Journal of The Korea Society of Computer and Information
Vol. 22 No. 12, pp. 101-108, December 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.12.101

A Strategy Study on Sensitive Information Filtering for Personal Information Protect in Big Data Analyze

Gun-Seo Koo*

Abstract

The study proposed a system that filters the data that is entered when analyzing big data such as SNS and BLOG. Personal information includes impersonal personal information, but there is also personal information that distinguishes it from personal information, such as religious institution, personal feelings, thoughts, or beliefs. Define these personally identifiable information as sensitive information. In order to prevent this, Article 23 of the Privacy Act has clauses on the collection and utilization of the information. The proposed system structure is divided into two stages, including Big Data Processing Processes and Sensitive Information Filtering Processes, and Big Data processing is analyzed and applied in Big Data collection in four stages. Big Data Processing Processes include data collection and storage, vocabulary analysis and parsing and semantics. Sensitive Information Filtering Processes includes sensitive information questionnaires, establishing sensitive information DB, qualifying information, filtering sensitive information, and reliability analysis. As a result, the number of Big Data performed in the experiment was carried out at 84.13%, until 7553 of 8978 was produced to create the Ontology Generation. There is considerable significan ce to the point that Performing a sensitive information cut phase was carried out by 98%.

▸Keyword: Big Data, Personal Information Protect, Sensitive Information, Deep Learning, Sensitive Information Filtering System

Ⅰ. Introduction

빅데이터가 21세기 원유로 불리며 관심을 모으고 있는 가운데, 엄격한 개인정보보호제도가 국내 빅데이터 활용을 지연시키고 있다는 지적과 국내 빅데이터 공급기업 및 수요기업을 대상으로 조사한 결과, 빅데이터산업 활성화에 가장 큰 걸림돌 중 하나로 개인정보보호법이 문제로 지적되고 있다[1]. 그러나 개인정보보호법 시행 이후 5년 동안 개인정보보호 환경은 급속히 변화했다. 빅데이터, 사물인터넷, 클라우드 컴퓨팅, 이동형 영상정보처리기기 등 신기술의 발달로 개인정보 활용이 일상생활 전반으로 급속히 확산되었다. 서비스 산업에서 개인정보 활용 증가로 생활은 편리해진 반면, 개인정보 유출 등 침해 사고가 증가하고 있다. 2011년 개인정보보호법 시행 이후에도 개인정보 침해사고는 지속되고 있는 반면, 국민의 개인정보보호 인식은 심화되고 있다. 2016년 행정자치부와 개인정보보호위원회가 공동으로 실시한 개인정보보호 실태조사 결과, 정보주체의 88.5%가 '개인정보를 중요하게 인식하고 있다'라고 답변했다. 반면, 정보주체의 45.4%는 '개인정보처리자가 개인정보보호의 중요성을 인식하고 실천하고 있다'고 답변했다. 이렇듯 정보주체의 개인정보보호 인식수준에 비해 개인정보처리자의 인식수준은 낮게 평가되고 있다[2].

Fig.1은 2016년 개인정보보호 실태 조사자료에서 나타났듯이 민감 정보 처리 제한에 대한 인지율(잘 안다+ 들어본적 있다)은 공공기관이 99.6%, 민간기업의 경우 72.1%로 나타났으

• First Author: Gun-Seo Koo, Corresponding Author: Gun-Seo Koo
*Gun-Seo Koo (gskoo@sewc.ac.kr.ac.kr), Dept. of Digital Media, Soongeui Women's College
• Received: 2017. 11. 01, Revised: 2017. 11. 10, Accepted: 2017. 11. 25.
• This work was supported by Research Grant of Soongeui Women's College in 2016.

며, 공공기관, 교육기관은 비교적 민감정보 처리 제안 인지 여부를 묻는 질문에 공공기관 91.5%, 교육기관 88.8%를 인지한 반면, 민간기업의 경우는 겨우 32.9%로 나타나 민감정보 의식에 많은 문제가 나타나는 것으로 평가되고 있다[3].

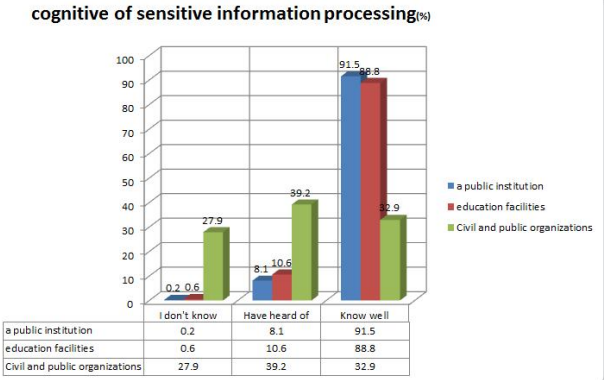

Fig. 1. Sensitive information processing restricted statistical data[1]

또한 개인과 관련된 데이터를 수집하고 분석하는 것이 사생활의 침해라는 염려에도 불구하고 국가와 개인의 경제적인 이익 창출에 기여할 것으로 예상된다. 빅데이터시장 활성화를 위해서는 다양한 데이터간 매쉬업을 통한 가치창출이 핵심인데, 이를 위한 데이터 거래 자체가 어려운 현실이다. 현재 국내에서는 개인정보 범위의 불명확성, 경직적 사전 동의 제도 등으로 인해 사실상 효율적 빅데이터 서비스가 곤란해외의 경우 개인정보의 보호 및 개인정보의 안전한 활용을 함께 추구하는 입장으로, 최근 IoT, 빅데이터 등의 신규사업 활성화를 위해 개인정보보호 관련 법제 개선을 추진 중이며, 실재 현업에서 개인정보관련 법제도로 인해 장애 요소 발굴을 통해 향후 정책방향을 모색 중이다. 따라서 개인정보 가운데 민감정보는, 민감정보에 대한 명확한 기준이 모호하여 사업추진에 장애요소가 발생하고, 또한 민감정보는 정보주체의 사생활을 침해할 우려가 있는 개인정보로 개인정보보호법에 의해 처리 제한하는 규정(개인정보보호법제23조)한다는 것이다[4].

II. An Analysis of Personal Information Processing Restriction

1. Trend Analysis of Personal Information Damage

개인정보는 개인에 관한 실질 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함한다)를 의미한다. 이러한 개인정보가 누군가에 의해 악의적인 목적으로 이용하거나 유출될 경우 개인의 사생활에 큰 피해를 줄 뿐만 아니라 개인 안전과 재산에 피해를 줄 수 있다. 또한 스팸문자나 보이스 피싱, 나를 사칭한 메신저 상의 금융사기 등이 모두 개인정보 유출과 관련되어 있는 경우가 많다. 아래 Table. 1은 개인정보의 유형과 범위를 나타낸 것으로 인적 정보, 신체적 정보, 정신적 정보, 신용정보, 사회적 정보 등이 포함되어 있다[5].

Table 1. Scope of Personal Information.

Type	Scope of personal information
Human intelligence	Name, resident registration number, address, telephone number, date of birth, birth date, email address, family member information, family member information, etc.
Physical information	Physical information (facial, fingerprint, iris, voice, gene information, height, weight, etc.) Medical care and health information (health status, medical records, physical disability, strength, etc.)
Mental information	Such as ideas, religion, values, etc.
Credit information	Financial transaction information (credit card number, account number, loan and savings details, real estate retention details, credit assessment information, etc.)
Social information	Education information (educational records, records, qualifications, details of criminal records, etc.), legal information (records, criminal records, details of fines, etc.), etc.
etc.	Location information such as phone calls, website access details, email or telephone messages, GPS, etc.

Table. 2는 2013년도~2016년까지 연도별 유형별 개인정보 침해신고 및 상담현황을 나타난 도표로 매년 15만 건의 개인정보 침해 신고 및 상담이 발생하고 있다는 것을 알 수 있다. 실제 통계자료에서는 2010년도~2013년까지 꾸준히 증가하다가 2014년부터 감소세가 이루진 것은 개인정보 보호법이 2011년 3월 29일 제정되어 같은 해 9월 30일부터 시행되었지만 세부 부칙은 2013년 8월 6일부터 시행된 것이 원인으로 분석된다[4].

Table 2. Reported reports of personal information violation and consultation by each type of year(Unit : Number)

Classification	'2013	'2014	'2015	'2016	Total
Report	2,347	2,992	2,316	1,559	9,214
Counseling	175,389	155,908	149,835	96,651	577,783
Total	**177,736**	**158,900**	**152,151**	**98,210**	**586,997**

아래 Table. 3의 경우는 개인정보침해 신고 접수 유형별 분석으로 침해 사안 중에 타인 개인정보 훼손침해 도용, 개인정보 안전성 확보 조치, 개인정보 수집 요건, 타법 관련 개인정보 사례 순으로 나타났다. 그밖에 침해 유형으로는 과도한 개인정보 수집, 목적외 이용 또는 제3자 제공, 개인정보취급자에 의한 훼손 침해 등으로 나타났다[4].

Table. 4는 개인정보 및 프라이버시 침해 유형을 나타낸 것으로 첫 번째 사업자의 관리 소홀로 개인정보가 유출된 사례로는 50대 91.7%로 가장 높게 나온 경우이며, 동의 없이 개인정보를 본래 목적이외의 용도로 이용 또는 제 3자 제공된 경우는 40대가 67.4%로 가장 높게 나온 경우이다[2].

Table 3. An Analysis of receipt of report by Type of Personal Information violation(Unit : Number)

Reception Type \ Year	'2013	'2014	'2015	'2016
Personal Information Collection Requirement	2,634	3,923	2,442	2,568
Notice of disclosure when collecting personal information	84	268	65	54
Collect excessive personal information	1,139	1,200	868	390
Use for purposes other than purpose or third party	1,988	2,242	3,585	3,141
Personal injury caused by personal information handlers	1,022	1,036	857	622
Personal information processing entrustment control	44	40	22	25
business transfer or Take over	47	54	41	41
Person responsible for personal information protection	51	39	48	123
Securing personal information safety measures	4,518	7,404	4,006	2,731
Personal information group	602	686	767	545
Right of information subject to information	674	792	957	855
Violation of inspection	510	352	381	286
Gather Children's Personal Information	36	33	34	33
Theft of personal information damage to other persons	129,103	83,126	77,598	485,57
Personal information related to stroke	35,284	57,705	60,480	38,239
Total	**177,736**	**158,900**	**152,151**	**98,210**

Table 4. Types of Personal Information and Privacy Infractions('06 ~'13: Rate(%))

Case		A	B	C	D	E	F	G
Age	10	79.9	50.0	30.0	0.0	20.0	0.0	0.0
	20	65.5	41.4	32.8	8.6	12.1	0.0	1.7
	30	74.0	66.0	28.0	8.0	12.0	2.0	0.0
	40	63.0	67.4	34.8	10.9	0.0	0.0	0.0
	50	91.7	41.7	25.0	8.3	0.0	0.0	0.0
Gender	male	70.2	52.4	35.7	14.3	7.2	0.0	0.0
	female	69.6	58.7	27.1	3.3	9.8	1.1	1.1
Total	Subtotal	69.9	55.7	31.2	8.5	8.5	0.6	0.6

A: Personal information leaks due to careless management of business operators
B: Use of personal information without consent for purposes other than consent
C: Collect personal information and use it for telemarketing purposes
D: No sign of website membership or economic damage due to resident registration
E: Theft items, cyber money, characters, etc. stealing game items with ID and password
F: Financial damage caused by theft of credit information / G: etc.

2. Rules for limiting sensitive information process

개인정보에는 이름, 주소, 전화번호 등과 같은 개인에 대한 객관적인 신상정보도 포함이 되지만, 개인의 감정이나 사상 또는 종교관 등 신상정보와 구별되는 개념의 개인정보도 있다. 이와 같은 정보들은 개인정보에 해당하지 않는 것으로 인식되기 쉬우며, 그만큼 정보주체의 프라이버시 침해 가능성도 높다. 이를 방지하기 위해 개인정보보호법 제23조에서는 '민감정보'에 대해 아래와 같은 규정을 두고 있다[1].

개인정보보호법 제23조(민감정보의 처리 제한) 개인정보 처리자는 사상·신념, 노동조합·정당의 가입·탈퇴, 정치적 견해, 건강, 성생활 등에 관한 정보, 그 밖에 정보주체의 사생활을 현저히 침해할 우려가 있는 개인정보로서 대통령령으로 정하는 정보(이하 "민감정보"라 한다)를 처리하여서는 아니 된다. 그밖에 시행령 제22조는 "유전정보, 범죄경력에 관한 정보도 민감정보에 해당한다." 라고 규정하고 있다. 이러한 민감정보는 다른 개인정보 항목과 비교하여 보다 민감하여, 침해나 유출 시 정보주체의 프라이버시에 보다 큰 영향을 미칠 수 있기 때문에 일반 개인정보와 구분하여 그 처리를 보다 엄격하게 규정한 것으로 타당하며, 민감정보는 원칙적으로 그 처리가 금지되있다. 다른 법령에서 민감정보의 처리를 구체적으로 언급하고 있다"[2]. 또 하나의 민감정보 처리 제한 규정으로 정보통신망법으로 "제23조(개인정보의 수집 제한 등) ① 정보통신서비스 제공자는 사상, 신념, 가족 및 친인척관계, 학력(學歷)·병력(病歷), 기타 사회활동 경력 등 개인의 권리·이익이나 사생활을 뚜렷하게 침해할 우려가 있는 개인정보를 수집하여서는 아니 된다." 라고 제한 규정이 명시되어 있다[1].

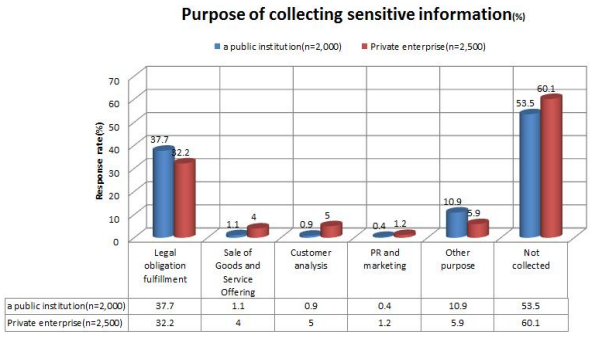

Fig. 2. Purpose of collecting sensitive information

그럼에도 불구하고 Fig. 2.와 같이 민감정보 수집을 하고 있으며, 수집 목적에 대해 공공기관(37.7%)과 민간기업(32.2%) 모두 '법적 의무 이행'이 가장 높게 나왔으며, 특히 공공기관의 경우 광역지자체(88.2%), 기초지자체(82.1%), 민간 기관의 경우 전기/가스/수도(50%), 보건/복지(45.8%)가 응답 비율이 높은 것으로 조사되었다[3]. 따라서 민감정보에 대해 개인정보보호법과 정보통신망법상의

범위가 달라 처리에 많은 혼란이 초래된다[4].

3. Comparative analysis by country for Personal information regulation

한국이외에 해외 각국마다 개인정보 규제에 대한 비교해보면 Table. 5와 같이 한국의 경우 개인정보 전체 처리과정에서 사전 동의(Opt-in) 방식을 취하고 있고, 정의에 있어서도 보다 포괄적이며, 기타 선진국에 비해 개인정보 활용이 엄격히 법으로 규제되고 있다[1].

Table 5. Apply Personal Information Regulatory per National

	Nation	Regulatory Information Regulatory Information	Note
Consent	Korea	Pre-processing Consensual method	Korea adopted Opt-In Approach in the current process of handling personal information in the current law.
	EU	Instructions are informed prior to pre-approval consent, consent arrangements, and pre-approval arrangements	
	Japan	Informed consent and follow-up arrangements	
	U.S.A.	identified cost method (Prior consent of third parties)	
Regulation	Korea	Mixed relations the criminal law and Administrative law	Korea discriminates against processes by separating them from the process of treatment.
	EU	There are no guidelines, however, the regulations are mandatory.	
	Japan	Violation of breach of corrective action	
	U.S.A.	identified cost method	

한국의 경우는 성명, 주민번호, 영상등과 같은 정보를 국한하고 다른 정보와 쉽게 결합해 개인 식별가능 정보를 개인정보로 포괄적으로 정의하고 있다. 일본의 경우는 '조합'(대조)이란 용어를 사용한 반면, 한국은 '결합' 단어를 사용하기 때문에 한국이 개인정보를 가장 넓게 규정한 것으로 볼 수 있다. 일본 생존한 개인의 정보(예: 성명, 생년월일 등)와 다른 정보와 쉽게 조합해 개인 식별 가능 정보를 개인정보로 정의하고 있다. EU는 식별되거나 식별가능한 자연인 정보(예: 신분증 번호 또는 신체·생리·정신·경제·문화·사회적 요소 정보로 직·간접적 알아볼 수 있는 사람)으로 정의하지만 규정에는 식별되거나 식별가능한 자연인 정보(예: 지침+이름, 지역정보, 고유식별자 또는 성적 정체성) 미국 개별법상 상이(프라이버시, 금융정보, 소비자정보, 의료정보, 통신정보 등)을 모두 포함하고 있다[4].

III. Sensitive information Filtering strategy in Big data analytics

1. Big Data Collection and Analysis Technology

빅 데이터 기술은 크게 보아, 빅데이터 저장 기술, 빅데이터 정제 기술, 빅데이터 분석/가시화 기술, 빅데이터 기반 예측 기술로 나눌 수 있다.

첫째 빅데이터 저장 기술은 대용량의 비정형화 된 데이터를 저장하고 처리, 통계 분석하는 기술을 의미한다. 클라우드 환경에서, 데이터를 수집, 저장하는 기술을 NoSQL이라고 하며 통상 카산드라DB, 몽고DB, Hbase, Redis를 많이 사용한다.

두 번째는 빅데이터 정제기술로 빅데이터 자체가 원체 크기 때문에 아무리 빅데이터가 전수조사를 전제로 하는 기술이라 해도 원시 데이터를 정제하지 않을 수 없다. 기계나 사람이 쏟아내는 데이터는 컴퓨터에 의해 바로 분석이 어렵기 때문이다. 대부분 빅데이터는 컴퓨터가 바로 처리할 수 없는 데이터, 컴퓨터 입장에서는 잘 정돈되지 않은 데이터이다. 이를 비정형 데이터(unstructured data, irregular formatted data) 라고 하며, 설사 정형적 데이터(structured data)라고 하더라도 측정값이 빠져 있다거나, 형식이 다르다거나, 내용 자체가 틀려있는 경우가 매우 많기 때문이다. 비정형 데이터를 분석 가공 가능한 형태로 만들거나, 컴퓨터가 바로 처리할 수 없는 데이터를 포함한 정형적 데이터를 분석 가공 가능한 형태로 만드는 데에는 방대한 컴퓨팅 능력이 필요하다.

세 번째로는 빅데이터 분석 방법으로는 의미를 발견하는 데이터 처리 및 분석기술, 즉 의미 분석 기술과 데이터마이닝기술 및 관련기법으로 알고리즘이 필요하다. 데이터마이닝(data mining)은 통계 및 수학적 기술뿐만 아니라 패턴인식 기술들을 이용하여 데이터 저장장소에 저장된 대용량의 데이터를 조사 분석하여 의미 있는 새로운 상관관계, 패턴, 트레드 등을 발견하는 과정으로 본 논문에서도 데이터마이닝 기법이 주요한 요소로 작용한다. 마지막으로 빅데이터 분석/가시화 기술은 대부분 통계학이나 데이터마이닝이나 OLAP(OnLine Analytical Processing)이라 부르는 분야의 일입니다. 원체 오래된 학문 분야이고 완성도가 높지만, 빅데이터가 전수분석을 암묵적으로 전제하는 기술이라서 시간을 단축하고 정확도를 높이기 위한 연구와, 축적된 학문적 연구들을 현실에 적용하는 컨설팅이나 상품화 자체에 많은 가능성이 열려 있다. 이러한 기술을 기반으로 본 논문에서는 빅데이터를 활용한 민감정보 필터링 시스템 구축을 위한 설계하고자한다.

2. Deep Learning for Big Data Classification

빅데이터 분류를 위해 적용하는 딥런닝은 컴퓨터가 여러 데이터를 이용해 마치 사람처럼 스스로 학습할 수 있게 하기 위해 인공신경망을 기반으로 구축한 한 기계 학습 기술[6]로 본 연구에서 민감정보 빅데이터 분류를 위해 적용하려는 딥러닝 모델은 SONN 모델을 기반으로 하고 있는 FE-SONN모델이다. SONN모델은 베즈덱(James Bezdek)의 퍼지 c-means 알고리즘의 퍼지 멤버십 등식을 신경망과 융합한 자율적인 자기조직화 신경망 모델이다[5,7]. 따라서 이 모델은 외부의 교사를 필요로 하지 않는다. 이 모델은 주어진 입력에 대한 클러스터의 수나 클러스터의 중심에 대한 사전 지식 없이 자율적으로 클러스터에 관한 정보를 제공하여 유사 패턴분류와 패턴인식 등에 적합하며 높은 신뢰도 결과를 보여주고 있다. SONN 모델에 쓰이는 SONN 알고리즘은 빅데이터

분류상으로 자율적인 학습을 하며 유사한 민감정보 입력 데이터 값을 분류 처리할 수 있는 알리즘으로 코호넨의 자기조직화 모델과 유사한 점이 있다[9].

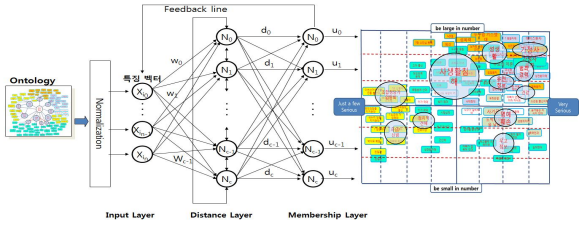

Fig. 3. The Model of FE-SONN

본 논문에서 적용한 빅데이터 분류를 위해 적용한 FE-SONN 모델은 Fig. 3에서 보는 바와 같이 입력벡터가 입력층으로 들어오고 거리층과 멤버십층에서 피드백 하면서 클러스터들의 정보를 제공하여 준다. 민감정보 빅데이터를 자기학습을 통해 Taxonomy를 생성하게 한다. 이 모델을 빅데이터에서 적용하는 장점은 자기조직화 기능이다. 이것은 입력데이터의 유사 클러스터 중심점 등에 관한 어떤 사전 정보도 없이 들어온 입력 민감정보에 대해 클러스터와 멤버쉽에 관한 정보를 제공한다.

3. Design of Sensitive Information Filtering System in Big Data

본 연구에서도 아래 그림 2와 같은 민감정보를 필터하기 위한 빅데이터 민감정보 필터 시스템을 제안했고, 빅데이터 민감정보 필터 시스템은 빅데이터 처리과정과 민감정보 처리 과정 등 2개 과정으로 구성되었다.

첫 번째 과정인 데이터 처리 과정은 통상적인 빅데이터 분석 과정 절차에 의한 처리된다. 외부 데이터는 Social Media(SNS, Blog, Community)와 Mass Media(News, BBS, RSS)를 통해 데이터 수집하고 수집된 데이터를 저장한다. 어휘 구분 문석 단계에는 문장 분리→단어분리→구문분석→어휘 분석을 거쳐 의미 분류를 시행한다. 의미 분류는 온톨로지 생성하고, 의미 정보 분류와 최종어휘 매칭 분류한다. 두 번째 과정인 민감정보 처리 과정은 민감정보 DB 생성→민감 정보 필터→신뢰도 분석→실무적용까지 수행한다. 즉, 1단계에서 어휘 매칭 분류된 자료와 민감정보와 비교하여 일치하면 필터링하는 방식이고, 민감정보를 필터하기 전 신뢰도 검사를 실시하여 정확도를 높였다는 것이 본 논문의 특징이다.

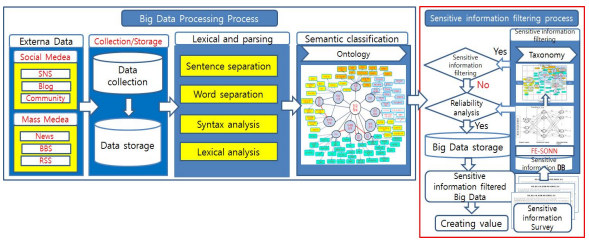

Fig. 4. System Design for Sensitive Information Filter in Personal Information

Fig.4에서 민감정보 설문 조사는 먼저 빅데이터 분석에서 민감정보에 관한 법률적 의미를 교육을 하고, 학생 본인의 경우 이러한 내용(단어, 문장 포함)이 SNS상에서 거론된다면 본인에게 있어서 그림 4와 같이 매우 심각, 심각, 보통, 심각도가 낮음 등 총 4단계로 작성하게 했으며, 한 항목 당 최대 10개까지 응답하게 했다. 민감정보를 적용하기 위해 실험 대상을 20대 대학생으로 제한했으며, 20대 대학생이 받아들이는 민감정보에 찾아내기 위한 다음과 같은 방법으로 적용했다. 설문조사에 참여학생이 민감정보 앙케이트에 응답학생은 모두 838명이며, 응답자 93%(779명)가 20대 여자이며, 7%(58명)가 30대 일반 성인이 응답하였으며, 대학생학력 소지자가 95%이다.

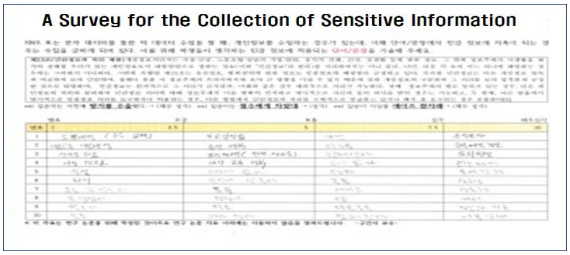

Fig. 5. Sensitive Information Survey Form

Fig. 5는 민감정보 설문조사 결과로, 매우심각, 심각, 보통, 심각도 낮은 총 4단계로 설문 조사 결과 여생들의 경우 민감정보에는 외모, 성형, 개인정보가 매우심각과 심각단계에서 민감정보로 느끼고 회피하고 싶은 단어로 조사되었다. 이렇게 조사된 민감정보 체감 앙케이트 조사 결과를 토대로 단어 및 문장으로 구성된 온톨로지를 구성하게 된다.

Fig. 6. Results of sensitive information questionnaire survey

Fig. 6은 민감정보에 대한 온톨로지로서 사람들이 세상에 대하여 보고 듣고 느끼고 생각하는 것에 대하여 서로 간의 토론을 통하여 합의를 이룬 바를 개념적이고 컴퓨터에서 다룰 수 있는 형태로 표현한 모델로, 개념의 타입이나 사용상의 제약조건들을 명시적으로 정의한 기술이다. 즉 언어로 표현된 개념 간 연관 관계 지식이 드러나는 망으로 표현했으며, 온톨로지를 통한 개념 간의 분석은, 정보시스템의 대상이 되는 자원의 개념을 명확하게 정의하고 상세하게 기술하여 보다 정확한 정보를 찾을 수 있도록 하는데 도움을 주는 한편, 용어와 용어 사이의 관

계를 컴퓨터가 이해할 수 있는 형태로 정의한 것으로 지식을 공유하고 새로운 지식이 추가되었을 경우 기존 지식과의 연계를 유연하게 할 수 있는 장점을 제공한다.

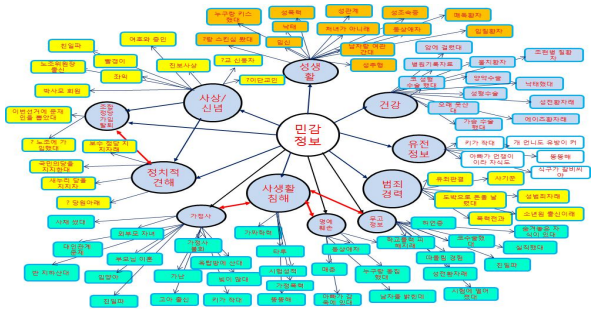

Fig. 7. Ontology for Creation and Data Classification

따라서 Fig. 7에서 나타난 데이터 생성 및 분류를 위한 온톨로지는 민감정보에 대한 빅데이터 분석을 위해 온톨로지 분석을 통해 검색 룰(rule)과 결합하여 연역적 논리 추론이 가능하다는 장점이 있다. 최근에는 대규모 소셜 미디어 및 동적 소셜 네트워크에서 분석을 수행하는 방법으로 시멘틱 기술을 사용하기도 한다. 이 방법의 특징은 복잡하고 이질적인 그래프 구조를 트리플로 구성된 RDF(Resource Description Framework)그래프를 표현하는데, RDF는 웹에 있는 자원에 관한 메타 정보를 표현하기 위한 언어로 상호 이질적 정보들을 단일적 표현체계로 통합 표현 및 연계가 가능하고, 각 노드의 특성정보를 통합 표현이 가능하다[5].

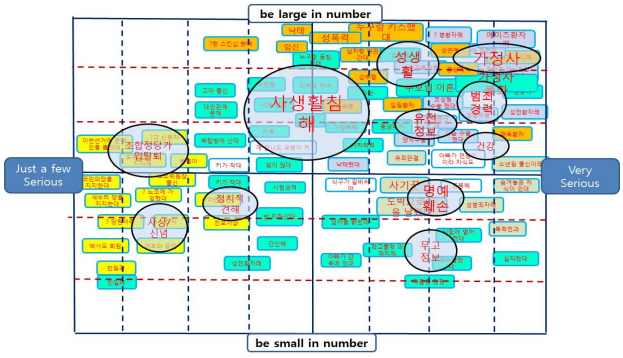

Fig. 8. Creating Taxonomy of Sensitive Information

Fig. 8은 Fig. 7에 의해 만들어진 온톨로지 생성 및 분류 정보를 기반으로 Taxonomy를 생성하게 한다. 이때 생성된 Taxonomy는 다층구조 형태의 신경망을 기반으로 하는 머신 러닝으로 빅 데이터의 자동 분류 방법인 딥런닝 기법을 적용하는 방법으로 FE-SONN모델을 제안했고, 딥러닝 모델은 SONN모델을 기반으로 하고 있는 FE-SONN모델이다. SONN모델은 베즈덱(James Bezdek)의 퍼지 c-means 알고리즘의 퍼지 멤버십 등식을 신경망과 융합한 자율적인 자기조직화 신경망 모델로 입력벡터가 입력층으로 들어오고 거리층과 멤버십층에서 피드백 하면서 클러스터들의 정보를 스스로 제공해주는 특징이 있다[6,7]. 따라서 민감정보 빅데이터를 자기학습을 통해 Taxonomy를 생성하는 역할을 한다.

Ⅳ. Reliability Factor and Performance Evaluation of Sensitive Information

본 연구는 빅데이터를 분석시 입력되는 데이터가 민감정보 인지를 판별하여 차단하고자 하는 부분에 있어서, 딥런닝의 부분에 FE-SONN 모델을 제안했고, 이를 적용했을 때 신뢰도 분석 및 성능평가를 통해 본 연구의 타당성을 입증하고자 한다.

1. Reliability Factor of Deep Learning

본 연구에서 신뢰도 분석은 동일한 측정대상에 대해 같거나 유사한 측정도구로 설문 방법을 사용하여 반복적으로 측정할 경우에는 동일하거나 비슷한 결과를 얻고자 하며, 신뢰도 분석은 안정성, 일관성, 예측가능성, 정확성을 의미한다.

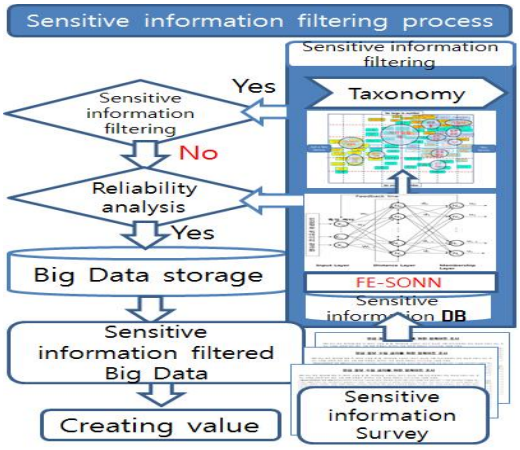

Fig. 9. Sensitive information interdiction process

따라서 Fig. 9에서 제시된 것은 빅데이터를 분석 시 입력되는 데이터가 민감정보 인지를 판별하여 차단하고자 하는 부분에 있어서, 딥런닝의 부분에 FE-SONN 모델을 제안했다. 딥러닝의 구조는 기존의 다층 인식기를 이용해서 인식 처리했을 때 입력 층의 노드 수는 전역적 특성을 갖고 있는 36개 특징 벡터와 지역적 특성(오프라인상에서 추출) 12개의 특징 벡터 정보 총 48개가 된다. 또한, 기존 신경망 알고리즘의 경우 목표값은 비트의 조합에 의해 출력한다. 그러나 목표값을 비트 조합으로 정의할 때 오인식 처리할 경우가 발생한다. 즉, 테스트 데이터의 결과가 임계치를 0.75로 두었을 때, 신경망이 학습된 목표값에서 어느 하나의 유니트 결과가 0.75를 초과하게 되면, 예측이 불가능한 인식 결과를 출력하게 된다. 이러한 인식 결과는 어떠한 경우라도 오인식 결과로 찾아 낼 방법이 없다. 이러한 단점을 해결하기 위해 본 논문에서 제안 방법은 다음과 같이 출력 유니트의 개수를 입력된 원본 서명 개수와 같게 함으로서 인식된 결과가 학습 목표값과 일치하는 지를 검색할 수 있으며, 이로 인해 오인식 요소를 줄일 수 있고, 이를 기반으로 감증할

Table 7. Results of Big Data Processing Step Experiment

Phase	Execute phase	Experimental data	Success data	Proportion Success rate(%)	Phased end Success rate(%)
Collection and storage	Data collection	8978	8972	99.93%	99.93%
	Big data storage	9872	8962	90.78%	99.82%
Lexical parsing	Sentence separation	8962	8956	99.93%	99.75%
	Word separation	8956	8931	99.72%	99.48%
	Syntax analysis	8931	7821	87.57%	87.11%
	Lexical analysis	7821	7735	98.90%	86.16%
Semantic classification	Semantic classification	7735	7652	98.93%	85.23%
	Ontology creation	7652	7553	98.71%	84.13%
Results of Big Data Processing Steps		8978	7553	84.13%	

Table 8. Test results of Sensitive Information Filtering Process Step

Phase	Execute phase	Experimental data	Success data	proportion Success rate(%)	Phased final Success rate(%)
Sensitive Information Learning Steps	Survey data analysis	3580	3580	100.00%	100.00%
	Lexical and Syntax analysis	3580	3575	99.86%	99.86%
	FE-SONN Deep Learning	3575	3560	99.58%	99.44%
	Taxonomy Execute phase	3560	3550	99.72%	99.16%
Result of Sensitive Information Prevention Steps		3580	3550	98.00%	

수 있다. 이것이 본 논문의 중요한 포인트이기도 하다. 그러나 시험 대상인 48개의 특징 추출 값 가운데 가장 큰 값은 0.9372이고 두 번째 큰 값은 0.1005이 된다. 이렇게 인식된 결과에 대한 신뢰도 측정을 하게 된다. 신뢰도 수식은 <식 1>과 같이 정의했다.

신뢰도 지수(Reliability Factor)

$$RF = \frac{MAX(bit\_array(i)) - 2nd\_MAX(bit\_array(i))}{MAX(bit\_array(i))}$$
$$= 1 - \frac{2nd\_MAX(bit\_array(i))}{MAX(bit\_array(i))} \quad \cdots\cdots(1)$$

따라서 FE-SONN망에 의해 처리된 목표값을 나타내는 빅데이터 분석 시 민감정보로 인식되는 값의 범주의 신뢰도에서 최대값 $MAX(bit(i))$은 0.8952이고 $2nd_MAX(bit(i))$는 0.2217 이므로 신뢰도 지수를 구하면 1-(0.2217/0.8952)= 0.752346이 된다. 따라서 본 논문에서는 Table. 6과 같이 RF의 값이 0.75이상이면 인식 결과의 신뢰성을 갖는 것으로 간주하여 인식 결과를 출력한다. 아래 표6에서 나타난 것과 같이, 신뢰도 지수를 0.75으로 나타난 신뢰도 경계 값으로 지정한 이유는 민감정보 빅데이터로 인식 결과의 비트 배열에서 가장 큰 $MAX(bit(i))$와 두 번째로 큰 $2nd_MAX(bit(i))$차에 대한 비율 값으로 신뢰도 지수로 평가 했을 때 0.75이상이면 정확도 부분에 만족하는 것으로 인정하는 것으로 했다.

Table 6. Compare table of Reliability Factor

$MAX(bit(i))$	$2nd\_MAX(bit(i))$	Reliability Factor
0.9372	0.1005	0.892766
0.9232	0.1409	0.847379
0.9092	0.1813	0.800594
0.8952	0.2217	0.752346
0.8812	0.2621	0.702565
0.8672	0.3025	0.651176
0.8532	0.3429	0.598101

2. Performance Evaluation and Experiment Results

본 논문은 빅데이터 수집 및 분석시 개인정보에서 민감정보에 관한 수집 및 분석을 제한하는 규정에 의해 이를 제거하기 위한 전략 구축에 관한 내용이다. 빅데이터 분석시 민감정보 차단을 위해 크게 두 개 단계로 나누었다. 첫 번째는 빅데이터 처리 단계와 민감정보 차단 단계로 나누어 수행 및 실험을 했다. Table. 7은 빅데이터 수행 과정에서 수집 및 저장 단계에는 데이터 수집 단계, 빅데이터 저장단계를 수행 했다. 어휘 구문 분석 단계에는 문장 분리단계, 단어 분리 단계, 구문 분석 단계, 어휘 분석 단계를 수행했고, 의미 분류 단계에는 의미 분류 단계, 온톨로지 생성하는 과정을 수행했다. 그 결과 실험에 수행된 빅데이터 수는 8978개 중 7553를 온톨로지 수행까지 최종 성공률은 84.13% 까지 수행 했다.

Table. 8에서는 민감정보 차단 단계에는 민감정보 학습 단계에서 민감정보 설문 자료 분석, 어휘 분석 및 구문 분석, FE-SONN 딥러닝 단계 Taxonomy 수행단계를 수행했을 때 실험 데이터 수 3580개중 성공 데이터 수는 3550개로 98.00%의 성공률을 보였다. 이렇게 높게 나온 원인은 이미지 데이터와 같이 모호한 데이터를 학습하는 것이 아니라, 문자 데이터 패턴매칭이기 때문에 성공률이 높은 것으로 나타났다.

최종 실험의 범위는 완벽한 의미 분석을 포함하고 있지 않고 단지 민감정보와 일치되는 단어, 문장으로 제한하여 인식하는 방식을 적용했기 때문에 숨은 뜻이나 뉘앙스, 억양에 관련한 정보는 배제했다.

Ⅴ. Conclusion

본 연구는 SNS, BLOG 등에서 빅데이터를 분석시 입력되는

데이터가 민감정보인지를 판별하여 필터링하는 시스템을 제안했다. 개인정보에는 이름, 주소, 전화번호 등과 같은 개인에 대한 객관적인 신상정보도 포함이 되지만, 개인의 감정이나 사상 또는 종교관 등 신상정보와 구별되는 개념의 개인정보도 있다. 이와 같은 정보들은 개인정보에 해당하지 않는 것으로 인식되기 쉬우며, 그만큼 정보주체의 프라이버시 침해 가능성도 높다. 이를 방지하기 위해 개인정보보호법 제23조에서는 '민감정보'에 수집 및 활용에 대한 규정을 두고 있다. 2016년 행정자치부와 개인정보보호위원회가 공동으로 실시한 개인정보보호 실태조사 결과, 정보주체의 88.5%가 '개인정보를 중요하게 인식하고 있다'라고 답변했다. 반면, 정보주체의 45.4%는 '개인정보처리자가 개인정보보호의 중요성을 인식하고 실천하고 있다'고 답변했다. 이렇듯 정보주체의 개인정보보호 인식수준에 비해 개인정보처리자의 인식수준은 낮게 평가되고 있다[10,11]. 따라서 본 논문에서는 빅데이터 민감정보 필터링 시스템을 제안했다. 제안된 시스템 구조는 빅데이터 처리과정과 민감정보 필터 처리과정 등 2단계 과정으로 나뉘어지며, 빅데이터 처리과정은 4단계로 빅데이터 수집에서 분석 및 적용하게 된다. 데이터 수집·저장→어휘 분석 및 구문 분석→의미 분류→민감 정보 필터→신뢰도 분석을 통해서 비로소 개인정보 민감정보가 필터링된 빅데이터를 저장하여 업무에 적용하게 되는 여러 단계별로 적용되는 형태로 설계 되어있다.

본 연구에서 민감정보 빅데이터 분류를 위해 적용하려는 딥러닝 모델은 SONN모델을 기반으로 하고 있는 FE-SONN모델이다. 이 모델을 빅데이터에서 적용하는 장점은 자기조직화 기능이다. 이것은 입력데이터의 유사 클러스터 중심점 등에 관한 어떤 사전 정보도 없이 들어온 입력 민감정보에 대해 클러스터와 멤버쉽에 관한 정보를 제공한다는 장점이 있다.

제안된 시스템의 실험결과, 실험에 수행된 데이터 수는 8978개 중 7553를 온톨로지 수행까지 최종 성공률은 84.13%까지 수행했고 민감정보 차단 단계 수행 결과 98% 정도 수행했다는 점에 대한 의의가 크다.

REFERENCES

[1] Oh, J. Y, Sun M. R, "A Case Study and Analysis of a Big Data make use caused by Personal Information Protect legislation", National Information Society Agency(NIA), No. 4., 2015.12.

[2] M. K. Kim, Personal Information Protect Consultation Cases, Korea Internet & Security Agency(KISA), 21pp. 2017. 1.

[3] B-Y Lee, "Utilization of Social Media Analysis using Big Data", Journal of the Korea Contents Association 13(2), 211-219. 2013.2,

[4] M. K. Kim, "Personal Information Protect a Fact-Finding Survey",National Information Society Agency(NIA), No. 4., 2015.12.

[5] J. S. Choi, Personal Information Protect How to Use & Protect, Personal Information Protect Practice guide Vol. 6., Security news, 2016.6.9.

[6] D. S. Park, etc, Bigdata Computing Technology, Published by Hanbit Academy, pp.30-35. 2015.1.19.

[7] G. S. Koo, "A Study on Customized Employment Strategy for Women's College Students Utilizing Big Data", Journal of The Korea Society of Computer and Information, Vol. 12 No. 5, pp. 73-81, 2015. 2.

[8] G. S. Koo, "The FE-MCBP for Recognition of the Titled New-Type Vehicle License Plate" Journal of The Korea Society of Computer and Information, Vol. 12 No. 5, pp. 73-81, 2007.

[9] G. S. Koo, "Edge Feature Extract CBIRS for Car Retrieval : CBIRS/EFI", Journal of The Korea Society of Computer and Information, Vol. 15 No. 11, pp. 75-82, 2010.

[10] G, B, Lee, "Implementation of Smart Government Using Bigdata", NISC, pp. 15-23, 2011.

[11] National Information Society Agency(NIA), "IT & Future Strategy," No. 3. 2012. 4.

Authors

Gun-Seo Koo received the M.S. and Ph.D. degrees in Computer Science and Engineering from Soongsil University, Korea, in 1990 and 1997, respectively. Dr. Koo is Currently a Professor in the Department of Digital Media from Soongeui Women's College. He is interested in Big Data Processing technique, Content Based Retrieval Technology,