이동 평균 기반 동적 시간 와핑 기법을 이용한 시계열 키워드 데이터의 분류 성능 개선 방안*

Enhancing Classification Performance of Temporal Keyword Data by Using Moving Average-based Dynamic Time Warping Method

정도헌 (Do-Heon Jeong)**

초 록

본 연구는 시계열 특성을 갖는 데이터의 패턴 유사도 비교를 통해 유사 추세를 보이는 키워드를 자동 분류하기 위한 효과적인 방법을 제안하는 것을 목표로 한다. 이를 위해 대량의 웹 뉴스 기사를 수집하고 키워드를 추출한 후 120개 구간을 갖는 시계열 데이터를 생성하였다. 제안한 모델의 성능 평가를 위한 테스트 셋을 구축하기 위해, 440개의 주요 키워드를 8종의 추세 유형에 따라 수작업으로 범주를 부여하였다. 본 연구에서는 시계열 분석에 널리 활용되는 동적 시간 와핑(DTW) 기법을 기반으로, 추세의 경향성을 잘 보여주는 이동 평균(MA) 기법을 DTW에 추가 적용한 응용 모델인 MA-DTW를 제안하였다. 자동 분류 성능 평가를 위해 k-최근접 이웃(kNN) 알고리즘을 적용한 결과, ED와 DTW가 각각 마이크로 평균 F1 기준 48.2%와 66.6%의 최고 점수를 보인 데 비해, 제안 모델은 최고 74.3%의 식별 성능을 보여주었다. 종합 성능 평가를 통해 측정된 모든 지표에서, 제안 모델이 기존의 ED와 DTW에 비해 우수한 성능을 보임을 확인하였다.

ABSTRACT

This study aims to suggest an effective method for the automatic classification of keywords with similar patterns by calculating pattern similarity of temporal data. For this, large scale news on the Web were collected and time series data composed of 120 time segments were built. To make training data set for the performance test of the proposed model, 440 representative keywords were manually classified according to 8 types of trend. This study introduces a Dynamic Time Warping(DTW) method which have been commonly used in the field of time series analytics, and proposes an application model, MA-DTW based on a Moving Average(MA) method which gives a good explanation on a tendency of trend curve. As a result of the automatic classification by a k-Nearest Neighbor(kNN) algorithm, Euclidean Distance(ED) and DTW showed 48.2% and 66.6% of maximum micro-averaged F1 score respectively, whereas the proposed model represented 74.3% of the best micro-averaged F1 score. In all respect of the comprehensive experiments, the suggested model outperformed the methods of ED and DTW.

키워드: 동적 시간 와핑, 이동 평균, k-최근접 이웃, 시계열 분석, 패턴 마이닝
dynamic time warping, moving average, k-nearest neighbor, temporal analysis, pattern mining

---

* 본 연구는 2018년도 덕성여자대학교 교내연구비 지원에 의해 이루어졌음(3000003047).
** 덕성여자대학교 문헌정보학과 조교수(doheonjeong@duksung.ac.kr)
■ 논문접수일자 : 2019년 11월 15일 ■ 최초심사일자 : 2019년 12월 11일 ■ 게재확정일자 : 2019년 12월 25일
■ 정보관리학회지, 36(4), 83-105, 2019. http://dx.doi.org/10.3743/KOSIM.2019.36.4.083

1. 서 론

수많은 데이터로부터 가치 있는 정보를 발굴하는 지식 발견(knowledge discovery)의 효과적인 기법을 찾는 일은 최근 많은 데이터 과학자들이 빅데이터와 인공지능 기술을 기반으로 도전하는 매우 중요한 과제 중 하나이다. 대량으로 빠르게 생산되는 빅데이터 환경에서는 일련의 데이터 처리 과정이 매우 빠르게 반복되므로 모든 데이터는 입력되는 즉시 처리되어 분석된 후 활용되어야 한다. 또한 이러한 처리 환경에 있는 대부분의 데이터는 시간의 흐름 속에서 생성되고 적재되는 시계열의 특성을 보인다. 다크 데이터(dark data)를 포함한 미활용 저장 데이터에서 각종 센서 환경(internet of things)으로부터 쏟아져 나오는 실시간 스트림 데이터까지, 다양한 시계열 데이터로부터 지식을 찾고 활용하는 기법의 연구가 더욱 중요해졌다고 할 수 있다.

본 연구는 대량의 데이터로부터 추출된 용어의 시계열 데이터를 생성하고 분석하는 지식 발견의 과정을 다룬다. 이러한 문헌의 추세 분석 연구와 관련하여 다양한 연구가 수행되어 왔다. 거시적 관점에서 문서 내 특정 개념의 시간에 따른 변화를 분석함으로써 국가 간, 학제 간 연구 동향의 변화를 관찰하거나(Glänzel & Schlemmer, 2007), 특정연구 분야에서 떠오르는 신규 기술과 유망 기술을 찾아내기도 한다(Aström, 2007; Doré & Ojasoo, 2001; Zhao & Strotmann, 2007), 또한, 텍스트 마이닝 기법을 이용해 말뭉치의 출현 빈도(tf-idf)를 기반으로 시계열 특성과 주제를 탐지하는 연구를 수행하기도 하며(Abe & Tsumoto, 2010), 기술 용어들의 발전 추세를 분석하고 기술의 미래를 예측함으로써 기업의 기술 경쟁력을 확보하고자 하는 연구가 수행되기도 한다(Daim, Rueda, Martin, & Gerdsri, 2006).

그러나 문헌 기반의 추세 분석 연구가 다수 수행되어 왔음에도 문헌 내 키워드의 추세 신호(trend signals)의 패턴 유사도에 기반하여 용어를 분류하고 분석한 연구는 거의 수행된 바가 없다. 본 연구는 웹 뉴스기사 내에서 유사한 신호 패턴을 갖는 키워드를 자동 분류하는 효과적인 방법을 제안하고자 한다. 이를 위해, 두 시계열 패턴의 유사도를 측정하는 대표적인 알고리즘인 유클리디언 거리 기법 대신 음성인식 분야를 비롯한 신호 처리 분야에서 널리 사용하는 동적 타임 와핑 기법을 활용하였다(이천주, 안원빈, 오경주, 2017; Juang, 1984; Keogh, 2005; Keogh & Pazzani, 2001; Salvador & Chan, 2007). 또한 파동이 심한 실측 데이터의 경향성을 효과적으로 측정하기 위해 이동 평균 기법(Tsokos, 2010; Zhuang, Chen, Wang, & Lian, 2007)을 적용하여, 용어의 패턴 식별 성능을 더욱 개선한 이동 평균 기반 동적 타임 와핑(moving average based dynamic time warping; MA-DTW) 기법을 제안하고자 하였다. 제안 모델을 통해 첫째, 특정 검색어의 추세 분석을 제공하는 현행 서비스가 있다면 특정 기간 동안 유사 추세를 갖는 키워드를 자동으로 찾아주는 기법으로 활용함으로써 기존의 도서관 시스템 및 정보 검색 시스템의 응용 서비스(최상희, 2017; 표순희, 김윤형, 김혜선, 김완종, 2015)를 고도화할 수 있을 것으로 기대한다. 둘째, 색인어의 연관성 분석과 관련된 다양한 텍스트 마이닝 연구(안주영, 안규빈, 송민, 2016; 정도헌,

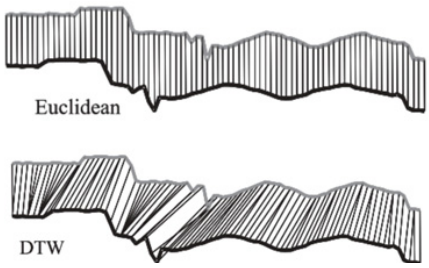
2017; 정도헌, 2018; 정도헌, 주황수, 2018)에 주요 가중치 요소로 활용함으로써 의미적인 분석 기법을 고도화할 수 있을 것이다. 제안한 기법이 향후 패턴 마이닝 분야와 추세 분석 관련 연구 분야에 기여할 수 있기를 기대한다.

본 논문의 구성은 다음과 같다. 2장에서 다양한 분야에서 수행된 관련 연구들을 간략히 소개한 후, 3장에서는 실험을 위한 데이터 수집, 처리의 과정을 소개한다. 대량의 데이터를 수집하고, 자연어 처리를 통해 주요 키워드를 추출하여 추세 데이터를 생산하는 단계, 제안 기법을 적용하여 다양한 모델을 생성하는 단계로 구성된다. 4장은 분석 단계로, 시계열 패턴의 자동 분류 실험을 통해 제안 모델의 성능을 평가하고 실험 과정에서 얻은 추가 분석 내용에 대해 논의한다. 마지막 결론에서는 연구의 성과와 제한점, 향후 연구 방향을 언급한다.

2. 관련 연구

2.1 동적 시간 와핑 기법

두 개의 시계열 데이터 간의 거리 측도로 많이 사용되는 유클리디언 거리(euclidean distance; ED) 알고리즘은 시간의 축이 뒤틀어진 상황에서는 유사도의 계산 정확도가 매우 낮아진다. 이에 반해 동적 시간 와핑(dynamic time warping; DTW) 알고리즘은 이러한 시간 축의 비틀림을 교정하여 유사한 추세를 보이는 시계열 데이터의 식별 성능을 개선하여 준다. 〈그림 1〉과 같이 ED 알고리즘은 기준 X좌표를 기준으로 두 점 간의 거리를 일대 일로 측정하는 데 비해, DTW는 시계열 데이터 간의 최소 거리를 탐색하여 최적의 경로를 도출하는 방법을 이용하는 동적인 해석 알고리즘을 제시한다. 이를 통해 시계열 데이터의 유사도 측정 시 시간의 기준점이 다르거나 시간 축의 스케일이 차이로 성능이 매우 저하되는 단점을 보완할 수 있다(Keogh, 2005).



〈그림 1〉 유클리디언 거리 기법과 동적 시간 와핑 기법의 신호 처리 방식 (Keogh, 2005)

DTW는 주로 자연어 처리(natural language processing; NLP)를 위한 음성 인식(voice recognition) 분야에서 활용되었다(Juang, 1984; Keogh & Pazzani, 2001). 이러한 패턴 마이닝 기반의 신호 유사성(signal similarities) 분석과 관련된 연구와 관련하여, 실시간 센서 데이터의 내용을 분석하거나(Ko, West, Venkatesh, & Kumar, 2005), 금융 정보의 유사 추세를 파악하고(이천주, 안원빈, 오경주, 2017), 다양한 모션 패턴의 유사성을 추출하는(ten Holta, Reinders, & Hendriks, 2007) 등 다양한 응용 분야에서 연구가 수행되고 있다. 또한 DTW 기법은 생명공학 분야에서 유전자 발현(gene expression)과 관련하여 DNA의 시계열 형성과정을 정렬하거나(Aach & Church, 2001), 정보 검색과 시스템 분야에서 유사 내용 분석에 응용하여 멀티미

디어 검색 및 분류 시스템의 성능 향상에 활용되기도 한다(김윤지, 박정희, 2014; Niennattrakul & Ratanamahatana, 2007).

또한, 기법의 성능 개선과 관련하여, 과도한 연산량으로 인한 DTW의 낮은 시간 비용 효율을 개선하고자 하는 알고리즘 연구도 다수 수행되고 있다(서장혁, 정우환, 심규석, 2019; Al-Naymat, Chawla, & Taheri, 2009; Salvador & Chan, 2007). 본 연구에서도 제안 모델의 성능 평가 실험에 막대한 연산량이 요구되었기 때문에, DTW 알고리즘의 지수적인(exponential) 연산량($O(n2)$)을 선형적인(linear) 연산($O(n)$) 알고리즘으로 변환하여 처리 성능을 개선한 FastDTW(Salvador & Chan, 2007) 모델을 활용하여 실험을 수행하였다.

2.2 DTW 관련 k-최근접 이웃 기법

k-최근접 이웃(k-nearest neighbor; kNN) 기법은 분류되지 않은 데이터에 분류된 데이터를 이용해 범주를 부여하는 지도 학습(supervised learning) 기법 중 하나이다. 분류 대상 데이터와 가장 가까운 거리의 데이터를 참고하여 가장 많이 투표된 범주를 선택하는 방식으로, 이때 참조하는 데이터의 수를 k로 표기한다. kNN 기법은 알고리즘이 간단하여 학습의 과정이 매우 빠르며, 모델을 생성하지 않으므로 속성들의 관계를 고려하지 않는다는 특징이 존재한다. 이러한 특징으로 인해 많은 시계열 데이터 기반의 마이닝 성능 평가 연구에서 kNN을 이용하는 다수의 연구 사례가 존재한다(Bagnall, Lines, Bostrom, Large, & Keogh, 2017; Geler, Kurbalija, Radovanović, & Ivanović, 2014; Hsu, Yang, & Lu, 2011; Yang & Shahabi, 2007).

kNN 알고리즘은 파라미터 K의 값이 작아지면 과대적합(over-fitting)되는 경향이 있으며, 너무 커지면 과소적합(under-fitting)이 일어나는 경향이 있는 것으로 알려져 있다. 이와 관련하여 Bagnall, Lines, Bostrom, Large, Keogh (2017)는 다양한 시계열 처리 알고리즘의 성능 비교 평가 연구를 통해 1-NN 기반의 DTW 모델이 실제로 높은 식별 성능을 보임을 밝히기도 하였다. 본 연구에서도 시계열 데이터의 유사도를 측정하고 범주를 부여하기 위해 kNN 기법을 이용하였으며 제안하는 모델에 대한 최적의 k 값을 도출하고자 다양한 성능 평가 실험을 수행하였다.

2.3 이동 평균 기법

많은 시계열 데이터의 추세 분석에 있어 시구간 별 빈도의 등락폭이 크게 나타나는 현상은 전체적인 추세의 파악을 어렵게 한다. 이동 평균(moving average; MA) 기법은 연속된 일정한 구간의 평균을 산출하여 보정된 추세 곡선을 생성함으로써 급격한 신호 파동으로 인한 해석의 어려움을 해결하는 방법이다. 이동 평균 기법의 종류는 단순 이동 평균(simple moving average; SMA), 가중 이동 평균(weighted moving average; WMA) 지수 이동 평균(exponential moving average; EMA) 등이 있으며, SMA가 모든 구간에 동일한 가중치를 주는 단순 모델인 반면, WMA와 EMA는 최신 데이터의 중요성을 더욱 강조하는 가중치 적용 모델이다(Tsokos, 2010; Zhuang, Chen, Wang, & Lian, 2007).

이동 평균 기법을 활용한 연구 사례로, WMA 기법을 기반으로 센서 데이터 처리와 정제에 활용하는 연구(Jeffery, Alonso, Franklin, Hong, & Widom, 2006; Zhuang, Chen, Wang, & Lian, 2007), 자기상관 함수(autocorrelation function) 기반의 추세 예측 기법인 자기회귀 이동 평균(autoregressive moving average; ARMA) 기법 등을 활용하여 대기 정보의 예측과 이상치 처리를 수행한 연구(Rajagopalan & Santoso, 2009), 금융 정보의 이동 평균 패턴을 분석하고 예측하는 시스템을 제안하거나(이재원, 2012), 시스템 이용자의 관심 패턴을 예측하여 이용자 패턴 분석을 하는(박기현, 유상진, 2003) 등의 다양한 연구가 수행되었다.

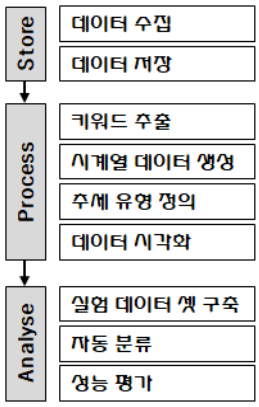
2.4 문헌 시계열 분석 연구

데이터 내에 존재하는 특정 개념이나 사물이 시간에 따라 보이는 특정한 양상을 분석하고자 하는 시계열 분석은, 센서로 부터 측정되거나 입력된 수치 정보의 데이터를 분석하는 정형 데이터 기반의 분석과 자연어 형태로 문헌 내 존재하는 특정 용어의 개념에 대한 추세를 분석하는 비정형 데이터 기반의 분석으로 구분할 수 있다. 앞서 소개한 많은 연구는 수치 데이터에 기반 한 통계적 분석 또는 데이터 마이닝에 의한 패턴 분석의 사례라 할 수 있다. 반면, 문헌에 기반 한 추세 분석은 계량정보학(informetrics) 분야에서 산학연의 연구 동향 변화 등을 관찰하거나(Glänzel & Schlemmer, 2007), 특정연구 분야에서 떠오르는 유망 기술을 찾아내기 위해 다수 수행되었다(Aström, 2007; Daim, Rueda, Martin, & Gerdsri, 2006; Doré & Ojasoo, 2001). 컴퓨터 공학을 비롯한 많은 분야에서 자연어 처리 기법을 기반으로 비정형 문헌 데이터에서 추출한 용어의 시계열 추세를 분석하고자 한 연구도 다수 존재한다. 특히, 문헌 내 말뭉치를 기반으로 하는 주제 탐지 및 추적 연구(topic detection & tracking; TDT)는 텍스트 마이닝을 이용하여 새롭게 부상하는 용어나 주제 등을 파악하고 분석하며(Mei & Zhai, 2005), 단어의 출현 빈도에 기반하여 해당 개념의 중요도를 측정하는 방식에 초점이 맞춰져 있다(Abe & Tsumoto, 2010). 또한, 기술 용어들의 발전 추세를 분석하고 기술의 현재 발전 정도와 발전 속도를 파악함으로써 미래를 예측하고자 하는 연구들이 기술기회의 발굴(technology opportunity discovery; TOD)이라는 측면에서 수행되기도 하였다(Daim, Rueda, Martin, & Gerdsri, 2006; Hwang, Cho, Hwang, Lee, & Jeong, 2011; Kim, Hwang, Jeong, & Jung, 2012).

이와 같이 문헌의 내용 분석 및 시계열 분석과 관련한 다양한 연구가 수행되었음에도 불구하고, 본 연구와 같이 비정형 데이터 환경에서 개별 키워드들의 추세 신호(trend signals)의 패턴 유사성을 효과적으로 파악하기 위한 패턴 마이닝 기법을 제안한 연구는 거의 수행된 바 없다. 앞서 언급한 바와 같이 대규모의 데이터로부터 빠르게 시계열적 패턴과 내용적 특성을 찾아내고 분석하는 기법의 개발은 지식 발견 분야에서 매우 중요한 도전 과제라 할 수 있다. 이후 3장에서는 본 연구가 제안하는, 대량의 문헌 내에서 유사한 시계열 특성을 갖는 키워드를 효과적으로 분류하는 패턴 마이닝 기법을 설명한다.

3. 데이터 수집 및 연구 방법

본 연구의 전체 데이터 프로세싱 과정은 〈그림 2〉와 같이 저장(store), 처리(process), 분석(analyse)의 세 단계로 크게 구분된다. 우선, 웹 뉴스 기사를 크롤링한 후 데이터베이스에 저장한다. 이후 자연어 처리를 통해 키워드를 추출하며 이 과정에서 적절한 정제 과정이 이루어진다. 추세 데이터를 이용해 시계열 패턴을 생성하고 추세 유형을 범주화하여 데이터 모델링을 수행한다. 처리의 마지막 단계로, 차트 이미지 생성을 위한 시각화를 수행한다. 분석 단계에서는 실험 데이터에 범주를 부여하여 데이터 셋을 만들고, 패턴 유사도 측정을 통해 자동 분류를 실시한다. 마지막으로 본 연구에서 제안하는 모델과 베이스라인 모델을 비교하여 최종 성능 평가를 수행한다.



〈그림 2〉 전체 실험 과정 개요

3.1 데이터 수집 및 전처리

추세 분석을 위한 대상 데이터 수집을 위해 네이버의 뉴스기사 서비스를 활용하였다. 정치, 경제, 사회, 생활/문화, 세계, IT/과학, 스포츠, 연예 등 주요 범주별로 서비스의 첫 페이지를 통해 매일 제공되는 TOP 30개의 기사를 웹 크롤러를 통해 수집하였다. 우선 2009년 4월부터 2019년 3월까지 10년간의 기사를 수집한 후, 키워드 별 발생빈도를 1개월 단위, 총 120개로 구간화하여 추세 데이터를 구축하였다. 수집된 기사의 제목에서 키워드를 추출하기 위해 자연어 처리 기법을 사용하였다. 명사 상당어구(noun phrases)를 추출하기 위한 형태소 분석기로 KoNLPy(https://konlpy-ko.readthedocs.io/)를 사용하였다. KoNLPy는 품사(part of speech; POS) 태깅을 위해 다양한 클래스를 임포트하여 사용할 수 있으며, 꼬꼬마(kkma), 코모란(komoran), 한나눔(hannanum), 트위터(twitter) 클래스에 대한 사전 성능 테스트를 통해 한나눔을 POS 태거로 최종 선정하였다. 키워드의 재현율을 높여 보다 정확한 추세를 얻기 위해 "*NN* + NN* + NN* +...*"과 같이 명사 계열(*NN**)로 태깅된 개체가 연속되는 패턴 "*〈NN*〉* 〈NN*〉*"이 나타나는 경우, 가능한 조합을 통해 명사구를 추가 생성하였다. 초기 수집된 웹 문서는 태그 정보를 비롯해 많은 오류 텍스트를 포함하는 경우가 많아 수차례 불용어 리스트를 갱신하여 반복 처리한 후 정제된 최종 색인어 집합을 데이터베이스로 구축하였다.

〈표 1〉은 최종 구축된 데이터 통계를 보여준다. 10년 치 기사 875,772건으로부터 자동 색인된 키워드의 총 수는 약 6백 20만 개이며, 이때 고유한 키워드는 약 170만 종 수준이었다. 120개 구간을 갖는 시계열 데이터로 구성하기 어려운 낮은 빈도의 데이터를 필터링하기 위해 장서 빈도(collection frequency) 10 미만인 키

워드를 제거한 결과, 대부분의 저빈도 키워드가 제거되어 약 4만 3천 여 종의 고유 키워드가 남게 되었다. 키워드 필터링 과정은 실험을 진행하기 위한 데이터 셋을 구축하는 과정에서 후보 색인어의 수를 줄이는 역할을 하므로, 이후의 실험 성능에는 영향을 끼치는 요소가 아니며 일정 빈도 이상으로 색인어 규모를 줄여 데이터 처리량을 줄이는 역할만을 한다.

〈표 1〉 구축된 기사 및 키워드 통계 (2009.4-2019.3)

종류	건수/종수
뉴스 기사	875,772
전체 추출 키워드	6,245,133
고유 키워드	1,693,650
CF 10 이상 키워드	43,082

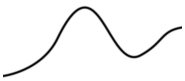
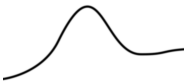
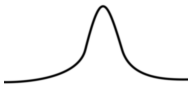
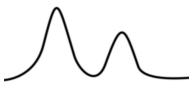
3.2 데이터 모델링과 데이터 셋 구축

주요 키워드를 선정한 후 추세를 해석하기 위해 시계열 데이터의 모델링을 수행하였다. 용어의 시계열 패턴 유형 정의와 관련하여, Abe와 Tsumoto(2010)는 기술용어의 추세를 파악하기 위해 *'emergent'*, *'popular'*, *'subsiding'* 유형으로 용어를 구분하였으며, Hwang, Cho, Hwang, Lee, Jeong(2011)은 대량의 IT 분야의 과학기술 논문으로부터 용어를 추출하고 7가지의 유형, 즉 *'growing'*, *'growth-continuing'*, *'slowing down'*, *'declining'*, *'extinctive'*, *'maintaining'*, *'reviving'*으로 용어 패턴을 분류하였다. 또한, Kim, Hwang, Jeong, Jung(2012)은 세계적인 시장조사 분석 기관인 Gartner 그룹이 매년 발표하는 유망 기술의 발전 주기 그래프인 하이프 사이클(hype cycle)과 유사한 개념을 도입하여 용어를 *'irruption'*, *'frenzy'*, *'turning point'*, *'synergy'*, *'maturity'*의 5단계의 발전 단계로 구분하기도 하였다.

본 연구에서는 추세의 기본형으로 '하강', '상승', '반복' 모형을 설정하고 다양한 변이가 나타나는 상승 모델을 좀 더 세분화하여 최종 8개의 모델을 확정하였다. 상승 모델은 앞서 소개한 시계열 분석 연구에서도 가치가 하락하는 하강 추세보다 중요한 추세로 간주하여 더욱 세분화하여 예측하고자 하였으며, 상승의 크기와 방향, 속도에 따라 그 용어(또는 기술)의 미래 가치를 다양하게 평가하고 있다. 〈표 2〉는 8가지 유형에 대한 설명과 해석을 나타낸다. 첫째, 상승점에서 시작하여 지속적인 하락 추세를 보이는 *'하강형(class1)'*, 둘째, 꾸준하게 지속적인 상승세를 보여주는 *'지속상승형(class2)'*, 셋째, 최근 급상승하는 *'급상승형(class3)'*, 넷째, 상승세가 꺾여 하강한 후 재상승 추세로 전환된 *'재상승형(class4)'*, 다섯째, 재상승 추세(class4)로 발전할 수 있는 가능성이 있는 *'재상승가능형(class5)'*이다. 여섯 번째로, 일시적으로 폭증 현상이 나타나 특정 시점에서 이슈가 되는 사건 또는 인물을 보여주는 *'이슈형(class6)'*이 있다. 마지막으로, 상승과 하강을 반복하는 추세가 규칙적으로 반복되는 패턴을 보이는가의 여부에 따른 *'불규칙형(class7)'*과 *'주기형(class8)'*으로 구분하였다.

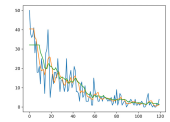
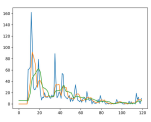
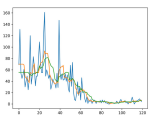
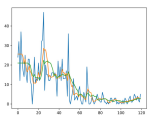
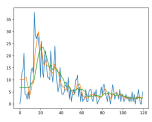
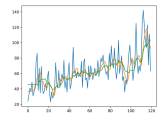
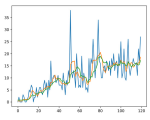
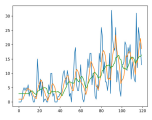
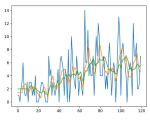
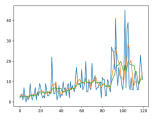
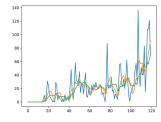
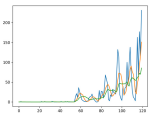
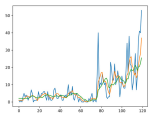
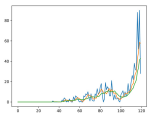
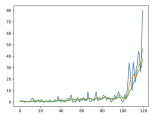
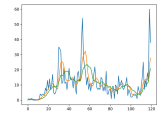
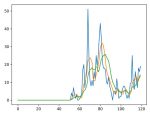
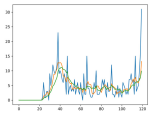
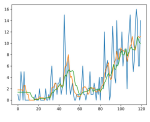
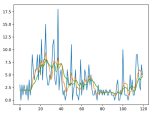
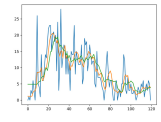
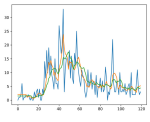
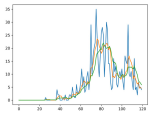
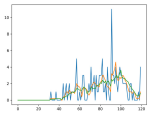
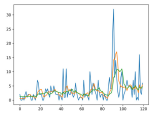
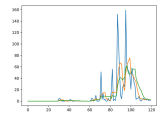
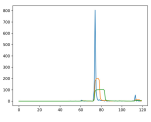
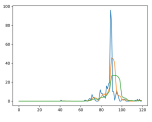
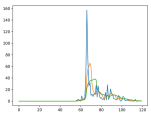
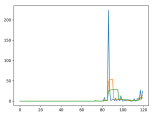
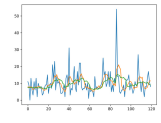
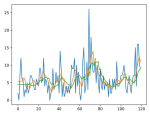
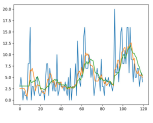
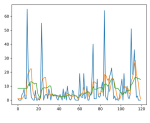
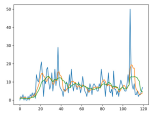
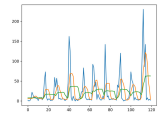
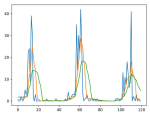
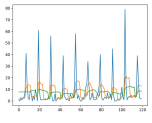
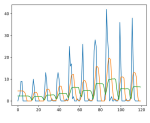
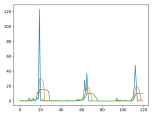
〈표 3〉은 구축된 데이터 셋의 추세 유형별 대표적 사례를 보여준다. 추세 특징을 잘 나타내는 키워드를 범주 당 5개씩 선별하였으나, 유명 정치인, 연예인 등 특정 인물에 대한 최근 인기도, 정치 및 사회적 이슈 현상을 직접 보여줄 수 있는 경우에는 사례에서 의도적으로 배제하였다.

〈표 2〉 키워드의 범주별 추세 유형 정의 및 해석

범주	추세	유형	해석(인물/이슈/사건)
Class1		하강형	특정 인물, 이슈 또는 사건에 대한 사회적 관심도 하락
Class2		지속상승형	인물, 이슈에 대한 관심도 상승
Class3		급상승형 (라이징스타)	최신 관심 인물, 이슈 또는 사건
Class4		재상승형	과거 관심이 높았던 인물이나 이슈가 재등장함
Class5		재상승가능형	특정인이나 사회적 이슈에 대한 관심이 다소 줄어들어 유지되는 상태
Class6		이슈형 (일시적 폭증)	특정인에 대한 일시적인 관심이나 큰 사회적인 이슈가 발생함
Class7		불규칙형 (패턴 없음)	특정 기간 동안 불규칙적으로 지속적으로 사회적 관심이 존재하는 인물이나 사회적 이슈
Class8		주기형 (반복 패턴)	특정 인물보다는 주로 일정한 주기로 발생하는 사회적 이슈, 사건이나 현상

추세를 통해 확인할 수 있는 몇 가지 유형별 특징은 다음과 같다. 최근 급격하게 증가하여 사회적 관심사를 잘 나타내는 모델(class3)을 통해 *'미세먼지'* 문제와 같은 사회 이슈, 급부상하는 기술(*'5G'*)이나 기업(*'화웨이'*)의 추세를 파악할 수 있다. 이슈형 모델(class6)은 일정기간 동안 급증하였다가 일정 시기가 지나 관심이 현저하게 사라지는 모형이다. *'메르스'*, *'사드'*, *'단통법'*, *'브렉시트'* 등이 이러한 유형으로 나타나고 있으며 향후 재상승 여부에 따라 다른 모델로 변화될 수 있다. 인기 전자 제품의 특정 모델들(*'아이폰5'*, *'갤럭시S7'* 등)은 모두 이러한 추세가 나타나 제품 모델별 출시시기 및 관심정도를 쉽게 비교할 수 있다. 또한 특정 인물 또는 사건과 밀접하게 연관된 키워드의 경우 유사한 패턴을 보이기도 한다. class1의 *'박지성'*과 *'퍼거슨(감독)'*, class5의 '손흥민'과 *'토트넘(축구팀)'* 등이 이러한 사례를 보여준다. 흥미로운 모델 중 하나인 주기형(class8)은 *'올림픽'*, *'아시안게임(AG)'*, *'WBC'*, *'총선'*, *'지방선거'*, *'수능'*, *'연말정산'* 등과 같이 일정한 주기로 나타나는 이슈 뿐 아니라 *'태풍'*, *'장마'*, *'황사'*, *'폭우'* 등의 계절 현상도 잘 설명하고 있다. 특히, *'폭염'*, *'무더위'* 등 더위 관련 키워드들은 공통적으로 주기형이면서 양적 상승세도 추가로 보이는 복합 추세를 보이기도 하였다.

〈표 3〉 대표적인 추세 유형별 사례(유형별 5개씩 표기)

Class1					
	프로야구	아이패드	박지성	퍼거슨	트위터
Class2					
	서울	SNS	기온	치매	배터리
Class3					
	손흥민	미세먼지	토트넘	화웨이	5G
Class4					
	갤럭시	샤오미	카톡	OLED	아나운서
Class5					
	태블릿	특허	드론	빅데이터	특혜
Class6					
	사드	메르스	아이폰5	단통법	브렉시트
Class7					
	영국	CCTV	아마존	강진	페이스북
Class8					
	태풍	지방선거	수능	무더위	AG(아시안게임)

수작업 레이블링 작업의 결과로 구축된 실험 데이터 셋의 범주 별 데이터 분포는 〈표 4〉와 같다. 레이블링 작업은 〈표 2〉의 범주별 추세 유형 정의 및 해석 기준을 숙지한 2인의 박사급(사회과학, 공학) 인력이 수행하였으며 지침에 따라 1차 분류를 한 후, 검증을 수행하는 방식으로 작업하여, 총 440개의 키워드로 구성된 데이터 셋을 구축하였다. 다른 범주에 비해 상대적으로 이슈형의 데이터가 다수 관찰되었으며, 특정한 추세가 보이지 않는 class7 유형의 키워드가 상당수임이 확인되었다. 추세 패턴 유형별 데이터의 수작업 분류를 실시하면서 추가적으로 확인된 사항으로, 첫째, 세분화된 유사 추세 모형들은 시간이 지나면서 곡선의 방향 전환에 따라 다른 패턴으로 전환될 수 있으며, 이러한 특성으로 패턴 인식 시 특정 클래스의 성능 저하의 원인이 될 수 있다는 점이다. 예를 들면 재상승 가능성이 있는 class5의 추세는 향후 진행 방향에 따라 class1의 하강형으로 변화하거나 class4의 재상승형으로 변화할 수 있다. 둘째, class7의 불규칙 패턴은 본 연구와 같이 측정 기간이 긴 거시적 분석 경우에는 규칙성을 보이지 않으나, 특정 시점에서 단기간을 분석할 경우에는 규칙적인 모형으로 해석될 수도 있다. 이와 같은 추세 분석과정에서 파악된 몇 가지 이슈는 후속 연구를 통해 심층 분석할 예정이다.

3.3 SMA 기반 DTW 알고리즘

본 연구는 시계열 데이터를 구성하는 모든 지점이 동일한 중요도를 갖는 단순이동 평균 모델을 사용하고 있다. 두 가지의 주요 이유로는, 첫째, 본 연구는 여러 방법론을 융합해 새로운 모델을 실험하고 그 결과를 제시하는 실험적인 성격이 있어, 되도록 가중치 요소가 포함되지 않은 기본 모델의 융합 결과를 살펴보고자 하였다는 점과 둘째, 본 연구는 최신 경향이 더 중요하다는 가중치 모델 개념이 아닌, 시계열을 구성하는 모든 지점에 같은 중요도가 있다는 기본적인 개념을 전제로 한다는 점이다.

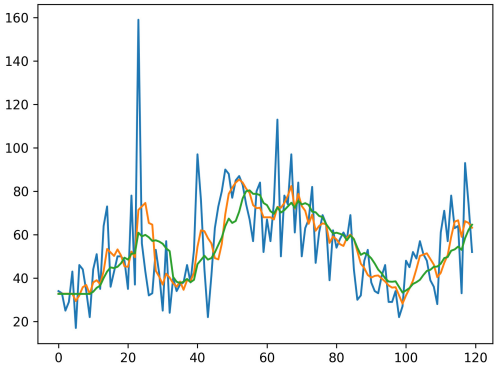
시계열 데이터 $\overline{x_t}$에 대해 k번째의 단순이동 평균값을 구하는 식은 공식 (1)과 같이 표현할 수 있다(Tsokos(2010)의 공식을 수정 및 보완하였음). 여기서, t는 $t=k, k+1, ..., n$의 위치 값을 가지며, k는 시계열 데이터에서 한 번에 가져오는 부분 집합의 크기 즉, 윈도우 사이즈를 의미한다. 이동 평균의 데이터 사이즈를 3으로 지정할 경우, 3개월 측정치의 평균값을 구하여 3번째 위치 값에 대체한다. 즉, $k=3$일 경우는 x_{t-2}, x_{t-1}, x_t의 데이터를 합하여 3으로 나눈 평균값을 x_t의 위치에 대응한다.

$$\overline{x_t} = \frac{(x_{t-k+1} + ... + x_t)}{k} = \frac{\sum_{j=0}^{k-1} x_{t-k+1+j}}{k} \quad (1)$$

〈표 4〉 구축된 실험 데이터 셋의 클래스별 데이터 분포

class1	class2	class3	class4	class5	class6	class7	class8	합계
20	33	38	20	61	80	165	23	440

〈그림 3〉은 실험에서 생성된 추세 차트 중 하나로, 가장 등락폭이 가파른 선(푸른 색)은 전체 구간 120개 지점에 월별 키워드 누적 빈도를 나타내고 있다. 실 데이터에 비해 다소 완만한 곡선은 이동 평균 윈도우 크기 5를 적용한 것(주황색)이며, 가장 부드럽게 움직이는 곡선은 이동 평균 10을 적용한 결과(녹색)이다. 등락폭이 매우 큰 실 데이터에 비해 구간 윈도우의 크기가 커질수록 곡선은 다소 완만해지며 보다 긴 시간동안 누적된 경향성을 보여줌을 확인할 수 있다.



〈그림 3〉 기본 추세 데이터와 이동 평균 기법이 적용된 추세 비교(이동 평균 윈도우 사이즈는 1, 5, 10을 각각 적용)

제안 모델의 기반이 되는 DTW 알고리즘은 다음과 같은 과정을 통해 측정된다. 우선, 길이가 각각 m, n인 두 시계열 데이터 X, Y에 대해, $m \times n$ 행렬을 생성하고 모든 지점에 대해 유클리디언 거리인 $|x_i - y_j|$를 측정한다. 이 때, 패턴의 시작점과 끝점을 일치시켜야 하는 경계 조건(boundary condition)과 최적 경로 탐색 과정은 단조 증가해야 한다는 단조성(monotonicity), 경로 탐색 시 연속되는 이웃 구간으로 진행되어야 한다는 연속성(continuity)을 만족해야 한다(Keogh & Pazzani, 2001). 이러한 조건 하에서 와핑(warping) 경로를 탐색하는 비용 매트릭스가 최소(min() 함수)가 되도록 warping 경로를 발견해야 한다. 따라서, 공식 (2)에서 k번째 와핑 경로인 w_k의 누적 와핑 거리 식 $D(i,j)$는 공식 (3)과 같이 정의할 수 있다. $D(i,j)$는 최종적인 유사도를 나타내는 값으로 $DTW()$에서 최단 경로를 결정하는 측정값이 된다(Keogh, 2005).

$$DTW(X,Y) = \frac{1}{K}\sqrt{\sum_{k=1}^{K} w_k} \quad (2)$$

$$D(i,j) = d(x_i, y_j) + \min\begin{Bmatrix} D(i-1, j-1) \\ D(i-1, j) \\ D(i, j-1) \end{Bmatrix} \quad (3)$$

앞서 설명한 MA와 DTW 알고리즘을 기반으로 작성한 전체 프로세싱 알고리즘은 〈그림 4〉와 같다. 웹에서 수집된 대량의 뉴스기사를 입력 받아(input), 1) 모든 키워드에 대한 추세 차트 그림, 2) 모든 키워드 쌍의 패턴 유사도 측정 결과, 3) 실험을 통한 모델별 분류 성능을 반환(output)하는 처리과정을 거친다. 전체 과정은 크게 3단계로 구분된다. 키워드별 시계열 데이터를 생성하는 단계, 제안 방법에 따라 다양한 분석 모델을 생성하는 단계, kNN 분류기를 통해 추세 패턴의 범주 별 분류 성능을 측정하는 단계이다.

프로세싱이 시작되면, 우선 웹으로부터 수집 저장한 뉴스기사를 자연어 처리하여 키워드를 추출하고 리스트를 만든다(line 1). 모든 키워드의 빈도를 계산하여 120개 구간의 시계열 데이터를 생성한다(line 2-5). 기본 추세 데이터

Input: large scale news data gathered from web
Output: 1) temporal trend charts of all keywords
2) similarity values between keywords
3) experimental result with *knn* classifier
Functions :
- *func* ***MA()*** *#simple moving average algorithm*
- *func* ***ED()*** *#euclidean distance algorithm*
- *func* ***FastDTW()*** *#fast dynamic warping time algorithm*

Main:

```
#step1 : building time series data according to generation options per keyword
build keyword_list from database storing web news data
for keyword in keyword_list:
    fetch all frequency of keyword
    make a baseline TS of keyword #TS = time series
    calculate MA3_TS, MA5_TS, MA10_TS, MA20_TS by MA(TS)  #window size
    store all generated temporal data
    visualize trend chart of each keyword

#step2: generating models applied by proposed method
for keyword in keyword_list:
    making all pairs of keyword
    if timeseries_type = baseline TimeSeries:
        calculate ED(TS_x, TS_y of pairs) and FastDTW(TS_x, TS_y of pairs)
    else if timeseries_type in [MA3_TS, MA5_TS, MA10_TS, MA20_TS]:
        for MTS in [MA3_TS, MA5_TS, MA10_TS, MA20_TS]:
            calcuate FastDTW(MTS_x, MTS_y of pairs)
    store sim_values_list of all pairs

#step3: evaluating the performance of pattern mining with knn classifier
#k-nearest neighbors algorithm
for keyword in keyword_list:
    for k_param in [1, 3, 5, 7]: #four models of knn
        get top k_param word sorted by sim_values_list:
        vote predicted_category of top k_param word
        get final_category from max(predicted_category)
        count results of final_category by comparing with origin_category
        return comprehensive values of TP, TN, FN, FP
```

〈그림 4〉 전체 데이터 처리 과정 알고리즘

를 이용해 4가지 윈도우 크기가 적용된 이동 평균 데이터를 생성한다(line 6). 범주별 레이블링을 위해 추세 차트를 시각화한다(line 8).

두 번째 단계에서, 모든 키워드 데이터에 대해 조합쌍(combination)을 생성한다(line 11-12). ED와 DTW 함수를 이용하여 두 키워드 패턴 간의 유사도를 반복 계산한다. 이때 기본 시계열 데이터는 ED와 DTW 기법을 적용하고, 이동 평균을 적용한 4종의 데이터에 대해서는 DTW를 적용한다(line 13-17). 모든 과정이 끝난 후 다양한 모델별 유사도 측정 결과를 메모리(또는 데이터베이스)에 저장한다(line 18).

세 번째 단계에서, kNN 분류기의 k(=1,3,5,7) 값에 따라 성능 평가를 반복 수행한다(line 22-23). 두 번째 단계에서 얻은 두 키워드 간 유사도 측정 결과를 이용해 상위 k개 만큼의 레이블 데이터를 이용해 다수결 투표한다(line 24-26). 분류기가 예측한 결과와 레이블된 범주의 정답을 비교한다(line 27). 정확률, 재현율 평가를 위한 항목(TP, TN, FN, FP)을 반환하여 성능 평가를 실시한다(line 28).

4. 성능 비교 실험

4.1 실험 환경 및 개요

실험을 위한 주요 환경으로, 인텔 CPU i7-7700k 4.20GHz, Ram 16GB 사양의 데스크탑 PC에서 Python 3.7.x 프로그래밍 언어를 기반으로 Numpy, Pandas 등 데이터 처리 및 분석을 위한 라이브러리, 대량의 추세 차트를 시각화하여 자동 생성하기 위한 Matplotlib 라이브러리, 신호 유사도 측정을 위한 fastDTW와 Scipy(euclidean) 라이브러리를 추가 사용하였다. 본 연구에서는 최적 모델을 도출하기 위해 파라미터를 변경하면서 모델별 반복 실험을 수행하였다. 유사도 측정 대상 수를 N이라 할 때 비교 프로세스의 연산 회수는 총 $N(N-1)/2$ 이다. 본 연구에서는 모델별로 다소 차이는 있으나 1종의 키워드 처리에 평균적으로 약 1분이 소요되었으며, 총 440개의 추세 차트의 상호 유사도를 모두 측정하는데 약 8시간이 소요되었다(96,580회 연산). 신호 유사도를 측정하는 6가지 모델에 대해 k값에 따른 4가지 kNN 분류 모델을 사용하였으므로 총 2,317,920회의 유사도 측정을 실시하였으며, 총 소요 시간은 약 190시간이었다. 앞서 설명한 〈그림 4〉의 알고리즘을 기반으로, 반복 실험의 전 과정을 제어하는 자동화 래퍼(wrapper)를 개발하여 멀티프로세싱을 수행하였다. 각 프로세스의 CPU 점유율이 100%에 근접하므로 속도 면에서 멀티프로세싱의 효과는 크지 않았으나 모든 공정을 제어하는 래퍼를 활용함으로써 막대한 시간이 소요되는 실험의 전 과정을 자동화하는 이점이 있었다.

DTW 알고리즘과의 성능 비교 연구 결과에 따르면, fastDTW 알고리즘은 100개의 포인트 수준에서는 상호 대등한 처리 성능을 보이며 약 10,000에서 100,000개의 포인트를 갖는 그래프에서 처리 속도가 높아지는 알고리즘의 특성이 있으므로(Salvador & Chan, 2007), 본 실험에서 사용하는 fastDTW의 처리 속도 향상은 크지 않을 것으로 판단된다. 향후 연구를 통해 보다 높은 헤르츠(Hz)의 신호 처리를 하는 실험에서 시간 비용 효과가 보다 높을 것으로

기대하며, 파티셔닝(partitioning) 기법 등을 통해 키워드 유사도 측정 시 불필요한 연산을 줄여 전체 프로세싱의 속도를 개선하는 방법에 대한 연구 역시 수행할 필요가 있다.

4.2 성능측정 방법 및 베이스라인 비교

앞서 언급한 바와 같이 수작업 레이블링 작업의 결과, 8종의 클래스 별 데이터 수의 편차가 존재하므로(〈표 4〉 참조), 실험 대상의 범주별 크기가 균등한 실험군의 평가에 적합한 accuracy 대신 데이터 불균등 환경의 평가에 적합한 F1 스코어를 성능 평가 지표로 선정하였다. 우선 데이터의 관측과 예측 결과를 이용하여 〈표 5〉와 같이 분류기가 긍정으로 올바르게 예측한 값을 TP(true positive), 긍정으로 예측했으나 오답인 경우의 값을 FP(false positive), 부정으로 예측한 결과가 옳은 경우를 TN(true negative), 부정으로 예측한 결과가 틀린 경우를 FN(false negative)로 구분하여 정확률(precision)과 재현율(recall)을 산출하였다(공식 (4), (5) 참조). 마지막으로 마이크로 평균(micro averaged) 정확률과 재현율, 매크로 평균(macro averaged) 정확률과 재현율을 각각 측정하고 두 성능 지표를 종합하여 나타내는 F1 점수를 각각 산출하여 최종 실험 결과를 요약 정리하였다(공식 (6) 참조). 자동 분류 실험에서의 매크로 평가는 범주별 재현율과 정확률을 각각 계산하여 합한 후, 이를 범주의 수로 나누어 평균을 구하는 방식이며, 마이크로 평가는 범주를 구별하지 않고 전체 범주 예측 수(TP+FP) 중 정확히 예측한 수(TP)로 정확률을 계산하거나 전체 정답 레이블(TP+FN) 중 정확히 예측한 수(TP)로 재현율을 측정하는 방식이다(Asch, 2013). 이후 모든 실험의 성능은 F1 스코어로 표기하였다.

〈표 5〉 예측 성능 평가용 2X2 분할표 (contingency table)

		Observation	
		YES	NO
Prediction	YES	TP	FP
	NO	FN	TN

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

실험 평가에 사용한 교차 검증(cross validation) 방법과 관련하여, k-fold 교차 검증에서는 일반적으로 k=5~10으로 설정하지만 본 연구에서는 k의 값이 데이터의 개수와 같은 leave-one-out 교차 검증(LOOCV) 방식을 사용하였다. k=440이므로 성능 측정 방식 별로 전체 실험을 440번씩 반복 수행하기 때문에 연산량은 매우 많지만 보다 정확한 평균 성능을 얻을 수 있었다.

본 연구에서는 DTW와 같은 신호 처리 분야의 연구에서 많이 사용하는 kNN을 분류 알고리즘으로 사용하였으며, 파라미터 k를 1, 3, 5, 7의 4가지 모델로 설정하고, 베이스라인 모델인 ED 알고리즘과 DTW 모델의 기본 성능 비교를 실시하였다. 이때 두 베이스라인 모델에 사용한 시계열 데이터는 이동 평균을

적용하지 않은 원본 데이터를 동일하게 사용하였다.

실험 결과, 매크로 평균(macro-averaged) F1의 경우, ED 알고리즘이 40% 중반에서 최고 50% 초반의 결과가 나오는데 비해, DTW 기법의 적용만으로도 50% 후반에서 60% 초반대의 성능 향상이 이루어졌다. 최저 7.3% 포인트에서 최대 14.4% 포인트의 차이가 나타났다. 마이크로 평균(micro-averaged) F1의 경우는 더욱 성능 차이가 나타나고 있다. ED는 분류 성능이 40% 대를 넘지 못하나, DTW는 모든 분류 모델에 대해 60% 이상이 나타나며 최고 66.6%까지 나타나는 것을 확인할 수 있었다(〈표 6〉, 〈그림 5〉 참조).

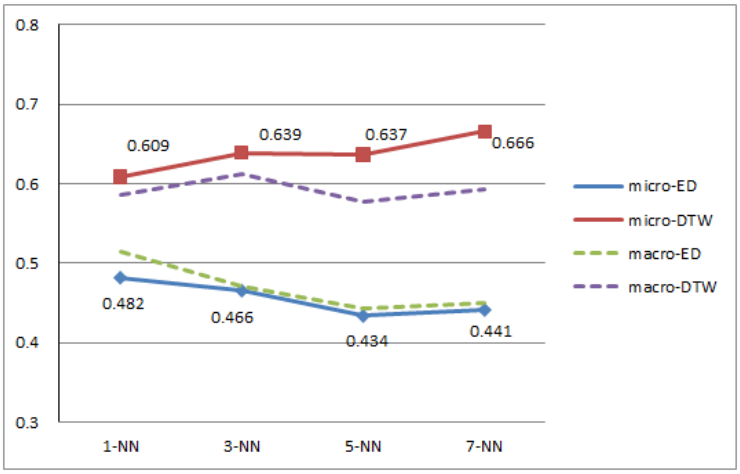
4.3 제안 모델의 성능 비교

두 번째 실험으로, ED와 DTW 베이스라인 모델의 비교 실험을 확장하여 본 연구에서 제안한 MA-DTW 모델의 성능 평가를 수행하였다. 추세 데이터에 이동 평균 윈도우 크기 n=3, 5, 10, 20를 적용한 4가지 유형의 시계열 데이터를 생성하고 DTW 기법을 적용한 DTW_M3, DTW_M5, DTW_M10, DTW_M20 4종 모델의 성능을 측정하였다. kNN 분류기의 파라미터 k는 앞선 실험과 동일하게 1, 3, 5, 7로 설정하여, leave-one-out 교차 검증을 수행하였다.

다양한 반복 실험의 측정 결과는 F1 종합 지

〈표 6〉 ED와 DTW 베이스라인 모델 간 성능 비교

		1-NN	3-NN	5-NN	7-NN
macro- averaged F1	ED	0.514	0.471	0.444	0.449
	DTW	0.587	0.612	0.577	0.593
micro- averaged F1	ED	0.482	0.466	0.434	0.441
	DTW	0.609	0.639	0.637	**0.666**



〈그림 5〉 ED와 DTW 베이스라인 모델 간 성능 비교 차트

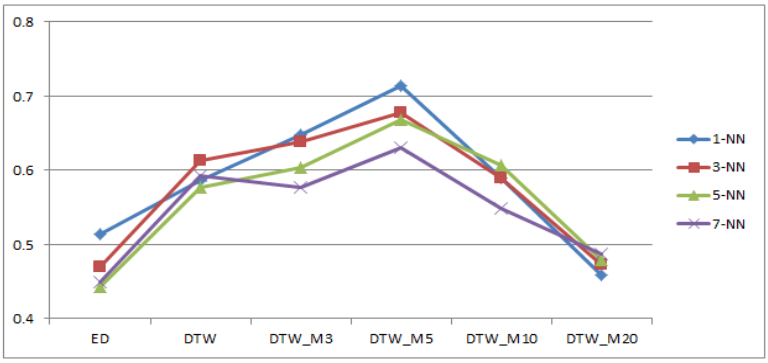
표로 요약하여 〈표 7〉과 〈그림 6〉, 〈표 8〉과 〈그림 7〉과 같이 정리하였다. 우선, 마이크로 평가와 매크로 평가의 전체적인 성능은 동일한 현상을 보이고 있음을 알 수 있다. 최고 분류 성능은 매크로 평균 F1 71.4%, 마이크로 평균 F1 74.3%로 모두 DTW_M5 알고리즘과 1-NN 분류기가 조합된 모델에서 가장 좋은 결과를 보여주었다. 반면에, 최저 성능은 거의 대부분 ED 알고리즘 기반의 모델에서 나타났으며, 그 중 5-NN 분류기를 적용한 경우 매크로 평균 F1 44.4%, 마이크로 평균 F1 43.4%로 최저 성능이 나타났다. 종합 실험을 통해 얻은 결론은 크게 다음과 같이 요약할 수 있다.

1) 베이스라인과 제안 모델의 성능 비교 결과, 제안 모델 DTW_M5가 가장 높은 성능을 보이고 있다. 적절한 이동 평균의 사이즈 적용이 성능 개선에 큰 영향을 주었음을 알 수 있다. 또한, DTW_M20과 같이 지나치게 큰 구간의 이동 평균을 계산한 모델을 제외하면, 적절한 사이즈의 이동 평균을 적용한 모델들이 모두 안정적으로 우수한 성능을 보이고 있음을 확인하였다.

2) DTW에 적용한 kNN 분류기의 모델별 성능 측정 결과, 가장 우수한 성능을 나타내는 DTW_M5와 1-NN이 적용된 모델의 성능이 모든 경우의 실험에서 우수함을 보여주었다. 이 결과는 앞서 언급했던 유사 연구의 실험 결과에서 1-NN, 3-NN 성능이 대체적으로 우수하게 나타난 것과 동일한 현상이었다. 또한, kNN 분류기의 알고리즘 특성을 잘 보여주어, k값이 작을수록 과적합 현상이 나타나며 k값이 높아질수록 성능 곡선이 완만하게 안정적인 경향을 보이고 있음을 동시에 확인 할 수 있었다. 본 실험에서는 ED보다는 DTW 계열 모델들이 이러한 경향성을 잘 보여주고 있다(〈그림 6〉, 〈그림 7〉 참조).

〈표 7〉 매크로 평균 F1 측정 결과

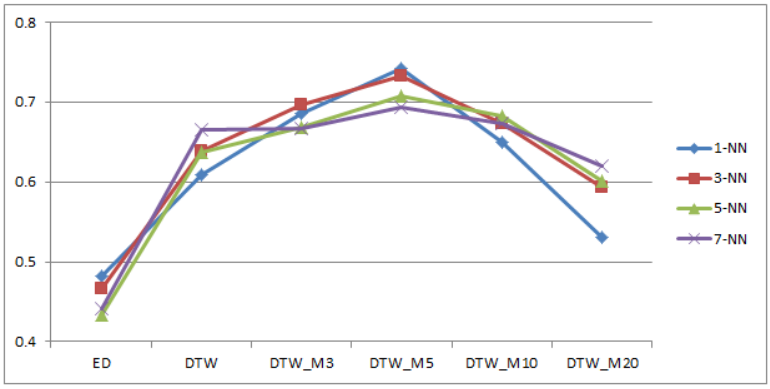
	ED	DTW	DTW_M3	DTW_M5	DTW_M10	DTW_M20
1-NN	0.514	0.587	0.648	**0.714**	0.589	0.460
3-NN	0.471	0.612	0.638	0.677	0.590	0.473
5-NN	**0.444**	0.577	0.604	0.669	0.607	0.479
7-NN	0.449	0.593	0.577	0.631	0.549	0.487



〈그림 6〉 매크로 평균 F1 측정을 통한 모델별 성능 비교

〈표 8〉 마이크로 평균 F1 측정 결과

	ED	DTW	DTW_M3	DTW_M5	DTW_M10	DTW_M20
1-NN	0.482	0.609	0.686	**0.743**	0.650	0.530
3-NN	0.466	0.639	0.697	0.733	0.673	0.593
5-NN	**0.434**	0.637	0.668	0.708	0.682	0.602
7-NN	0.441	0.666	0.666	0.693	0.674	0.620



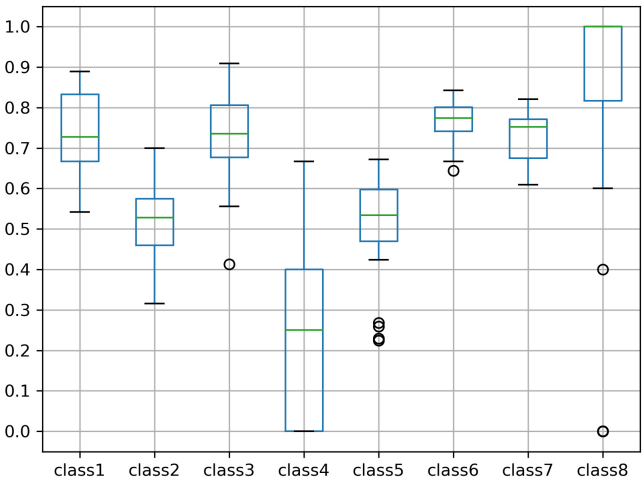
〈그림 7〉 마이크로 평균 F1 측정을 통한 모델별 성능 비교

4.4 범주별 성능 편차 분석

본 연구에서 수행한 총 24가지 모델에 대한 성능 평가 실험을 통해 얻은 매크로 측정 데이터를 이용해, 각 추세의 유형별 성능의 차이를 추가 분석해 보았다. 〈표 9〉는 매크로 평균 정확률에 대한 측정 값을 기술 통계(descriptive statistics)로 요약한 것이며, 〈그림 8〉은 4분위수(quartiles)를 기반으로 데이터 집합의 모습과 이상치를 쉽게 확인할 수 있도록 박스플롯(boxplot) 차트로 표현한 것이다. 이를 통해 범주별 성능 편차의 정도를 확인할 수 있었다. 우

〈표 9〉 범주(class)별 정확률(precision)의 기술통계

	class1	class2	class3	class4	class5	class6	class7	class8
cnt	24	24	24	24	24	24	24	24
mean	0.746	0.514	0.735	0.255	0.503	0.763	0.729	0.775
std	0.101	0.095	0.111	0.214	0.134	0.052	0.06	0.381
min	0.542	0.316	0.413	0	0.224	0.644	0.609	0
25%	0.667	0.46	0.677	0	0.469	0.741	0.675	0.817
50%	0.727	0.528	0.735	0.25	0.534	0.774	0.752	1
75%	0.833	0.574	0.806	0.4	0.598	0.801	0.772	1
max	0.889	0.7	0.909	0.667	0.672	0.842	0.821	1



〈그림 8〉 클래스별 정확률의 boxplot 분석

선, class8(주기형)의 경우 성능 편차가 극단적으로 나타났음을 확인할 수 있다. 이러한 결과는 ED, DTW 베이스라인 모델들이 전혀 해석하지 못한 패턴들을 본 연구에서 제안한 MA-DTW 기법 기반의 모델들이 대체로 정확하게 분류를 해냈기 때문으로 해석할 수 있다. 또한, 상승 계열로 구분되는 유형들 중 class4(재상승형)은 평균 식별 성능이 가장 저조하며, 모델 별 성능의 편차 역시 크게 나타나고 있다. 제안한 방법론의 개선을 위하여, 평균 성능이 저조한 추세 유형에 대한 심층 분석 및 식별 성능 개선을 위한 마이닝 기법의 연구 등이 필요할 것으로 보인다.

5. 결 론

본 연구는 시계열 특성을 갖는 데이터의 신호 패턴 유사도 비교를 통해 유사한 추세를 보이는 키워드를 자동 분류하기 위한 효과적인 방법을 제안하는 것을 목표로 한다. 이를 위해, 웹 사이트로부터 뉴스 기사 10년 치를 수집하고 키워드를 추출한 후 주요 키워드에 대해 120개 구간의 월 단위 시계열 데이터를 생성하였다. 다양한 추세를 보이는 440개의 키워드를 선별하고 대표적인 유형으로 구분하여 총 8가지의 범주를 설정한 후, 수작업으로 범주 레이블을 작성하여 성능 평가를 위한 테스트 셋을 구축하였다.

본 연구에서는 시계열 분석에 널리 활용되는 기법인 동적 시간 와핑 기법(DTW)을 기반으로 추세의 전체적인 경향성 파악에 효과적인 이동 평균(MA) 기법을 적용한 응용 모델인 MA-DTW를 제안하였다. 또한, 시계열 데이터의 패턴 인식 성능 평가를 위해 많은 연구에서 사용하는 k-최근접 이웃 (kNN) 알고리즘을 이용하여 제안 모델의 식별 성능 평가를 수행하였다. 베이스라인 모델과 제안 모델을 포함하여 총 24개 모델에 대한 평가를 수행한 결과, 마이크로와 매크로 평균 F1 점수의 모든 항목에서 제안 모델이 베이스라인 모델인 ED와 DTW에 비해

우수한 성능을 보여주었다. 특히 ED와 DTW가 각각 마이크로 평균 F1 기준 48.2%와 66.6%의 최고 점수를 보인 데 비해, 제안한 모델은 최고 74.3%의 식별 성능을 보여주었다.

본 연구는 주로 음성 인식과 같은 신호처리 분야에서 시계열 데이터의 유사도 측정을 위해 활용하는 DTW 기법을 기반으로, 경제 분야에서 금융 정보의 추세 분석을 위해 사용하는 기법인 MA의 개념을 적용한 새로운 모델을 제안하고, 이를 비정형 문헌 데이터 환경에서 키워드의 특성 및 추세 분석에 활용하였다는 점에서 의의가 있다. 연구의 한계점으로, 본 연구는 새로운 기법과 모델을 제안하고 성능을 측정하는 과정의 수많은 반복 실험으로 인해 막대한 DTW의 연산 시간이 필요하였다. 또한, 실험 결과의 매크로 성능 평가를 통해 추세 범주별 성능의 편차가 존재함을 확인할 수 있었다. 이러한 점을 보완하기 위해 향후에는, 첫째, 분류 속도를 향상할 수 있도록 클러스터링을 이용한 파티셔닝 기법을 적용하는 응용 알고리즘 연구를 수행해야 하며, 둘째, 모델 유형을 통합하거나 세분화하는 등의 최적화 연구를 통해 성능 편차를 감쇄하여 마이크로 정확률과 재현율의 더욱 향상시키는 내용 분석 연구를 수행해야 할 것으로 보인다.

참 고 문 헌

김윤지, 박정희 (2014). 허밍 질의기반 음악 검색을 위한 동적 타임 워핑의 성능향상 방법. 정보과학회논문지: 소프트웨어 및 응용, 41(4), 318-326.

박기현, 유상진 (2003). 이동평균 개념을 이용한 웹 사이트 사용자 관심도 예측 시스템. 경영과학, 20(1), 25-36.

서장혁, 정우환, 심규석 (2019). 희소하고 긴 시계열 데이터의 동적 시간 워핑 거리 상계값 개선. 정보과학회논문지, 46(6), 570-576. http://dx.doi.org/10.5626/JOK.2019.46.6.570

안주영, 안규빈, 송민 (2016). 텍스트 마이닝을 이용한 매체별 에볼라 주제 분석: 바이오 분야 연구논문과 뉴스 텍스트 데이터를 이용하여. 한국문헌정보학회지, 50(2), 289-307. https://doi.org/10.4275/KSLIS.2016.50.2.289

이재원 (2012). 이동 평균선 패턴과 전환점 행렬에 기반한 주식 거래 시스템. 정보과학회논문지: 컴퓨팅의 실제 및 레터, 18(7), 528-532.

이천주, 안원빈, 오경주 (2017). DTW를 이용한 패턴기반 일중 price momentum 효과 분석. 한국데이터정보과학회지, 28(4), 819-829. http://dx.doi.org/10.7465/jkdi.2017.28.4.819

정도헌 (2017). 자동 분류 기법과 지적 구조 분석 기법을 융합한 처방적 분석 시스템 구현 방안 연구. 정보관리학회지, 34(4), 33-57. https://dx.doi.org/10.3743/KOSIM.2017.34.4.033

정도헌 (2018). 데이터 활용률 제고를 위한 기술 용어의 상호 네트워크 생성과 통제. 정보관리학회지,

35(1), 157-182. https://dx.doi.org/10.3743/KOSIM.2018.35.1.157

정도헌, 주황수 (2018). 토픽 모델링 기반 내용 분석을 통한 학제 간 융합기술 도출 방법. 정보관리학회지, 35(3), 77-100. http://doi.org/10.3743/KOSIM.2018.35.3.077

최상희 (2017). 독자 추천도서 정보를 이용한 작가 이미지 분석 연구. 정보관리학회지, 34(4), 153-171. https://doi.org/10.3743/KOSIM.2017.34.4.153

표순희, 김윤형, 김혜선, 김완종 (2015). 도서관 빅데이터 서비스 모형 개발에 관한 연구: 공공도서관을 중심으로. 정보관리학회지, 32(2), 63-86. https://doi.org/10.3743/KOSIM.2015.32.2.063

Aach, J., & Church, G. M. (2001). Aligning gene expression time series with time warping algorithms. Bioinformatics, 17(6), 495-508. http://dx.doi.org/10.1093/bioinformatics/17.6.495

Abe, H., & Tsumoto, S. (2010). Trend detection from large text data. 2010 IEEE International Conference on Systems Man and Cybernetics (SMC), 310-315. http://dx.doi.org/10.1109/ICSMC.2010.5641682

Al-Naymat, G., Chawla, S., & Taheri, J. (2009). SparseDTW: a novel approach to speed up dynamic time warping. Proceeding of the Eighth Australasian Data Mining Conference, 101, 117-127.

Asch, V. V. (2013). Macro- and micro-averaged evaluation measures [BASIC DRAFT].

Aström, F. (2007). Changes in the LIS research front: time-sliced cocitation analyses of LIS journal articles, 1990-2004. Journal of the American Society for Information Science and Technology, 58(7), 947-957. http://dx.doi.org/10.1002/asi.20567

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances, 31(3), 606-660.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. Technological Forecasting and Social Change, 73(8), 981-1012. http://dx.doi.org/10.1016/j.techfore.2006.04.004

Doré, J. C., & Ojasoo, T. (2001). How to analyze publication time trends by correspondence factor analysis: Analysis of publications by 48 countries in 19 disciplines over 12 years. Journal of the American Society for Information Science and Technology, 52(9), 763-769. http://dx.doi.org/10.1002/asi.1130

Geler, Z., Kurbalija, V., Radovanović, M., & Ivanović, M. (2014). Impact of the sakoe-chiba band on the DTW time-series distance measure for kNN classification. International Conference on Knowledge Science, Engineering and Management (KSEM 2014): Knowledge Science,

Engineering and Management, 105-114.

Glänzel, W., & Schlemmer, B. (2007). National research profiles in a changing europe (1983-2003): An exploratory study of sectoral characteristics in the Triple Helix. Scientometrics, 70(2), 267-275. http://dx.doi.org/10.1007/s11192-007-0203-8

Hsu, H. H., Yang, A. C., & Lu, M. D. (2011). KNN-DTW based missing value imputation for microarray time series data. Journal of Computers, 6(3), 418-425. http://dx.doi.org/10.4304/jcp.6.3.418-425

Hwang, M. N., Cho, M. H., Hwang, M., Lee, M., & Jeong, D. H. (2011). Application of trend detection of technical terms to technology opportunity discovery. Communications in Computer and Information Science (CCIS), 264, 258-262. http://dx.doi.org/10.1007/978-3-642-27210-3_33

Jeffery, S. R., Alonso, G., Franklin, M. J., Hong, W., & Widom, J. (2006). Declarative support for sensor data cleaning. International Conference on Pervasive Computing (LNCS 3968), 88-100.

Juang, B.-H. (1984). On the hidden markov model and dynamic time warping for speech recognition - A unified view. AT&T DELL LAB Technical Journal, 63(7), 1213-1243. http://dx.doi.org/10.1002/j.1538-7305.1984.tb00034.x

Keogh, E. (2005). Exact indexing of dynamic time warping. Knowledge and Information Systems, 7(3), 358-386. http://dx.doi.org/10.1007/s10115-004-0154-9

Keogh, E. J., & Pazzani, M. J. (2001). Derivative dynamic time warping. Proceedings of the 2001 SIAM International Conference on Data Mining, 1-11. http://dx.doi.org/10.1137/1.9781611972719.1

Kim, J., Hwang, M., Jeong, D.H., & Jung, H. (2012). Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. Expert Systems with Applications, 39(2012), 12618-12625. http://dx.doi.org/10.1016/j.eswa.2012.05.021

Ko, M. H., West, G., Venkatesh, S., & Kumar, M. (2005). Online context recognition in multisensor systems using dynamic time warping. 2005 International Conference on Intelligent Sensors, Sensor Networks and Information Processing. http://dx.doi.org/10.1109/ISSNIP.2005.1595593

Mei, Q., & Zhai, C. X. (2005). Discovering evolutionary theme patterns from text: An exploration of temporal text mining. The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 198-207. http://dx.doi.org/10.1145/1081870.1081895

Niennattrakul, V., & Ratanamahatana, C. A. (2007). On clustering multimedia time series data

using K-Means and dynamic time warping. 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07). http://dx.doi.org/10.1109/MUE.2007.165

Rajagopalan, S., & Santoso, S. (2009). Wind power forecasting and error analysis using the autoregressive moving average modeling. 2009 IEEE Power & Energy Society General Meeting. http://dx.doi.org/10.1109/PES.2009.5276019

Salvador, S., & Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis, 11(5), 561-580. http://dx.doi.org/10.3233/IDA-2007-11508

ten Holt, G. A., Reinders, M. J. T., & Hendriks, E. A. (2007). Multi-dimensional dynamic time warping for gesture recognition. Thirteenth annual conference of the Advanced School for Computing and Imaging

Tsokos, C. P. (2010). K-th Moving, Weighted and exponential moving average for time series forecasting models. European Journal of Pure and Applied Mathematics, 3(3), 406-416.

Yang, K., & Shahabi, C. (2007). An efficient k nearest neighbor search for multivariate time series. Information and Computation, 205(1), 65-98. http://dx.doi.org/10.1016/j.ic.2006.08.004

Zhuang, Y., Chen, L., Wang, X.S., & Lian, J. (2007). A weighted moving average-based approach for cleaning sensor data. 27th International Conference on Distributed Computing Systems (ICDCS '07). http://dx.doi.org/10.1109/ICDCS.2007.83

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

An, Juyoung, Ahn, Kyubin, & Song, Min (2016). Text mining driven content analysis of ebola on news media and scientific publications. Journal of the Korean Society for Library and Information Science, 50(2), 289-307. https://doi.org/10.4275/KSLIS.2016.50.2.289

Choi, Sanghee (2017). Analysis of author image based on book recommendation from readers. Journal of the Korean Society for information Management, 34(4), 153-171. https://doi.org/10.3743/KOSIM.2017.34.4.153

Jeong, Do-Heon (2017). Prescriptive analytics system design fusing automatic classification method and intellectual structure analysis method. Journal of the Korean Society for information Management, 34(4), 33-57. https://dx.doi.org/10.3743/KOSIM.2017.34.4.033

Jeong, Do-Heon (2018). Generating and controlling an interlinking network of technical terms

to enhance data utilization. Journal of the Korean Society for information Management, 35(1), 157-182. https://dx.doi.org/10.3743/KOSIM.2018.35.1.157

Jeong, Do-Heon, & Joo, Hwang-Soo (2018). Discovering interdisciplinary convergence technologies using content analysis technique based on topic modeling. Journal of the Korean Society for information Management, 35(3), 77-100. http://doi.org/10.3743/KOSIM.2018.35.3.077

Kim, Yunji, & Park, Cheong Hee (2014). An improved dynamic time warping method for query by humming. Journal of Korean Institute of Information Scientists and Engineers(KIISE): Software and Applications, 41(4), 318-326.

Lee, Chunju, Ahn, Wonbin, & Oh, KyongJoo (2017). Analysis of intraday price momentum effect based on patterns using dynamic time warping. Journal of the Korean Data & Information Science Society, 28(4), 819-829. http://dx.doi.org/10.7465/jkdi.2017.28.4.819

Lee, Jae Won (2012). A stock trading system based on moving average patterns and turning point matrix. Journal of KIISE: Computing Practices and Letters, 18(7), 528-532.

Park, KeeHyun, & Yoo, Sangjin (2003). A prediction system on user interest degree to web sites using the concept of the moving averages. Korean management science review, 20(1), 25-36.

Pyo, Soon Hee, Kim, Yun Hyung, Kim, Hye Sun, & Kim, Wan Jong (2015). A study on the developing of big data services in public library. Journal of the Korean Society for information Management, 32(2), 63-86. https://doi.org/10.3743/KOSIM.2015.32.2.063

Seo, Janghyuk, Jung, Woohwan, & Shim, Kyuseok (2019). Improving the upper bound of the dynamic time warping for sparse and long time sequences. Journal of Korean Institute of Information Scientists and Engineers(KIISE), 46(6), 570-576. http://dx.doi.org/10.5626/JOK.2019.46.6.570