마코프 체인을 이용한 컴퓨터 바이러스 발생 빈도수 예측 모델링

정영석°, 박구락*, 안우영**
°공주대학교 컴퓨터공학과
*공주대학교 컴퓨터공학과
**대전보건대학교 바이오정보과
e-mail:merope@kongju.ac.kr°, ecgrpark@kongju.ac.kr*, wyahn@hit.ac.kr**

Computer virus occurrence frequency Predictive modeling using Markov chains

Young-Suk Chung°, Koo-Rack Park*, Woo-Young Ahn**
°Dept. of Computer Science & Engineering, Kongju national University
*Dept. of Computer Science & Engineering, Kongju national University
**Dept. of Bio Information, Daejeon Health Sciences College

● 요 약 ●

최초의 컴퓨터 바이러스인 브레인 바이러스가 만들어진 이후로, 현재까지 컴퓨터 바이러스로 인한 피해는 늘어나고 있다. 이에 따라 컴퓨터 바이러스를 막기 위한 여러 가지 노력이 현재도 진행 중에 있다. 컴퓨터 바이러스로 인한 피해 방지와 예방을 위한 대책을 수립하기 위해서는 컴퓨터 바이러스의 발생 빈도수를 예측 하는 것이 필요하다. 본 논문은 다양한 예측 연구에 활용되고 있는 마코프 체인을 적용하였다. 본 논문은 마코프 체인을 적용하여 컴퓨터 바이러스 빈도수를 예측하는 모델링을 제안한다.

키워드: 마코프 체인(Markov chains), 시뮬레이션(Simulation), 예측 모델링(Predictive modeling), 컴퓨터 바이러스(Computer Viruses)

I. 서 론

SF소설 속에서 존재하던 컴퓨터 바이러스는 1985년 파키스탄에서 만들어진 브레인 바이러스를 시초로 현재까지 계속 생성되고 있다. 컴퓨터 바이러스로 인한 피해의 예를 보면 2003년 발생한 슬래머 웜으로 인한 피해규모는 1,675억 원에 이른다[1]. 또한 한국 인터넷 진흥원의 인터넷 통계 정보 시스템의 바이러스 신고 접수 현황을 보면 2005년부터 꾸준히 증가함을 알 수 있다[2]. 그러므로 컴퓨터 바이러스를 예방하기 위한 대책을 실행하기 위해서는 컴퓨터 바이러스가 얼마나 발생할지, 발생 빈도를 예측하여 대비하는 예측 자료가 필요하다. 그래서 본 논문에서는 컴퓨터 바이러스의 발생 빈도를 예측 할 수 있는 컴퓨터 바이러스 예측 모델링을 제안한다. 2장 관련연구로 컴퓨터 바이러스와 마코프 체인에 대해 논의한다. 3장에서는 컴퓨터 바이러스 예측 모델링에 대해 논의한다. 마지막으로 4장에서 결론 및 향후 연구과제에 대해 논의한다.

II. 관련 연구

1. 컴퓨터 바이러스

컴퓨터 바이러스는 컴퓨터 내에 침투하여 프로그램을 파괴하여 작동할 수 없도록 하는 컴퓨터 프로그램을 의미한다. 1949년 J 폰 노이만이 발표한 논문에서 프로그램이 자기 자신을 복재함으로서 증식할 수 있다는 가능성을 제시한 것에서 유래한다. 컴퓨터 바이러스란 용어는 1983년 개최된 보안 세미나(Security Seminar)에서 미국 남가주대학의 프레드릭 코헨(Fredrick Cohen) 박사의 논문「컴퓨터 바이러스 : “Computer Virus : Theory and Experiment”」이라는 논문에서 처음 언급되었다[3]. 전달경로는 보통 전자 메일 메시지 첨부 파일이나 인스턴트 메시징 메시지를 통해 확산된다. 또한 오디오, 비디오 파일과 불법복제 소프트웨어나 기타 다운로드 한 파일, 프로그램에 포함되어 있을 수 있다. 컴퓨터 바이러스에 감염된 컴퓨터는 메모리 부족, 프로그램응답 부재, 파티션의 사라짐, 컴퓨터 부팅 실패 등 다양한 증상을 나타낼 수 있다[4].

2. 마코프 체인

마코프 과정은 과학, 공학 및 경영 모델링에서 널리 사용되고

있다. 어떤 랜덤 과정$\{X(t)|t \geq 0\}$이 임의의 시점 $t_0 < t_1 < ... < t_n$에 대해 $X(t_0), X(t_1), ..., X(t_{n-1})$이 주어진 경우에 대한 $X(t_n)$의 조건부 누적함수가 $X(t_{n-1})$에만 의존한다면, 즉

$$P[X(t_n) \leq x_n | X(t_{n-1}) = x_{n-1}, X(t_{n-2}) = x_{n-2}, ..., X(t_0) = x_0] = P[X(t_n) \leq x_n | X(t_{n-1}) = x_{n-1}]$$

이라면 랜덤과정 $\{X(t)|t \geq 0\}$는 마코프 과정이라 한다. 마르코프 과정 중 이산상태를 마코프 체인이라 한다[5]. 마코프 체인은 다음의 세 가지로 구성된 것을 말한다[6].

◎ 상태집합 : 가능한 상태들을 집합으로 생성한 것
◎ 초기 확률: 정의된 상태들이 초기 상태에 가질 수 있는 발생 확률
◎ 전이행렬 : 각 상태들 간의 전이되는 확률

미래를 예측 할 수 있는 마코프 체인 특성상 다양한 분야에 활용되고 있다. 공동주택내의 쾌적한 살내 환경을 위해 재실자의 움직임을 확률적으로 예측한 연구에 사용되었고[7], 전력 계통 시스템의 고장 건수를 예측하기 위한 모델로 마코프 체인이 적용되었다[8]. 또한 마코프 체인 기법을 이용하여 범죄 발생 예측 확률을 높이고, 도시공간에 포함된 여러 특성을 이용하여 범죄가 발생할 상대적 위험지수를 산출하여, 분석 대상 지역에 적용하여 시간의 흐름에 따라 위험도 확률 지도를 생성하는 모델링 연구에 적용되었다[9].

Ⅲ. 본 론

컴퓨터 바이러스를 예측하기 위한 모델링은 그림 1과 같다.

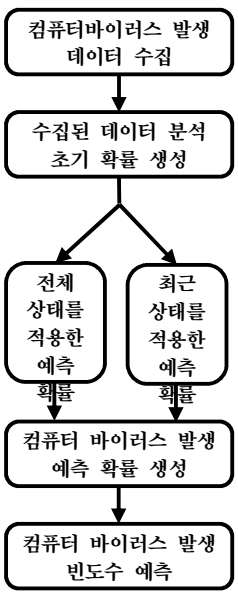

그림 1. 컴퓨터 바이러스 발생 빈도수 예측 모델링
Fig. 1.Computer viruses, the frequency of occurrence predictive modeling

각 모델링의 단계는 다음과 같다.

첫 번째, 컴퓨터 바이러스의 발생 데이터를 수집한다. 컴퓨터 바이러스 발생 빈도수 데이터는 한국 인터넷진흥원 홈페이지에 공개 하고 있다.

두 번째, 수집된 데이터를 분석하여 임계값, 상태를 정의한다. 본 논문에서는 컴퓨터 바이러스 발생 빈도수를 바탕으로 임계값을 설정한 후, 임계값에 따라 상태를 정의한다. 그리고 상태집합, 초기 확률, 전이행렬을 생성 한다[6].

◎ 상태 집합 : 상태 집합은 컴퓨터 바이러스의 발생 빈도수를 의미하며, 컴퓨터 바이러스 통계 자료를 이용하여 적절한 임계값을 설정한 후 임계값에 따라 상태들을 설정하고 집합으로 정의 하였다.
◎ 초기 확률 : 컴퓨터 바이러스가 초기 상태에서 가질 수 있는 발생 확률로서 최근에 발생한 바이러스의 발생 상태를 이용하여 정의하고 식(1)로 정의 한다.

$$P(S_1, S_2, \cdots, S_n) = P(\frac{a}{F}, \frac{b}{F}, \cdots, \frac{c}{F}) \quad (1)$$

위의 식(1)에서 $a, b, c,$는 각 상태$(S_1, S_2, \cdots, S_n)$의 컴퓨터 바이러스 발생 횟수이고 F는 $a, b, c,$의 합이다.

◎ 전이 행렬: 컴퓨터 바이러스 발생 데이터들을 정의된 상태들과 매칭하여 상태들을 열거한 후, 열거된 하나의 상태가 다른 상태로 전이되는 횟수를 구하여, 정의된 상태 사이에 전이되는 상태들을 확률로 나타낸 것이다. 각 열은 한 상태에서 다른 상태로의 확률을 의미하며, 각 행의 합은 1이다. 식(2)의 P는 전이행렬이다.

$$P = \begin{pmatrix} P_{11} & P_{12} & P_{13} & \cdots & P_{1n} \\ P_{21} & P_{22} & P_{23} & \cdots & P_{2n} \\ \cdots & \cdots & \cdots & P_{ij} & \cdots \\ P_{n1} & P_{n2} & P_{n3} & \cdots & P_{nn} \end{pmatrix} \quad (2)$$

$$\sum_{j=1}^{n} P_{1j} = 1, \ \sum_{j=1}^{n} P_{2j} = 1, \ \cdots, \ \sum_{j=1}^{n} P_{nj} = 1, \ P_{ij} \geq 0$$

$$\sum_{j=1}^{n} P_{ij} = 1, \ i = 1, 2, \ldots, n \quad (3)$$

위의 식(2)은 조건(3)을 만족한다.

세 번째, 초기 확률과 전이행렬을 마코프 체인의 식(4)에 적용하여 컴퓨터 바이러스 예측 확률을 생성한다.

$$P(S_k) = \sum_{i=1}^{n} P(S_i) P_{ik} \quad (4)$$

$P(S_i)$: 초기 확률 P_{ik}:전이행렬

전이 행렬에 사용될 상태를 두 가지 상태로 나눈다.

◎ 전체 상태 : 전체 상태는 예측에 사용된 기간 전체를 상태로 적용한 전이 행렬

◎ 최근 상태 : 최근 상태는 예측에 사용된 기간 중 가장 마지막 기간 예를 들면, 최근 1년을 대상으로 한 전이 행렬을 적용하여 만든 전이행렬

본 논문에서는 컴퓨터 바이러스를 예측하기 위해 전체 상태와 최근 상태로 두 가지로 나누었다. 바이러스의 발생은 시간에 따라 많이 변화한다.

상태를 나눈 이유는 상태를 전체 상태로만 적용했을 때보다, 최근의 변화가 적용된 최근 상태를 같이 적용하였을 때 예측 확률이 높을 것으로 예상되기 때문이다.

다음의 그림 2의 바이러스 신고 접수 빈도수 자료의 연도별 그래프를 보면 2001년부터2004년까지는 증감치의 변화가 심하고 2005년부터는 꾸준히 증가함을 알 수 있다[2].

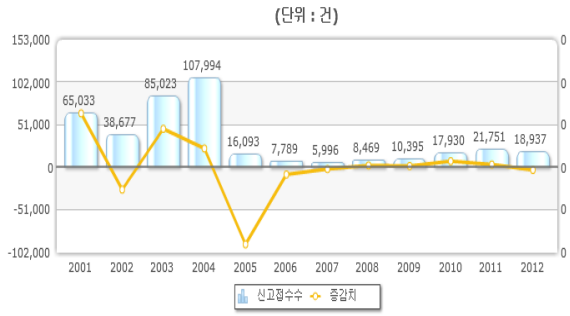

그림 2. 바이러스 신고 접수 빈도수

Fig. 2. Report frequency of computer viruses

초기 확률은 전체 상태, 최근상태에 동일하게 사용 한다. 그리고 마코프 체인 식에 적용하여 전체 상태를 적용했을 때 예측확률과 최근 상태를 적용하였을 때 예측 확률을 각각 구한다.

네 번째, 전체 상태를 적용하였을 때의 예측확률과 최근 상태를 적용했을 때의 예측 확률의 평균값을 구하여, 컴퓨터 바이러스 예측 확률로 적용한다.

마지막으로 예측 확률에 예측 기간 중 발생한 컴퓨터 바이러스의 발생 빈도수 중 최근 발생 빈도수의 평균값을 적용하여 컴퓨터 바이러스 발생 빈도수를 예측 한다.

IV. 결 론

컴퓨터와 인터넷의 결합으로 각 기업체 관공서에서 업무의 효율은 증가하고 있고, 개인의 여가 생활 등 현대 사회의 모든 분야에 활용 되고 있다. 그리고 그 활용 범위는 점점 늘어나고 있다. 그러나 이와 동시에 해킹, 컴퓨터 바이러스 등 정상적인 컴퓨터와 인터넷의 활동을 막고, 파괴하는 행위도 늘어나고 있고 이에 대한 대책이 필요하다. 본 논문에는 컴퓨터 바이러스의 발생 빈도를 예측하는 연구를 수행하였다.

컴퓨터 바이러스의 발생 빈도를 예측하기 위해 다양한 예측 연구에 활용되고 있는 마코프 체인을 이용하여 컴퓨터 바이러스 발생 빈도수를 예측할 수 있는 모델링을 제안하였다. 본 논문이 기존의 마코프 체인이 적용된 연구와 차이점은 전체 상태와 최근상태 상태를 두 가지로 나눈 것이다. 최근 상태를 추가함에 따라 최근 상태의 변화를 더 반영하여 예측 확률을 높일 것으로 예상되기 때문이다. 제안된 모델링을 이용하면 컴퓨터 바이러스 발생 빈도수를 예측할 수 있다. 컴퓨터 바이러스의 발생 빈도수를 예측할 수 있다면, 컴퓨터 바이러스 예방 정책 수립에 도움이 될 것으로 예상된다. 향후 연구과제로 실제 컴퓨터 바이러스 발생 데이터를 이용하여, 컴퓨터 바이러스 발생 빈도수를 예측할 예정이다.

참고문헌

[1] Jinho Yoo, Sangho Gee, Hyein Song, Kyungho Chung, Jongin Lim, “Estimating Economic Damages from Internet Incidents,” *http://www.nia.or.kr/*, Vol15, No.1, March 2008.

[2] Computer viruses Report Status, *http://isis.kisa.or.kr/sub07/index.jsp?pageId=070402*

[3] Computer viruses, *http://terms.naver.com/*

[4] Computer viruses : Description, prevention, and recovery, *http://support.microsoft.com/kb/129972/ko*

[5] Oliver C. Ibe, “Fundamentals of Applied Probability and Random Processes,” ACADEMIC PRESS, pp371-376, 2008.

[6] Young-Gab Kim, Young-kyo Baek, Hoh Peter In, Doo-Kwon Baik “A Probabilistic Model of Damage Propagation based on the Markov Process,” Journal of KIISE, Vol. 33, No. 8, pp.524-535, 8. 2006.

[7] Kim Young-Jin, Park Cheol-Soo, “Prediction of Occupant’s Presence in Residential Apartment Buildings using Markov Chain,” Korea Institute of Architectural Sustainable Environment and building System. 2008 autumn conference, pp116~121, 2008.

[8] Hee Tae-Lee, Jae-Chul Kim, “A Study on The Prediction of Number of Failures using Markov Chain and Fault Data,” KIIEE Annual Autumn Conference 2008, pp. 363-366, 10. 2008.

[9] Chan-Sook Noe, Dong-Hyun Kim, “Crime Occurrence Risk Probability Map Generation Model based on the Markov Chain,” Journal of Korean Institute of Information Technology, Vol. 10, No. 8, pp.89-98, 10, 2012.