웹 문서 분석에 근거한 유해 웹 문서 검출

김 광 현[†] · 최 정 미[††] · 이 준 호[†††]

요 약

인터넷에 공개된 수 많은 웹 문서들에는 유익한 정보를 제공하는 웹 문서들뿐만 아니라, 음란 정보와 관련된 불건전한 유해 웹 문서들이 다수 포함되어 있으며, 본 연구에서는 이러한 유해 웹 문서들을 효과적으로 검출할 수 있는 방법을 제안한다. 즉, 유해 웹 문서들의 분석을 통하여 유해 웹 문서 선정을 위한 평가 항목들을 도출하고, 각 평가 항목별 유해 점수 부여를 위한 평가 기준을 제시한다. 그리고, 유해 점수들의 총합이 임계값 이상인 웹 문서를 유해 웹 문서로 검출한다. 본 연구의 결과는 유해 웹 문서들로부터 이용자를 보호하고 인터넷 사용의 안전성을 향상시키는데 기여할 것으로 기대된다.

키워드 : 정보 검색, 필터링, 유해 웹 문서

Detecting Harmful Web Documents Based on Web Document Analyses

Kwang Hyun Kim[†] · Joung Mi Choi[††] · Joon Ho Lee[†††]

ABSTRACT

A huge amount of web documents, which are published on the Internet, provide to users not only helpful information but also harmful information such as pornography. In this paper we propose a method to detect the harmful web documents effectively. We first analyze harmful web documents, and extract factors to determine whether a given web document is harmful. Detail criteria are also described to assign a harmfulness score to each factor. Then the harmfulness score of a web document is computed by adding the harmfulness scores of all factors. If the harmfulness score of a web document is greater than a given threshold, the web document is detected as harmful. It is expected that this study could contribute to the protection of users from harmful web documents on the Internet.

Key Words : Information Retrieval, Filtering, Harmful Web Document

1. 서 론

인터넷의 사용과 보급이 급격히 증가함에 따라 수 많은 정보들이 웹 문서의 형태로 공개되고 있으며, 이용자들은 웹 문서들을 통해서 원하는 정보를 쉽게 획득할 수 있게 되었다. 그러나, 이러한 웹 문서들 중에는 음란 정보와 관련된 문서들이 다수 포함되어 있으며, 현재에도 지속적으로 증가하고 있다[1]. 이러한 유해 웹 문서들의 무분별한 공개는 어린이 또는 청소년들의 정서에 악영향을 주기 때문에, 정부, 기관, 기업 등에서는 유해 웹 문서로부터 이용자들을 보호하기 위해 많은 노력을 기울이고 있다[2, 3].

일반적으로 인터넷 이용자들은 원하는 정보의 발견을 위하여 검색 포탈들이 제공하는 검색 서비스를 이용하고 있기 때문에, 유해 웹 문서를 포함한 웹 검색 결과는 이용자들이 유해 웹 문서에 접근할 수 있는 경로로 활용될 수 있다[4]. 따라서 인터넷 검색 포탈들은 웹 로봇을 이용하여 수집된 웹 문서들로부터 유해 웹 문서들을 제거하기 위하여 노력하고 있으며[5], 또한 웹 검색 결과에 포함된 유해 웹 문서들의 화면 노출을 방지할 수 있는 검색 환경 설정 기능을 제공한다[6, 7].

유해 웹 문서들을 검출하기 위한 대표적인 방법은 과거에 발견된 유해 사이트들의 목록을 작성하고 그 목록에 포함된 사이트들의 웹 문서 모두를 유해 웹 문서로 간주한다[8, 9, 10, 11]. 그러나 이 방법은 급격히 증가하는 신생 유해 사이트들이 목록에 포함되지 않을 가능성이 높으며, 또한 비유해 사이트에 존재할 수 있는 유해 웹 문서들에 대한 검출도 불가능하다. 이러한 문제점을 해결하기 위해서는 각각의 웹 문서에 대한 유해성 여부의 결정이 요구되나[3, 12], 지금까지 이에 대한 연구는 매우 미흡한 실정이다[13].

본 연구에서는 웹 문서 내용 및 특성의 분석에 근거하여 유해 웹 문서를 효과적으로 검출하는 방법을 제안한다. 이를 위해 유해 웹 문서들과 비유해 웹 문서들의 특성을 비교 분석하여 유해 웹 문서 선정을 위한 평가 항목들을 도출하고, 각 평가 항목별 유해 점수를 부여하기 위한 평가 기준

† 준 회 원 : 숭실대학교 대학원 컴퓨터학과 박사과정
†† 정 회 원 : 숭실대학교 대학원 컴퓨터학과 석사
††† 종신회원 : 숭실대학교 컴퓨터학부 부교수
논문접수 : 2005년 6월 20일, 심사완료 : 2005년 9월 29일

을 제시한다. 그리고 유해 점수들의 총합이 임계값 이상인 웹 문서를 유해 웹 문서로 검출한다. 또한 본 연구에서는 국내 웹 검색 포탈인 네이버에서 사용하고 있는 유해 웹 문서 검출 방법과 본 연구에서 제안한 방법과의 성능을 비교 평가한다.

본 논문의 구성은 다음과 같다. 2장에서는 유해 웹 문서들을 분석한다. 3장에서는 유해 웹 문서 선정을 위한 평가 항목들과 각 평가 항목별 평가 기준을 기술하고, 이를 이용한 유해 웹 문서 검출 방법을 제안한다. 그리고 4장에서는 제안된 유해 웹 문서 검출 방법에 대한 성능을 평가하며, 마지막으로 5장에서 결론을 맺는다.

2. 유해 웹 문서 분석

인터넷 이용자들은 웹 브라우저 주소창에 URL(Uniform Resource Locator)을 입력함으로써 원하는 웹 문서에 접근할 수 있으며, 일반적으로 웹 문서들은 HTML로 작성된다. 이러한 웹 문서는 머리글(head)과 본문(body)으로 구성되며, 머리글에는 제목 또는 메타 정보, 그리고 본문에는 내용이 기술된다. 또한 본문에는 다른 웹 문서로 이동할 수 있는 링크를 삽입할 수 있다. 이들 중 메타 정보는 웹 브라우저에 노출되지 않을 지라도, 웹 검색 시스템은 메타 정보로부터 색인어들을 추출하여 검색에 활용한다[14].

인터넷에 공개된 웹 문서들로부터 유해 웹 문서를 검출하기 위해서 유해 웹 문서 선정을 위한 평가 기준이 필요하며, 이러한 평가 기준의 선정을 위해서 유해 웹 문서에 대한 분석이 선행되어야 한다. 따라서 본 연구에서는 2004년 2월부터 3월까지 두 달 동안 수집된 26,927건의 유해 웹 문서들과 25,721건의 비유해 웹 문서들의 특성을 비교 분석하였으며, 다음에서는 그 결과로서 얻어진 유해 웹 문서들의 특성에 대하여 기술한다.

2.1 유해 단어

일반적으로 유해 웹 문서의 제목, 메타 정보, 본문 내용은 다수의 유해 단어를 포함하고 있다. 또한 유해 웹 문서 작성자는 유해 웹 문서가 검색 결과에 노출되는 수를 증가시키기 위하여 의도적으로 본문 내용과 관련 없는 인기 검색어 등의 비유해 단어들을 추가하기도 한다. 그리고 이러한 유해 단어 및 비유해 단어들은 유해 웹 문서에 반복적으로 출현하는 경향이 있다[15, 16]. <표 1>은 유해 웹 문서에 자주 출현하는 단어들의 유형을 상세히 보여준다.

본 연구에서는 <표 1>에 기술된 유형의 단어들 중에서 유해 웹 문서에 빈번히 출현하는 약 11,000개의 단어들로 유해 단어 사전을 구축하였으며, 본 논문의 이후에서 유해 단어는 이러한 유해 단어 사전에 수록된 단어를 의미한다. <표 2>는 유해 웹 문서 집합과 비유해 웹 문서 집합 내에서 제목, 메타 정보, 본문 내용에 유해 단어를 15% 이상 포함하고 있는 웹 문서들의 수를 보여준다. 이 표로부터 유해 웹 문서는 비유해 웹 문서보다 많은 수의 유해 단어들을 포함하고 있음을 알 수 있다.

〈표 1〉 유해 웹 문서에 자주 출현하는 단어 유형

구 분		설 명
유해 단어		음란, 폭력 등 유해한 의미로 사용되는 단어, 예: "섹스", "포르노", "야동", "야설", "몰카"
비유해 단어	인기 검색어	유명사이트, 연예인 이름 등 검색 횟수가 높은 단어 예: "싸이월드", "네이버", "이효리", "전지현"
	영문 모드 한글 입력	유해 단어, 인기 검색어 등을 영문 모드에서 입력한 단어 예: "tprtm(섹스)", "ditjf(야설)", "dlgyfl(이효리)"
	오타	검색 횟수가 높은 입력 오류 단어 예: "댜음", , "넹이버", "얗후", "사이월드"

〈표 2〉 유해 단어 비율 조사

구분	유해 웹 문서		비유해 웹 문서	
	문서 수	비율(%)	문서 수	비율(%)
제목	24,209	89.91	854	3.32
메타 정보	20,003	81.71	4	0.02
본문 내용	24,156	89.71	2	0.01

한편, 많은 경우에 유해 웹 문서를 지시하는 URL은 "sex", "porno", "molka" 등과 같은 유해 단어를 포함하고 있다. 본 연구에서는 1개 이상의 유해 단어가 포함된 URL을 유해 URL이라 정의하고, 유해 URL에 의해 지시되는 웹 문서들의 수를 조사하였다. 그 결과는 유해 웹 문서 집합에서 7,918개(29.41%)의 문서들이 유해 URL에 의해 지시되고, 비유해 웹 문서 집합에서 171개(0.66%)의 문서들이 유해 URL에 의해 지시되었음을 보여준다.

2.2 사전 순서 나열

(그림 1)은 질의 "쌀보리"의 검색 결과로서 브라우저에 노출된 유해 웹 문서를 보여주며, 이 유해 웹 문서에는 "쌀보", "쌀보리", "쌀부", "쌀부대" 등의 본문 내용과 관련 없는 단어들이 사전 순서로 나열되어 있음을 알 수 있다. <표 3>은 유해 웹 문서 집합과 비유해 웹 문서 집합 내에서 제목, 메타 정보, 본문 내용에 5개 이상의 단어들이 사전 순서로 나열된 웹 문서들의 수를 보여준다. 이 표로부터 유해 웹 문서, 특히 메타 정보에 다수의 단어들이 사전 순서로 나열되어 있음을 알 수 있다.

anftotkwlsdPtnfdusrnghl
... 매용 매용제 매우 매우고단한 매우기 매우외롭소 매욱 매욱스럽다 매욱하다 매운 쌀랑쌀랑 쌀랑쌀랑하다 쌀랑하다 쌀래 쌀래쌀래 쌀말 쌀목 쌀목탁 쌀무 쌀무리 쌀미 쌀밥 쌀벌 쌀벌레 쌀보 **쌀보리** 쌀부 쌀부대 쌀북 쌀새 ...

(그림 1) 사전 순서 나열

〈표 3〉 사전 순서 나열 조사

구분	유해 웹 문서		비유해 웹 문서	
	문서 수	비율(%)	문서 수	비율(%)
제목	437	1.62	0	0.00
메타 정보	1,068	4.03	13	0.05
본문 내용	2,541	9.44	976	3.79

vid vie vif vig vih
... dpfhvhxhroffjfl dprtmwlsanfy dpavktm durhtod durhtodqhwl durhtodtkwls duqodn duqltj dkqdls dkqdns dkqdus dkqeh dkqehdfl dkqek dkqfb dkqfhrfl dkqfud dkqfur dkqgud dkqh dkqj dkqja dkqk dkqkd dkqks **dkqksxp** dkqkxk ...

(그림 2) 영문 모드 한글 입력

〈표 4〉 영문 모드 한글 입력 조사

구분	유해 웹 문서		비유해 웹 문서	
	문서 수	비율(%)	문서 수	비율(%)
제목	1,001	3.71	1	0.00
메타 정보	1,487	5.52	9	0.03
본문 내용	4,347	16.14	10	0.04

2.3 영문 모드 한글 입력

검색창에 질의를 입력할 때, 웹 검색 이용자들은 영문 입력 모드로 설정된 키보드에서 한글을 입력하는 오류를 종종 범한다. 이러한 질의에 대한 검색 결과에 유해 웹 문서를 노출시키기 위하여 유해 웹 문서 작성자는 웹 문서의 머리말이나 본문 내용에 영문 입력 모드에서 입력된 한글 단어들을 나열하기도 한다. (그림 2)는 "아반테"를 영문 입력 모드에서 입력한 질의 "dkqksxp"의 검색 결과로서 브라우저에 노출된 유해 웹 문서를 보여 주며, 이 유해 웹 문서에는 "dkqksxp(아반테)", "dkqks(아반)", "dkqkxk(아바타)" 등의 영어 입력 모드에서 입력된 한글 단어들이 나열되어 있음을 알 수 있다. <표 4>는 유해 웹 문서 집합과 비유해 웹 문서 집합 내에서 제목, 메타 정보, 본문 내용에 영어 입력 모드에서 입력된 한글 단어가 3개 이상 포함된 웹 문서들의 수를 보여준다.

2.4 머리글 길이

2.2절과 2.3절에서 설명된 바와 같이, 일부 유해 웹 문서들은 사전 순서로 나열된 단어들과 영문 입력 모드에서 입력된 한글 단어들을 포함하고 있으며, 이는 유해 웹 문서의 머리글 길이가 비유해 웹 문서보다 길어지는 요인이 될 수 있다. <표 5>는 유해 웹 문서들과 비유해 웹 문서들의 제목 및 메타 정보 길이에 대한 통계를 보여준다. 이 표로부터 유해 웹 문서의 머리글, 특히 메타 정보에 더욱 많은 단어들이 나열되어 있음을 알 수 있다.

〈표 5〉 머리글 길이

(단위: 바이트)

구분	유해 웹 문서			비유해 웹 문서		
	최소	최대	평균	최소	최대	평균
제목	0	10,357	98	0	12,218	30
메타 정보	2	37,345	1,632	1	14,445	82

2.5 유해 URL

다수의 유해 웹 문서는 다른 유해 웹 문서로 이동할 수 있는 링크들을 많이 포함하고 있으며, 또한 리다이렉션 기능에 의해 다른 유해 웹 문서로 자동 이동되기도 한다. 유해 웹 문서들과 비유해 웹 문서들을 분석한 결과, 3개 이상의 유해 URL들을 링크한 유해 웹 문서 및 비유해 웹 문서들의 수는 각각 5,672개(21.06%)와 105개(0.41%) 이었다. 또한, 1개 이상의 유해 URL로 리다이렉션하는 유해 웹 문서 및 비유해 웹 문서들의 수는 각각 6,308개(23.43%)와 523개(2.03%)로 조사되었다. 이는 유해 웹 문서들이 비유해 웹 문서보다 많은 수의 유해 URL들을 링크하며, 또한 많은 수의 유해 URL로 리다이렉션함을 의미한다.

2.6 동일 사이트 링크

현재 다수의 웹 검색 서비스들은 검색된 문서들의 순위를 결정하는 요소들 중의 하나로서 페이지 순위(pagerank)를 사용하고 있으며, 이러한 웹 검색 서비스에서는 인링크(inlink)가 많은 웹 문서가 검색 결과의 상위에 노출되는 경향이 있다[17, 18]. 이러한 특성을 이용하기 위하여 유해 사이트 운영자는 동일 사이트 내의 유해 웹 문서들을 지시하는 링크들을 유해 웹 문서에 가능한 많이 포함시키기도 한다. 본 연구에서는 동일 사이트 내의 웹 문서를 지시하는 링크가 100개 이상인 웹 문서들의 수를 살펴보았으며, 그 결과 유해 웹 문서 및 비유해 웹 문서들의 수는 각각 3,342개(12.41%)와 353개(1.38%)로 조사되었다.

2.7 은닉 유해 단어

은닉 유해 단어는 유해 웹 문서의 HTML 소스에는 포함되어 있으나, 웹 브라우저에 노출되지 않는 유해 단어를 의미하며, 이러한 은닉 유해 단어는 HTML의 색 지정 또는 글자 크기 지정 기능을 이용하여 구현될 수 있다. 본 연구에서는 유해 웹 문서 집합과 비유해 웹 문서 집합 내에서 1개 이상의 은닉 유해 단어가 포함된 웹 문서들의 수를 조사하였다. 그 결과 17,669개(65.62%)의 유해 웹 문서들과 851개(3.30%)의 비유해 웹 문서들에서 1개 이상의 은닉 유해 단어들이 발견되었다.

3. 유해 웹 문서 검출 방법

본 장에서는 2장에서 기술된 유해 웹 문서들의 분석 결과에 기초하여 유해 웹 문서를 효과적으로 검출하기 위한 평가 항목으로서 유해 단어, 사전 순서 나열, 영문 모드 한글 입력, 머리글 길이, 유해 URL, 동일 사이트 링크, 은닉 유해 단어를 선정하였다. 또한 각 평가 항목별 유해 점수와 이러한 유해 점수를 부여하기 위한 평가 기준을 <표 6>에서 제시하였다. 이 표는 제목, 메타 정보, 본문 내용에 포함된 유해 단어들의 수가 유해 웹 문서 검출에 중요한 요소임을 보여준다. 본 연구에서는 <표 6>을 기반으로 웹 문서에 평가 항목별 유해 점수를 부여하고, 이러한 유해 점수들의 총합이 임계값 이상인 웹 문서를 유해 웹 문서로 검출하였다.

〈표 6〉 평가 항목 및 평가 기준

평가 항목		평가 기준	유해 점수
유해 단어	URL	1개 이상	1
	제목	15% 이상 또는 5개 이상	2
	메타 정보	15% 이상 또는 10개 이상	2
	본문 내용	15% 이상 또는 50개 이상	2
사전 순서 나열	제목	단어 5개 이상	1
	메타 정보	단어 5개 이상	1
	본문 내용	단어 5개 이상	1
영문 모드 한글 입력	제목	단어 3개 이상	1
	메타 정보	단어 3개 이상	1
	본문 내용	단어 3개 이상	1
머리글 길이	제목	100 바이트 이상	1
	메타 정보	200 바이트 이상	1
유해 URL	링크	3개 이상	1
	리다이렉션	1개 이상	1
동일 사이트 링크		100개 이상	1
은닉 유해 단어		1개 이상	1

4. 성능 평가

제안된 유해 웹 문서 검출 방법의 성능을 평가하기 위하여 2004년 5월부터 6월까지 두 달 동안 각 1만 건의 유해 웹 문서 집합 3개와 비유해 웹 문서 집합 3개를 수집하였다. 이때 총 6만개의 웹 문서들이 특정 유형으로 편중되는 것을 방지하기 위하여, 국내 웹 사이트들 중에서 6만개의 사이트들을 무작위로 추출한 후, 각 사이트로부터 하나의 웹 문서를 무작위로 선택하였다.

<표 7>은 유해 및 비유해 웹 문서 집합 내에서 웹 문서들에 부여된 유해 점수 총합의 분포를 보여준다. 이 표로부터 유해 점수 총합이 9 이상인 비유해 문서는 존재하지 않으며, 유해 점수 총합이 0인 유해 웹 문서가 존재함을 알 수 있다. 또한 유해 점수 총합이 높아짐에 따라 비유해 웹 문서들의 수는 급격히 감소하고, 유해 웹 문서들의 수는 증가함을 알 수 있다.

제안된 유해 웹 문서 검출 방법은 유해 점수들의 총합이 임계값 이상인 웹 문서를 유해 웹 문서로 검출한다. 따라서 임계값의 결정은 제안된 방법의 성능을 결정하는 매우 중요한 요소이며, <표 8>은 다양한 임계값에 대한 유해 웹 문서들의 검출 결과를 보여 준다. 이 표로부터 임계값을 1로 결정할 경우, 제안된 방법은 약 99%의 유해 웹 문서들을 정확히 검출하나, 동시에 약 45%의 비유해 웹 문서들을 유해 웹 문서들로 오검출함을 알 수 있다. 따라서 검출되는 유해 웹 문서들의 수를 최대화하고 오검출되는 비유해 웹 문서들의 수를 최소화하는 임계값의 선정이 요구된다.

일반적으로 웹 검색 서비스 업체들은 웹 검색 결과의 질을 높이기 위하여 유해 웹 문서 검출 방법을 자체적으로 개발하여 사용하고 있으나, 이러한 방법들을 공개하지 않고 있다. 본 연구에서는 NHN㈜의 협조를 얻어 검색 포탈 사이트 네이버에서 사용하고 있는 유해 웹 문서 검출 방법을 성능 평가를 위해 구축된 6만 건의 웹 문서에 적용하였다. 네이버는 웹 문서를 수집하는 과정에서 유해 웹 문서를 검출하여 삭제하며, 매우 엄격한 기준에 의해 유해 웹 문서를 검출한다[5].

<표 9>는 임계값 2의 제안된 유해 웹 문서 검출 방법과 네이버의 유해 웹 문서 검출 방법을 유해 웹 문서들과 비유해 웹 문서들의 집합에 적용한 결과를 보여준다. 네이버 방법과 제안된 방법 모두는 98% 이상의 유해 웹 문서들을 정

〈표 7〉 웹 문서들의 유해 점수 총합의 분포

유해 점수	비유해 웹 문서 집합			유해 웹 문서 집합		
	1	2	3	1	2	3
0	5,477	5,217	5,542	30	24	41
1	3,248	3,015	3,002	121	88	125
2	923	1,247	959	772	363	594
3	268	372	372	944	835	888
4	59	100	99	953	985	945
5	20	34	21	1,601	1,790	1,674
6	5	10	2	1,146	1,353	1,187
7	0	3	3	781	864	826
8	0	2	0	1,621	1,675	1,513
9	0	0	0	1,015	1,050	1,123
10 이상	0	0	0	1,016	973	1,084
합 계	10,000	10,000	10,000	10,000	10,000	10,000

〈표 8〉 다양한 임계값에 대한 유해 웹 문서 검출 결과

임계값	비유해 웹 문서 집합			유해 웹 문서 집합		
	1	2	3	1	2	3
1	4,523	4,783	4,458	9,970	9,976	9,959
2	1,275	1,768	1,456	9,849	9,888	9,834
3	352	521	497	9,077	9,525	9,240
4	84	149	125	8,133	8,690	8,352
5	25	49	26	7,180	7,705	7,407

〈표 9〉 유해 웹 문서 검출 방법의 성능 비교

구 분		네이버 방법		제안된 방법 (임계값=2)	
		검출된 문서수	검출률 (%)	검출된 문서수	검출률 (%)
유해 웹 문서 집합	1	9,847	98.47	9,849	98.49
	2	9,872	98.72	9,888	98.88
	3	9,881	98.81	9,834	98.34
비유해 웹 문서 집합	1	1,761	17.61	1,275	12.75
	2	2,381	23.81	1,768	17.68
	3	1,972	19.72	1,456	14.56

확히 검출하였다. 그러나, 네이버 방법이 약 20%의 비유해 웹 문서들을 유해 웹 문서들로 오검출한 반면에, 제안된 방법은 약 15%의 비유해 웹 문서들을 유해 웹 문서들로 오검출하였다. 즉, 제안된 방법은 유해 웹 문서의 검출에 있어서 네이버 방법과 유사한 성능을 제공하며, 비유해 웹 문서의 오검출에 있어서 네이버 방법 보다 우수한 성능을 제공한다.

5. 결 론

인터넷에 공개된 수 많은 웹 문서들 중에는 음란 정보를 포함하는 웹 문서들이 다수 포함되어 있으며, 이러한 유해 웹 문서들의 무분별한 공개는 성인들에게 불쾌감을 일으키며, 특히 어린이나 청소년들의 정서에 악영향을 줄 수 있다. 따라서 정부, 기관, 기업 등에서는 유해 웹 문서에 대한 이용자의 접근을 차단하기 위해 많은 노력을 기울이고 있다. 그러나, 유해 웹 문서들이 매우 다양한 방법으로 작성되기 때문에, 이들에 대한 이용자들의 접근을 차단하는데 많은 어려움을 겪고 있다.

본 연구에서는 유해 웹 문서들에 대한 분석을 기반으로 이들을 효과적으로 검출하는 방법을 제안하였다. 즉, 유해 웹 문서 선정을 위한 평가 항목들을 도출하고, 각 평가 항목별 유해 점수 부여를 위한 평가 기준을 제시하였다. 그리고, 유해 점수들의 총합이 임계값 이상인 웹 문서를 유해 웹 문서로 검출하였다. 또한, 제안된 방법이 유해 웹 문서의 검출에 있어서 검색 포탈 네이버에서 사용하는 유해 웹 문서 검출 방법과 유사한 성능을 제공하며, 비유해 웹 문서의 오검출에 있어서 네이버 방법 보다 우수한 성능을 제공함을 실험을 통하여 입증하였다.

본 연구의 결과는 인터넷 관련 기업이나 기관에서 유해 웹 문서들로부터 이용자들을 보호하고 인터넷 사용의 안정성을 향상시키는데 기여할 것으로 기대되며, 향후에는 본 연구에서 제시한 유해 웹 문서들의 평가 항목들을 나이브 베이즈 모델[19], 최대 엔트로피 모델[20], 지지 벡터 기계[21] 등의 문서 분류 방법에 적용하여 보다 이론적이고 체계적인 유해 웹 문서 검출에 대한 연구를 진행할 예정이다.

참 고 문 헌

[1] 조동욱, 최병갑, 김지영, "음란 유해 사이트에 대한 현황과 신호 처리에 기반한 차단 방법의 제안", 한국정보처리학회 추계학술대회, 제10권 제2호, 2003.

[2] 정보통신윤리위원회, http://www.icec.or.kr

[3] 인터넷내용등급서비스, http://www.safenet.ne.kr

[4] 한국정보보호진흥원, "2003년 개인 인터넷 이용자의 정보화 역기능 실태 조사 보고서", 2003.

[5] 김광현, 이준호, "웹 로봇의 성능 평가를 위한 방법론", 정보처리학회논문지D, 제11-D권 제3호, 2004.

[6] Google, SafeSearch Filtering, http://www.google.com/help/customize.html#safe

[7] Yahoo, SafeSearch Filter, http://search.yahoo.com/search/preferences

[8] 육헌규, 유병전, 박명순, "페이지 그룹 검색 모델: 음란성 유해 정보 색출 시스템을 위한 인터넷 정보 검색 모델", 정보과학회 논문지, 제26권 제12호, 1999.

[9] 심재권, 김귀복, 박기홍, "유해 정보의 경향과 유해 정보 차단 소프트웨어의 문제점에 관한 연구", 한국정보과학회 가을학술발표논문집, 제27권 제2호, 2000.

[10] 김성운, 김인홍, 강현석, "유해 정보 차단을 위한 데이터 관리 에이전트들의 설계 및 구현," 한국정보처리학회 추계학술발표논문집, 제6권 제2호, 1999.

[11] 정희, 이은애, 이우선, 정성환, 하석운, "청소년 유해 사이트 검색 및 차단을 위한 검색 시스템의 설계와 구현," 한국멀티미디어학회 추계학술발표논문집, 제2권 제2호, 1999.

[12] 이은애 정명숙, 김재건, 하석운, "웹 문서의 내용등급화 알고리즘에 관한 연구," 한국정보처리학회 춘계학술발표논문집, 제6권 제1호, 1999.

[13] 이승만, 장영헌, 임정환, "형태소 분석과 Skin Color 분포의 Human Detection 알고리즘을 이용한 유해 사이트 자동 분류 시스템의 구현," 한국정보과학회 춘계학술대회, 제31권 제1호, 2004.

[14] Ricardo Baeza Yates, and Berthier Ribero Neto, Modern Information Retrieval, Addison Wesley Longman, 1999.

[15] Bill Hunt, "What, Exactly, is Search Engine Spam," http://searchenginewatch.com/searchday/article.php/ 3483601

[16] Search Engine Secrets.Net, "What is Search Engine Spam," http://www.search engine secrets.net/search_engine_spam.htm

[17] S. Brin and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine," In Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 1998.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing Order to the Web," Technical report, Stanford University Database Group, 1998.

[19] Mitchell, T. M., Machine Learning, Chapter 6: Bayesian Learing, McGraw Hill, 1997.

[20] A. Berger, S. D. Pietra, and V. D. Pietra, "A Maximum Entropy Approach to Natural Language Processing," Computational Linguistics, 1996.

[21] Joachims, T, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," European Conference on Machine Learing, 1998.

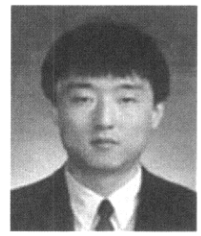
김 광 현

e-mail : iamkkh@naver.com
1999년 숭실대학교 컴퓨터학부(학사)
2001년 숭실대학교 대학원 컴퓨터학과 (석사)
2002년~현재 숭실대학교 대학원 컴퓨터학과 박사과정
관심분야 : 정보검색, 웹로봇

최 정 미

e-mail : jmichoi@naver.com
2000년 숭실대학교 컴퓨터학부(학사)
2004년 숭실대학교 대학원 컴퓨터학과 (석사)
2000년~2004년 서치솔루션㈜ 근무
관심분야 : 정보검색

이 준 호



e-mail : joonho@naver.com
1987년 서울대학교 컴퓨터공학과(학사)
1989년 한국과학기술원 전산학과(석사)
1993년 한국과학기술원 전산학과(박사)
1993년~1994년 한국과학기술원 인공지능연구센터 연구원
1994년~1995년 코넬대학교 전산학과 방문연구원
1994년~1997년 연구개발정보센터 선임연구원
2003년~2005년 매사츄세츠대학교 전산학과 방문교수
1997년~현재 숭실대학교 컴퓨터학부 부교수
관심분야 : 정보검색