웹에서 동물영역 관련문서 필터링

김상모 김원우 변영태[0]
홍익대학교 전산학과
{smkim, wwkim, byun}@cs.hongik.ac.kr

A Web Document Filtering System for Animals

Sang-Mo Kim Won-Woo Kim Young-Tae Byun[0]
Dept. of Computer Science, Hongik University

요 약

인터넷에 돌아다니는 정보의 양은 무한정에 가까워지고 있고 이용자는 필요한 정보들을 얻을 수 있게 되었으나 검색 가능한 정보의 양이 폭발적으로 증가함에 따라 이용자는 정보검색을 하는데 있어 어려움이 따랐고, 이는 원하는 정보만을 필터링하여 보여주는 정보검색방법이 필요하게 되었다. 본 연구에서는 웹 사용자들이 정보검색을 하는데 원하는 정보를 정확하게 찾아주기 위해 웹 문서에 대한 TAG가중치와 관련용어 영역지식의 구축 및 웹 문서 평가작업을 통한 Term의 웹 문서 DF테이블의 구축을 이용한 필터링 방법을 제안하고 그 유효성을 확인하였다

1. 서론

인터넷은 상호 정보 교환 및 최신 정보를 획득하기 위한 수단으로 이용되어 왔으며 인터넷의 사용자의 증가와 함께 인터넷에서 얻을 수 있는 정보의 양도 증가 되었다. 그러나 인터넷에서 정보의 양이 폭발적으로 늘어남에 따라 사용자가 원하는 정보를 얻기 위하여 검색을 하는 데 있어 시간비용이 늘어날 뿐 아니라 정확한 정보를 얻기가 힘들게 되었다. 전문 검색엔진의 개발로 인해 사용자는 정보검색이 용이하게 되었으나 이직까지도 원하는 정보를 정확하게 제공해주지는 못하고 있는 실정이다. 이용자의 부담을 덜기 위한 정보검색 기법들중 가운데 필터링 기법은 이용자가 원하는 정보만을 검색하도록 하는 도구이다

웹 문서 정보검색을 위한 필터링에 대한 방법으로는 Salton의 TF/IDF방법을 이용한 필터링 기법[1]이 대표적이다. TF/IDF방법은 검색된 문서의 연관성의 정도를 검색된 문서강에서 나타나는 검색어의 빈도수(Term Frequency)와 전체 문서상에서 검색어가 나타나는 비율(Document Frequecy)을 통해 계산하게 된다. 위 방법은 필터링 뿐아니라, 전통적으로 문서에 대한 정보검색으로 방법으로도 널리 쓰여지고 있다. 그러나 위 방법으로는 웹 기반 정보검색 환경하에서는 근본적인 결함을 가지고 있다

우선 웹 상에서 존재하는 전체 문서의 개수를 파악하기 어려우며 인터넷상에 존재하는 웹 문서들은 한 가지 주제를 대상으로 한 문서들의 집합체가 아닌 다양한 주제들을 포함하고 있는 문서들의 집합체이다. 이는 웹 문서에서 ' bear' 란 단어가 동물로서의 ' bear' 뿐 아니라 ' teddy bear' 를 지칭하는 상호명이나 상품, 단체명 등 다양한 의미로 사용되어지는 어휘 의미 중의성 문제[2]를 유발하게 된다. 이는 사용자가 검색하고자 하는 질의어가 어떠한 영역에 속한 것이라는 것을 염두해 두고 있지 않기 때문이고, 이러한 문제점을 해결하는 방법중하나로 특정 도메인 영역에 대한 검색방법이 있고 본 연구에 있어서는 동물 도메인을 두고 검색하는 방법에 대하여 알아보기로 하겠다.

2. 관련연구

인터넷 문서 검색에 있어서 여러 방면으로 연구가 되고 있다. 웹 문서 필터링 방법으로 TAG Weight[3], 개념들을 그 개념들 사이의 관계로서 표현해두고 정보 검색시에 개념들의 관계를 이용하여 정확한 색인어가 아니더라도 사용자로 하여금 원하는 정보를 찾을 수 있도록 도와주는 시소러스 시스템을 이용한 지식 기반의 정보 검색 시스템[4], 인터넷 사용자의 경향분석등에 대한 연구가 활발하다.

.본 연구는 뇌과학 연구 사업의 지원으로 진행 되었슴

한편 인터넷 이용자의 대부분은 검보검색에 대한 전문 지식을 갖고 있지 못하며 검색식을 작성하는데 있어 미리 전략을 세우기 보다는 필요에 따라 단순히 키워드를 입력하는 형태를 나타내는데, 이는 초보자의 정보검색행동 경향분석[5]에서도 잘 나타난다. 실험 결과에 의하면 과반수에 해당하는 49.4%가 입력한 검색어의 수는 1개였으면 검색어 수가 4개 이상 입력한 경우는 거의 없었다. 또한 Excite의 log file에 기록된 50만여개의 질의를 대상으로 사용자 성향을 분석한 결과[6]에서도 사용자의 평균 입력 검색어의 수는 2~3개에 불과하고 사용자의 95%가 검색엔진의 구절(Phrase)검색기능을 사용하지 않는다고 한다. 이로 미루어 보면 대부분의 웹 문서 검색 이용자들은 찾고자 하는 정보문서에 대해 단일 질의어를 상황에 따라 바꿔가며 검색한다고 볼 수 있겠다. 이는 위에서 언급한 어휘의미중의성으로 인해 검색결과가 정확하지 못하고 재검색으로 인한 시간의 비용이 많아지게 된다. 일반적으로 어휘의미중의성으로 인한 정확률의 저하를 최소화하려면 중복의미를 갖는 단어에 대해 별도의 단어를 추가해 갖은 뜻을 갖는 그룹을 만들어야 하는데 입력할 키워드의 개수와 정확률과 재현률의 비율을 조사한 바에 의하면[7] 검색 키워드의 수가 3개에 적합율과 재현율이 가장 좋은 비율로 나타나는 것을 알 수 있다. 그럼에도 불구하고 일반 사용자들에 대해 보다 정확한 정보 검색 결과를 제공하기 위하여 새로운 방법을 시도해 보고자 하였다

관련실험의 기본이 되는 것은 동물키워드 6 개를 가지고 검색엔지 Altavista 를 이용하여 2000 개의 URL과 또 다른 동물키워드 6 개를 가지고 200 개의 URL을 평가한 결과를 이용하였다

3. 필터링 기법

본 연구에서는 크게 3가지 방법을 사용한 필터링을 제안하고 있다. 첫번째로 웹 문서의 소스를 분석하여 중요단어의 위치를 찾아 그에 따른 Term Weight를 부여하는 TAG Weight 방법과 두번째로 관련서적과 참고 문헌등을 보고 계층정보나 속성 정보등 필요한 정보들을 추출하고 영어어휘 데이터베이스 WordNet[8]을 이용하여 관련 정보와 비관련 정보를 추출하여 관련용어 영역지식의 구축하고 이를 이용한 방법과 마지막으로 일정수의 문서를 평가작업을 통해 관련 문서와 비관련 문서로 분류한 후 관련 문서에 나온 Term과 비관련 문서에 나온 Term을 DF에 관련하여 비교하여 만든 WebTerm 테이블을 이용한 방법이 있다

3-1. TAG가중치

WWW에서 대다수를 차지하는 문서의 언어인 HTML의 표시방법은 명령 부분을 부등호 '<>' (TAG)로 묶어 명령을 실행하는 범위를 지정하는 형식의 작성법을 택하고 있다. 이 TAG들이 하는 역할이 모두 틀리듯이 TAG들 안에 포함되어 있는 단어들 역시 그 문서를 알리는 비중이 다를 것이다.

본 실험에서는 8 개의 동물키워드를 검색어로 ' HIIA-1a' [9]에서 사용된 WebDB를 이용하여 관련문서 312개 URL을 검색하였고 각 키워드별로 검색된 URL에 대하여 키워드가 속해있는 TAG의 DF(Document Frequency)를 조사하였다. 그리고 TAG의 DTF에 대해 상위 3 ~ 10 개에 대한 DTF의 합을 구한 후 내림차순 정렬로 10개를 선정하여 <표 1>과 같이 각TAG에 대한 비중을 구했다

TAG NAME		TAG WEIGHT
A		14
TITLE		12 (%)
B		9
IMG		8
META		8 (%)
I		8
H1 ~ H2		7
H3 ~ H6		3
FONT	SIZE	3
	FACE	4
	COLOR	4
DEFAULT		1

<표 1> TAG Weight Table

title, meta TAG 는 문법의 특성상 특정위치에 나타나기 때문에 웹 문서에서 추출한 Term Kind 수에 대한 비율로 정했다(단 Term Kind 가 100 개 이상일 때)

3-2. 관련용어 영역지식의 구축

' HIIA-1a' 에서 사용되었던 계층정보와 특성정보등 동물관련 지식정보를 일부 이용하였고 동물관련서적에서 동물과 관련된 전문 정보를 추출하였고 WordNet을 이용하여 동물관련 단어 그룹과 그 외의 단어 그룹을 추출하여 Positive Term과 Negative Term 그룹으로 이루어진 관련용어 영역지식을 만들어 관련문서 평가에 이용하였다.

3-3. WebTerm테이블 구축

웹 문서를 검색하는 데 있어 찾고자 하는 관심 있는 관련문서와 그렇지 않은 비관련 문서가 있을 수 있다. 그래서 관련문서와 비관련 문서의 차이를 알기 위해 관련문서에서 나오는 Term들과 비관련 문서에서 나오는 Term들을 비교해 보고자 했다.

위에서 언급한 동물관련 URL 2000 개를 평가하여 관련문서와 비관련 문서로 나누고 각 URL 에 나온 Term 들중 Yahoo 영어사전(금성출판사)에 있는 Term들과 앞에서 구한 관련용어에 대해 TF와 DF를 구한 후 다음 <표 2>와 같은 WebTerm 테이블을 구축하였다.

Term	Relevant URL		Irrelevant URL	
	TF	DF	TF	DF
t_i	RTF_i	RDF_i	ITF_i	IDF_i

<표 2> WebTerm Table

구축한 Term 테이블을 이용하여 새로 검색된 문서가 관련문서인가 비관련 문서인가를 판별하는 방법으로

1. 판별할 문서 Di 에 대해 모든 Term을 추출한 후 각 Term t_d 에 대하여 $\frac{RDF-IDF}{RDF+IDF}$ ($t_i = t_d$)의 값을 구한다.
2. 1에서 구한 값을 모두 더한 후 구한 값의 개수로 나눈다.

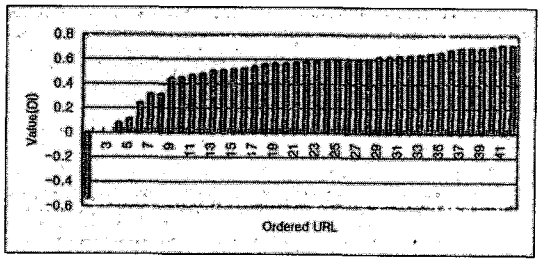
$$\text{Value}(Di) = \frac{\sum_{i=1}^{n} \frac{RDF-IDF}{RDF+IDF}(t_i = t_d)}{n}$$

4. 결론 및 향후계획

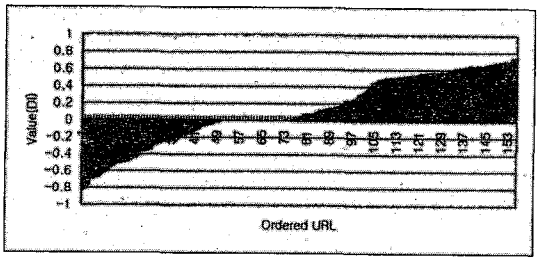
새로운 웹 문서 200개를 수집한 후 WebTerm테이블을 이용하여 문서의 적합성 여부를 평가해 봤다.

WebTerm Table에서 $\frac{RDF-IDF}{RDF+IDF}$ 을 사용하여 나온 값의 범위 □1 ~ 1 사이에서 $0.5 \le t_i < 1$ 과 $-1 < t_i \le -0.5$ 인 Term t_i 에 대해서만 적용시켰을 때 가장 좋은 변별력을 갖는 것을 알 수 있었다.

다음 <표 3>과 <표 4>는 새롭게 수집한 웹 문서 200개에 대하여 미리 관련 문서와 비관련 문서로 분류한 후 각 문서들에 대해 적절성 평가 공식을 적용시켜 본 것이다.



<표 3> 관련문서 DF 0.5 AVG



<표 4> 비관련문서 DF 0.5 AVG

위 표를 보면 Threshold를 0.4로 주었을 때 관련문서의 손실을 최소로 하면서 비관련 문서를 최대로 필터링 해주는 것을 알 수 있다.

웹 문서 필터링을 하는데 필요한 WebTerm을 구축하는 데 있어서 다수의 웹 문서에 대해 사람의 평가작업이 요구 되었고 특정 도메인에 대하여 정보를 얻어내는 것도 인위적으로 해야만 하였으나 WordNet을 통하여 특정 영역에 대한 정보를 얻을 수 있고 WebTerm구축의 기준이 되는 웹 문서의 관련성 평가 작업은 도메인을 대표하는 특정 키워드의 TF에 대하여 높은 Threshold를 적용하여 확실한 문서 판단을 함으로서 보다 쉽게 WebTerm을 구축하도록 하고있다.

동물 영역에서 필터링의 효과를 더욱 향상시키기 위하여 보다 정확한 Term Set을 찾아내는 것에 대한 연구가 필요하며 동물 도메인 영역에서의 필터링 효과를 높힌 후 TAG Weight를 이용한 순위작업과 Usage Mining을 적용시키고 이후 다른 도메인 영역에서의 적용으로 확대할 예정이다

5. 참고 문헌

[1] Marko Balabanovic and Yoav Shoham, " Learning Information Retreival Agetn: Experiments with Automated Web Browsing " , NSF IRI -9411306 with NSF/ ARPA/NASA Digital Library project, 1994

[2] 황상규, 변영태, " HIIA-F를 위한 지식베이스와 질의어의 의미적 확장" , 한국정보과학회 99년 봄 학술발표논문집, 1999

[3] Michal Cutler, Yungming Shih, Weiyi Meng, " Using the Structure of HTML Documents to Improve Retrieval" , Dept. of Computer Science,State of NewYork at Binghamton, 1997

[4] 박영몽,김민구,이정태," 지식 기반의 정보 검색 시스템" , 한국정보과학회논문지 제21권 제11호, 1994

[5] 박창호,박민규,이정모," 가이드라인이 인터넷 정보검색 수행에 미치는 영향" , 한국심리학회지: 실험 및 인지, 10권 2호, 1998

[6] D. Cutting, " Industry Panel Discussion " , SIGIR,1997(A sample of Excite! queries form September 16, 1997)

[7] 황상규,김상모,변영태," 지식기반 웹 문서 펑터링" , 한국정보과학회 99년 가을 학술발표논문집, 1999

[8] WordNet , Princeton University Cognitive Science Laboratory. WordNet □ A Lexical Database for English. http://www.cogsci.princetion.edu/~wn/

[9] 이용현," 정보통신망에서 지능형 정보 에이전트와 특정영역에서의 구현" , 홍익대학교 박사학위논문, 1999