Apriori 알고리즘 기반의 N-그램 선정과 이를 이용한 파일 조각 타입 결정

원종훈○*, 강민지*, 박지성, 김지홍
서울대학교 컴퓨터공학부
{barber, kyang}@snu.ac.kr, {jspark, jihong}@davinci.snu.ac.kr

File-Fragment Type Identification using Selected N-grams by Apriori Algorithm

Jonghoon Won○, Minji Kang, Jisung Park, Jihong Kim
Department of Computer Science and Engineering, Seoul National University

요 약

파일 조각의 타입 정보는 안티 바이러스, 디지털 포렌식 등의 많은 응용에서 유용하게 활용된다. 타입 정보가 담긴 파일시스템 메타데이터나 파일 헤더는 쉽게 변조되며 대부분의 파일 조각에 포함되지 않으므로, 파일 조각의 내용에 기반한 타입 결정 기술이 제안되었다. 제안된 기술은 파일 조각을 바이트 값 빈도 분포로 요약하고 다층 신경망을 사용하여 타입을 결정하지만, 큰 크기의 n-그램 분포를 활용할 수 없다는 한계점을 갖는다. 본 논문에서는 Apriori 알고리즘을 이용하여 타입 결정에 이용할 n-그램을 효과적으로 선정하고, 이를 이용하여 파일 조각의 타입을 결정하는 기법을 제시한다. 실험 결과, 제시한 기법은 7개 타입의 1,000바이트 파일 조각의 타입을 79.75%의 높은 정확도로 결정하였으며, 기존 기법에 비해 7.22배의 속도 향상을 보였다.

1. 서 론

파일의 타입을 결정하는 기술은 운영체제의 작동, 웹 브라우저, 방화벽, 침입 감지 시스템, 안티 바이러스, 디지털 포렌식 등의 응용에서 유용하게 활용된다 [1]. 특히 디지털 포렌식이나 웹 브라우저, 안티 바이러스 등과 같은 응용에서는 전체 파일이 아닌 파일 조각의 타입을 결정하는 것이 중요하다. 이는 파일이 저장 장치에 저장되거나 네트워크를 통해 전송될 때, 파일이 페이지와 패킷 같은 작은 단위로 나누어져 무작위적인 순서로 저장되거나 전송되기 때문이다 [2].

파일의 타입을 결정하기 위해서 파일시스템의 메타데이터에 확장자를 기록하거나, 파일의 헤더 부분에 타입 구분자를 기록하는 방식이 주로 사용된다 [1, 3]. 이러한 방식은 파일을 직접 열지 않고도 파일시스템의 메타데이터나 헤더 부분만을 참조하여 빠른 속도로 파일의 타입을 확인할 수 있다는 장점이 있다. 그러나 확장자나 타입 구분자는 사용자의 실수, 운영체제나 응용 프로그램의 오류, 악의적인 프로그램의 작동 등으로 인하여 변경될 수 있기 때문에 파일의 타입이 쉽게 변조될 수 있다. 게다가 대부분의 파일 조각은 파일시스템 메타데이터나 파일 헤더를 포함하지 않으므로, 파일 조각들의 타입을 결정할 때 활용될 수 없다는 문제점이 존재한다 [2].

이러한 문제점을 해결하기 위하여 파일의 내용에 기반한 타입 결정 방법이 제안되었다 [3]. 파일의 내용에 기반하여 타입을 결정하는 경우 파일 타입이 변조되었거나, 메타데이터나 파일 헤더를 포함하지 않는 대부분의 파일 조각에 대해서도 타입을 결정할 수 있기 때문이다 [2]. 제안된 방법에서는 파일의 타입을 결정하기 위해 파일을 바이트 단위로 읽어 바이트 값 빈도 분포(Byte Frequency Distribution, BFD)로 요약한다. 여러 파일 타입의 평균적인 바이트 값 빈도 분포를 조사한 후, 결정 대상 파일의 바이트 값 빈도 분포를 이와 비교하여 파일 타입을 결정하였다 [3]. 나아가, 바이트 값 빈도 분포를 다층 신경망(Multi-Layer Perceptron, MLP)을 통해 학습시켜 타입 결정의 정확도를 높이는 기법이 제안되었다 [1, 2]. 그러나 바이트 값 빈도 분포는 파일의 내용을 바이트 단위로 분석하므로, n개의 연속된 바이트인 n-그램 값이 갖는 의미를 분석할 수 없다는 한계점이 존재한다.

1-그램과 2-그램을 함께 사용하여 파일 타입을 결정하는 연구에서는 1-그램만 사용한 경우에 비해 높은 타입 결정 정확도를 보였다 [4]. 그러나 가능한 n-그램의 종류는 256^n으로 n이 커질 때 기하급수적으로 증가하기 때문에, 3-그램 이상의 n-그램을 타입 결정에 이용하는 연구는 미미한 실정이다 [5].

본 논문에서는 큰 크기의 n-그램을 파일 조각의 타입 결정에 활용할 수 있도록 하는 기법을 제시한다. Apriori 알고리즘을 이용하여 모든 n-그램 중 파일 조각들에 자주 등장하는 n-그램을 선정한 후, 파일 타입 결정에 도움을 줄 수 있는 소수의 n-그램만을 파일 타입 결정 시에 고려한다. 우선 n-그램을 이용하여 파일 조각의 타입을 결정하고, n-그램을 통해 파일 조각의 타입을 결정할 수 없는 경우에만 바이트 값 빈도 분포와 다층 신경망을 이용한다.

실험 결과, 제시한 기법은 2-그램부터 7-그램까지의 모든 n-그램 중 261개의 n-그램을 선정하였다. 이를 이용하여 7개

* 공동 제1저자.

타입의 1,000 바이트 파일 조각의 타입을 79.75%의 높은 정확도로 결정하였다. 특히, 파일 조각의 바이트 값 빈도 분포와 다층 신경망만을 타입 결정에 이용한 기존의 기법에 비해, 파일 조각의 타입 결정 속도가 7.22배 향상되었다.

본 논문의 구성은 다음과 같다. 2장에서는 Apriori 알고리즘을 이용하여 n-그램을 선정하고, 선정한 n-그램과 다층 신경망을 이용하여 파일 조각의 타입을 결정하는 방법에 대하여 설명한다. 이어 3장에서 제안한 기법의 타입 결정 결과에 대하여 보이고, 4장에서 결론을 맺는다.

2. 파일 조각 타입 결정 기법

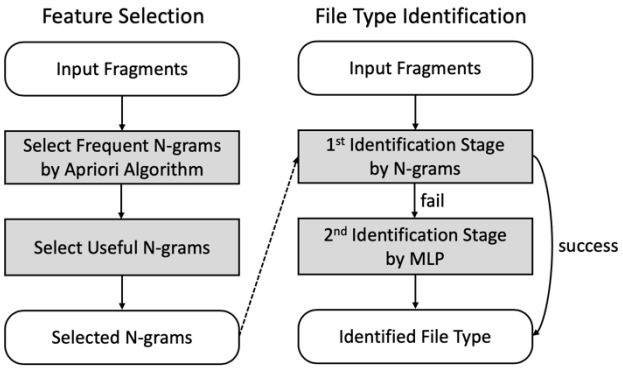

그림 1. n-그램 선정 및 파일 조각 타입 결정 과정

n-그램 선정 및 파일 조각의 타입 결정은 그림 1과 같은 과정으로 진행된다. 먼저 파일 타입 결정을 위한 n-그램을 선정한다. Apriori 알고리즘을 통해 자주 등장하는 n-그램을 선정한 다음, 그 중에서 파일 타입 결정에 유용하게 사용될 수 있는 n-그램을 선정하는 방식으로 진행된다.

선정한 n-그램을 바탕으로 파일 조각의 타입을 결정하는 과정은 두 단계로 이루어진다. 첫 번째 단계에서는 앞서 선정한 n-그램을 통해 파일 조각의 타입을 결정한다. 만약 n-그램을 통해 파일 타입 결정이 불가능한 조각이라면 두 번째 단계로 넘어가며, 바이트 값 빈도 분포와 다층 신경망을 이용하여 파일 조각의 타입을 결정한다.

2.1. Apriori 알고리즘을 이용한 n-그램 선정

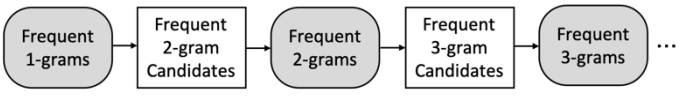

그림 2. Apriori 알고리즘을 이용한 n-그램 선정

먼저, Apriori 알고리즘을 통해 자주 등장하는 n-그램을 선정한다. Apriori 알고리즘은 그림 2와 같이 1-그램부터 단계적으로 진행되며, n-그램이 등장하는 파일 조각의 수를 계산하여 그 값이 경곗값 θ 이상인 n-그램을 선정한다. 이때, 이전 단계에서 선정된 (n-1)-그램으로만 이루어진 n-그램만을 후보로 고려한다. 이는 특정 n-그램이 자주 등장한다면, 그 n-그램을 이루는 (n-1)-그램 또한 자주 등장해야 한다는 Apriori 특성에 기반한다. 이를 통해, n이 증가함에 따라 기하급수적으로 증가하는 탐색 공간을 효과적으로 제한할 수 있다.

다음으로 Apriori 알고리즘으로 선정한 n-그램들 중 파일 타입을 결정하는 데 유용한 n-그램만을 선정한다. 타입을 결정하는 데 유용한 n-그램은 등장하는 모든 파일 조각 중에서 타입 하나가 차지하는 비율이 경곗값 φ 보다 큰 n-그램으로 정의하였다. 이를 통해 대상 파일 조각에 선정된 n-그램이 포함될 경우, 높은 정확도로 타입을 결정할 수 있도록 하였다.

2.2. 파일 조각의 타입 결정

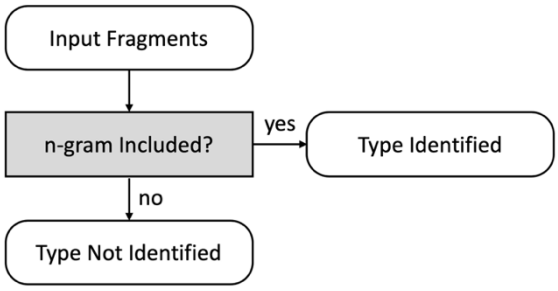

그림 3. n-그램을 이용한 파일 조각 타입 결정

파일 조각의 타입을 결정하는 과정은 n-그램을 이용한 첫 번째 단계와 다층 신경망을 이용한 두 번째 단계로 이루어진다. 첫 번째 단계는 그림 3과 같다. 대상 파일 조각에 앞서 선정한 n-그램이 등장하는지 확인하여, 만약 파일 조각이 특정 파일 타입에서 자주 등장하는 n-그램을 포함할 경우 대상 파일 조각을 해당 타입으로 결정한다. 대상 파일 조각이 선정한 n-그램을 아무것도 포함하지 않아 타입 결정이 불가능할 경우, 두 번째 단계로 넘어간다.

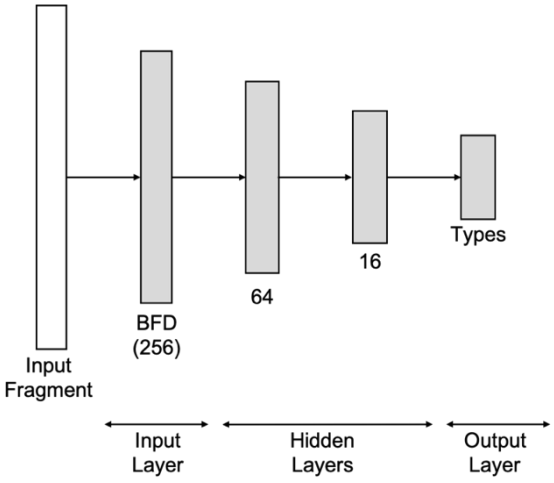

그림 4. 바이트 값 빈도 분포를 이용한 타입 결정 신경망

두 번째 단계에서는 그림 4와 같이 다층 신경망을 이용하여 타입을 결정한다. 파일 조각을 바이트 값 빈도 분포로 요약한 후, 이를 3층 신경망에 입력하면 신경망은 파일 타입을 출력한다.

이와 같이 타입 결정 과정을 두 단계로 나눔으로써 속도가 향상될 것으로 기대하였다. 기존 연구에서는 다층 신경망만을 사용하여, 항상 신경망의 덧셈 및 곱셈 연산을 수행하였다. 반면, 본 연구에서는 파일 조각에 n-그램이 등장하는지 먼저 확인하여 일차적으로 파일 타입을 결정하고, 타입을 결정하지 못한 경우에 대해서만 다층 신경망을 이용하여 필요한 연산을 최소화하였다.

3. 실험 및 결과

3.1. 실험 환경

실험에는 exe, html, hwp, jpg, mp3, pdf, png의 7개 파일 타입이 사용되었다. 각 파일 타입 당 1000 바이트 길이의 파일 조각을 60,000개 수집하여 50,000개는 학습 데이터로, 10,000개는 시험 데이터로 이용하였다.

분류기는 Python과 Caffe 라이브러리를 이용하여 구현하였다. 다층 신경망의 학습은 Adam 최적화 방법을 이용하였으며, 학습률 10^{-4}, 드롭아웃 비율 0.5로 500 epoch 동안 진행되었다.

3.2. 실험 결과

n-그램 선정 시 경곗값을 각각 $\theta = 0.05$, $\varphi = 0.99$ 로 설정하여 5% 이상의 파일 조각들에서 나타나는 n-그램 중 99%의 정확도로 파일 조각의 타입을 결정할 수 있는 n-그램을 선정하였다. 2-그램부터 7-그램까지의 가능한 n-그램은 약 2^{56} 개로, 이들 모두를 고려하는 것은 현실적으로 불가능하다. 한편, Apriori 알고리즘을 통해, 894개의 n-그램이 자주 등장하며, 그 중에서도 261개만이 파일 타입을 결정할 때 유용함을 확인하였다. 따라서, Apriori 알고리즘을 이용한 n-그램 선정 기법이 큰 크기의 n-그램 선정에 효과적임을 확인하였다.

선정한 n-그램과 다층 신경망을 이용하여 파일 조각의 타입을 결정한 결과는 표 1과 같다. 전체 파일 조각 중 39.98%의 조각은 n-그램을 이용한 첫 번째 단계에서 타입이 결정되었으며, 60.02%의 파일 조각은 두 번째 단계에서 다층 신경망을 통해 타입이 결정되었다. 결정 정확도는 첫 번째 단계에서 95.66%, 두 번째 단계에서 69.14%였으며, 총 정확도는 79.75%였다.

파일 조각의 타입 별 결정 정확도는 표 2와 같다. 멀티미디어 데이터를 포함하는 hwp, pdf는 각각 44.35%, 65.32%의 비교적 낮은 결정 정확도를 보였다. 한편, html, mp3, png와 같은 타입은 96.29% 이상의 높은 결정 정확도를 보였다.

파일 조각의 타입을 결정하는 데 걸리는 소요 시간은 표 3과 같다. 다층 신경망만을 타입 결정에 이용하는 기존의 기법은 초당 142.9개 파일 조각의 타입을 결정하였다. 선별된 n-그램을 이용하여 파일 조각의 타입을 결정하는 경우, 초당 1030.1개 파일 조각의 타입을 결정하여, 7.22배의 속도 향상을 보였다.

4. 결 론

파일 조각의 타입 결정 기법은 방화벽, 디지털 포렌식 등 여러 응용에서 활용된다. 특히 실시간으로 파일 조각의 타입을 확인할 필요가 있는 안티 바이러스 등의 응용에서는 빠른 타입 결정 속도가 보장되어야 한다. 기존 연구에서 파일 조각의 바이트 값 빈도 분포 기반의 다층 신경망 활용 기법이 제안되었으나, 큰 크기의 n-그램 분포를 활용하지 못하는 문제점이 있었다. 본 논문에서는 큰 크기의 n-그램을 활용하기 위한 Apriori 알고리즘 기반의 n-그램 선정 기법과, 선정된 n-그램을 이용한 파일 조각 타입 결정 기법을 제안하였다. 실험 결과, 2-그램에서 7-그램까지 총 2^{56}개의 n-그램 중 261개의 n-그램을 선정하였고, 이를 이용하여 7개 타입의 파일 조각에 대하여 79.75%의 높은 정확도로 타입을 결정하였다. 또한, 제안한 기법은 n-그램을 우선적으로 이용하여, 다층 신경망만을 타입 결정에 이용하는 기존의 기법보다 7.22배의 속도 향상을 보였다.

표 1. 결정 단계 별 결정 정확도

Identification Stage	Proportion (%)	Accuracy (%)
1st stage (by n-gram)	39.98	95.66
2nd stage (by MLP)	60.02	69.14
Total	100	79.75

표 2. 파일 조각의 파일 타입 별 결정 정확도

Detected Type (%)	Actual Type						
	exe	html	hwp	jpg	mp3	pdf	png
exe	**80.12**	0.14	0.67	0.07	0.32	0.32	0.08
html	9.20	**99.65**	0.15	0.19	0.17	1.81	0.22
hwp	5.95	0.15	**44.35**	0.65	0.17	2.58	0.73
jpg	0.26	0.19	1.60	**75.09**	0.06	3.45	0.91
mp3	0.14	0.17	0.95	0.01	**97.40**	0.45	1.45
pdf	2.28	1.81	3.50	18.77	0.37	**65.32**	0.32
png	2.05	0.22	48.78	4.22	1.51	26.07	**96.29**

표 3. 타입 결정 소요 시간

Identifier	Speed (fragments / sec)	Comparison
MLP Only	142.9	1x
n-gram + MLP	1030.1	7.22x

감사의 글

이 연구를 위해 연구장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에 감사드립니다. 이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018R1A2B6006878).

참 고 문 헌

[1] M. C. Amirani, M. Toorani, and A. Beheshti Shirazi, "A New approach to Content-based File Type Detection," IEEE Symposium on Computers and Communications, pp. 1103 – 1108, 2008.

[2] M. C. Amirani, M. Toorani, and S. Mihandoost, "Feature-based Type Identification of File Fragments," *Security and Communication Networks*, vol. 6, no. 1, pp. 115 – 128, 2012.

[3] M. McDaniel and M. H. Heydari, "Content Based File Type Detection Algorithms," Hawaii International Conference on System Sciences, pp. 10 – 19, 2003.

[4] S. Gopal, Y. Yang, K. Salomatin, and J. G. Carbonell, "Statistical Learning for File-Type Identification," IEEE Symposium on Computers and Communications, pp. 68 – 73, 2011.

[5] N. L. Beebe, L. A. Maddox, L. Liu, and M. Sun, "Sceadan – Using Concatenated N-Gram Vectors for Improved File and Data Type Classification.," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 9, pp. 1519 – 1530, 2013.