KIPS Tr. Comp. and Comm. Sys.
Vol.3, No.2 pp.43~46 pISSN: 2287-5891

http://dx.doi.org/10.3745/KTCCS.2014.3.2.43

Instance-Level Subsequence Matching Method based on a Virtual Window

Sun-Young Ihm† · Young-Ho Park††

ABSTRACT

A time-series data is the collection of real numbers over the time intervals. One of the main tasks in time-series data is efficiently to find subsequences similar to a given query sequence. In this paper, we propose an efficient subsequence matching method, which is called Instance-Match (I-Match). I-Match constructs a virtual window in order to reduce false alarms. Through the experiment with real data set and query sets, we show that I-Match improves query processing time by up to 2.95 times and significantly reduces the number of candidates comparing to Dual Match.

Keywords : Subsequence Matching, Time-Series Data, Instance-level Query Processing

가상 윈도우 기반 인스턴스 레벨 서브시퀀스 매칭 방안

임 선 영† · 박 영 호††

요 약

시계열 데이터는 시간에 따라 변화되는 실수 값을 저장한 것이다. 시계열 데이터에서 사용자 질의 시퀀스가 주어졌을 때, 유사한 서브시퀀스를 가지는 데이터 시퀀스를 검색하는 서브시퀀스 매칭은 매우 중요한 문제이다. 본 논문에서는 인스턴스 레벨의 새로운 서브시퀀스 매칭 방법인 I-Match (Instance-Match)를 제안한다. I-Match는 인스턴스 레벨에서 가상 윈도우를 생성하여 질의 시퀀스와 데이터 시퀀스를 비교하여 착오 해답을 줄이는 방법으로 기존 방법인 Dual Match에 비해 후보의 개수를 줄임으로써 성능을 향상시켰다. 실험을 통해 I-Match의 질의 처리 시간이 Dual Match와 비교하여 최대 2.95배 빠르며, 후보의 개수를 줄임을 보인다.

키워드 : 서브시퀀스 매칭, 시계열 데이터, 인스턴스 레벨 질의 처리

1. 서 론

시계열 데이터(time-series data)는 시간에 따라 변화되는 데이터의 값을 저장한 것이다. 예를 들어, 주식 데이터, 음악 데이터 등과 같이 정보통신산업, 유비쿼터스 환경, 모바일 컴퓨팅 등 다양한 분야에서 시간에 따라 값이 변화하는 특성을 가지는 데이터들이 시계열 데이터가 될 수 있다. 이런 시계열 데이터를 데이터 시퀀스라고 부르며, 사용자의 질의는 질의 시퀀스라고 부른다. 사용자 질의 시퀀스가 주어졌을 때, 유사한 시퀀스를 찾는 방법을 시퀀스 매칭이라고 하는데, 시퀀스 매칭은 질의 시퀀스가 데이터 시퀀스와 길이가 같은 전체 매칭과 서로 길이가 다른 서브시퀀스 매칭으로 나눌 수 있다 [1]. 전체 매칭에서는 질의 시퀀스와 사용자가 정의한 허용치 t가 주어졌을 때 유사한 데이터 시퀀스를 모두 검색한다. 서브시퀀스 매칭에서는 질의 시퀀스와 허용치 t가 주어졌을 때, 유사한 데이터 서브시퀀스를 모두 검색한다. 이 때, 유사한 시퀀스란 두 시퀀스 간의 거리가 t보다 작음을 뜻하며, 본 논문에서는 시퀀스 간의 거리를 구하기 위해 유클리디안 방법을 사용한다. 이러한 서브시퀀스 매칭 방법은 패턴 매칭, 모션 예측, 규칙 발견 프로그램 등 시계열 데이터의 패턴을 검색하는데 많이 사용 된다 [3, 5].

서브시퀀스 매칭 방법은 먼저 데이터시퀀스와 질의 시퀀스를 저차원 변환하여 요약된 정보를 저장한 후, 요약된 정보를 먼저 비교하여 후보를 찾아낸다. 다음으로 후보들에 대하여 유클리디안 방법으로 거리 계산을 함으로써 계산 비용을 줄인다. I-adaptive [1]는 서브시퀀스 매칭을 위하여 데이터 시퀀스를 슬라이딩 윈도우로 나누고, 질의 시퀀스를 디스조인트 윈도우로 나누는 방법을 제안하였다. 하지만, 이 방법은 데이터 시퀀스를 슬라이딩 윈도우로 나누므로 각각의 서브시퀀스의 개별 점을 저장하는 오버헤드가 매우 커지며, 후보의 개수가 매우 많다는 문제가 있다. Moon 등 [4]

※ 이 논문은 숙명여자대학교 SRC여성질환연구센터 특별연구비 지원으로 수행되었음(2011).
† 준 회 원 : 숙명여자대학교 멀티미디어과학과 박사과정
†† 종신회원 : 숙명여자대학교 멀티미디어과학과 부교수
논문접수 : 2014년 2월 20일
심사완료 : 2014년 2월 21일
* Corresponding Author : Young-Ho Park(yhpark@sm.ac.kr)

은 이 문제를 해결하기 위하여 I-adaptive의 이원적인 방법으로 데이터 시퀀스를 디스조인트 윈도우로 나누고 질의 시퀀스를 슬라이딩 윈도우로 나누는 방법인 Dual Match를 제안하였다. Dual Match는 I-adaptive와 비교하여 후보의 개수를 줄여 질의 처리 시간과 성능을 개선하였다. 하지만 여전히 많은 착오 해답(*false alarm*, 후보이나 실제로는 질의 시퀀스와 유사하지 않은 서브시퀀스)을 발생한다는 문제점이 있다.

E-Dual Match [8]는 Dual Match의 착오 해답을 줄이기 위하여 인덱스 레벨의 새로운 방법을 제안하였다. E-Dual Match는 먼저 저차원 변환한 요약된 정보를 순차적으로 저장하는 인덱스를 추가로 생성한 후, Dual Match와 마찬가지로 데이터 시퀀스를 디스조인트 윈도우로 나누고, 질의 시퀀스를 슬라이딩 윈도우로 나눈다. 그리고 k-윈도우 정책에 따라 서브시퀀스 안의 모든 윈도우를 새로운 인덱스를 사용하여 비교함으로써 Dual Match와 비교하여 착오 해답을 줄여 성능을 개선하였다. E-Dual Match는 인덱스 레벨에서의 착오 해답은 개선하였지만, 여전히 인스턴스 레벨에서의 착오 해답이 많이 발생하기 때문에 계산 비용이 크다는 문제점이 있다. 따라서 본 논문에서는 이를 개선하기 위하여 인스턴스 레벨에서의 착오 해답을 줄이는 방법인 I-Match (*Instance-Match*)를 제안한다. 인덱스 레벨에서 k-윈도우 정책에 따라 시퀀스 안의 모든 윈도우를 비교하고 나면, 비교하지 못하고 남는 부분이 생기게 된다. 이렇게 남는 부분으로 인하여 착오 해답이 많이 발생하게 되는데, 본 논문에서는 이러한 남는 부분을 사용하여 가상 윈도우를 생성한 후, 인스턴스 레벨에서 가상 윈도우를 비교함으로써 착오 해답을 줄여서 계산 비용을 줄임으로써 성능을 향상시키는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문과 관련된 기존의 연구들을 살피고, 3장에서는 본 논문에서 제안하는 새로운 인스턴스 레벨의 서브시퀀스 매칭 방법인 I-Match를 설명한다. 4장에서는 실험을 통해 제안하는 방법의 성능상의 이점을 보이고, 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

본 장에서는 시계열 데이터에서의 서브시퀀스 매칭 방법들 중 본 논문과 직접적으로 관련이 있는 연구들에 대해 설명한다. 서브시퀀스 매칭 방법에서 검색 시간을 줄이는 것은 매우 중요한 문제이다. 이러한 서브시퀀스 매칭 방법은 많은 연구에서 다루어져왔다 [1, 3-4, 6-8].

I-adaptive [1]는 인덱스를 생성하는 단계와 서브시퀀스 매칭을 하는 단계로 이루어져 있다. 먼저 인덱스 생성 단계에서는 데이터 시퀀스를 저차원 변환한 후 요약된 정보들을 가지고 인덱스를 생성한다. 다음으로 질의 처리 단계에서는 사용자 질의와 허용치 t가 주어지면 질의 시퀀스와 서브시퀀스 간의 유클리디안 거리가 t보다 작은 서브시퀀스를 가지는 데이터 시퀀스를 모두 검색한다. 이 때, I-adaptive는 데이터 시퀀스를 슬라이딩 윈도우로 나누고, 질의 시퀀스를 디스조인트 윈도우로 나눈다. 하지만 데이터 시퀀스를 슬라이딩 윈도우로 나누면서 후보가 많이 생겨 착오 해답을 많이 발생시킨다.

Dual Match [4]는 이러한 문제를 해결하기 위하여 데이터 시퀀스를 디스조인트 윈도우로, 질의 시퀀스를 슬라이딩 윈도우로 나누었다. 하지만 후보의 개수가 많기 때문에 착오 해답이 여전히 많이 발생한다. E-Dual Match [8]는 순차적인 값을 저장하는 인덱스를 추가로 생성하고, k-윈도우 정책을 사용함으로써 인덱스 레벨에서 후보의 개수를 줄였다. 하지만 k개의 윈도우를 비교한 후 남는 부분이 발생하면서 착오 해답이 발생하는 문제가 있다.

3. 인스턴스 레벨의 서브시퀀스 매칭 방법

본 장에서는 본 논문에서 제안하는 인스턴스 레벨의 서브시퀀스 매칭 방법인 I-Match에 대하여 설명한다. 먼저 3.1 절에서는 I-Match의 생성 단계에 대하여 설명하고, 3.2 절에서는 I-Match 방법 중 가상 윈도우를 생성하는 방법에 대하여 자세히 설명한다.

3.1 I-Match 생성 단계

본 절에서는 I-Match의 생성 단계에 대하여 설명한다. I-Match는 총 네 단계로 구성된다. 먼저 첫 번째 인덱스 생성 단계에서는 E-Dual Match [8]과 마찬가지로 데이터 시퀀스를 저차원 변환한 요약된 정보로 R*-tree 인덱스와 순차적인 파일 인덱스를 생성한다. 두 번째 윈도우 필터링 단계에서는 k-윈도우 정책 [8]에 따라 서브시퀀스 안의 k개의 윈도우를 모두 비교한다. 세 번째 가상 윈도우 생성 단계에서는 두 번째 단계에서 k개의 윈도우를 비교하고 남은 부분을 하나의 가상 윈도우로 만들어 비교한다. 마지막으로 실제 거리 계산 단계에서는 질의 시퀀스와 데이터 시퀀스 사이의 실제 거리를 비교하여 유사 시퀀스인지 판별한다. 각 단계 별 자세한 설명은 다음과 같다.

a) 단계 1 (*인덱스 생성 단계*) : I-Match의 첫 번째 단계인 인덱스 생성 단계에서는 E-Dual Match [8]와 같은 방법으로 저차원 변환된 데이터 시퀀스 값들을 저장하여 R*-tree와 순차적인 정보를 가지는 파일 인덱스를 생성한다. 또한, 데이터 시퀀스는 디스조인트 윈도우로 나누어 저장한다.
b) 단계 2 (*윈도우 필터링 단계*) : 윈도우 필터링 단계에서는 먼저 단계 1에서 생성된 인덱스를 검색하여 데이터 서브시퀀스와 질의 시퀀스를 비교한다. 이 때, k-윈도우 정책 [8]을 적용하여 시퀀스 내의 모든 윈도우를 비교한다. 두 시퀀스 간의 거리가 허용치 t보다 작다면 그 데이터 서브시퀀스는 후보로 설정된다.
c) 단계 3 (*가상 윈도우 생성 단계*) : 단계 2에서 k개의

윈도우를 비교하고 나면 데이터 서브시퀀스와 질의 시퀀스에는 앞, 뒤로 남는 구간이 생기게 된다. 이렇게 남는 구간을 하나의 가상 윈도우로 구성하여 거리를 비교했을 때, 허용치 t보다 크다면 후보에서 빠지게 된다.

d) 단계 4 (*실제 거리 계산 단계*) : 마지막 단계에서는 후보 데이터 시퀀스와 질의 시퀀스간의 실제 거리를 유클리디안 방법으로 계산하여 결과인지 판별한다.

3.2 I-Match의 가상 윈도우 생성 방법

본 절에서는 I-Match의 가상 윈도우 생성 방법에 대하여 자세히 설명한다. Fig. 1은 3.1절에서 설명한 단계 2의 윈도우 필터링 단계를 마친 후에 데이터 서브시퀀스에 남는 구간을 보여주고 있다. 파란색 그래프는 시간에 따른 시그널 값을 나타내는 시계열 데이터 시퀀스 중 i번째 데이터 서브시퀀스이다. 빨간색 점선으로 표시된 부분은 디스조인트 윈도우이며, 단계 2에서는 서브시퀀스 안에 있는 두 개의 디스조인트 윈도우가 질의 시퀀스와 비교되었다. 하지만 이 때, 초록색 점선으로 표시된 것처럼 앞, 뒤로 비교하지 못하고 남는 구간이 생기게 된다.

본 논문에서는 이렇게 남는 구간을 하나의 가상 윈도우로 생성하여 비교한다. E-Dual Match에서는 이 남는 구간을 비교하지 못했기 때문에 남는 구간에서 질의 시퀀스와 거리 차이가 많이 발생함에도 불구하고 후보로 설정되어 착오 해답이 많이 발생되었다. 하지만 I-Match는 이를 가상 윈도우로 구성하여 비교하기 때문에 착오 해답을 줄일 수 있다는 장점이 있다.

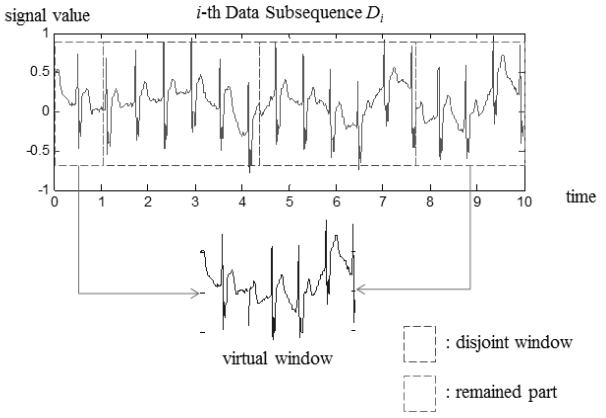

Fig. 1. Creating process of virtual window

4. 실 험

I-Match는 후보들의 개수를 줄여 착오 해답의 발생을 줄이고, 서브시퀀스 매칭의 성능을 매우 향상시켰다. 실험을 통해 I-Match가 Dual Match와 비교하여 최대 2.95배 질의 처리 시간을 단축 한다는 것을 보였고, 후보들의 수를 줄일 수 있음을 보였다. 본 장에서는 5.1절을 통해 실험에서 사용된 데이터 및 환경을 설명한 다음 5.2절을 통해 I-Match의 실험 결과를 설명한다.

4.1 실험 데이터 및 실험 환경

실험에는 329,112 엔트리로 구성된 실제 데이터인 주식 데이터 [4]가 사용되었으며 실험에 사용된 질의는 총 10개이다. 각 질의와 유사한 25개의 서브시퀀스를 얻기 위하여 사용자 허용치 t를 설정하였다. Table 1은 실험에 사용된 허용치 t의 값을 요약한 것이다.

실험에서는 I-Match와 Dual Match의 질의 처리 시간과 착오 해답을 발생시키는 후보의 수를 비교하였다. I-Match는 C언어를 사용하여 구현되었으며, 실험은 16GB의 메모리를 가진 2.80GHz 리눅스 PC에서 실행되는 인텔 I5-760 쿼드 코어 프로세서에서 수행한다.

Table 1. Summary of notation

질의 처리 번호	주식 데이터 실험에 사용되는 허용치 t
1	0.0449966
2	0.065017
3	0.052359
4	0.052812
5	0.052937
6	0.053231
7	0.037903
8	0.041119
9	0.043788
10	0.051532

4.2 실험결과

실험에서는 I-Match와 Dual Match 간의 질의 처리 시간과 후보의 수를 비교한다. 질의 처리 시간의 척도로써 wall clock time을 사용한다. 질의 처리 시간과 후보의 수를 대한 쿼리 숫자 1부터 10까지 측정한다.

Fig. 2는 I-Match와 Dual Match 간의 평균 질의 처리 시간을 보여준다. 어두운 회색 막대 그래프는 Dual Match의 질의 처리 시간을 나타내며 밝은 회색 막대 그래프는 I-Match의 질의 처리 시간을 나타낸다. I-Match가 Dual Match보다 질의 처리 시간이 2.95배 빠름을 보인다.

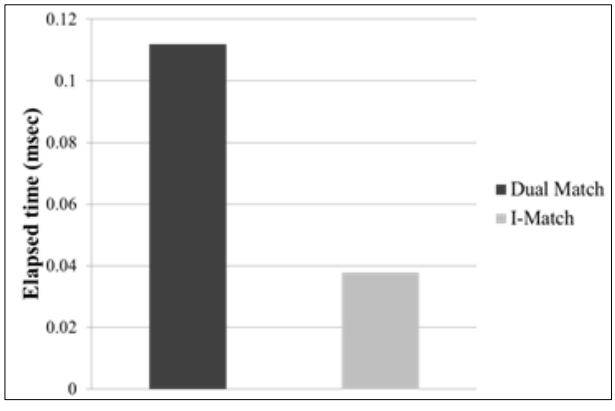

Fig. 2. The comparison of the query processing time.

Fig. 3은 I-Match와 Dual Match의 후보의 수를 보여준다. 어두운 회색 막대 그래프는 Dual Match의 후보 수를 나타내며 밝은 회색 막대 그래프는 I-Match의 후보 수를 나타낸다. 10개의 사용자 질의에 대하여 실험을 진행하였고, 질의의 길이는 384이다. I-Match가 Dual Match에 비해 후보의 수를 87-96% 줄인 것을 볼 수 있다.

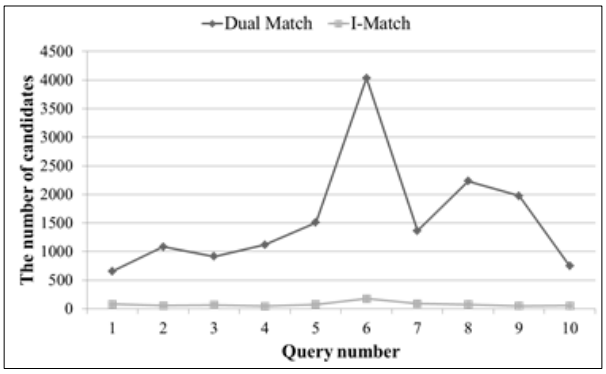

Fig. 3. The comparison of the number of candidates.

5. 결 론

본 논문에서는 시계열 데이터에 대한 서브시퀀스 매칭을 효과적으로 하기 위한 가상 윈도우를 사용하는 인스턴스 레벨의 서브시퀀스 매칭 방법인 I-Match를 제안하였다. 기존의 서브시퀀스 매칭 방법들은 시퀀스 내에서 비교하지 못하는 구간이 발생하여 착오 해답을 많이 발생시키는 문제점을 가지고 있었다. 하지만 I-Match는 이러한 남는 구간을 하나의 가상 윈도우로 구성하여 비교함으로써 후보의 수를 매우 줄였으며, 질의 처리 성능을 개선하였다. 실제 데이터 셋에 대한 실험 결과, I-Match의 질의 처리 시간은 Dual Match보다 2.95배 빨랐으며, 후보의 수는 최대 96% 줄어들었다.

참 고 문 헌

[1] C. Faloutsos, M. Ranganathan and Y. Manolopoulos, "Fast Subsequence Matching in Time-Series Databases," In Proceeding of International Conference on Management of Data (ACM SIGMOD), 1994, pp.419-429.

[2] S. H. Lim "Using Multiple Indexes for Efficient Subsequence Matching in Time-Series Databases," In Proceedings of the Database Systems for Advanced Applications (DASFAA), 2006, pp.65-79.

[3] H. Wu, "Structured Time Series Stream Data," Dissertation, Northeastern University, 2005.

[4] Y. S. Moon, K. Y. Whang and W. K Loh, "Duality-based subsequence matching in time-series databases," In Proceedings of the 17th International Conference on Data Engineering (ICDE), 2001, pp.263-272.

[5] M. H. Pandi, O. Kashefi and B. Minaei, "A Novel Similarity Measure for Sequence Data," Journal of Information Processing Systems, Vol.7, No.3, pp.413-424, 2011.

[6] S. H. Lim, H. Park and S. W. Kim, "Using multiple indexes for efficient subsequence matching in time-series databases," Journal of Information Science, Vol.170, No.24, pp.5691-5706, 2007.

[7] Y. S. Moon, K. Y. Whang and W. K. Loh, "General Match: A Subsequence Matching Method in Time-Series Database Based on Generalized Windows," In Proceedings of International Conference on Management of Data (ACM SIGMOD), 2002, pp.382-393.

[8] S. Y. Ihm, A. Nasridinov, J. H. Lee and Y. H. Park, "Efficient duality-based subsequent matching on time-series data in gree computing," Journal of Supercomputing, 2013, [DOI] 10.1007/s11227-013-1028-2.

임 선 영

e-mail : sunnyihm@sm.ac.kr

2011년 숙명여자대학교 멀티미디어과학과 (이학사)

2013년 숙명여자대학교 멀티미디어과학과 (이학석사)

2013년~현 재 숙명여자대학교 멀티미디어과학과 박사과정

관심분야 : 데이터베이스, IR(정보검색), Top-k 질의처리, 서브시퀀스 매칭, 머신러닝

박 영 호

e-mail : yhpark@sm.ac.kr

1992년 동국대학교 컴퓨터공학과(공학석사)

2005년 한국과학기술원 전산학과(공학박사)

1993년~1999년 한국전자통신연구원 교환전송연구단 선임연구원

2005년~2006년 한국과학기술원 첨단정보기술연구센터 연구원

2005년~2006년 동국대학교 컴퓨터멀티미디어학과 겸임교수

2006년~현 재 숙명여자대학교 멀티미디어과학과 부교수

관심분야 : 데이터베이스, XML, IR(정보검색), 멀티미디어 데이터베이스, Bio정보공학, 영상미디어, 예술&공학인터페이스, 데이터베이스 관리시스템, 머신러닝, 빅데이터, 데이터분석, Telecommunication System