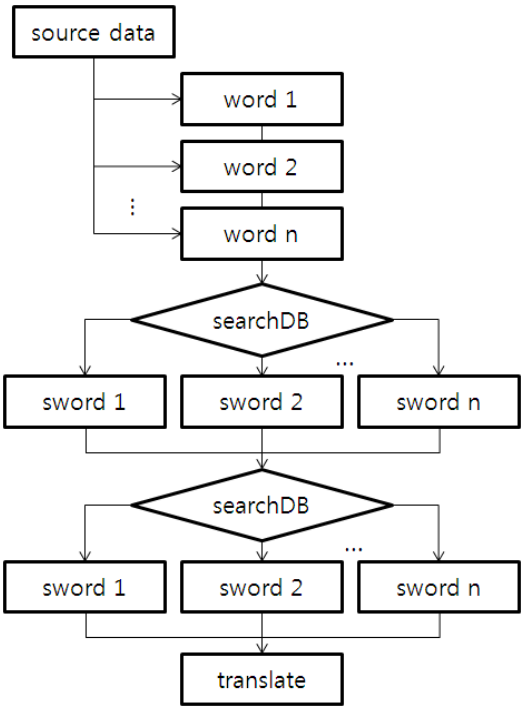


그림 3. 유사성 연결 알고리즘 [7]
Fig. 3 Algorithm for similarity connection [7]

빅데이터는 빠르게 변화하는 정보통신 기술과 이용자들의 요구에 의해 미래에 대한 예측이나 사용자의 선호도 추이 등의 상업적으로 유용한 데이터로써 인식되고 있다. "인간의 사고 과정이 진행되면서 시간의 요소가 적용되면 데이터는 변형의 과정을 밟게 된다. 이러한 변형의 과정을 통해 창의적인 사고를 낳게 된다는 의미에서 봤을 때 변형 데이터의 유용성은 빅데이터 시대에서 더욱 그 가치를 발하고 있다. 그림 3은 이러한 데이터 변형성에 과정에서 선택된 최종 데이터에서 유용한 데이터를 선택하는 과정을 나타내고 있다. 원본 데이터(source)를 기준으로 추출한 의미어(word)를 기반으로 유사어 데이터베이스를 통한 검색(searchDB)를 통하여 추출된 유사어(sword)를 기준으로 설문 기반의 선택 단어를 유의미한 통계의 결과를 기반으로 기준(mid)을 결정하고 해당 기준 이상의 유사어를 선택(sel)하는 과정을 거친다. 이러한 유사어 선택의 과정에서 원본 데이터와 선택된 유사어와의 밀도를 기준으로 한 선택을 통하여 변형된 데이터를 추출(translate) 할 수 있겠다."[7]

본 연구에서는 유사성 연결 알고리즘의 연결선 상에서 빅데이터의 데이터 변형성을 기반으로 정보 선택의 방법에 대해 진행한다.

III. 제안 시스템

데이터 변형성의 특징을 적용한 유의어 리스트를 구성하기위해서는 우선 원본 데이터를 선정하는 과정이 필요하다.

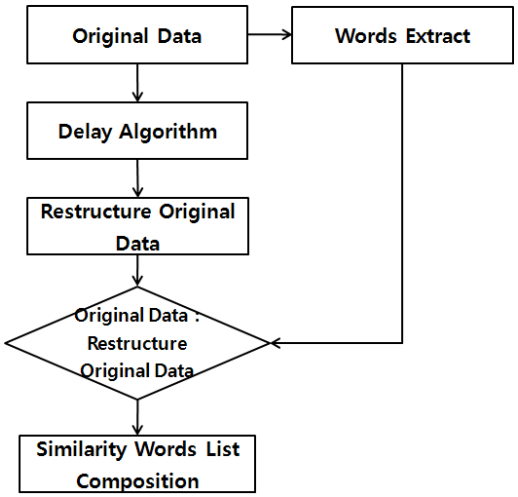


그림 4. 유의어 리스트 알고리즘
Fig. 4 Similarity words list algorithm

본 연구는 특정 원본 데이터가 있는 상태에서 진행되었다. 이러한 원본 데이터는 시간 알고리즘에 의해 일정 시간이 경과된 후 기억의 변화가 진행된 상태를 기준으로 한다. 그 다음으로 원본 데이터의 내용을 재구성하는 과정을 거친다. 이 결과로 재구성 원본을 취합하고, 원본과 재구성 원본을 비교하여 변형 데이터를 추출한다. 이때 원본데이터를 기준으로 추출한 단어를 기준으로 비교를 진행한다. 이러한 변형 데이터를 취합하여 유의어 리스트를 구성한다.

데이터 변형성을 적용하여 구성한 유의어 사전을 이용하여 단어 추천 시스템에서 이용할 단어를 추출한다. 그리고 관련 유사어를 유사도 분석 결과에 의거하여 저장한다. 이러한 과정을 거쳐 데이터 변형 과정에서 선택되어진 정도에 따라 우선순위의 근거로 적용하여 유사어를 정렬한다. 실제 단어 추천 시스템의 동작 상황에 따라서 추천 단어의 리스트를 제시한다. 기존의 메뉴가 아닌 변화하는 단어의 상황에 따른 추천 메뉴이므로 팝업 메뉴의 형태로 제시될 수 있다.

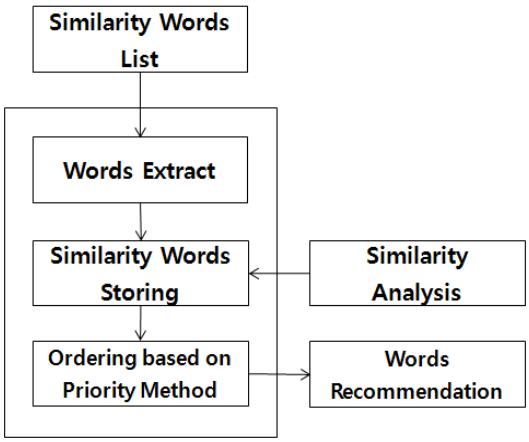


그림 5. 단어 추천 과정
Fig. 5 Words recommendation process

유사도 분석은 식 1과 같은 기준으로 이루어진다. 유사도 분석의 기존 방법들은 다양한데, 본 논문에서는 유클리디언 거리를 기반으로 분석하였다. 그림 5의 단어 추천 과정은 유사도 분석 과정이 들어간 절차를 거쳐 정렬과 단어 추천이라는 흐름에 초점을 맞추어 제안되었다.

원 문자(A)와 유사문자(B)와의 거리(Length)는 2차원 공간상에서의 원 문자(A)의 위치값과 유사문자와의 거리 개념으로 적용했을 때 식 (1)과 같다. 거리의 차이가 원 문자와 유사문자 간의 유사한 정도를 나타낸다.

$$Length(A,B) = \sqrt{W1^2 + W2^2}$$
$$, W1 = |X2 - X1|$$
$$, W2 = |Y2 - Y1| \quad (1)$$

거리간의 차이(DT)는 식 (2)와 같이 원 문자(A)의 유사문자(B)와의 거리값과 원 문자(A)의 유사문자(C)의 거리값의 차이를 기준으로 정하게 된다.

$$DT = Length(A,B) - Length(A,C) \quad (2)$$

저장된 유사어들은 우선순위 기반으로 정렬되고, 이러한 순위를 기반으로 단어 추천이 이루어진다. 그림 6에서 정렬의 방법은 현재 우선순위에 기반하고 있다.

```
Words_Recommendation_Algo
Initialization : input of similarity value
Begin Words_Recommendation_Algo
Begin ordering
Do
  If (previous data comparison) then
  ordering execution;
  priority index change;
While (similarity words)
End ordering
Begin recommendation
For (list size)
  pop up menu creation;
  list component add;
pop up menu visualization;
If (word selection) then
  selection word index storing;
  text area application;
End recommendation
End Words_Recommendation_Algo
```

그림 6. 단어 추천 알고리즘
Fig. 6 Words recommendation algorithm

각 단어들은 순위를 저장하는 레코드에 변경된 순위 인덱스를 저장한다. 이 작업은 유사어의 유무를 점검하는 과정과 함께 수행된다. 단어 리스트의 전체 크기를 반복의 횟수로 하여 추천 단어를 가시화할 팝업 메뉴를 생성한다. 그리고 해당 팝업 메뉴에 단어 리스트의 컴포넌트를 추가한다. 해당 작업이 마무리되면 실제 팝업 메뉴를 가시화 하는 작업이 수행된다. 단어 추천 시스템의 이용자에 의해 단어가 선택되면, 선택 단어에 대해서 선택과 관련된 레코드에 선택의 횟수를 나타내는 인덱스를 저장한다. 그리고 이용자가 선택한 단어를 실제 문장의 한 요소로 반영하기위해 텍

스트 에어리어에 적용하는 작업을 수행한다.

정렬의 방법은 현재 우선순위에 기반하고 있는데, 시스템 이용자의 실제 선호도를 반영한 가중치 부여 방법이 요구된다. 우선순위 정렬 방법을 기반으로 한 단어 추천과 가중치를 부여한 정렬 방법을 기반으로 한 단어 추천 방법을 기반으로 비교한 결과는 다음과 같다. 단어 별 우선순위의 정확성은 일반 우선순위 부여 방법(GAP1)과 가중치 부여 방법(GAP2)이 비슷한 결과를 보이는 경우도 많았으나 전반적으로 가중치를 부여한 방법에서 우수한 결과를 보였다.

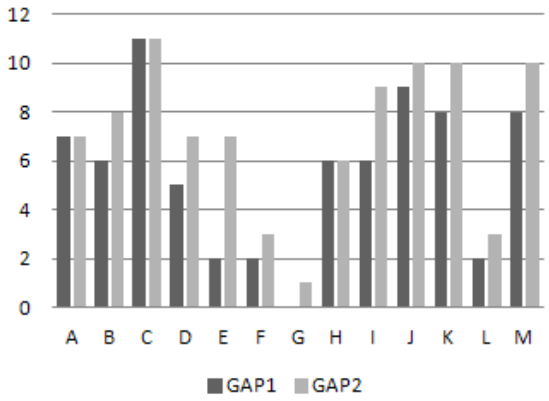


그림 7. 비교 결과
Fig. 7 Results of comparison

IV. 결론 및 향후 연구

빅데이터는 데이터 수집, 데이터 처리, 데이터 분석의 과정을 통해 발생된 결과 정보를 제공하는 과정에서 대용량 데이터를 저장하고 분석하는데 오픈소스 기반의 하둡, NoSQL, R등을 툴을 활용할 수 있다. 이러한 오픈소스 기반의 분석이 가능하다는 환경적인 장점과 더불어 데이터를 저장할 스토리지의 저가격화의 경향으로 빅데이터를 활용한 연구가 활발히 진행되고 있다. 빅데이터 관련 연구 가운데 인간의 사고 과정에서 발생하는 데이터 변형성은 분석의 질을 결정할 수 있어 매우 중요한 분야이다. 빅데이터 처리에 있어서 원본 데이터를 활용하기 위해서는 데이터 변형성이 진행된 단어를 어떻게 다루느냐에 따라 데이터 분석의 질이 결정된다. 따라서 의미 있는 빅데이터 분석이 이루어지려면 데이터 변형성과 원래 데이터와의 상관관계에 정의가 중요하다. 본 연구에서는 빅데이터 기술에서 요구되는 요소 기술로써 데이터 변형성의 특징을 기반으로 우선순위 기반 단어 추천 시스템을 제안하였다.

향후 연구로는 2차 데이터에 1차 데이터의 사용자 선택 사항을 반영하여 유사도의 정밀성을 높이는 과정이 필요하겠다. 여기서 2차 데이터는 시스템에 의해 추천된 단어를 사용자에 의해 선택된 결과로 시스템에서 제안한 우선순위와는 사뭇 다른 결과를 낳을 수 있고, 이는 현재의 데이터 해석 경향을 반영하는 사항으로 이와 관련한 연구가 지속적으로 진행되어야 할 필요가 있다. 그리고 다양한 유사도 분석 방법을 적용하여 유의미한 값을 밝히는 연구가 더해져 유사도 분석에 초점을 맞춘 진행이 필요하다. 또한 유비쿼터스 환경에서의 많은 연구[8][9][10][11][12] 환경에서 이러한 단어 추천 알고리즘의 적용성에 대한 연구도 진행될 필요가 있겠다.

참고 문헌

[1] TDWI Research, “Big Data Analytics Report,” 2011.
[2] Woo-Seung Kim, "Big Data," Big Data Conference, 2012. http://kimws.wordpress.com/2012/07/19/
[3] Answer Tough Big Data and Big Analytics Questions with Google BigQuery, 2012. http://www.enor.com/blog/web-analytics/answer-tough-big-data-and-big-analytics-questions-with-google-bigquery
[4] R. Spence, Information Visualization: Design for Interaction, 2nd ed., Prentice Hall, 2007.
[5] L.H. Boyd, W.L. Boyd, and G.C. Vanderheiden, “The Graphical User Interface: Crisis, Danger, and Opportunity”, J. Visual Impairment Blindness, Vol. 84, No. 10, pp. 496-502, 1990.
[6] H.S. Shin, J.M. Lim, and J.S. Park, "Information Visualization and Information Presentation for Visually Impaired People", 2013 Electronic and Telecommunications Trends, Vol. 28, No. 1, 2013.
[7] Boon-Hee Kim, "Selection Algorithm for

Similarity Connection based on Data Transmutability", The Journal of The Korea Institute of Electronic Communication Sciences, Vol. 7, No. 1, pp. 234-235, 2013.
[8] Hee-hoon Kang, Young-jong Lee, Won-ok Han, "Energy-Efficient Hierarchical Cluster-Based Routing for Ubiquitous Sensor Networks", The Journal of The Korea Institute of Electronic Communication Sciences, Vol. 4, No. 3, pp. 243-246, 2009.
[9] Hyeon-jae Lee, Houn-taek Lee, Hyun-sik Shin, "A Study On Ubiquitous Sensor Network Technologies", The Journal of The Korea Institute of Electronic Communication Sciences, Vol. 4, No. 1, pp. 68-74, 2009.
[10] Tae-Su Jang, Jae-Hyun Kwon, Yong-Kab Kim, Choon-Bae Park, "A LED Light Communication Transceiver Module for Ubiquitous Sensor Networks", The Journal of The Korea Institute of Electronic Communication Sciences, Vol. 4, No. 3, pp. 243-246, 2009.
[11] Han-Young Lee, "High-Tag anti-collision algorithm to improve the efficiency of tag Identification in Active RFID System", The Journal of The Korea Institute of Electronic Communication Sciences, Vol. 7, No. 2, pp. 243-246, 2012.
[12] Hyeon-jae Lee, Houn-taek Lee, Hyun-sik Shin, "A Study On Ubiquitous Sensor Network Technologies", The Journal of The Korea Institute of Electronic Communication Sciences, Vol. 4, No. 1, pp. 68-74, 2009.

저자 소개

김분희(Boon-Hee Kim)

2005년 2월 중앙대학교 컴퓨터공학과(공학박사)
1999년～(주)CEDAR.com 연구원
2005년～현재 동명대학교 미디어공학과 조교수
※ 관심분야 : 분산시스템, P2P 검색 기법, HCI 응용