

# 문자 기반 유해사이트 판별 기법

정규철<sup>†</sup> · 이진관<sup>\*\*</sup> · 이태현<sup>\*\*\*</sup> · 박기홍<sup>\*\*\*\*</sup>

## 요 약

본 논문에서 청소년들의 정신 건강을 해치는 유해 정보 사이트를 차단하기 위해 기존 방식이 아닌 내용 기반을 중심으로 하여 중요도가 가장 높은 한 개의 복합 키워드와 정보통신윤리 위원회에서 제시한 유해단어의 가중치를 이용하여 가중치 평균을 더해 유해도를 판단하여 유해 사이트와 일반 사이트를 구별하는 시스템을 구현하였다. 예비 실험을 통해 구해진 유해도의 값 3.5를 유해 정보 사이트를 판단하는 기준으로 정한 다음 유해 정보 차단 시스템의 성능 실험을 위해 유해 정보 사이트와 일반 사이트를 각각 무작위로 100개씩 추출해 접속해 본 결과 유해 사이트를 유해 정보 사이트로 판명된 비율이 78%를 보였고 일반 사이트를 일반 사이트로 판명된 비율이 96%가 되어 본 시스템의 유효성을 확인 할 수가 있었다.

키워드 : 키워드 추출, 복합명사, 의존관계, 유해도, 유해 사이트

## A Harmful Site Judgement Technique based on Text

Kyu-Cheol Jung<sup>†</sup> · Jin-Kwan Lee<sup>\*\*</sup> · Taehun Lee<sup>\*\*\*</sup> · Kihong Park<sup>\*\*\*\*</sup>

### ABSTRACT

Through this research, it was possible to set up classification system between 'Harmful information site' and 'General site' that badly effect to teenagers emotional health. To intercept those entire harmful information sites, it using contents basis isolating. Instead of using existing methods, it picks most frequent using composed key words and adds all those harmful words' harmfulness degree point by using 'ICEC(Information Communication Ethics Committee)' suggested harmful word classification. To testify harmful information blocking system, to classify the harmful information site, set standard harmfulness degree point as 3.5 by the result of a fore study, after that pick up a hundred of each 'Harmful information site' and 'General site' randomly to classify them through new classification system. By this classification could found this new classification system classified 78% of 'Harmful Site' to 'Harmful information site' and 96% of 'General Site' to 'General site'. As a result, successfully confirm validity of this new classification system.

**Keywords** : Extracting keyword, Compound noun, Dependence, Harmful degree, Harmful site

### 1. 서 론

1950년대로부터 정보검색[1]~[6]의 연구가 번창하게

되어 인터넷은 정보의 바다라고 할 만큼 많은 정보와 지식을 사람들에게 제공해서 삶을 풍요롭게 하는 도구가 되었다. 최신 정보의 검색이나 컴퓨터 데이터 처리 및 전송을 빠르게 할 수 있는 장점이 있다. 그러한 장점에 도 불구하고 많은 단점들 가지고 있다[5].

인터넷이 가진 단점들 중에서는 인터넷을 통해 음

† 정회원: 군산대학교 컴퓨터정보과학과 IT교수  
 \*\* 정회원: 군산대학교 컴퓨터정보과학과 박사과정  
 \*\*\* 정회원: 군산대학교 컴퓨터정보과학과 IT교수(교신저자)  
 \*\*\*\* 정회원: 군산대학교 컴퓨터정보과학과 교수  
 논문접수: 2004년 5월 11일, 심사완료: 2004년 8월 25일

란, 폭력, 마약, 성 매매, 도박, 자살 및 살인 등의 사회적으로 받아들일 수 없는 내용들도 포함하고 있고 또한 개인 사생활의 침해 및 불법행위의 조장 등 유해한 정보들이 이성적인 성인들뿐만 아니라 아직은 감수성이 예민한 어린 청소년들에게까지 무차별적으로 제공되어지고 있다.

인터넷의 단점들을 해결하기 위해 유해 정보를 분류하고 차단하기 위해 여러 가지 방법들이 동원되고 있다. 정보검색사 또는 개발자가 직접 조사해서 등록하는 방법, 당사자가 직접 등록해서 차단하는 방법, 차단소프트웨어를 이용해 등록하는 방법과 인터넷서비스사업자(ISP : Internet Service Provider) 서비스 제공업체 자체적으로 차단하는 방법들이 이용되고 있다[8].

인터넷의 단점 해소를 위한 방법에도 문제점이 있다. 하루에도 수십에서 수백의 음란 사이트가 개설되고 있는데, 그 음란 사이트를 다 찾아낼 수 없다는 작업의 한계성, 장비의 고장이나 사고로 데이터가 손실되는 것을 막기 위해 데이터를 하나 이상의 장치에 중복 저장하는 미러링 기법을 유해 정보 사이트의 파일이나 전체를 복사해 다른 웹 사이트로 저장·운영하는 미러사이트를 이용해 등급을 설정한 사이트의 정보를 아무런 제약조건 없이 접근이 가능한 점, 당사자의 등록은 주관적인 성향으로 음란 사이트를 일반 사이트로 등록할 수도 있다는 점과 그리고 유해 차단 소프트웨어의 차단의 한계성이다. 국내의 유해 정보 차단 소프트웨어들은 대부분이 정보검색사 또는 개발자에 의해서 유해 정보 사이트의 목록을 입력하거나 갱신하게 되는데 그 목록에 모든 유해 정보 사이트를 다 입력할 수도 없어 유해 사이트를 차단하는 실효성이 떨어져 청소년들을 유해 정보로부터 보호하는데 한계가 있다[9].

본 연구에서는 기존의 유해 정보 차단의 문제점을 해결하기 위해 내용 기반에 중점을 두고 자동 판별하여 청소년들을 유해 사이트로부터 접근을 자동 차단할 수 있는 시스템을 구현하여 제시하고자 한다. 유해성 여부의 판단 기준을 위해 정보통신 윤리위원회의 등급 판정 기준을 토대로 각 등급에 해당하는 단어의 등급에 따라 가중치를 부여한 후 계산하는 방법과 의존 규칙을 기반으로 복합 키워드를 추출하여 중요도가 가장 높은 한 개의 키워드를 계산하는 방법을 이용하여 유해성 여부를 판별하기 위한 유해도를 계산하여 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 유해 정보 차단 방법을 살펴보고, 3장에서는 유해정보 차단 시스템 설계 및 구현과 등급판정의 의한 가중치 평균과 복합키워드의 중요도를 고려한 유해성 여부를 계산 방법을 제안한다, 4장에서는 실험을 통한 시스템의 평가하여 제안방법의 유효성을 확인하고, 마지막으로 5장에서는 결론과 향후 연구 과제를 말한다.

## 2. 기존 유해 정보의 차단

### 2.1 유해 정보 차단 방법

유해 정보를 차단하는 방법에는 여러 가지가 있으나 대표적으로 몇 가지를 다음의 각 절에서 설명하고자 한다.

#### 2.1.1 등급 설정에 의한 차단

제3자가 직접 조사해서 음란 사이트로 등록해서 차단하는 경우와 음란 사이트를 운영하는 당사자가 직접 성인 정보임을 나타내는 등급을 매기는 경우가 있다. 일정 기준의 정의된 등급에 의해 차단되는 인터넷 내용 선별 기술 체계(PICS : Platform for Internet Contents Selection)에 의해 등급 표시 방법을 표준화해서 차단하는 방법이 있다. 또한 국내 정보에 대해서는 정보 제공자가 자율적으로 등급을 표시하고, 보다 쉽고 편리하게 등급표시를 할 수 있도록 인터넷 내용등급 서비스 홈페이지(<http://www.safenet.net.kr>)에서 자율 등급표시 웹 서비스를 제공하고 있다. 유럽 경제사회 이사회에서도 유럽의 인터넷 사이트가 아동보호를 위해 의무 등급 표시를 하도록 하고 있다.

#### 2.1.2 인터넷서비스사업자(ISP)를 통한 근본적 차단

기술적으로는 ISP의 인터넷 접속장치에 음란물 차단프로그램을 설치하고 음란사이트의 주소를 차단프로그램에 입력시켜, 음란 사이트로의 연결을 원천적으로 차단하자는 것이며, 음란사이트 목록 데이터베이스는 청소년보호위원회, 정보통신 윤리위원회, 시민감시단체 등의 지속적 정보 수집을 통하여 수시로 갱신한다는 것이다.

#### 2.1.3 차단 소프트웨어를 통한 자율적 규제

유해 정보 차단제품은 지난 95년 7월 미국에서 개발된 사이버패트롤(Cyber Patrol)을 시작으로 서프워치(Surf Watch), 사이버시터(Cyber Sitter), 넷내니(Net Nanny) 등이 개발되어 판매되고 있고 서버용 제품으로는 스마트필터, 리틀브라더 등이 있다. 국내에서는 한국전산원이 개발 중이던 개인용 유해 정보차단 소프트웨어인 NCA패트롤의 핵심 기술을 바탕으로 수호천사, 안티-X, 녹스2, 인터넷지킴이 등 개인용 제품과 수호천사×NET, 넷아거스, 스펀, 보더매니저 등 서버용 제품이 개발되어 있다.

## 2.2 유해 정보 차단 방법에 있어서의 문제점

### 2.2.1 기존 유해 정보 차단의 전반적인 문제

제 3 자에 의한 등급 설정이나 차단 목록 기반의 차단 소프트웨어는 차단 목록 데이터베이스를 주기적으로 갱신해 주어야하며, 포괄적 기준에 의해서 무차별적인 차단이 이루어질 수 있다. 제 3 자 또는 차단 도구 개발자나 차단 목록 입력자에 의한 사전 검열의 반발이 있을 수 있다. 정보의 제공자에 의한 자율적인 등급 책정은 객관성이 결여되고 주관적인 판단으로 정보의 명확한 등급 책정이 어렵다. 또한 정보의 제공 장소가 일반 이용자가 접속 가능한 인터넷 불법·유해 정보의 대부분이 해외에서 전송되는 현실을 감안할 때, 사이트 운영자의 등급 설정은 실효성을 기대하기 어렵다. 전체 인터넷 콘텐츠의 약 80%가 미국의 서버에 저장되어 있는 것으로 추산되며, 국내로 유입되는 해외의 인터넷 불법·유해 정보들에 대한 법적·제도적 규제 방법이 전무한 상태이다. 현 상황에서 등급 책정에 대한 법적·제도적 규제를 강화 한다하더라도 해외사이트에는 적용되지 않기 때문에 국내 이용자 및 서비스 제공자들의 반발이 일어나고 있다.

### 2.2.2 '건전 위장' 성인사이트 개설의 문제

인터넷 닷컴 기업들의 경영난이 가중되면서 유명 사이트나 국내 게임 정보를 제공하던 사이트 등 10대 청소년들에게 인기가 있었던 사이트들이 음란 포르노 사이트로 둔갑되어있는 경우가 늘고 있다. 유명 사이트나 게임 사이트가 성인 사이트

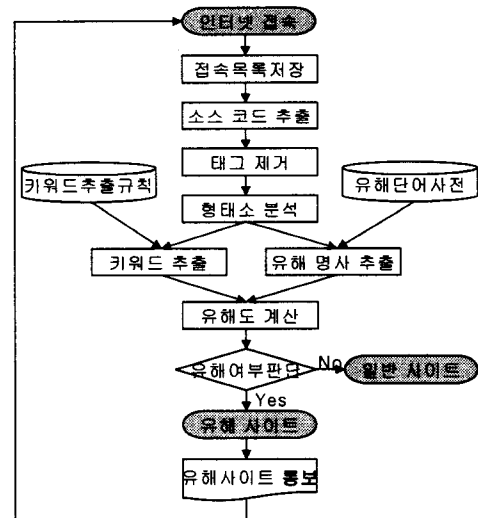
로 둔갑하는 경우가 늘고 있는 것은 단속을 피하기 쉽고 성적 호기심이 왕성한 어린 청소년들을 유혹하기 쉽기 때문이다. 기존의 차단 방법으로 성인 사이트를 차단하던 것을 교묘하게 따돌릴 수 있어서 업자들이 건전한 게임 사이트나 유명 사이트로 10대 청소년들에게 인기를 끌던 사이트를 성인 사이트로 바꾼다. 이럴 경우 기존의 차단 방법으로는 게임 정보를 제공하는 사이트로 인식해서 차단이 이루어지지 않는다는 것이다.

본 논문에서는 기존 유해 정보 차단의 문제와 인터넷 불건전 정보 차단의 기술적 문제, '건전 위장' 성인 사이트 개설 문제를 해결하는 방안으로 내용 분석에 의해 청소년 유해 사이트를 차단하는 시스템의 구현에 대해서 다루고자 한다.

## 3. 시스템의 설계와 구현

본 시스템은 제 3자나 당사자의 등급 책정과 차단 목록 입력 방법은 배제하고 내용 판별을 자동으로 수행하여 복합 키워드를 추출하여 중요도를 계산하고 유해문자 가중치를 계산하여 계산된 복합키워드와 유해문자 가중치에 의해 유해도를 계산하여 유해성을 판별한다.

### 3.1 시스템 개요



<그림 1> 유해 사이트 차단 시스템

<그림 1>는 내용 분석에 의해 청소년 유해 사이트를 차단하는 시스템의 구현을 위한 시스템 구성도로서

개요를 설명하면 시스템을 처음 실행해서 인터넷을 접속하면 기본 홈페이지가 접속되고 접속 주소가 파일에 저장된다. 저장된 주소의 소스를 추출해서 소스로부터 태그를 제거하고 텍스트로 저장한다. 저장된 텍스트를 형태소 분석을 한 수 키워드 추출 규칙을 이용하여 키워드를 추출하고 유해 단어사전을 이용하여 유해 명사를 추출한다. 추출된 키워드의 중요도가 가장 높은 상위 1의 중요도[1]와 유해 단어의 가중치를 계산하여 유해도를 추출하여 유해성 여부를 판단하고 유해 정보 사이트인 경우에 유해 정보 사이트임을 공지하고 유해 정보 사이트의 내용을 읽을 수 없도록 접근을 막았으며 확인 버튼을 누르면 일반 사이트로 전환되도록 구성되어 있다. 다음 절부터는 시스템 구성도 중에서 키워드의 추출 규칙과 유해명사 추출방법에 대해서 설명한다.

### 3.2 키워드 추출 규칙

인간이 문서 중에 떨어진 문자열을 합성하고, 보다 더 문장의 뜻에 따른 키워드를 추출하는 점에 주목하여 단어 정보(표기, 품사)[1]와 구문 정보를 이용한 규칙 베이스의 의존 관계에 기초한 복합어 키워드 추출 방법을 제안한다.

#### 3.2.1 의존 관계에 기초하는 복합명사 생성 규칙

[10]은 의존 규칙을 구축하여 고정밀도의 복합명사 자동 분할을 실현했다. 본 절에서는 의존 규칙을 개선하여 복합 키워드를 추출하기 위한 복합명사 생성 규칙[1]을 제안한다. 아래에 규칙의 예를 보여 준다.

[규칙1]: x(보통 명사+) 을 또는 를 y(서술형 명사) → xy

예: 음성을 인식하다 → 음성 인식

[규칙2]: x(보통 명사+) 의 y(보통 명사+) → xy

예: 정보의 검색은 → 정보 검색

[규칙3]: x(보통 명사+) 의 y(보통 명사+) 및 z

(보통 명사+) → xy, xz

예: 음성의 인식 및 합성은 → 음성 인식

예: 음성의 인식 및 합성은 → 음성 합성

여기에서 기호 x(a+), y(b+)는 품사 a, b가 1회 이상 연속적으로 구성되는 단어 x, y를 의미하

고, 왼쪽의 품사패턴과 적합한 문중의 형태소에서 오른쪽의 복합명사 xy를 생성하는 것을 표현한다. 또 동일 장소에 대한 규칙의 적용은 1회만 한다.

#### 3.2.2 의존 규칙의 적용에 대한 제약

문장의 뜻에 맞지 않는 키워드의 추출을 억제하기 위해 생성 규칙에 대한 제약을 이하에 정의한다.

##### 1) 구문에 대한 제약

생성 규칙은 구문에서 애매성이 있을 때 의미적으로 올바른 복합명사를 생성할 수 없는 경우가 있다. 예를 들면 “남성과 여성을 비교하다”는 “비교하다”가 남성과 여성을 비교하는 것이 아니라 규칙1에 의해 여성을 비교하기 때문에 의미적 부적절한 복합명사 “여성 비교”를 생성한다. 그래서 규칙1에 대해서는 규칙을 적용하는 형태소의 직전에 격조사 “와, 과”가 존재하는 경우 키워드를 추출할 수 없다는 제약이 생겨나게 된다. 그밖에 수식 등 의존 관계가 애매성을 생기는 구문에 대해 규칙을 적용하여 규칙의 직전 직후에 존재하는 형태소의 품사에 주목하고 키워드의 추출을 제한한다. 이하에 제약의 예를 보여 준다.

[규칙1의 제약]: 앞에 격조사 “와, 과”가 있을 경우 추출 안 함

[규칙2의 제약]: 앞에 격조사 “와, 과”, “의”, 또는 뒤에 격조사 “와, 과”, “의”가 있을 경우 추출 안 함

[규칙3의 제약]: 뒤에 격조사 “의”가 있을 경우 추출 안 함

##### 2) 서술형 명사에 의한 제약

[11]도 주목하고 있는 것처럼 인간은 복합명사를 이해할 때 복합명사의 구성 어간에 생략된 조사를 보충하면서 서술형 명사로 구문을 구성하여 전체의 의미를 파악하고 있다고 생각할 수 있다. 또 복합명사 중에 알 수 없는 단어가 존재하는 경우에서도 서술형 명사가 포함되어 있다면 알 수 없는 단어의 의미를 추측할 수 있기 때문에 복합명사를 대강 파악하는 것이 가능하다.

이상에 의해, 요약 키워드는 동사 특히 서술형 명사를 포함한 단어만 추출한다.

### 3.3 유해 명사 추출

1) 규칙의 기술에는 [10]이 제안한 다독성 규칙과 조합 엔진을 확장하여 이용했다.

정보통신 윤리위원회의 등급 판정 기준을 참고해 각 등급에 해당하는 사이트에서 사용되는 유해 단어들을 조사해 유해 단어 데이터베이스를 구축하였다. 특정 단어들에 대해서 아래에 제시한 표1을 기준으로 가중치를 부여했다.

다음은 유해성 여부를 판단하는 근거자료로 사용한 <정보통신 윤리위원회 SafeNet 등급기준 (<http://www.safenet.ne.kr>)>을 나타낸 표이다.

<표 1> 정보통신윤리위원회 SafeNet 등급기준

	노출	성행위	폭력	언어
4등급	성기노출	성범죄 또는 노골적인 성행위	잔인한 살해	노골적이고 외설적인 언어
3등급	전신노출	노골적이지 않은 성행위	살해	심한 비속어
2등급	부분노출	착의상태의 성적접촉	상해	거친 비속어
1등급	노출복장	격려한 키스	격투	일상 비속어
0등급	노출없음	성행위 없음	폭력 없음	비속어 없음

상기 표 1의 기준을 적용해서 0등급에 해당하는 단어는 가중치를 1로 부여하고 1등급은 가중치를 2, 2등급은 가중치를 3, 3등급은 가중치를 4, 5등급은 가중치를 5로 부여했다.

### 3.4 유해도 계산 방법

#### 3.4.1 추출된 키워드에 대한 중요도 계산 방법

중요도 결정법으로써 키워드를 구성하는 단어는 정규화빈도<sup>1)</sup>의 총화를 중요도라고 하는 방법이 생각할 수 있다.

전문가법으로는 키워드 후보와 문중에 있는 단어의 관련도를 형태소의 중복 비율로 하고 관련도에 따른 가중치를 키워드 후보에게 부가하는 방법을 제안한다. 문서 중에 출현하는 단어 w의 키워드 후보 k에 대한 관련도를 R(w, k)라고 했을 때 k의 중요도S(k)를 이하의 식에서 표현한다.

$$S(k) = \sum\{N(w) \times R(w, k)\}$$

1) 단어 w의 정규화 빈도 = w의 출현빈도 / 문서 중의 전체 보통명사 및 서술형 명사의 총 빈도

$$R(w, k) = \left\{ \frac{C(w, k)}{L(w)} \right\}$$

N(w) : w의 정규화 빈도

C(w, k) : k에 대한 w의 중복 형태소 수

L(w) : w의 형태 소수

예를 들면, N(정보)=0.4, N(추출)=0.3, N(정보 검색)=0.2이라고 가정하면, 이 때 키워드 후보 「정보 추출」에 대한 중복 형태소 수는 C(정보, 정보 추출)=1; C(추출, 정보 추출)=1; C(정보 검색, 정보 추출)=1이 된다. 또 「정보 추출」에 대한 관련도 R은 R(정보, 정보 추출)=1/1=1; R(추출, 정보 추출)=1/1=1; R(정보 검색, 정보 추출)=1/2=0.5가 된다. 이상에 의해 중요도S는 아래와 같이 된다.

$$\begin{aligned} S(\text{정보 추출}) &= \{N(\text{정보}) \times R(\text{정보, 정보 추출})\} \\ &+ \{N(\text{추출}) \times R(\text{추출, 정보 추출})\} \\ &+ \{N(\text{정보 검색}) \times R(\text{정보 검색, 정보 추출})\} \\ &= 0.4 \times 1 + 0.3 \times 1 + 0.2 \times 0.5 = 0.8 \end{aligned}$$

이들 중 중요도가 가장 높은 즉 상위 1의 복합 키워드만을 추출한다.

#### 3.4.2 추출된 유해명사의 가중치를 이용한 계산 방법

유해도 가중치 평균S'(k)을 구하기 위해서 정보통신 윤리위원회 SafeNet 등급기준의 가중치를 이용하여 다음과 같이 식을 표현했다.

$$S'(k) = \frac{\sum\{N'(w) \times \alpha\}}{L'(w)}$$

N'(w) : w의 유해단어

L'(w) : w의 추출된 단어 총 개수

α : 유해단어의 가중치

#### 3.4.3 추출된 키워드와 유해명사의 유해도 계산 방법

유해 판단을 위해서 의존관계에 기초한 키워드를 추출하여 중요도S(k)가 가장 높은 키워드 상위 1에 위치한 키워드가 문장에서 제일 중요한 키워드[1]를 인용하여 상위 1로 추출된 키워드와 유해 명사를 가중치 평균S'(k)을 더하여 유해도(I)를 다음과 같은 식으로 표현했다.

$$I = S(k) + S'(k)$$

## 4. 실험 및 평가

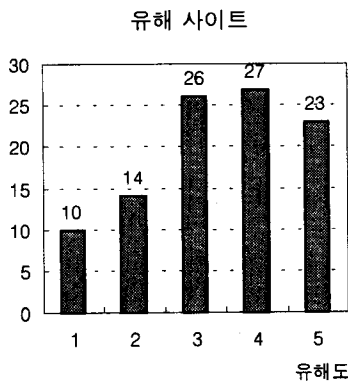
### 4.1 예비 실험

유해 사이트의 유해성 판단 기준 설정을 만들기 위해 유해 단어들의 가중치의 분포를 알아보기 위해서 유해 정보 사이트와 일반 사이트, 성교육 사이트를 대상으로 유해도의 분포를 추출하는 예비 실험을 실시했다.

#### 4.1.1 실험 데이터

실험 데이터로는 첫 번째 실험에서는 무작위로 유해 정보 사이트를 100개를 추출해서 이용했다. 두 번째 실험에서는 일반 사이트의 가중치 분포를 알아보기 위해서는 정보통신 윤리위원회에서 청소년 권장 사이트 목록을 참조해서 무작위로 100개를 추출해서 이용했다. 세 번째 실험은 성교육 사이트에 대한 가중치의 분포를 알아보기 위해 검색 사이트를 접속해 '성교육'이란 검색어로 검색해본 결과 나온 사이트를 무작위로 100개를 추출해서 이용했다.

#### 4.1.2 기준을 평가하기 위한 실험



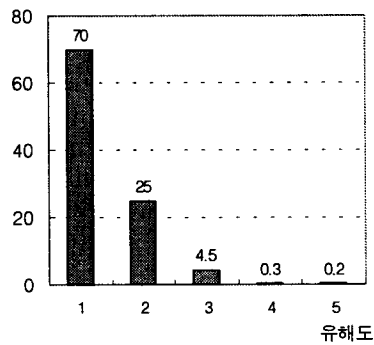
<그림 2> 유해 사이트의 유해도별 평균 분포도

<그림 2>는 유해 사이트에서 사용되는 단어들을 유해도별로 나타낸 분포도이다. 각 유해도별 분포는 유해도 1이 10%, 유해도 2가 14%, 유해도 3이 26%, 유해도 4가 27%, 유해도 5가 23%로 나타났다. 청소년들이나 성인들에게 거부감을 일으키지 않는 유해도 1과 2의 단어들은 겨우 24%를 차지하고 거부감이 일어나게 되는 유해도 3 이상

이 76%를 차지하고 있다. 모든 유해 정보 사이트들이 가중치 3 이상인 단어들이 많이 사용되고 있다는 것을 알 수 있다. 가중치 평균이 3.5 이상으로 나왔다.

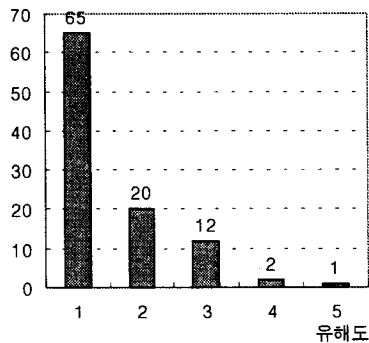
<그림 3>은 일반 사이트의 유해도 분포는 대체로 1과 2로 편중되고 있다. 그러나 유해도가 3 이상인 유해 단어들이 약 5%정도 포함되어 있는 경우가 있었는데 그것은 그 사이트 내에서 욕설이나 비방하는 글, 또는 불법 복사 CD를 판매한다는 광고성의 글, 음란 사이트를 광고하는 글, 음란 시디 판매를 홍보하는 글 등이 게시되어 있는 경우에 유해도가 4 또는 5가 나타났다. 일반 사이트에서는 청소년이나 성인이 거부감을 느낄 만한 단어들이 거의 사용되지 않고 있다. 유해도의 평균이 1.7 정도로 나왔다.

일반 사이트



<그림 3> 일반 사이트 접속시 유해도 분포

성교육 사이트



<그림 4> 성교육 사이트 접속시 유해도의 분포도

<그림 4>는 성교육 사이트의 유해도별 평균

분포도를 살펴보면 성교육 사이트는 성에 관련한 단어들 나오기는 하나 직접적인 표현보다는 의학적 신체 용어를 사용하고 있다. 유해도의 평균을 계산해보면 1.8 ~ 2.1 사이의 값을 갖는다. 교육적인 내용을 가진 경우는 비속어나 음란성이 강한 단어를 거의 사용하지 않는다는 것을 알 수 있었다.

상기의 예비 실험을 통해 경험적으로 유해 정보 사이트는 유해도 평균이 3.5 이상인 값으로 나타났고 일반 사이트는 유해도 평균이 2 미만으로 나타났다. 유해 정보 사이트임을 판별하는데 유해도 평균이 3.5 이상일 경우에 유해 정보 사이트로 판별하도록 했다.

## 4.2 평가

### 4.2.1 실험 데이터

실험을 위해서는 유해 정보 목록 데이터베이스로부터 성 관련 유해 사이트 중 자율 등급 표시를 하지 않은 사이트 100개의 주소를 추출해서 구현 시스템의 실험 데이터로 이용했고, 또한 일반 사이트 100개를 무작위로 구현 시스템의 실험 데이터로 이용했다.

#### 4.2.2 유해 사이트의 타당성 평가

5.2.1절의 실험 데이터를 이용해서 시스템의 성능을 실험해서 유해 정보 사이트로서 타당할 것인가 판단하도록 한다.

먼저 5인의 피험자에 의해서 100개의 유해 사이트가 타당한가의 여부에 대해 이하의 4단계로 평가하게 했다.

- A. 유해 사이트이다.
- B. 유해 사이트로 적절하다.
- C. 일반 사이트로 적절하다.
- D. 일반 사이트이다.

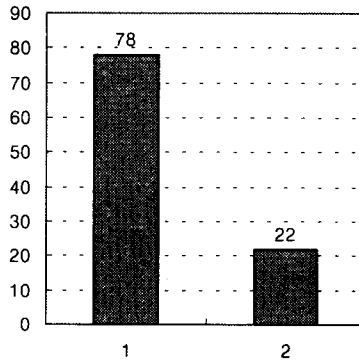
그리고 5인 전원이 A 평가를 주었던 사이트만이 유해 정보 사이트로 인정하고, 5인 전원이 D 평가를 주었던 사이트만이 일반 정보 사이트로 인정한다.

#### 4.2.3 유해 사이트에 대한 실험

<그림 5>는 유해 사이트를 무작위로 100개를 선정하여 접속 해 본 결과이다. 유해 사이트를 유해 사이트로 판별 78%, 유해 사이트를 일반사

트로 판별 22%로 나왔다. 유해 사이트를 일반 사이트로 판별한 경우에는 사이트 내에서 문자를 사용하지 않고 이미지로 처리했기 때문에 유해성이 없는 것으로 판별한 것이다.

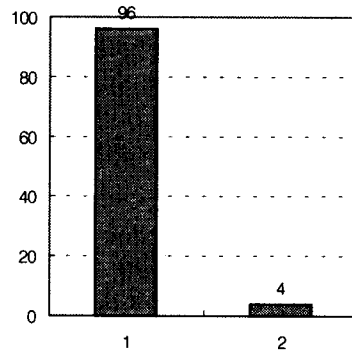
유해사이트의 실험결과



<그림 5> 유해 사이트 무작위 100개 접속 실험 결과

#### 4.2.4 일반 사이트에 대한 실험

일반사이트 실험 결과



<그림 6> 일반사이트 무작위 100개 접속 실험 결과

<그림 6>은 일반 사이트 100개를 무작위로 접속해 본 결과 일반 사이트를 유해 사이트로 인식하는 경우 4% 정도 발생했다. 일반 사이트를 유해 사이트로 인식한 경우는 일반 사이트 내에 유해 정보에 속하는 욕설이나 비방 또는 음란 사이트를 홍보하는 내용의 글이 들어있었다. 그 외의 일반 사이트로 판별한 사이트들은 청소년들이 접

속했을 경우 아무런 해를 끼치지 않는 사이트를 알 수 있었다.

## 5. 향후과제 및 결론

나날이 범람하고 있는 인터넷상의 청소년의 정신 건강을 해치는 유해 정보들로부터 청소년을 보호하기 위해 웹 문서 속에 사용된 단어들의 유해성을 판별하여 유해 사이트를 차단시켜 제공되지 않도록 하는 위해 내용분석을 먼저 실시하고 내용에 따른 중요도가 가장 높은 복합 키워드를 추출하여 중요도를 계산하고 또한 정보통신윤리 위원회에서 제시한 유해 단어의 가중치를 이용하여 가중치 평균을 더해 유해도를 판단하여 유해 사이트와 일반 사이트를 구별하는 시스템을 구현하였다.

유해 사이트와 일반 사이트, 성교육 사이트의 경우로 나누어 실시한 유해 정보의 기준을 정하는 예비 실험을 통해 유해 정보 사이트에서 유해도를 구했다. 예비 실험을 통해 구해진 값 3.5을 유해 정보 사이트를 판단하는 기준으로 정했다. 유해 정보 차단 시스템의 성능 실험을 위해 유해 정보 사이트와 일반 사이트를 각각 무작위로 100개씩 추출해 접속해 본 결과 유해 정보 사이트로 판명한 비율이 78%를 보였고 일반 사이트의 경우는 96% 정도로 나타나 대체적으로 바람직한 결과를 얻을 수 있었다. 그러나 유해 사이트 중 이미지만으로 구성된 사이트인 경우 일반 사이트로 오인되는 경우가 발생했고, 일반 사이트 중에는 욕설·비방·음란사이트 홍보의 글로 인해서 유해 사이트로 판별되는 경우가 발생하였다. 이미지로 구성된 유해 사이트에 대한 정확한 분석을 위해 이미지 분석기가 필요하겠다.

향후 연구과제로써는 텍스트 위주의 유해 정보 사이트를 차단하는 시스템을 음란 이미지를 서비스하는 사이트에 대해 유해 여부를 판단할 수 있도록 성능 향상을 시킬 것이다. 추가로 익스플로러나 넷스케이프와 같은 웹 브라우저 상에서의 구현과 한글 유해 정보 사이트 이외에 영문이나 프랑스어, 독일어를 사용한 유해 정보 사이트에 대해서 유해성 여부를 판단해 차단할 수 있도록 시스템의 성능을 향상시킬 것이다.

## 참고문헌

- [1] Taehun . et al.: "A Method of Deriving Keywords Using Generation Rules of Compound Word", The Information and Systems Society, The Institute of Electronics, Information and Communication Engineers(Japan) pp.812-821, Apr. 2001.
- [2] Taehun L. et al.: "키워드 구성의 분석과 그의 응용", 정보처리학회 자연언어처리연구회, 99-NL-133, pp.99-106, 1999(in Japan).
- [3] 이태현, 정규철, 박기홍, "의존관계에 기초한 일본어 키워드 추출방법", 한국정보처리 논문지 제10-B권 제1호. pp.41-46, Feb. 2003.
- [4] 이태현, 이진관, 정규철, 박기홍, "의존 관계에 근거한 키워드 추출 방법", 한국 정보처리학회 추계 학술 대회 논문집, pp.95-98, 2001. 10.
- [5] Morohashi, M. et al.: "자동 색인 첨가 연구의 동향", 정보처리학회, Vol.25, No.9, pp.918-925, Sep. 1984(in Japan).
- [6] Tokunaga, T. et al.: "정보 검색과 언어 처리", 동경대학 출판회, 동경, 1999(in Japan).
- [7] 류광재, "침입탐지에 의한 실시간 인터넷 해킹방지 연구", 명지대학교 컴퓨터공학부, 석사학위 논문, 1996.
- [8] 김현중, "HTML을 지원하는 라이브러리를 이용한 웹 문서 생성 시스템의 설계 및 구현", 이화여자대학교 대학원 전자계산학과, 석사학위 논문, 1996.
- [9] 정희, "유해 정보 차단을 위한 검색 시스템의 설계와 구현", 창원대학교 대학원 전자계산학과, 석사학위 논문, 1999.
- [10] Andou K. et al.: "일본어 정형 표현의 패턴 기술 규칙과 효율적인 조합 알고리즘", 전자정보 통신학회 논문지, Vol.J80-DII, No.7, pp.1860-1869, 1997(in Japan).
- [11] Miyazaki M. et al.: "의존해석을 이용한 복합어의 자동 분할", 정보처리학회 논문지, Vol.25, No.6, pp.970-979, 1984(in Japan).



### 정 규 철



1995 군산대학교 컴퓨터과학과(학사)  
1999 군산대학교 컴퓨터과학과(석사)  
2001~현재 군산대학교 컴퓨터과학과  
박사과정  
2003~현재 정보통신연구진흥원 IT  
교수요원

※관심분야 : 자연어처리, 정보검색, 음성인식, 유  
해 차단, 텔레메틱스

E-mail: kcjung@kunsan.ac.kr

### 박 기 흥



1982 숭실대학교 전자계산학과(학사)  
1986 숭실대학교 전자계산학과(석사)  
1995 일본 토쿠시마대학 지능정보과  
학과(박사)  
1997~1998 영국 Middlesex Univ 객  
원교수

1987~현재 군산대학교 컴퓨터과학과 교수  
2004~현재 NURI사업 텔레메틱스 인력양성 사업  
단(군산대) 단장

※관심분야 : 자연어처리, 정보검색, 소프트웨어통  
합, 텔레메틱스

E-mail: spacepark@kunsan.ac.kr

### 이 진 관



1995 군산대학교 컴퓨터과학과(학사)  
1999 군산대학교 컴퓨터과학과(석사)  
2002~현재 군산대학교 컴퓨터과학과  
박사과정  
2002~현재 군산대학교 정보전산원 조교

※관심분야 : 자연어처리, 정보검색, 음성인식, 유  
해 차단, 텔레메틱스

E-mail: leejinkwan@kunsan.ac.kr

### 이 태 현



1993 군산대학교 컴퓨터과학과(학사)  
1998 군산대학교 컴퓨터과학과(석사)  
2001 일본 토쿠시마대학 지능정보과학  
과(박사)  
2003~현재 정보통신연구진흥원 IT교수요원

※관심분야 : 지식 사전검색, 지적 문서관리 및 검  
색기술, 스트링 패턴 매칭, 정보검색 및 추출의 기초  
연구 및 응용, 자연언어처리의 기초해석(형태소, 구  
문해석 등)과 응용, 텔레메틱스

E-mail: thlee@kunsan.ac.kr