퍼지 및 지능시스템학회 논문지 2001, Vol. 11, No. 5 pp. 419-425

퍼지추론을 적용한 웹 음란문서 검출

Detection of Porno Sites on the Web using Fuzzy Inference

김병만 · 최상필* · 노순억 · 김종완**
Byeong Man Kim, Sangphil Choi, Sunok Rho and Jongwan Kim

금오공과대학교 컴퓨터공학부
*케이티프리텔 통신망관리팀
**대구대학교 컴퓨터정보공학부

요 약

본 논문은 인터넷 상에서 무수히 많은 음란 문서를 검출하는 방법을 제시한다. 제시된 방법은 정보검색 기술에 퍼지추론을 적용시킨 것이다. 먼저 음란 사이트 주소를 몇 개 선정하고 이 문서들로 부터 어휘분석과 스테밍과정을 통하여 음란 사이트를 대표하는 후보단어들을 추출한다. 추출된 후보단어가 음란문서를 대표할 중요도를 계산하기 위해, 각 후보 단어별로 용어 빈도수(TF), 문서 빈도수(DF), 휴리스틱 정보(HI)를 계산하고 이 값들을 이용하여 퍼지추론을 수행한다. 이렇게 계산된 후보용어의 중요도들이 주어진 사이트가 음란사이트인지 아닌지를 판별하는데 최종적으로 사용된다. 소규모 테스트 데이터를 갖고 실험한 결과, 본 논문에서 제시한 방법이 음란 사이트 자동 검출시 유용함을 알 수 있었다.

Abstract

A method to detect lots of porno documents on the internet is presented in this paper. The proposed method applies fuzzy inference mechanism to the conventional information retrieval techniques. First, several example sites on porno are provided by users and then candidate words representing for porno documents are extracted from these documents. In this process, lexical analysis and stemming are performed. Then, several values such as the term frequency(TF), the document frequency(DF), and the Heuristic Information(HI) is computed for each candidate word. Finally, fuzzy inference is performed with the above three values to weight candidate words. The weights of candidate words are used to determine whether a given site is sexual or not. From experiments on small test collection, the proposed method was shown useful to detect the sexual sites automatically.

Key Words : Automatic Query Construction, Keyword Extractio, Detection of Parno Sites

1. 서 론

인터넷이 발달하고 등록되는 웹 페이지의 수가 많아지면서 삶의 유용한 정보를 인터넷에서 얻을 수 있으나, 이와 반대로 유해한 정보에 노출될 기회도 많아지고 있다. 최근에 기하급수적으로 증가하고 있는 음란사이트는 사회적 파급 효과가 대단하다. 청소년들을 이러한 사이트로부터 보호하기 위해서는 이러한 사이트들에 대한 접근을 차단할 수 있는 장치가 필요하다.

미로처럼 복잡한 구조를 가지고 있는 무수한 웹 페이지 속에서 사용자들이 자신이 관심을 가지는 정보를 정확히, 쉽게, 빨리 찾기란 점점 어려워지고 있다[1]. 이에 최근에 정보를 쉽게 찾을 수 있는 검색도구들이 속속 등장하고 있으며, 대표적인 것으로 AltaVista, Yahoo, MetaCrawler, 네이버, InfoSeek, HanMir 등이 있다[2-3]. HanMir에서는 기존의 검색 외에 전화번호 검색 및 지도 검색까지도 제공하고 있다. 이러한 도구들을 이용하여 음란 사이트를 찾을 수 있으나, 음란사이트를 찾기 위한 정확한 질의를 구성하기가 어렵다는 문제점을 갖고 있다.

이에 본 논문에서는 몇 개의 예제 음란사이트를 이용하여 자동으로 다른 음란사이트를 탐지하는 시스템을 제시하였으며, 자동으로 음란사이트 검색 후 주소를 데이터베이스화하는 실험을 하였다. 제안 시스템의 특징은 기존의 텍스트 검색 기술에 퍼지추론을 적용시켜 키워드별 음란 웹 페이지를 나타낼 가능성을 계산하는 것이다. 먼저 음란 사이트 주소를 몇 개 선정하고 이 문서들로부터 어휘분석과 스테밍과정을 통하여 음란 사이트를 대표하는 후보단어들을 추출한다. 추출된 후보단어가 음란문서를 대표할 중요도를 계산하기 위해, 각 후보 단어별로 TF(Term Frequency : 용어 빈도수), DF(Document Frequency : 문서 빈도수), HI(Heuristic Information : 휴리스틱 정보)를 계산하고 이 값들을 이용하여 퍼지추론을 수행한다. 이렇게 계산된 후보용어의 중요도들이, 주어진 사이트가 음란사이트인지 아닌지를 판별하는데 최종적으로 사용된다.

접수일자 : 2001년 2월 21일
완료일자 : 2001년 8월 10일
본 연구는 한국과학재단 목적기초연구(2000-1-51200-008-2) 지원으로 수행되었음.

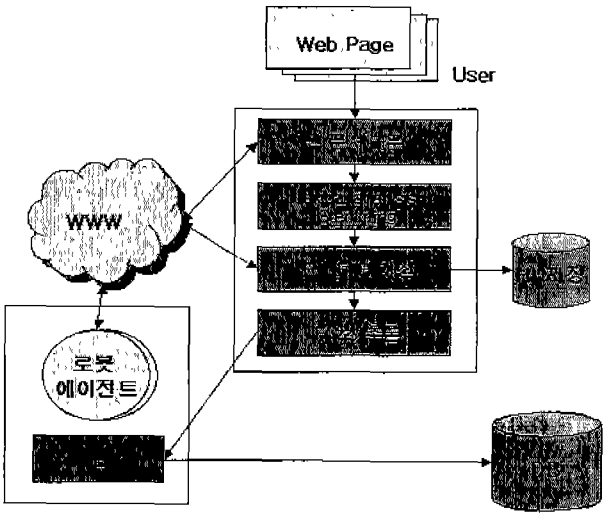
2. 제안된 시스템의 설계 및 구현

2.1 시스템 구성

기존의 검색엔진을 사용한 정보 검색은 부정확한 질의로 인하여 상당히 많은 문서가 출력되며, 출력된 문서 중에서 필요한 것을 찾기가 점점 어려워진다. 따라서, 본 논문에는 음란 사이트를 탐지하기 위하여, 몇 개의 예제 사이트로부터 음란 사이트를 대표하는 키워드와 이의 중요도를 추출하고 이러한 정보를 이용하여 음란 사이트를 보다 정확히 찾는 방법을 제시하고자 하였다. 전체적인 시스템은 그림 1과 같이 구성된다.

그림 1의 전체 시스템 구성도를 기반으로 작업 흐름을 설명하면 다음과 같다. 먼저 사용자는 음란 웹 문서의 주소를 줄 수 있으며(2.2절 참조), 이 웹 문서에 의해서 사용자 관심 정보가 추출되게 된다. 이 단계에서는 제공된 URL을 이용하여 웹에서 문서를 전송 받은 후 태그를 제거하고, 어휘분석 단계와 스테밍 단계(불용어 제거를 포함)를 거쳐 순수 단어들의 리스트를 뽑아낸다(2.3절 참조).

다음 단계로, 단어들의 리스트에서 각 단어별로 퍼지변수들인 TF, DF, HI을 계산한다(3.1절 참조). 계산된 퍼지변수 값을 퍼지추론에 적용하여 단어별로 사용자가 찾고자 하는 사이트에 얼마나 적합한지를 계산한다(3.2절 참조). 이를 기반으로 실질적으로 사용자가 원하는 웹 페이지를 탐지하는 것이다(4장 참조).



TF : Term Frequency, DF : Document Frequency,
HI : Heuristic Information

그림 1. 전체 시스템의 구성도
Fig. 1 Overall System Architecture

2.2 관심 웹 페이지 제시

사용자는 자신이 원하는 사이트를 효율적으로 찾기 위해서 기존의 검색엔진에서 사용하는 몇몇 단어나 단어들의 조합을 사용하지 않고, 관심분야와 비슷한 사이트의 주소를 질의로 제시한다. 질의되는 사이트의 수는 기본적으로 3개 이상으로 잡았다. 질의되는 사이트의 수가 3개 이하일 경우에 질의된 사이트에서 사용자가 원하는 관심정보를 자동으로 추출하기가 어렵게 되고 정확도가 높은 특징 값의 추출이 어려워진다. 이에 반해 너무 많은 사이트의 입력을 받는 것은 사용자에게 많은 부담을 줄 수 있으므로 본 논문에서는 3개에서 7개 사이의 사이트를 입력으로 받아들여 실험을 하였다.

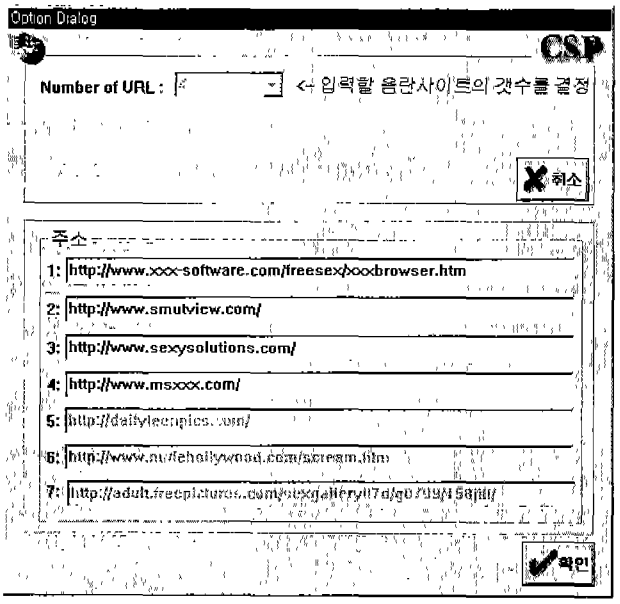


그림 2. 사용자 관심 웹 페이지 제시
Fig. 2 User presents some web pages

사용자 질의는 그림 2에서 보는 것과 같이 관심분야가 비슷한 URL로 주어진다. 그림 2에서는 4개의 음란 사이트 주소가 지정되었다. 입력이 완료되고 확인 버튼이 눌려지면 제시된 웹사이트에 접근하여 문서를 전송 받는다. 각 문서는 HTML코드로 되어있으므로 코드 내에서 태그를 제거하고 텍스트만 뽑아내는 파싱 작업이 이루어진다. 또한 인터넷에서 원하는 사이트의 웹 문서를 전송 받기 위해서는 HTTP 프로토콜을 사용하여야 하는데, Visual C++ 6.0 MFC 클래스에서 웹 문서를 쉽게 전송 받을 수 있도록 특정 URL과 연결을 시도하는 클래스인 CInternetSession과 연결된 세션을 컨트롤하는 인터넷용 파일 핸들러 CInternetFile 클래스를 사용한다[4]. 이 클래스를 사용하면 기본적으로 HTTP 프로토콜을 숙지하지 않아도 URL만으로 쉽게 웹 문서를 다운받아 사용할 수 있다.

2.3 어휘분석(Lexical analysis)과 스테밍(Stemming)

태그가 제거된 웹 문서는 어휘분석 단계를 거치면서 색인에 있어 불필요한 문자를 제거하고 문자들을 단어 또는 토큰 열로 변환하게 된다. 일반적으로 어휘분석 단계에서 불용어 처리가 되나 본 논문에서는 스테밍 과정에 포함되어 있다[5].

어휘 분석단계에서는 먼저 DFA(Deterministic Finite Automata : 결정적 유한 오토마타)를 생성하고 다음으로 DFA를 이용하여 토큰 열로부터 단어를 생성하게 된다. 생성된 단어는 검색의 효율을 높이고 색인 파일의 크기를 줄이기 위하여 스테밍 과정을 수행한다. 대표적인 스테밍 알고리즘에는 단순 스테머, Porter 스테머, Krovetz 스테머 등이 있다. 각 스테머마다 장점이 있으나 본 논문에서는 정보검색 분야에서 많이 사용하는 Porter 스테머를 사용하였다[5]. 또한 스테밍 과정 내에서 불용어 삭제가 이루어지게 된다. 불용어는 모든 문서에서 빈도수가 높으며 특징 추출시 영향을 주지 못하는 "a"나 "the" 같은 단어들이다. 불용어 제거는 불용어 사전을 구축하고, 스테밍 과정에서 각 단어가 불용어

사전에 존재하면 제거하는 방식으로 수행된다.

3. 퍼지추론을 사용한 단어별 중요도 계산

3.1 퍼지 변수 및 정규화

어휘분석과 스테밍을 마치고 생성된 단어는 사용자가 원하는 사이트를 찾을 질의를 생성하기 위하여 TF, DF, HI을 추출한다.

3.1.1 TF

TF는 단어의 빈도수로서 사용자 입력으로부터 인터넷에서 전송 받은 문서에 대해 어휘분석과 불용어 처리, 스테밍에 의해 생성된 단어 각각의 개수를 계산한 값이다. 단어의 빈도수가 클수록 이 단어가 사용자 관심 사이트에 나타날 가능성도 높게된다. 어휘 분석단계에서 불용어가 제거되었고, 이 단계에서 다시 한번 여러 문서를 통틀어 한번밖에 나오지 않는 단어는 제거된다. 이런 단어들은 계산속도를 느리게 만들고 질의 형성에 별 다른 영향을 미치지 못한다.

각 단어의 빈도 수는 퍼지 계산에서 적용되기 위해서 정규화되어야 하는데 식 (1)이 적용된다.

$$NTF_i = 1 - \frac{DF_i}{TF_i} \quad (1)$$

TFi : 예시 문서 내에서의 i번째 단어의 TF값
DFi : i번째 단어의 DF값(3.1.2절 참조)

NTFi는 i번째 단어의 정규화된 TF값을 나타낸다. 이 값은 0과 1사이의 값이 나타나며 DF값이 작을수록 또한 TF값이 커질수록 NTF값은 커지게 된다.

3.1.2 DF

DF는 각 단어가 몇 개의 문서에서 나타나는지 수를 세는 것으로, 많은 문서에서 나타날수록 사용자가 원하는 사이트를 나타낼 가능성이 커진다. 본 논문에서는 정보검색에서 용어의 가중치 계산을 위해서 자주 사용하는 IDF(Inverse Document Frequency : 역 문헌 빈도수) 대신에 문헌의 빈도수를 사용한다[7]. 왜냐하면 주어진 예제문서들이 공통된 주제에 관한 것이므로 일반 용어를 제외하고 전체 문서를 통해서 많이 출현하는 용어일수록 그 단어가 사용자가 원하는 사이트를 나타낼 가능성이 커지기 때문이다. i번째 단어에 대한 문서 빈도 수 또한 정규화되어야 한다. 정규화는 식 (2)에서 보듯이 각 단어의 문서 빈도 수를 전체 문서의 수(사용자 지정 웹사이트 수)로 나눈 값을 취한다.

$$NDF_i = \frac{DF_i}{TD} \quad (2)$$

DFi : i번째 단어의 DF
TD : 사용자 지정 웹 페이지 수

3.1.3 HI

TF와 DF를 사용하여 많이 사용되고 여러 문서에 존재하는 단어를 찾아 그 단어에 가중치를 증가시켰다. 일반 용어의 경우 많은 문서에서 높은 빈도 수를 보이는 경향이 있다. 이러한 일반적인 용어의 가중치를 낮추려면 위에서 계산된 TF와 DF 만으로는 불가능하므로 HI를 사용한다.

일반적인 용어가 웹 문서 내에 많이 나타나므로, 검색엔진에 질의로 줄 경우 많은 문서가 검색되는데 착안하여, 단어를 검색엔진에 질의할 경우 검색되는 문헌의 수 HI를 이용하여 식 (3)과 같이 정규화된 NHI_i를 계산한다.

$$NHI_i = 1 - \frac{HI_i}{AV} \quad (3)$$

HIi : i번째 단어의 HI값
AV : HI값 중 상위 10개의 평균

NHI를 구하기 위해서 먼저 각 단어별로 인터넷 검색엔진에 질의하여 검색되어지는 문헌의 수 HI를 받아온다. 여기서 사용한 검색엔진은 Altavista이며 Altavista로부터 각 단어의 검색 결과를 받아오는 부분에 많은 시간이 소요된다. 시간을 줄이기 위하여 HI의 값을 DB(데이터베이스)로 구축하게 되는데, HI의 값을 계산할 경우 먼저 DB에서 찾고, 존재하지 않은 경우 검색엔진에 질의로 주고 계산한 후, 이를 DB에 저장하게 된다. 정규화된 HI를 구하기 위해 모든 단어들에 대한 HI 가운데 최대값을 사용하면 값의 편차가 너무 커서 정규화의 의미가 없어질 수 있으므로, 상위 10개 값의 평균(AV)을 계산하고 이 값으로 HI를 나누면 각 단어의 일반 용어 정도를 알 수 있다. 따라서 (식 3)에 의해 얻어진 NHI 값이 크면 단어가 일반용어가 아니라 전문 용어이며, 작으면 일반용어가 됨을 알 수 있다.

실례로, Altavista에 "xxx"를 사용하여 연관 웹 문서를 검색한 결과를 가지고 NHI를 계산하는 방법을 설명하겠다. Altavista에 질의를 주기 위한 URL의 형태는 다음과 같다.

http://www.altavista.com/cgi-bin/query?pg=q&q=(?)&kl=XX&stype=stext

위 URL에서 (?) 부분에 엔진에 질의할 단어를 삽입하면 된다. 단어 "xxx"를 사용하여 검색된 총 웹 페이지의 수는 2992081개이다. 물론 이 값은 검색엔진이 업데이트 됨에 따라 바뀌게된다. 앞에서 예제로 제시한 사이트에서 추출한 단어들의 HI 중 상위 10개의 평균을 구하였더니 64815566 이었다. 따라서, 단어 "xxx"의 NHI는

$$NHI_{xxx} = 1 - \frac{HI_{xxx}}{AV} = 1 - \frac{2992081}{64815566} = 0.9538$$

가 된다. 짐작한 바와 같이 단어 "xxx"는 일반 용어, 즉 많은 문서에서 발생하는 단어가 아님을 알 수 있다.

표 1은 한국어 문서 집합인 KT SET 2[8]를 대상으로 단어별 HI와 NHI를 계산한 결과의 일부를 발췌한 것이다. 표 1은 단어 "Disk"나 "반도체" 등은 높은 NHI 값을 갖지만, "기술" 같은 일반 용어는 낮은 값을 가짐을 보여준다. 이 결과로부터 제안된 휴리스틱 정보 HI가 의미있는 전문용어 판별 기준임을 확인할 수 있다.

표 1. 단어별 HI 및 NHI (KT SET 2 대상)
Table 1. HI and NHI for words (KT SET 2)

단 어	HI	NHI	단 어	HI	NHI
Disk	4380	0.984	PC	170330	0.394
기술	264034	0.061	기억	66007	0.765
대한	416266	0.000	데이타	34819	0.876
디스크	28962	0.897	메모리	21585	0.923
반도체	28575	0.898	발전	141339	0.497
설계	80170	0.714	장치	70573	0.749

전원	32173	0.886	펄요	332983	0.000
486	4570	0.984	IC	8220	0.971
IC카드	1059	0.996	LCD	8420	0.970
개발	236260	0.159	계획	184772	0.343

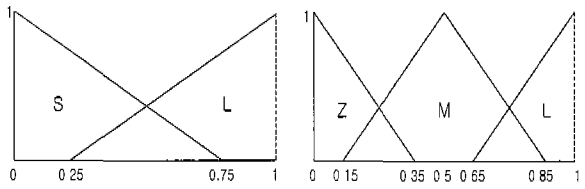
3.2 단어별 중요도 계산

본 논문의 핵심적인 부분으로서 구해진 각각의 TF, DF, HI를 대상으로 퍼지추론을 적용하여 단어별 가중치를 계산함으로써 각 단어가 사용자가 원하는 웹 페이지를 나타낼 가능성을 구하게 된다.

3.2.1 중요도 계산 과정

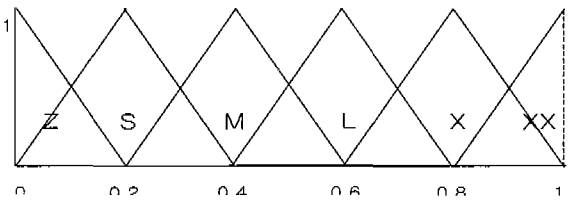
우선 구해진 NTF, NDF, NHI의 명확한 값으로 측정된 입력변수 값을 적절한 퍼지 값으로 바꾸어야 하는데 본 논문에서는 그림 3(a)와 같은 삼각형 형태의 퍼지 수를 사용하였다. NTF는 S와 L의 두 부분으로 나누었고, NDF와 NHI에 대해서는 Z, M, L 세 부분으로 나누었다. 퍼지 출력변수 SIM(Similarity : 유사도)은 그림 3(b)에서 보듯이 6개의 구간으로 나누고 각 삼각형의 밑변의 범위를 0.4 크기로 설정하였다.

퍼지 추론규칙은 표 2에서 보이고 있으며, 18개의 규칙을 만들었다. 연관 정도를 Z, S, M, L, X, XX의 6단계로 나누었고, Z의 결과를 갖는 규칙이 4개, S의 결과를 갖는 규칙이 5개, M의 결과를 갖는 규칙이 2개, L의 결과를 갖는 규칙이 3개, X의 결과를 갖는 규칙이 3개, XX의 결과를 갖는 규칙이 1개 있다.



NTF NDF, NHI

(a) 입력변수



SIM

(b) 출력변수

Z : Zero, S : Small, M : Middle, L : Large, X : X large, XX : XX large

그림 3. 퍼지 입출력 변수

Fig. 3 Fuzzy Input and Output Variables

퍼지 추론규칙 생성 과정을 살펴보면 다음과 같다. 만약 NDF가 Z(문서 빈도수가 작다)이고, NTF가 S이면 각 문서에 단어의 빈도수가 낮다. 그리고 NHI 또한 Z(휴리스틱 정보가 작다 : 일반용어일 가능성이 크다)이면 전체적인 관련 정도는 Z(관련 정도가 거의 없다)이다. NDF가 L(여러 문서에서 많이 출현)이고, NTF가 S이면 각 문서에 단어의 빈도수가 낮다. 그리고 NHI가 Z(일반용어 가능성이 큼)이면 관련정도는 S이다. NDF가 L이고 NTF가 L이면 각 문서에 걸쳐 단어의 빈도수가 높으며, NHI가 만약 Z이면 일반용어일 가능성이 크므로 관련정도가 S이고, NHI가 L이면 단어의 관련정도는 상당히 큰 것이다. 나머지 퍼지 추론규칙들도 이러한 논리로 설정되었다.

표 2. 퍼지 추론규칙

Table 2. Fuzzy Inference Rules

NTF = S

NDF \ NHI	Z	M	L
Z	Z	Z	S
M	Z	M	L
L	S	L	X

NTF = L

NDF \ NHI	Z	M	L
Z	Z	S	M
M	S	L	X
L	S	X	XX

NTF : Normalized Term Frequency,
NDF : Normalized Document Frequency,
NHI: Normalized Heuristic Information

기본적인 준비가 끝나면 이제 퍼지 규칙을 적용하여 퍼지 추론을 하게된다. 먼저 NTF, NDF, NHI의 퍼지 입력 값을 그림 3의 퍼지함수에 적용시켜서 퍼지값을 계산한 후 이 값들 중 min의 값을 취한다. 모든 규칙들에 적용하여 계산된 결과는 18개가 생성된다. 이 값들을 퍼지 출력변수인 SIM에 따라 6개 그룹으로 분류한다. 형성된 분류 각각에서 max값을 취한다. 다음으로 6개의 결과가 생성되는데 6개를 이용하여 퍼지값을 비 퍼지값으로 변환하는 과정은 가장 많이 사용되는 무게중심법을 사용하여 비퍼지화 하였다[9]. 이렇게 해서 각 단어에 대해서 구해진 SIM값이 그 단어가 사용자가 원하는 문서를 나타낼 가능성을 보인다.

표 3은 4장 실현부분에서 사용한 4개의 음란사이트 주소에서 추출한 207개의 단어중 상위 20개에 대한 중요도를 보여준다.

표 3. 4개의 음란사이트를 질의어로 주고 생성된 각 단어별 중요도 (전체 207단어 중 상위 20단어)

Table 3. Weights of higher 20 words among 207 ones extracted from the given 4 porno web sites

단 어	가능성	단 어	가능성
movie	0.96613	hardcore	0.96218
lesbian	0.95987	free	0.91360
amateur	0.89000	download	0.85000
teen	0.85000	fuck	0.85000
sex	0.85000	asian	0.85000
video	0.85000	saint	0.80116
celebrity	0.80000	lolita	0.80000
amp	0.80000	playmate	0.80000
live	0.80000	update	0.80000
webcam	0.80000	cam	0.80000

4. 실험 결과의 분석 및 고찰

본 논문은 Windows NT 기반에서 Visual C++ 6.0을 사용하여 구현되었다[10]. 먼저 음란사이트 주소 4개를 질의로 주고 여기서부터 필요한 단어별 가중치를 계산한다. 이 값을 사용하여 웹 로봇에 의해 수집된 웹 페이지를 대상으로 표 3의 단어들이 나타나는 빈도를 계산하여 해당 문서가 음란사이트를 나타내는지를 판단한다.

예시로 제시된 주소는 "http://xxx.pornports.com",
"http://www.sexcamheaven.com/",
"http://www.celebritypixx.com/freevideos",
"http://porn-movies.net/main.html"이며, 각 웹 문서에서 추출된 단어의 개수는 33, 94, 413, 424개이다. 단어 추출 시 TF가 1인 값과 TF는 2이고 DF가 1인 값을 제거한 후 계산을 하였다. 그 이유는 TF가 1이나 2인 값은 문서 전체에서 한두 번밖에 나오지 않는 별 영향을 미치지 못하는 단어이지만, 이런 단어들이 문서 내에 상당히 많기 때문이다. 그 결과 실제로 문서들에서 추출된 단어는 모두 207개이며, 이 중 상위 20개는 표 3과 같다. TF와 DF계산 후 HI를 계산하게 되는데, HI에서는 각 단어별로 인터넷 검색엔진에 질의를 주고 결과를 받아오는 과정을 수행하여야 하기 때문에 빈도수가 적으면서 영향을 미치지 못하는 단어를 삭제함으로써 수행속도를 향상시킬 수 있다.

실제로 실험할 경우에는 웹 로봇이 수집한 페이지들에서 추출된 모든 단어를 사용하여야 한다. 비교 대상이 되는 웹 문서에서 단어를 추출하고 스테밍 과정을 수행한다. 단어 각각에 대해서 먼저 추출한 값과 비교를 행한다. 비교 후 같은 단어가 존재하면 단어의 가중치(즉 표 3에서 구한 바와 같이 각 단어가 음란사이트를 나타낼 가능성)를 더하고 없으면 무시한다. 추출된 단어들에 대해서 이런 과정을 행한 후 추출된 전체 단어의 개수로 나눔으로써 유사도를 계산할 수 있다.

표 3과 같이 구해진 207개의 음란사이트를 나타내는 단어들을 벡터 V=(t_i, w_i)로 표현한다. 여기서 i는 예시 사이트로부터 추출된 207개의 단어를 나타내는 첨자이다.
또한 비교 대상이 되는 웹 페이지에서 추출한 단어를 e_j라 두면(여기서 j는 웹 페이지에서 추출한 단어를 나타내는 첨자이다), 식 (4)와 같이 웹 페이지가 음란사이트 인지를 판단하는 기준이 되는 유사도가 계산된다.

$$Similarity = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{207} W_i \quad \text{for } e_j = t_i \qquad (4)$$

t_i : 사용자 제시 문서로부터 추출한 단어들
w_i : 추출한 단어의 가중치(음란사이트를 나타낼 가능성)
e_j : 웹 페이지에서 추출한 단어들
n : 비교대상이 되는 웹 페이지에서 추출한 단어의 갯수

본 논문에서는 성능 평가를 위하여 117개의 웹 사이트로 구성된 테스트 집합을 구성하였다. 117개 중 55개는 음란 사이트이며 나머지 62개 중 41개는 음란 사이트는 아니지만 성범죄, 성교육 등과 같이 성과 관련된 사이트이다. 나머지 21개는 전혀 관련 없는 사이트들이다. 수집한 웹 페이지의 음란사이트 여부를 판별하기 위해서는 계산된 유사도를 대상으로 임계치를 주면 된다.

본 논문에서 제안한 방법의 유용성을 비교하기 위해 사용자가 직접 질의를 구성한 경우와 비교 실험하였다. 음란 문서에 나타나는 주요 단어 3개로 구성된 2 개의 질의 - "movic sex pussy", "movie fuck hardcore" - 를 실험에 사용하였다. 질의에 나타나는 용어의 중요도는 벡터 모델 기반 정보검색 분야에서 많이 사용하는 아래와 같은 공식을 사용하였다. 원래는 NHI 대신에 IDF를 사용하여야 하지만 웹 상에 존재하는 모든 문서에 대한 정확한 정보를 알기 힘들고, 또한, 본 논문에서 이러한 정보가 없는 환경을 고려하기 때문에 IDF 대신에 본 논문에서 제안한 NHI를 대신 사용하였다.

$$w_{i,q} = (0.5 + \frac{0.5 freq_{i,q}}{\max_l freq_{l,q}}) \times NHI_i$$

여기서, w_{iq}는 질의 q에서 용어 i의 중요도를, $freq_{i,q}$는 질의 q에서 용어 i 가 나타나는 횟수를 의미한다.

본 논문에서는 임계값을 변화시키면서 실험을 하였는데 그 결과는 표 4~6과 같다. 표 4와 5는 사용자 질의에 대한 결과이며, 표 6은 본 논문에서 제안한 방법에 의해 자동으로 구성된 질의를 사용한 경우의 실험 결과이다. 여기서, 대상 사이트 수는 유사도가 주어진 임계값보다 큰 사이트의 수를, M은 대상 사이트 중 비음란 사이트면서 성관련 사이트의 수를, N은 전혀 관련 없는 사이트의 수를, T는 전체 음란사이트의 개수 (=55)를 의미한다.

표 4. 임계값에 따른 실험 결과 - 첫 번째 질의
Table 4. Experiment result according to the threshold - the first query

임계값		0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05
대상사이트수(R)		16	19	22	27	32	43	49	55
음란사이트수(S)		16	19	22	27	32	37	37	39
비음란 사이트수	M	0	0	0	0	0	6	12	16
	N	0	0	0	0	0	0	0	0
정확률(S/R)		1.00	1.00	1.00	1.00	1.00	0.86	0.76	0.71
재현률(S/T)		0.29	0.35	0.40	0.49	0.58	0.67	0.67	0.71

표 5. 임계값에 따른 실험 결과 - 두 번째 질의
Table 5. Experiment result according to the threshold - the second query

임계값		0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05
대상사이트수(R)		7	11	14	20	22	30	33	35
음란사이트수(S)		7	11	14	20	22	30	33	34
비음란 사이트수	M	0	0	0	0	0	0	0	1
	N	0	0	0	0	0	0	0	0
정확률(S/R)		1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
재현률(S/T)		0.13	0.20	0.25	0.36	0.40	0.55	0.60	0.62

표 6. 임계값에 따른 실험 결과 - 예제 문서에서 구성된 질의
Table 6. Experiment result according to the threshold - the query constructed from example documents

임계값		0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05
대상사이트수(R)		31	38	44	46	52	71	89	112
음란사이트수(S)		31	37	43	45	49	52	53	53
비음란 사이트수	M	0	0	0	0	1	12	23	39
	N	0	1	1	1	2	7	13	20
정확률(S/R)		1.00	0.97	0.98	0.98	0.94	0.73	0.60	0.47
재현률(S/T)		0.56	0.67	0.78	0.82	0.89	0.95	0.96	0.96

실험 결과, 사용자 질의인 경우는 정확률은 좋지만 재현률은 좋지 않음을 알 수 있다. 정확률이 높은 것은 당연한 결과이다. 왜냐하면, 질의가 주로 음란 문서에만 나타나는 단어들로만 구성되었고 따라서 이러한 단어를 포함하는 문서는 음란문서인 가능성이 상당히 높기 때문이다. 하지만, 사용자 질의인 경우는 질의를 구성하는 단어가 음란문서뿐만 아니라 비음란문서에서도 많이 나타나는 경우에는 정확률이 떨어진다. 또한, 사용자 질의인 경우는 실험결과에서 보듯이 질의를 구성하는 단어들이 전혀 나타나지 않는 음란 문서들인 경우에는 이를 검출할 수 없다. 즉, 임계값을 0.05로 낮추더라도 70% 정도의 음란문서만 검출됨을 알 수 있다.

본 논문에서 제안한 방법인 경우는 보다 많은 음란문서를 검출함을 알 수 있다. 임계값을 0.15로 하면 95%의 음란사이트를 검출함을 알 수 있다. 하지만, 비음란 사이트도 음란 사이트로 오판하는 비율 (27%)이 큼을 알 수 있다. 임계값이 0.20인 경우 정확률과 재현률 측면에서 좋은 결과를 보임을 알 수 있다. 이 경우, 음란 사이트인데 비음란 사이트로 판별된 사이트들은 대부분 텍스트는 거의 없고 대부분 사진인 사이트이고, 다른 하나는 음란 사이트로 보기에는 애매한 사이트였다. 반대로, 3개의 비음란 사이트가 음란사이트로 판별되었는데 이러한 사이트는 주로 성생활, 성폭력, 성과 건강 등에 관련된 내용이었다. 또한, 성과 관련없는 사이트는 거의 정확하게 0.25 값을 기준으로 분류되었는데 이는 본 논문에서 제안한 방법이 음란 문서를 대표하는 단어들에 대해서 어느 정도 정확하게 중요도를 산정하였기 때문이다.

5. 결론 및 향후 연구방향

기존의 많은 검색엔진에서는 기본적인 부울 연산에서 벗어나지 못하는 서비스를 제공하므로 현대와 같이 인터넷 환경에서 사용자들이 자신이 원하는 웹 페이지를 정확하게 발견하기 어렵다. 또한 검색 결과 페이지가 너무 많아서 원하는 웹 페이지를 찾는 것은 사용자의 부담으로 존재한다.

본 논문에서는 사용자로부터 웹 페이지를 제시받아 음란 사이트 검색을 위한 효율적인 단어별 가중치를 계산하는 시스템을 제안하였다. 음란사이트 및 일반적인 사이트를 대상으로 테스트하였다. 실험 결과를 통해서 제안된 방법이 음란 사이트를 효과적으로 판별하는 것으로 입증되었다.

현재 후보용어가 일반용어인지를 판단하기 위해 사용하고 있는 단어의 휴리스틱 정보 추출 방법은 단순하며 일반용어를 제거하기에 아직 미흡하다. 앞으로, 이에 대한 추가의 연구가 필요하다. 또한, 성관련 사이트와 음란 사이트의 구별이 정확하지 않는데 이에 대한 해결책으로 이미지 검색[11] 기술을 도입하는 연구도 필요하다.

참 고 문 헌

[1] Daniel D, Adele E. H, An Information Gathering Agent for Querying Web Search Engines, Technical Report CS-96-111, Colorado State Univ., 1996.

[2] 신봉기, "인터넷 정보검색 서비스 동향," *정보과학회지* vol. 16, no. 8, pp. 16-20, 1998.

[3] Gravano, L., et al, STARTS: Stanford Proposal for Internet Meta-Searching, Proc. of SIGMOD 97, 1997.

[4] 이상엽, Visual C++ Programming Bible Ver 6, (주) 영진 출판사, 1999

[5] William B. Frakes, Ricardo Baeza-Yates, Information Retrieval : Data Structures & Algorithms, Prentice Hall, pp. 102-160, 1992.

[6] Kiduk Yang, Denqi Song, Wooseob Jeoung, Rong Tang, Nice Stemmer, INLS161 Final Project, http://ils.unc.edu/iris/irisnstem.htm.

[7] Sparck Jones, K., "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," J. Documentation, vol. 28, no. 1, pp. 11-20, 1972.

[8] 김재군외 2인, "한국어 정보 검색 연구를 위한 시험용 데이터 모음 (KTSET)," *제6회 한글 및 한국어정보처리학술대회*, 1998.

[9] 이광형, 오길록, 퍼지 이론 및 응용 II권: 응용, 개정4쇄, 홍릉과학출판사, pp. 5-1~5-91, 1997.

[10] 김용성, Visual C++ 6 완벽 가이드, (주)영진출판사, 1999

[11] S. Mehrotra, Young Rui, M. Ortega-B., and T. S. Huang, Supporting content-based queries over images in MARS", in Proc. of IEEE Int. Conf. on Multimeia Computing and Systems, 1997.

저 자 소 개

김병만 (Byeong Man Kim)

1987년 : 서울대학교 컴퓨터공학과 학사
1989년 : 한국과학기술원 전산학과 공학석사
1992년 : 한국과학기술원 전산학과 컴퓨터 공학박사
1992년~현재 : 금오공과대학교 부교수
1998년~1999년 : 미국 Univ. of California, Irvine Post Doc.

관심분야 : 인공지능, 정보검색, 소프트웨어 검증

최상필 (Sangphil Choi)

1995년 : 금오공과대학교 컴퓨터공학부 학사
2000년 : 금오공과대학교 컴퓨터공학과 공학석사
2000년~현재 : KTF 근무

관심분야 : 정보검색, IP 통신

노순억 (Sunok Rho)
1999년 : 금오공과대학교 컴퓨터공학과 학사
2000년 ~ 현재 : 금오공과대학교 컴퓨터공학과 석사과정

관심분야 : 정보검색, 인공지능

김종완 (Jong-Wan Kim)
1987년 : 서울대학교 컴퓨터공학과 학사
1989년 : 서울대학교 대학원 컴퓨터공학과 공학석사
1994년 : 서울대학교 대학원 컴퓨터공학과 공학박사
1995년 ~ 현재 : 대구대학교 컴퓨터정보공학부 부교수
1999년 ~ 2000년 : 미국 U. of Massachusetts Post Doc.

관심분야 : 지능형 에이전트, 퍼지시스템, 인공지능.