디지털산업정보학회 논문지 제14권 제2호-2018년 6월
http://dx.doi.org/10.17662/ksdim.2018.14.2.019

텍스트마이닝을 이용한 정보보호 연구동향 분석

김 태 경* · 김 창 식**

Research Trends Analysis of Information Security using Text Mining

Kim Taekyung · Kim Changsik

〈Abstract〉

With the development of IT technology, various services such as artificial intelligence and autonomous vehicles are being introduced, and many changes are taking place in our lives. However, if secure security is not provided, it will cause many risks, so the information security becomes more important. In this paper, we analyzed the research trends of main themes of information security over time.

In order to conduct the research, 'Information Security' was searched in the Web of Science database. Using the abstracts of theses published from 1991 to 2016, we derived main research topics through topic modeling and time series regression analysis. The topic modeling results showed that the research topics were Information technology, system access, attack, threat, risk management, network type, security management, security awareness, certification level, information protection organization, security policy, access control, personal information, security investment, computing environment, investment cost, system structure, authentication method, user behavior, encryption. The time series regression results indicated that all the topics were hot topics.

Key Words : Information Security, Text Mining, Topic Modeling, Time Series Regression Analysis

Ⅰ. 서론

IT 기술의 발전에 따라 인공지능, 자율 주행 자동차 등 IoT와 연관된 다양한 서비스들의 등장하고 있으며, 우리의 생활에 많은 변화를 가져오고 있다. 이러한 서비스들은 우리의 생활에 많은 편의를 제공해 주고 있지만, 안전한 보안이 제공되지 않는다면 여러 가지 많은 위험을 야기하게 된다. 따라서 기술이 발전될수록 정보보안의 중요성은 더욱 높아지고 있다. 이러한 정보보안에 대한 연구는 시간의 흐름에 따라 계속적으로 다양한 주제로 변화하고 있다.

따라서 본 연구에서는 정보보안에 대한 연구들이 어떻게 변화하고 있는지 연구동향을 분석하고자 한다.

일반적으로 텍스트마이닝을 활용하여 연구동향을

* 명지전문대학 인터넷응용보안공학과 교수
** 배화여자대학교 글로벌관광과 교수(교신저자)

분석하는 방법들은 많이 사용되는 단어들에 가중치를 부여하고 군집도를 파악하는 방법으로 분석을 수행하고 있으나 이 방법의 단점으로는 전문가의 전문성에 따라 그 결과가 달라질 수 있다는 한계점을 가지고 있다[1]. 따라서 이러한 한계점을 극복하기 위해서는 토픽모델링을 수행한 후에 시계열회귀분석 기법을 적용하여 결과에 대한 타당성을 통계적 기법을 적용하여 제시할 수 있다[2]. 본 논문에서는 정보보안 분야의 연구동향 분석을 위해 토픽 모델링과 시계열 분석 기법을 적용하였으며, 분석은 해외 저널 논문들의 초록을 대상으로 보안과 관련된 주요 토픽들을 분석하였다. 분석대상의 선정은 웹오브사이언스(Web of Science) 데이터베이스에서 'Information Security' 을 검색하여 1991년부터 2016년까지 게재된 논문의 초록들을 대상으로 하였다.

본 논문의 2장에서는 텍스트마이닝 관련연구에 대해서 제시하였으며, 3장에서는 토픽 모델링과 시계열 회귀 분석을 이용해서 정보보호 연구 동향을 분석하는 절차와 분석결과를 제시하였다. 마지막으로 4장에서는 본 연구의 결론으로 구성하였다.

Ⅱ. 관련연구

텍스트 마이닝은 문자와 텍스트로 구성되어 있는 비정형 데이터를 분석하는 기법으로, 대표적인 분석 방법으로는 시멘틱 웹 및 온톨로지, 오피니언 마이닝, 토픽 모델링 등이 있으며, 이 방법들은 텍스트를 기반으로 분석을 수행하는 기법이다[3]. 이중에서 토픽 모델링은 비정형 데이터에서 특정 토픽을 추출하는 알고리즘을 적용하고, 확률적 기법을 활용하여 문서에서 어떤 단어들 간에 연관성이 있으며, 어떤 주제어들을 가지고 있는가를 확인하는 알고리즘이다[4]. 토픽 모델링은 많은 수량의 문서들을 그 주제에 따라 그룹핑한다는 측면에서 문서의 군집화와 유사하다고 할 수 있지만, 여러 토픽에 하나의 문서가 동시에 대응될 수 있다는 측면이 현실 세계의 모델링에 적합하다고 할 수 있다. 토픽 모델링은 많은 양의 문서들에 대한 통찰을 제공한다는 측면 이외에도, 토픽 모델링 결과를 활용하여 다양한 분석을 수행할 수 있다는 점에서 그 활용가치가 높다고 할 수 있다[5].

토픽 모델링은 다양한 분야에서 활용되고 있는데, 새로운 이슈의 도출 및 그 이슈를 추적하기 위한 연구, 여러 온라인 리뷰들을 분석하여 사용자의 경험을 파악하는 연구 등 다양한 분야에서 이용되고 있다. 특히 토픽 모델링은 특정 분야의 연구동향을 파악하는 연구에서도 다양하게 활용되고 있다. 그 사례로는 이슈가 되고 있는 주제와 점차 감소하는 추세에 있는 주제를 파악하기 위해 PNAS (Proceedings of the National Academy of Sciences of the United States of America)의 초록 분석을 통해 주제어에 대한 연구 현황을 분석하는 연구가 수행되었으며[6], 국내의 문헌정보학 분야에서는 연구동향을 분석하기 위하여 토픽 모델링을 수행한 연구도 있다. 이 연구에서는 문헌정보학 분야의 주요 학술지들의 1970년도부터 2012년도까지 발표된 논문의 초록들을 기반으로 토픽 모델링을 수행하였다[7].

또한 국내 교통·ICT 융합 분야의 연구기회를 발견하기 위해 토픽 모델링을 이용한 연구가 수행되었다. 이 연구에서는 주로 교통 관련 연구 동향을 정량적으로 분석하였고, 교통 분야에서 IoT 활성화를 위해 필요한 연구주제를 파악하였다[8].

이외에도 시뮬레이션 연구 동향을 분석하기 위해 토픽 모델링을 이용한 연구도 있으며, 이 연구에서

는 KCI 등재된 논문 중에서 '시뮬레이션' 키워드가 있는 논문들을 수집하여 논문의 분야와 트렌드 분석, 토픽분석을 진행하였다[9].

Ⅲ. 토픽 모델링 및 시계열 회귀 분석

3.1 분석 대상

정보보안 분야 연구트렌드에 대한 분석을 위하여, 웹오브사이언스(Web of Science) 데이터베이스에서 'Information Security' 을 검색하여 1991년부터 2016년까지 1,446편 논문의 초록을 분석의 대상으로 선정하였다. 본 연구는 데이터 전처리 후 1,096편의 논문 초록을 대상으로 기간별 연구동향을 파악하였다. 다음의 <표 1>은 5년 단위로 분석을 수행한 연도별 논문의 수를 나타낸 것이다.

<표 1> 연도별 논문 수

기간	논문 수
1991-1996	64
1997-2001	69
2002-2006	189
2007-2011	247
2012-2016	527
합계	1,096

정보보호에 관련된 논문은 시간의 흐름에 따라 그 수가 계속적으로 증가하는 추세에 있다. IT 기술이 우리의 생활과 더욱 밀접하게 연관될수록 정보보호에 관련된 논문을 더욱 증가할 것으로 예상된다. 다음의 <그림 1>은 분석을 수행한 논문들의 연도별 논문의 수를 그래프로 나타낸 것이다.

<그림 1> 연도별 논문 수

3.2 분석 방법 및 절차

분석의 절차는 데이터의 전처리, 토픽모델링, 시계열회귀분석 3단계로 분석을 수행하였다[2].

▪ 전처리: 연구대상 논문의 분석을 위한 전처리 작업에는 Excel을 주로 활용하였으며, Excel에서 전처리가 완료된 데이터는 SAS Enterprise Guide 7.2를 통하여 SAS 분석용 파일로 변환함

▪ 토픽모델링: SAS Enterprise Miner(SAS EM)을 활용한 토픽모델링은 ① 분석용 텍스트 데이터 업로드 ② 파싱 ③ 필터 ④ 토픽모델링의 순으로 진행함

▪ 시계열 회귀분석: 토픽모델링의 결과 값이 임계치 값(Cutoff value)을 상회하는 경우 '1', 그렇지 않을 경우 '0'으로 변환하였으며, 연도별로 합산한 값을 기준으로 SPSS를 활용하여, 시계열회귀분석을 진행함

3.3 분석 결과

본 연구에서의 토픽 모델링은 SAS EM을 활용하였으며, 토픽 모델링의 빈도는 각 토픽이 얼마나 많이 등장하였는지를 나타낸다. 토픽은 핵심 키워드를 대표하는 용어로 정의하였다. 다음의 <표 2>는 1991년에서 2016년까지의 정보보호에 대한 토픽 모델링

결과를 나타낸 것이다.

<표 2> 토픽 모델링 결과

Topic		Keyword	Freq
T19	정보기술	technology, management, business, information technology, policy	157
T15	시스템 접근	system, research, approach, paper, issue	154
T17	공격	network, threat, computer, system, attack	154
T08	위협	government, threat, article, policy, level	152
T04	위험관리	risk, assessment, asset, analysis, risk assessment	134
T18	네트워크 형태	protocol, communication, network, message, scheme	115
T07	보안관리	ism, management, organization, standard, information, security management	112
T10	보안인식	awareness, user, student, behaviour, information security awareness	111
T11	인증수준	evaluation, method, model, criterion, level	108
T12	정보보호 조직	culture, employee, asset, information security culture, organization	100
T01	보안정책	behavior, compliance, policy, employee, theory	99
T06	접근제어	flow, information, flow, control, access, policy	84
T05	개인정보	health, care, privacy, patient, hospital	82
T13	보안 투자	investment, firm, security investment, decision, attack	82
T14	컴퓨팅 환경	cloud, computing, service, cloud computing, environment	82
T03	투자비	market, investor, firm, price, debt	80
T16	시스템 구조	requirement, framework, standard, system, process	75
T09	인증방법	scheme, authentication, password, user, attack	66
T20	사용자 행동	decision, problem, user, behavior, method	66
T02	암호화	image, phase, encryption, technique, decryption	58

이 기간 논문에서는 정보기술, 시스템 접근, 공격, 위협, 위험관리, 네트워크 형태, 보안관리, 보안인식, 인증 수준, 정보보호 조직, 보안정책, 접근제어, 개인정보, 보안 투자, 컴퓨팅 환경, 투자비, 시스템 구조, 인증 방법, 사용자 행동, 암호화 등의 순으로 주제가 나타났다.

<표 3> 토픽 시계열 회귀분석

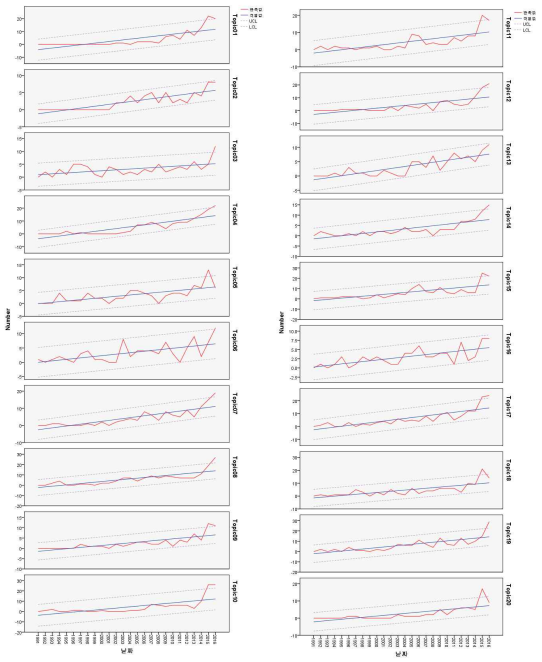
Topic		추정값	p-value	Hot/Cold
T01	보안정책	.633	.000	Hot
T02	암호화	.273	.000	Hot
T03	투자비	.171	.006	Hot
T04	위험관리	.719	.000	Hot
T05	개인정보	.260	.000	Hot
T06	접근제어	.258	.001	Hot
T07	보안관리	.543	.000	Hot
T08	위협	.649	.000	Hot
T09	인증방법	.319	.000	Hot
T10	보안인식	.635	.000	Hot
T11	인증수준	.492	.000	Hot
T12	정보보호 조직	.541	.000	Hot
T13	보안 투자	.360	.000	Hot
T14	컴퓨팅 환경	.375	.000	Hot
T15	시스템 접근	.612	.000	Hot
T16	시스템 구조	.211	.000	Hot
T17	공격	.676	.000	Hot
T18	네트워크 형태	.470	.000	Hot
T19	정보기술	.656	.000	Hot
T20	사용자 행동	.377	.000	Hot

시간의 흐름에 따른 정보보호 분야의 핵심 토픽의 변화를 파악하기 위하여, 토픽모델링 결과를 토대로 SPSS를 활용한 시계열회귀분석을 수행하였다. 이러한 시계열회귀분석의 결과 값을 통해, 각 토픽들의 26년간 연도별 추세를 파악할 수 있다.

시계열회귀분석 결과 회귀계수가 양수이면서, 유의확률 값이 통계적으로 유의미하면, 상승을 의미하는 핫 토픽(Hot Topic), 회귀계수 값이 음수이면서,

유의확률 값이 유의하면, 하락을 의미하는 콜드 토픽(Cold Topic), 회귀계수 값이 유의미하지 않은 토픽은 중립토픽(Neutral Topic)으로 구분한다[6].

<그림 2> 는 각 토픽 별 트렌드 분석결과를 나타낸 것이다.



<그림 2> 토픽 트렌드 분석 결과

토픽모델링을 통해 도출된 주제어들과 정보통신망의 안정성, 신뢰성 확보를 위한 정보보호 관리체계의 인증심사 항목을 비교해 보면 다음의 <표 4>와 같다. 정보보호 관리체계 인증심사 항목은 관리과정 및 정보보호대책 104개로 구성되어 있다. 정보보호대책 통제항목으로는 정보보호 정책, 정보보호 조직, 외부자 보안, 정보자산 분류, 정보보호 교육, 인적보안, 물리적 보안, 시스템 개발보안, 암호통제, 접근통제, 운영보안, 침해사고 관리, IT재해복구 등으로 구성되어 있다[10].

<표 4> 토픽과 ISMS 인증심사 항목 비교

토픽	인증심사 항목
보안정책	정보보호정책
암호화	암호통제
투자비	관리과정
위험관리	관리과정
개인정보	정보보호조직, 인적보안, 외부자보안, 시스템개발보안, 암호통제, 접근통제, 운영보안
접근제어	접근통제
보안관리	운영보안
위협	관리과정
인증방법	시스템개발보안
보안인식	정보보호교육
인증수준	운영보안
정보보호 조직	정보보안조직
보안 투자	관리과정
컴퓨팅 환경	접근통제
시스템 접근	접근통제
시스템 구조	운영보안
공격	침해사고 관리
네트워크 형태	운영보안
정보기술	관리과정
사용자 행동	인적보안

논문에서 주로 연구되고 있는 분야와 인증심사 항목과의 비교를 수행하면, 정보자산 분류, 물리적 보안, IT재해복구 등 3개의 정보보호대책 통제항목을 제외하고는 대부분의 항목들이 연관되어 있는 것을 알 수 있다. 따라서 정보보호와 관련되어 주로 연구되고 있는 연구의 주제들은 IT 확산과 패러다임

의 변화, 사이버 침해 증가와 관련되어 있다.

Ⅳ. 결론

IT 기술의 발전에 따라 다양한 서비스들이 도입되고 있으며 인공지능, 자율 주행 자동차 등 새로운 서비스들의 등장으로 우리의 생활에 많은 변화를 가져오고 있다. 그러나 안전한 보안이 제공되지 않는다면 많은 위험을 야기하게 되므로 기술이 발전될수록 정보보안의 중요성은 더욱 높아지고 있다. 따라서 본 논문에서는 시간의 흐름에 따른 정보보안의 주요 주제들의 연구동향을 분석하였다.

연구의 수행은 웹오브사이언스(Web of Science) 데이터베이스에서 'Information Security' 을 검색하여 1991년부터 2016년까지 게재된 논문의 초록들을 대상으로 텍스트마이닝과 시계열회귀분석을 통해 기간별 연구동향을 파악하였으며, 도출된 주제들은 IT 확산과 패러다임의 변화, 사이버 침해 증가에 관련된 연구가 중요하게 다루어지고 있음을 의미한다.

참고문헌

[1] 이기헌, 정효정, 송민식, "문헌정보학 분야 핵심 학술지들의 가중 주제-방법 네트워크 분석," 한국문헌정보학회지, 49(3), 2015.8, pp.457-488.

[2] 김창식, 최수정, 곽기영, "토픽모델링과 시계열회귀분석을 활용한 정보시스템분야 연구동향 분석," 디지털콘텐츠학회논문지, 18(6), 2017.10, pp.1143-1150.

[3] 김성근, 조혁준, 강주영, "학술연구에서의 텍스트마이닝 활용 현황 및 주요분석기법," 정보기술아키텍처연구, 13(2), 2016, pp.317-329.

[4] Blei, D., Ng, A. and Jordan, M., "Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol. 3, 2003, pp.993- 1022.

[5] 김남규, 이동훈, 최호창, William Xiu Shun Wong, "텍스트 분석 기술 및 활용 동향," 한국통신학회논문지, 42(2), 2017.2, pp.471-492.

[6] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in Proc. National Academy Sci., 101(1), Apr. 2004, pp.5228-5235.

[7] 박자현, 송민, "토픽모델링을 활용한 국내 문헌정보학 연구동향 분석," 정보관리학회지, 30(1), 2013, pp.7-32.

[8] 오준석, "텍스트마이닝 방법을 통한 국내 교통·ICT 융합 분야 연구기회 발견," 교통연구, 22(4), 2015, 12. pp.93-110.

[9] 나상태, 김자희, 정민호, 안주언, "토픽 모델링을 이용한 시뮬레이션 연구 동향 분석," 한국시뮬레이션학회 논문지, 25(3), 2016.9, pp.107-116.

[10] 임효창, 권윤화, 박소희, 한혜중, "7S 모델과 경쟁가치 모델을 활용한 기업보안 조직문화 진단 도구 개발," 경영컨설팅연구, 17(3), 2017, pp.183-192.

■ 저자소개 ■

김 태 경
(Kim Taekyung)

2017년 9월~현재
명지전문대학 인터넷응용보안공학과
교수
2008년 3월~2017년 8월
서울신학대학교 교수
2006년 3월~2008년 2월
서일대학 정보전자과 교수
2005년 8월 성균관대학교
전기전자및컴퓨터공학과(공학박사)
2001년 8월 성균관대학교
정보통신공학과(공학석사)

관심분야 : 네트워크보안, IoT 보안,
개인정보보호
E-mail : tkkim@stu.ac.kr

김 창 식
(Kim Changsik)

2018년 3월~현재
배화여자대학교 글로벌관광과 교수
2015년 3월~ 2018년 2월
국민대 BIT전문대학원
BK21 플러스 사업팀 계약교수
2013년 8월 국민대 BIT전문대학원
비즈니스IT전공(경영정보학박사)
2002년 2월 경희대학교 산업정보대학원
경영정보학과(경영학석사)

관심분야: 관광정보, 텍스트마이닝, 지식경영,
데이터 애널티틱스, 기술경영,
소셜네트워크 분석 및 응용
E-mail: solo21solo@naver.com

논문접수일 : 2018년 03월 16일
게재확정일 : 2018년 05월 14일