음란 사이트 탐지 시스템의 설계 및 구현

최상필 김병만 이숙희 김주연 김경호
금오공과대학교 컴퓨터공학부
{spchoi,bmkim,shlee,jykim,ghkim}@cespc1.kumoh.ac.kr

The Design and Implementation of Lewdness Site Detection System

Sang Pil Choi, Byeong Man Kim, Suk Hee Lee, Ju Yeoun Kim, Gyung Ho Kim
Dept. of Computer & Software Engineering, Kumoh National University of Technology

요 약

본 논문에서는 음란사이트를 효과적으로 탐지하기 하기 위하여 퍼지 추론을 이용한 방법을 제안한다. 사용자로부터 몇 개의 음란 사이트 URL을 질의로 입력받아, 해당 URL로부터 수집된 웹 문서들에서 웹 태그와 불용어를 제외한 모든 용어들을 추출한 후, 용어의 DF, TF, HI(Heuristic Information) 정보들을 퍼지 추론에 적용하여 사용자가 제시한 음란 사이트에서 용어의 중요도를 산정한다. 또한, 웹 로봇은 인터넷에서 웹 문서를 수집하고, 퍼지 추론에 의해 산정된 용어의 중요도를 이용하여 수집된 웹 문서가 음란 문서일 가능성을 판별한다.

1. 서론

인터넷의 발달과 WWW의 대중화로, 삶에 유용한 다양한 정보들을 누구나 쉽게 얻을 수 있는 반면에, 청소년들에게 유해한 정보들도 아무런 제약 없이 폭넓게 전파되고 있다. 최근에는 이러한 유해 정보가 사회적 문제를 유발하게 됨에 따라, 청소년들로부터 누드사진, 포르노 그라피 같은 음란 정보를 제공하는 웹 페이지로의 접근을 막을 수 있는 시스템이 요구되고 있으며, 이러한 시스템들은 음란 사이트에 대한 URL을 미리 수집함으로써 가능해 진다. 그러나, 음란 정보를 제공하는 사이트를 수집하기 위하여 기존의 AltaVista, Yahoo, InfoSeek등의 검색엔진을 이용할 경우, 이러한 검색엔진들은 제한된 양의 질의어로서 단순한 용어 매칭 혹은 불리언 연산만을 수행하므로 용어 불일치 문제가 발생할 수 있으며, 또한, 용어의 의미를 파악할 수 없음으로 동일한 용어지만 의미가 전혀 다른 용어일 경우에는 사용자가 원하는 결과와 전혀 상반된 결과를 얻게되는 문제점이 있다. 그러므로, 기존의 검색엔진을 이용하여 음란 사이트를 수집하는 것은 비효율적이며, 용어의 의미를 파악하여 질의를 자동으로 생성할 필요가 있다.

본 논문에서는 사용자로부터 "xxx","sex"같은 음란 관련 용어가 아니라 음란 사이트의 URL을 입력받아, 질의된 웹 문서들에서 용어가 차지하는 중요도를 퍼지 추론을 이용하여 산정하고, 용어의 중요도를 이용하여 문서를 검색함으로서 음란, 비음란 사이트를 구분할 수 있는 시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 시스템의 전체 구조와 퍼지 추론을 이용한 용어의 중요도 산정 방법에 대해여 기술하고, 3장에서 음란 사이트를 대상으로 실험한 결과를 설명한다. 또한, 4장에서는 결론 및 향후 연구과제를 제시한다.

2. 음란 사이트 탐지 시스템의 설계 및 구현

2.1 시스템 구조

그림 1에서는 본 논문에서 제안하는 시스템의 전체 구조를 보인다. 사용자가 음란 사이트 URL을 질의어로 제출하면, 사용자가 지정한 웹 문서에서 어휘분석과 스테밍(불용어 제거)을 거쳐 사용자의 관심 정보를 나타내는 용어들이 추출된다. 추출된 용어들이 사용자가 지정한 웹 문서들에서 차지하는 중요도를 산정하기 위하여, 지정된 전체 웹 문서들에서 웹 태그와 불용어를 제외한 모든 용어들을 추출하고, 용어 발생 수(Term Frequency), 해당 용어가 발생한 문서 수(Document Frequency), 기존의 검색엔진에 해당 용어를 질의어로 사용하여 검색할 경우 검색되는 문서의 수(Heuristic Information)등의 정보를 얻어 퍼지 추론에 적용한다. 이때, 각 용어의 HI 정보를 획득하는데 소요되는 시간을 최소화하기 위하여 획득한 HI정보는 DB(DataBase)에 저장된다. 또한, 웹 로봇은 인터넷에서 웹 문서를 수집하고, 퍼지 추론에 의해 산정된 용어 중요도를 이용하여 수집된 웹 문서와의 유사도를 산정함으로서 음란 문서일 가능성을 판별하게 된다.

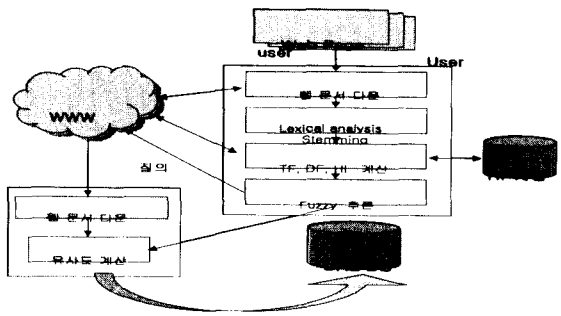

그림 1 전체 시스템 구성도

2.2. 사용자의 관심 웹 문서 제시

사용자는 음란 사이트의 URL을 질의어로 입력한다. 본 논문에서는 시스템의 정확성과 사용자의 편리성을 고려하여 입력되는 사이트의 수를 3 - 7개로 제한하였다.

2.3 퍼지 추론을 이용한 중요도 산정

2.2절에서 사용자로부터 제시된 웹 문서들에서 발생하는 용어들을

대상으로 퍼지 추론을 이용하여 해당 용어의 중요도를 산정 한다. 이때, 용어의 중요도를 퍼지 추론을 이용하여 산정한 이유는 퍼지 이론을 이용할 경우 각 요소 값들을 인간의 직관적인 사고에 반영하기 위하여 쉽게 해석할 수 있고, 인간의 직관적인 사고를 퍼지 규칙으로 간단하게 작성할 수 있기 때문이며, 이러한 퍼지 이론에서 제어 값의 추론은 퍼지 추론을 이용하기 때문이다.

1) 퍼지 변수

그림 2에서는 본 논문에서 사용한 퍼지 입출력 변수와 소속 함수들을 나타내고 있다. 그림 2의 (a)는 사용자가 지정한 웹 문서들에서 용어의 평균 발생 빈도수(TF)를 정규화한 NTF 입력 변수의 소속 함수를 나타내고, (b)는 특정 용어를 기존의 검색 엔진에 질의하여 검색되는 문서 수(HI)를 0.0-1.0으로 정규화한 NHI와 사용자가 지정한 웹 문서들 중에서 특정 용어가 발생한 문서 수(DF)를 정규화한 NDF 입력 변수의 소속함수를 나타낸다. 또한 (c)는 출력 변수 S의 소속 함수로써 6개의 소속 함수들로 구성되고, 최대 1.0의 중요도를 가지도록 소속 함수 구간의 차를 0.2로 설정하였다.

NHI 입력 변수는 특정 용어가 질의어로서 중요하지 않은 일반 용어일 가능성을 평가한다. 일반적으로 정보검색에서는 용어의 중요성을 평가하기 위하여 역문헌빈도수(Inverse Doc ument Frequency)를 자주 사용하나, IDF는 검색 대상이 되는 문서 전체에 대한 정보를 이용할 수 있을 경우에만 사용될 수 있으며, 현재의 웹 환경에서는 사용할 수가 없다. 그러므로, 본 논문에서는 기존의 검색 엔진에 일반용어를 이용하여 질의를 할 경우 중요용어와 비교하여 많은 문서들이 검색된다는 점에 착안하여 일반 용어를 판별하였다. Altavista에 각 용어에 대한 질의를 수행하고, 가장 많은 문서를 검색한 용어 10개를 선택하여 평균 검색 문서 수(AV)를 산정하며, (식 1)을 이용하여 퍼지 추론에 이용할 수 있는 정규화된 NHI를 산정하였다.

$$NHI_i = 1 - \frac{HI_i}{AV} \qquad \text{(식 1)}$$

NHI_i : 용어 i의 HI 소속 함수 값
HI_i : 용어 i의 검색 문서 수
AV : 최대로 검색된 상위 10개의 용어에 대한 평균 검색 문서 수

NDF 입력 변수는 지정된 웹 문서 내에서 특정 용어의 발생 분포를 평가하기 위하여 사용되었으며, NHI 정보와 결합하여 전체 웹 문서에서 용어의 중요도를 판단하는 중요한 근거가 된다. 예를 들면, NHI 정보에 의해 중요용어로 판단되어 지고, 많은 문서에서 발생할 경우 이러한 용어는 사용자의 관심을 나타내는 중요한 용어로서 판단할 수 있다. (식 2)에서는 DF를 정규화하기 위하여 사용된 식을 나타내고 있다.

$$NDF_i = \frac{DF_i}{TD} \qquad \text{(식 2)}$$

NDF_i : 용어 i의 DF 소속 함수 값
DF_i : 지정된 웹 문서내에서 용어 i의 발생 문서 수
TD : 지정된 전체 웹 문서 수

NTF 입력 변수는 사용자가 지정한 웹 문서들에서 용어의 평균 발생 빈도수(TF)를 정규화한 것으로, 특정 용어의 상대적인 중요도를 평가하기 위하여 사용하였다. (식 3)에서는 0.0 - 1.0의 값으로 정규화하기 위하여 용어의 평균 발생 빈도수(TF)의 역수로 산정하였다.

$$NTF_i = 1 - \frac{DF_i}{\sum_{k=1}^{n} TF_{ik}} \qquad \text{(식 3)}$$

NTF_i : 용어 i의 평균 발생 빈도수(TF) 소속 함수 값
DF_i : 지정된 웹 문서내에서 i의 발생 문서 수
TF_i : 지정된 웹 문서 k에서 용어 i의 발생 빈도수
n : 지정된 웹 문서 수

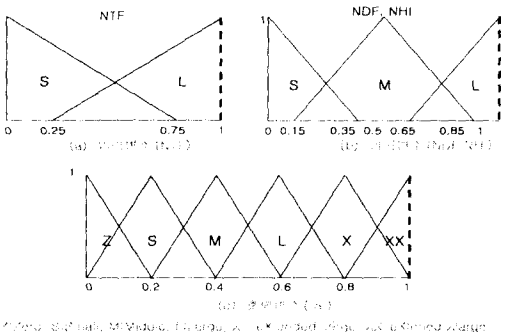

그림 2 퍼지 변수 및 소속함수

2) 추론 규칙

그림 3은 본 논문에서 사용한 18개의 퍼지 추론 규칙을 나타낸다. 규칙은 용어의 중요도에 따라 6단계의 소속 함수로 구분할 수 있으며, Z 소속함수 결과를 갖는 규칙이 4개, S는 5개, M은 2개, L은 3개, X는 3개, XX는 1개의 추론 규칙이 있다. 규칙에서는 용어가 특정 문서에서만 발생하고(S), 평균 발생 빈도 수가 적으며(S) 일반용어의 가능성이 클 경우(S) 이러한 용어는 사용자 관심을 나타내는 용어로서는 의미가 전혀 없기 때문에 Z 소속 함수를 가지도록 하였으며, 많은 웹 문서에서 발생하고(L), 평균 발생 빈도 수가 높으며(L) 일반용어일 가능성이 클 경우(S)에는 중요도를 결정하기 어렵기 때문에 S 소속 함수를 가지도록 하였다. 또한, 많은 웹 문서에서 출현하고(L), 평균 빈도 수가 높으며(L) 중요용어일 가능성이 클 경우(L) 사용자의 관심을 나타내는 용어일 가능성이 매우 높으므로 XX의 소속 함수를 가지도록 하였다.

NTF - S

NHI / NDF	S	M	L
S	Z	Z	S
M	Z	M	L
L	S	L	X

NTF - L

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

그림 3 퍼지 추론규칙

3) 비퍼지화

퍼지 추론 규칙에 의하여 생성된 출력 변수(S)의 소속 함수 값들을 단일한 값으로 비퍼지화 하기 위하여 본 논문에서는 (식 4)와 같이 무게중심(center of gravity) 법을 사용하였다. 무게중심법은 합성된 출력 퍼지 변수들의 무게중심을 구하여, 해당하는 제어 값을 제어입력으로 사용하는 방법이다.

$$S_i = \frac{\sum_{j=1}^{n} \mu(\mu_{ij}) \times \mu_j}{\sum_{j=1}^{n} \mu(\mu_{ij})} \qquad \text{(식 4)}$$

S_i : 용어 i의 중요도
$\mu(\mu_{ij})$: 용어 i가 소속 함수 j에 소속된 정도
μ_j : 소속 함수 j의 구간 값
n : 출력 변수(S)의 소속 함수 수

2.4 웹 문서의 유사도 산정

웹 로봇에 의해 탐색된 문서는 2.3절에서 퍼지 추론을 이용하여 산정된 용어의 중요도(비퍼지화 값)와의 유사도를 산정함으로서 음란 사이트 여부를 판별한다. (식 5)에서는 유사도 산정에 사용된 식을 나타내고 있으며, 중요도가 부여된 용어가 해당 웹 문서에 발생할 경우, 그 평균값으로 유사도를 계산하게 된다.

$$Smt(S,D) = \frac{\sum_{i=1}^{n} S_i}{n} \qquad \text{(식 5)}$$

$Smt(S,D)$: 사용자 관심 분야와 웹 문서의 유사도

S_i : 용어 i 의 중요도

n : 사용자 지정 문서와 특정 문서에서 동시에 발생한 용어의 수

3.실험 및 결과

본 논문에서는 실험을 위하여 표 1과 같이 4개의 URL을 사용하였으며, 표 2에서는 지정된 웹 문서에서 발생하는 용어들중 중요도가 가장 높은 상위 50개의 용어들을 보여주고 있다. 또한, (식 5)를 이용하여 100개의 웹 문서에 대하여 음란성을 평가한 결과 표 3과 같았다. 본 실험에서는 유사도의 threshold 값이 0.12 이상이면 음란사이트로 판별된다.Threshold를 너무 크게 잡으면 많은 수의 음란 사이트가 검색에서 제외되게 되며, 작게 잡을 시에는 관련 없는 사이트가 검색되는 결과를 초래한다.

표 1 사용자 제시 URL의 예

http://www.xxx-software.com/freesex/xxxbrowser.???
http://www.smutview.???/
http://www.sexysolutions.???/
http://www.msxxx.???/

표 2 음란 사이트에서 발생하는 용어의 중요도

단 어	가능성	단 어	가능성	단 어	가능성
xxx	1.000000	pics	1.000000	adult	1.000000
hot	0.900000	content	0.900000	click	0.868657
girl	0.868657	material	0.868657	pussy	0.851163
fuck	0.832258	feature	0.800000	link	0.800000
sex	0.796054	hardcore	0.785714	love	0.777778
cum	0.777778	amateur	0.766667	system	0.766667
day	0.766667	live	0.744558	totally	0.740000
include	0.740000	fun	0.740000	latin	0.740000
join	0.710009	video	0.704906	address	0.700000
membership	0.700000	pay	0.700000	require	0.700000
bookmark	0.700000	trial	0.700000	theme	0.700000
sexual	0.700000	lesbian	0.700000	offensive	0.700000
sexy	0.700000	lot	0.700000	guy	0.700000
image	0.700000	ass	0.700000	tit	0.700000
picture	0.688572	chat	0.683040	real	0.674727
nude	0.672681	gallery	0.645793	age	0.633333

4. 결론 및 향후 연구 방향

본 논문에서는 사용자로부터 관심 웹 페이지를 제시받아 퍼지 추론을 이용한 사용자 관심정보를 구함으로써 효율적인 음란 사이트 검색 시스템을 설계하였다. 음란 사이트는 물론 다른 응용에도 사용가능하며, 실험결과 효율적으로 음란 사이트가 판별되는 것이 검증되었다. 본 연구에서는 퍼지 추론을 이용한 용어의 중요도 산정에 용어의 발생 빈도수와 DF만을 고려하였다. 앞으로, 문서 내에서 용어의 발생 위치에 대한 정보를 파악하여 제목이나 강조에 포함된 용어에 대해 높은 가중치를 주는 방법에 관한 연구와, 시스템을 통하여 구축되는 URL 데이터베이스 정보를 기반으로 학습 기능을 첨가함으로써 보다 정확도가 높은 사용자 관심정보 추출에 대한 연구가 필요하다. 또한 음란 사이트의 경우 이미지의 비중이 많이 차지하므로 이미지 특징 추출방법에 대한 연구가 필요하다.

표 3 음란 사이트 탐지 예

	주 소	배치	총단어	유사도
비음란사이트	www.altavista.com	23	183	0.0631
	www.javasoft.com	36	470	0.0362
	www.palmhotel.com	16	221	0.0369
	www.geocities.com/Wellesley/7746/Pregnancy.html (임신관련 웹 페이지)	11	78	0.0504
	lynx.uio.no/catfolk/cnissues/cn08-13.htm (불임관련 웹 페이지)	1	108	0.0027
	www.ageing.org/prog_166.htm (성 호르몬 관련 웹 페이지)	18	165	0.0603
	caag.state.ca.us/megan/notice.htm (성 범죄 관련 웹 페이지)	19	109	0.1073
	www.capitolhillblue.org/Aug1998/sexaddictaug27.htm (성 문제 관련 웹 페이지)	19	127	0.0924
	www.howtohavegoodsex.com/chat.htm (성 생활 관련 웹 페이지)	27	129	0.1151
	www.epigee.org/guide/tl.html (피임 관련 웹 페이지)	5	29	0.0656
음란사이트	www.xxx.???	15	36	0.2410
	www.xxx-gay-sex-porn.???	40	116	0.1824
	www.ifriends.net/warning/altavista/???	14	65	0.1312
	www.xxxcounter.com/???/	33	98	0.2102
	www1.dormvideo.com/mtp/p/dormvideo.htm?mci-???	21	59	0.2085
	www.junglegirls/6cb96bb7/???	12	33	0.1824
	www.dirtog.???	16	36	0.2805
	www.absoluteaccess.com/???/	63	273	0.1468
	www.adultorigin.com/???	196	437	0.3057

5. 참고 문헌

[1] Gerald Kowalski, "Information Retrieval Systems -Theory and Implementation",KLUWER ACADEMIC PUBLISHERS,p125,148, 1997.

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

[3] Willianm B.Frankes, Ricardo Baeza-Yates, Information Retrieval Data Structures&Algorithms, Prentice Hall, p102-160, 1992.

[4] Q. Kong and G.Chen, On Deductive Databases with Incomplet Information, ACM TODS, Vol. 13, pp. 167-196, 1988.

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

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

[7] G. J. Klir and B. Yuan, "Fuzzy Sets and Fuzzy Logic : Theory and Application," Prentice Hall PTR, 1995.

[8] Willianm B. Frakes, Ricardo Baeza-Yates. Information Retriev al. pp102-160. Prentice Hall. 1992.