벡터 간의 유사도 측정을 위한 방법론 비교

강정경
중앙대학교 컴퓨터공학과

Comparison of methodologies for finding similarities between vectors

Jeong-Kyeong Kang
Dept of Computer Science, Chung-Ang University

요 약

텍스트 마이닝 기술은 비정형 텍스트 데이터에서 의미 있는 정보를 찾아내는 기술이다. 텍스트 마이닝 기술은 주로 문서 분류, 군집, 요약이나 정보 추출에 이용된다. 해당 기술을 사용하기 위해, 문서나 단어를 벡터로 표현한 후(word2vec), 벡터 간의 유사성을 이용하는 방법을 활용한다. 본 논문에서는 이때 벡터 간의 유사성을 이용하는 방법에는 어떤 것들이 있으며 각각에 대해 비교를 한 후, 실제 웹 상의 데이터를 이용하여 키워드 군집을 통한 결과를 분석하고자 한다.

1. 서론

텍스트 마이닝 기술은 자연어 처리(Natural Language Processing)기술에 기반한다. 정형화 되지 않은 텍스트 데이터에서 가치 있는 의미와 정보를 찾아내는 기술이다. 텍스트 마이닝 기술을 통해 문서 분류, 문서 군집, 키워드 분석(분류 및 군집) 등 여러가지 다양한 영역에 활용될 수 있다.

해당 기술을 사용하기 위해서는 텍스트를 벡터로 표현하는 방법인 word2vec 을 이용한다. word2vec 은 텍스트를 처리하는 두개의 층을 가진 인공 신경망이며, 이를 이용하면 문장의 의미를 이해하거나 단어 간의 유사성을 구할 수 있다.

본 논문에서 이용할 데이터 셋은 페이스북의 대나무숲 페이지다. 이는 대부분의 대학교마다 존재하고 매일 새로운 글이 게시된다. 시기에 따라 학교 내외의 이슈들, 개인적인 고민 등 다양한 카테고리와 내용들로 이루어져 있다.

벡터 간 유사도를 계산하기 위해 사용할 measure 는 5 가지로 Minkowski distance 부류인 Euclidean distance 와 Manhattan distance, Cosine similarity, Tanimoto Coefficient 그리고 TS-SS 를 이용하고 각각의 결과를 비교하고 평가한다.

2. 관련연구

벡터의 유사도 계산하는 방법은 여러가지가 있으며 다른 특징을 가진다. 논문[1]에서는 11 가지의 벡터 유사도 측정 함수를 Intersection family, Dot product family, Fidelity family 3 가지로 분류하고 각각의 대수적 특성을 보여준다.

벡터 유사도 측정은 텍스트 마이닝 뿐 아니라 전자상거래 시스템이나 사용자 맞춤 추천 서비스 등 여러 분야에 활용된다. 유사도 측정 함수의 특성과 적용시킬 분야 그리고 데이터에 따라 사용해야할 방식이 달라진다. 가장 대표적으로 사용되는 방식은 Cosine similarity 이다. 논문[2]에서와 같이 Cosine similarity 는 영화 데이터를 이용해 사용자에게 알맞은 추천을 해주는 경우에도 사용할 수 있다.

3. 본문

본 논문에서는 페이스북 대나무숲 페이지의 텍스트 데이터에서 키워드를 분석하기 위해 필요한 여러 벡터 유사도 측정 방식 중 5 가지를 사용하여 한글에 적합한 방식은 무엇인지 비교하고 각각 어떤 결과를 보이는지 평가한다.

3.1 Measures

3.1.1 Euclidean Distance

두 벡터 사이의 거리를 피타고라스 정리를 이용해서 계산하는 방법이 Euclidean distance 이다. 결과값이 작을수록 유사도가 높다고 판단한다.

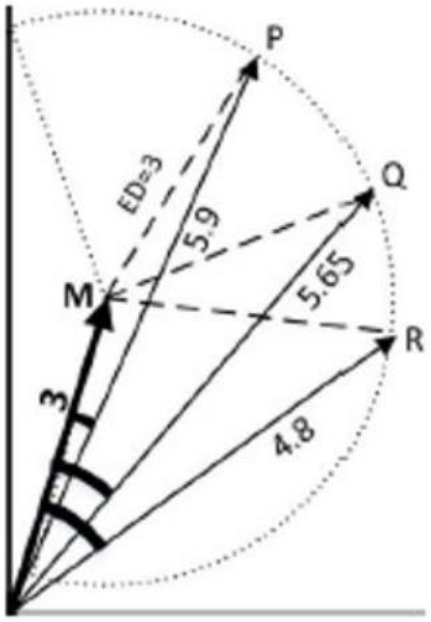

Figure 1

이때 Figure1 처럼 두 벡터가 이루는 각이 다르더라도, 같은 거리상에 있는 모든 벡터 사이의 유사도는 같다고 평가하는 단점이 있다.

Euclidean Distance 를 수식으로 나타내면 다음과 같다.

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

Equation 1

3.1.2 Manhattan Distance

두 벡터 사이의 거리를 두 좌표의 절대값 차를 이용해서 계산하는 방법이 Manhattan distance 이다. Euclidean Distance 와 마찬가지로 결과값이 작을수록 유사도가 높다고 판단한다.

Manhattan Distance 를 수식으로 나타내면 다음과 같다.

$$d = |a_1 - b_1| + |a_2 - b_2| + \ldots + |a_n - b_n|$$

Equation 2

3.1.3 Cosine Similarity

Fig3 과 같이, 두 벡터 사이 각도의 cosine 값을 이용하여 유사한 정도를 계산한다. 0 에서 1 사이의 값을 가지며, 1 에 가까울수록 유사도가 높다고 판단한다.

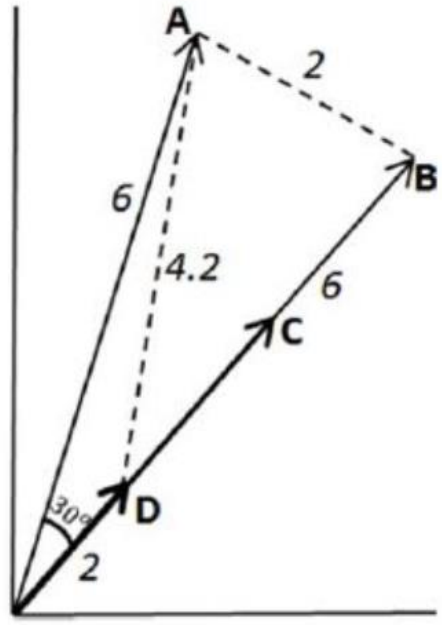

Figure 2

이때 Figure2 처럼 벡터의 크기는 다르지만, 같은 각을 갖는 모든 벡터 사이의 유사도는 같다고 평가하는 단점이 있다.

Cosine Similarity 를 수식으로 나타내면 다음과 같다.

$$s = \frac{\sum_{i=1}^{n}(A_i B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Equation 3

3.1.4 Tanimoto Coefficient

Extended Jaccard Coefficient 와 같다. 두 벡터 간의 유사성을 평가한다. Cosine Similarity 와 마찬가지로. 0 에서 1 사이의 값을 가지며, 1 에 가까울수록 유사도가 높다고 판단한다.

Tanimoto Coefficient 를 수식으로 나타내면 다음과 같다.

$$T = \frac{\sum_{i=1}^{n}(A_i B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} + \sqrt{\sum_{i=1}^{n}(B_i)^2} - \sum_{i=1}^{n}(A_i B_i)}$$

Equation 4

3.1.5 TS-SS

2016 년 논문[3]에서 제안하는 TS-SS 는 Cosine Distance 와 Euclidean Distance 의 단점을 각각 보완한 방법이다.

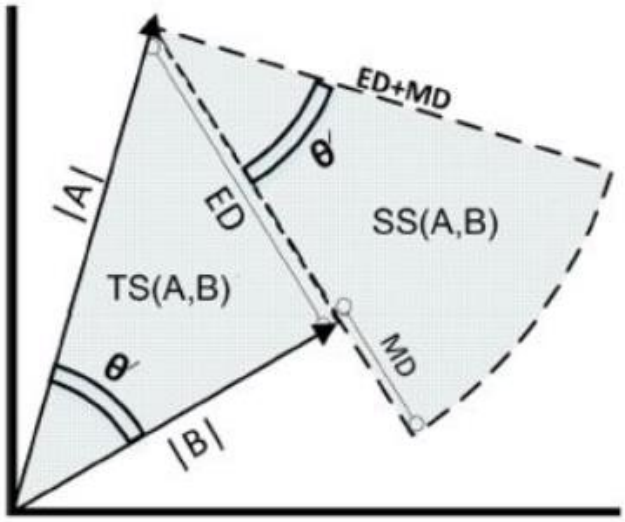

Figure 3

Figure3 에서, TS 는 A 와 B 벡터로 이루어진 삼각형의 넓이이며, SS 는 A, B 벡터의 Euclidean Distance 와 새롭게 정의한 MD 와의 합을 반지름으로 갖는 부채꼴의 넓이이다.

TS 와 SS 를 수식으로 나타내면 다음과 같다.

$$TS(A,B) = \frac{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2} \cdot \sin\theta}{2}$$

Equation 5

$$MD(A,B) = \left| \sqrt{\sum_{i=1}^{k} A_i^2} - \sqrt{\sum_{i=1}^{k} B_i^2} \right|$$

Equation 6

$$SS(A,B) = \pi \cdot (ED(A,B) + MD(A,B))^2 \cdot \left(\frac{\theta}{360}\right)$$

Equation 7

TS-SS 는 TS 와 SS 를 곱한 것으로, 다음과 같이 나타낼 수 있다.

$$TS\_SS(A,B) = \frac{|A| \cdot |B| \cdot \sin\theta \cdot \theta \cdot \pi \cdot (ED(A,B) + MD(A,B))^2}{720}$$

Equation 8

3.2 실험

실험은 페이스북의 대나무숲 페이지 중, 중앙대학교로 한정 지었다.

데이터의 개수의 경우 10000 개와 25000 개일 때 실험을 진행했다. 실험을 진행하기 전에, 데이터는 모두 전처리를 수행한 후, 실험하였다. 이때 전처리는 형태소분석과 불용어제거 등을 의미한다.

3.2.1 10000 개 데이터셋 실험

1. 사랑

Euclidean	Manhattan	Cosine	Tanimoto	TS-SS
행복	행복	바보같은	행복	이모
내	내	어여쁜	그대	일
세상	누구	행복	상처	월
누구	세상	우정과	당신	엄마
만큼	만큼	내	감정	전공
…	…	…	…	…
눈빛	이제	만큼	이제	누나
모습	모습	마지막	마음	아빠
마지막	가슴	길고양이	상대방	공학부
첫사랑	마지막	상처	연애	복수
마음	눈물	다짐	확신	장학금

2. 수업

Euclidean	Manhattan	Cosine	Tanimoto	TS-SS
교수	교수	교수	교수	이모
강의	강의	물시	전공	엄마
교양	교양	강의	수강	오빠
과제	강의실	과제	학생	아빠
강의실	영어	피에르	월	당신
…	…	…	…	…
갈색	노트북	김형진	학년	이기
기숙사	시간표	구기	공학부	대나무
노트북	갈색	학생	여성	사랑
오전	자취	세계경제	복수	월
시간표	학기	우호	영어	학번

3. 이야기

Euclidean	Manhattan	Cosine	Tanimoto	TS-SS
사실	사실	사실	친구	이모
이상	이상	고민	고민	일
하나	하나	하나	행동	엄마
처음	처음	처음	얘기	월
외모	외모	이상	사실	학번
…	…	…	…	…
거리	절대	원래	상대방	누나
생각하는	호구	그건	때문	입학
너	실망	얘기	아니	그대
외로움	할거	건지	이해	집중
호구	한가지	주위	주변	신입생

3.2.2 25000 개 데이터셋 실험

1. 사랑

Euclidean	Manhattan	Cosine	Tanimoto	TS-SS
내	내	내	내	일
순수	순수	번째	그대	삭제
진심	진심	선문답	행복	요청
존재	존재	집시	진심	월
자격	자격	순수	순수	전공
…	…	…	…	…
한때	한때	영원	스스로	누나
다짐	다짐	향란	당신	이모
오랫동안	오랫동안	만큼	아픔	중복
운명	세상	한때	가치	새터
내사랑	권태	스스로	연애	아빠

2. 수업

Euclidean	Manhattan	Cosine	Tanimoto	TS-SS
강의	강의	교수	교수	삭제
교수	교양	강의	강의	일
교양	교수	교양	교양	요청
강의실	강의실	강의실	과목	그대
교시	교시	엘리	강의실	월
…	…	…	…	…
조교	조교	출석	월요일	이별
중간	기말	허연정	출석	전공
범위	과학	미생물	노트북	신청
개론	책상	한국사	학기	미련
과학	개론	정우영	신청	너와

3. 이야기

Euclidean	Manhattan	Cosine	Tanimoto	TS-SS
공감	공감	얘기	얘기	일
여기	여기	공감	카테고리	삭제
최근	최근	여기	공감	요청
주제	비밀	카테고리	친구	전공
공유	주제	기분	이성	월
…	…	…	…	…
일기장	결론	이유	이해	아빠
결론	절대	사실	여기	시험
성별	성별	친구	반응	종일
하소연	사심	낙심	이유	새터
사심	한계	이성	경험	공학부

4. 결론

위 실험은 대나무숲 페이지에서 자주 나오는 키워드를 구한 뒤, 각 키워드와 유사성이 높은 단어 순으로 나열한 것이다. 본 실험으로 알 수 있는 결과는 다음과 같다.

1. TS-SS 가 의외로 결과가 좋지 않다. 논문[3]에 따르면, 문서의 수가 많을 때 성능이 좋다고 알려져 있다. 그러나 10000, 25000 개의 문서에서도 좋은 성능을 보이지 못하고 있다.
2. '수업'이나 '사랑' 키워드를 보면, Cosine similarity 의 경우 문맥상 함께 나올 수 있는 단어 혹은 키워드에서 파생된 단어들이 분포 되어있는 반면, Euclidean Distance 이나 Manhattan Distance, Tanimoto Coefficient 의 결과를 보면 의미상 비슷한 단어가 분포 되어있다.
3. 따라서 한국어에서 실험을 할 때, 데이터가 1M 이상일 경우, TS-SS 를 고려를 하고 그 이하일 경우에는 Cosine Similarity 나 Euclidean Distance 를 사용하는게 좋다.
4. 만약, 의미적 유사성을 포함해서, 연관된 단어까지 분석하고자 한다면, Cosine Similarity 를 이용하는게 좋다.

5. 참고

[1] 이동주, 심준호 (2012), "대수적 특성을 고려한 벡터 유사도 측정 함수의 고찰", 한국전자거래학회지, 17(4), 209-210.

[2] 신동걸, 이승형, 조진성 (2016), "대용량 영화 데이터를 활용한 벡터 유사도 기반 영화 추천 시스템", 한국정보과학회 학술발표논문집, 1857-1859.

[3] Arash Heidarian, Michael J. Dinneen (2016), "A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering", IEEE Second International Conference on Big Data Computing Service and Applications, 978-1-5090-2251-9

[4] Wikipedia, "tanimoto coefficient", https://en.wikipedia.org/wiki/Jaccard_index