주변정보 분할을 이용한 주제 중심 웹 문서 수집기

조 창 희[†] · 이 남 용[††] · 강 진 범[†††] · 양 재 영[††††] · 최 중 민[†††††]

요 약

주제 중심 웹 문서 수집기는 검색엔진에서 최신의 웹 문서 색인을 유지하는 대안방안으로 부상하고 있다. 그러나 주제 중심 웹 문서 수집기는 비 관심문서에서 연결된 관심문서들을 수집할 수 없는 문제점을 가지고 있다. 이러한 문제점은 문서의 구조적 특징을 고려하지 않아서 발생한다. 특히 문서분석 방법인 문서의 발생 횟수 및 역문헌 발생빈도는 이러한 문제를 야기하는 주요 원인이 된다.

주제 중심 웹 문서 수집기의 성능을 향상하기 위해서 본 논문에서는 국소정보기반의 문서 분할법을 제안한다. 본 논문에서는 문서를 하이퍼링크 주변의 문맥을 고려한 특징 정보들을 사용하여 여러 조각의 문서로 나눈다. 본 논문에서 제안하는 주제 중심 웹 문서 수집기는 나누어진 문서들을 이용하여 하이퍼링크가 관심문서를 가리키는 것인지를 판단하여 문서를 수집할 것인지를 판단한다.

키워드 : 주제 중심 웹 문서 수집기, 문서분류, 특징정보 추출, 문서분할

A Focused Crawler by Segmentation of Context Information

Cho Chang Hee[†] · Lee Nam Yong[††] · Kang Jin Bum[†††] · Yang Jae Young[††††] · Choi Joong Min[†††††]

ABSTRACT

The focused crawler is a topic-driven document-collecting crawler that was suggested as a promising alternative of maintaining up-to-date web document indices in search engines. A major problem inherent in previous focused crawlers is the liability of missing highly relevant documents that are linked from off-topic documents. This problem mainly originated from the lack of consideration of structural information in a document. Traditional weighting method such as TFIDF employed in document classification can lead to this problem.

In order to improve the performance of focused crawlers, this paper proposes a scheme of locality-based document segmentation to determine the relevance of a document to a specific topic. We segment a document into a set of sub-documents using contextual features around the hyperlinks. This information is used to determine whether the crawler would fetch the documents that are linked from hyperlinks in an off-topic document.

Key Words : Focused Crawler, Document Classification, Feature Extraction, Document Segmentation

1. 소 개

인터넷 상의 정보는 매일 방대한 양이 생성과 소멸을 반복하고 수시로 변경된다. 이러한 인터넷 정보의 유연성 때문에 사용자에게 효율적인 정보제공을 위한 웹 문서 수집기 개발 및 관련 연구가 많이 이루어지고 있다.

1993년 전세계 웹 서버의 숫자를 파악하기 위해 만들어진 Mathew K. Gray의 'World Wide Web Wanderer'로부터 기초적인 웹 문서 수집기에 대한 연구가 시작되었다. 당시의 웹 문서 수집기는 인터넷 정보를 보다 빠르고, 정확히 찾을 수 있는 검색엔진의 한 부분으로써 "관리자의 개입 없이 중복되지 않는 URL을 자동으로 찾아가 인터넷 문서 정보를 축적하여 사용자가 정보를 검색할 수 있도록 도움을 주는 소프트웨어"라고 정의되었다[1].

근래의 폭발적인 인터넷 정보의 증가는 이러한 검색엔진의 중요성을 더욱 부각시켰다. 하지만 무분별한 웹 문서의 수집으로 검색엔진은 사용자의 질의에 대해 적합한 결과물을 제공해 주지 못하고, 방대한 데이터로 인해 검색엔진의 부하가 생기게 되었다. 이것은 검색엔진이 최신의 정보를 수집하는데 오랜 시간을 소요하게 됨을 의미한다. 이러한 문제점을 해결하기 위해서 특정한 주제만을 검색 대상으로 하는 전문분야 검색엔진들이 발생하게 되었다. 전문분야 검색엔진들은 기존 검색엔진이 사용하는 웹 문서 수집기가 아닌 특정 분야의 데이터만을 여과하고 수집하는 주제 중심

† 정 회 원 : 법제처 법령정보화 총괄담당 사무관
†† 정 회 원 : 숭실대학교 컴퓨터학과 교수
††† 준 회 원 : 한양대학교 대학원 석사과정
†††† 정 회 원 : 동부정보기술주식회사 컨설팅사업팀
††††† 정 회 원 : 한양대학교 컴퓨터공학과 교수
논문접수 : 2005년 5월 3일, 심사완료 : 2005년 7월 18일

웹 문서 수집기가 필요하게 되었다[2].

기존의 주제 중심 문서 수집기는 웹 문서 수집시 현재 방문한 웹 문서가 수집기가 다루는 특정주제에 적합한 문서인지 평가를 하고 관련 문서(on-topic)로 평가되면 그 문서가 포함하는 하이퍼링크들이 문서 수집의 후보로 등록한다. 따라서 관련 없는 문서(off-topic)로 평가가 된 문서에 포함되는 하이퍼링크들은 문서 수집의 대상에서 제외한다. 이것은 관련문서로 평가된 웹 문서를 이용해 수집 경로를 구성 함으로써 관련 없다고 평가된 웹 문서에 관련성 있는 정보가 존재할 때에도 문서 전체의 관련성 여부에 따라 웹 문서 수집기의 수집 경로가 결정된다. 하지만 문서 수집 시 관련 없는 문서로 평가가 되었지만 특정주제에 적합한 웹 문서를 하이퍼링크로 보유할 수 있고, 관련 있는 문서로 평가된 웹 문서의 하이퍼링크들이 관련 없는 문서를 가리킬 수 있다. 하지만 수집기는 해당 웹 문서들을 모두 방문해 평가 후 이를 결정하는 구조를 가진다.

본 논문에서는 이러한 문제점을 해결하고 수집기의 성능 향상을 위해 문서 분할법을 제안하다. 문서 분할법은 문서의 관련성 여부를 판단할 때 문서단위의 판단이 아닌 하이퍼링크의 수만큼 분할된 문서를 바탕으로 관련성 여부를 판단하기 때문에 기존 주제기반 문서 수집기들이 직면한 문제점을 보안할 수 있다.

2. 관련 연구

주제 중심 웹 문서 수집기의 연구는 크게 두 방향으로 나누어 볼 수 있다. 첫 번째 방향은 웹 문서 수집을 위한 순서를 재배열하여 관심되는 주제부터 수집하는 방법이다. 이 연구는 주로 초창기에 이루어졌으며 문서 분류기를 사용하지 않아 문서의 손실이 없다. FishSearch[3]는 Seed URL을 시발점으로 주어진 질의에 적합한 내용을 가진 문서를 고려해 수집이 이루어졌다. 그리고 방문할 URL의 목록에서 우선순위에 따른 수집하는 형태로 접근하였다. SharkSearch[4]는 문서의 연관성을 판단하기 위해 코사인 유사도 측정을 위한 TF(Term Frequency) 또는 IDF(Inverse document Frequency)와 같은 전형적인 정보 검색(Information Retrieval) 알고리즘을 사용 하였다. 그리고 [5]의 논문에서 문서가 다른 문서를 연결하고 있는 수(in-links), 문서 링크(page link)와 같은 경험적인(heuristic) 방법이 제시되었다. [5]는 문서 수집시 우선순위에 따른 수집 순서를 재정의하여 연관성 있는 문서를 먼저 수집 될 수 있도록 하였다. FishSearch와 SharkSearch, [5]의 알고리즘들은 분류기를 이용하지 않은 초기 모델이며, 단순히 유사도를 측정하기 위해 정보검색(Information Retrieval) 기반 기술들을 이용하였다.

두 번째 방향은 문서 분류기를 사용하여 불필요한 웹 페이지 방문을 차단하여 문서를 수집하는 시간을 줄이면서 효율을 높이는데 목적을 두고 있다. Charabati[2]는 주제 중심 웹 문서 수집기에 처음으로 분류기를 이용하였다. 고려하는 주제에 연관성이 있는 문서들을 나타내기 위해 문서의 점수를 할당하였다. 하지만 Charabati 역시 관련 있는 문서로 평가된 문서로부터 하이퍼링크를 추출해 수집 경로를 형성함으로써 관련 없는 문서로 평가된 웹 문서들의 하이퍼링크들은 고려하지 않고 있다.

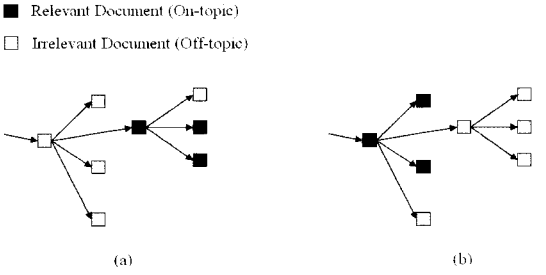
문맥그래프(context-graph)기반 수집기[6]는 문맥그래프를 생성하여 그래프를 통해 수집 경로를 형성하였으며 Cora[7]는 최적화 기법인 강화학습 알고리즘을 이용하여 수집경로를 구성하였다. 그러나 이러한 연구들도 수집 경로를 구성할 때 관련 있는 문서로 평가된 웹 문서를 통해 작업이 이루어짐으로써 관련 없는 문서로 평가되었지만 문서에 내재된 연관성 있는 하이퍼링크들은 고려하지 않고 있다. Mukherjea[8]의 WTMS 또한 관련 있는 문서로 평가된 웹 문서에서 정보검색 기반의 방법을 이용하여 문서를 수집하였다.

3. 기존 주제 중심 웹 문서 수집기의 문제점

주제 중심 웹 문서 수집기는 관련 있는 문서로 평가된 문서의 하이퍼링크들을 통해 문서를 수집한다. 최근의 주제 중심 웹 문서 수집기는 문서를 평가하고 관련 문서로 판단이 될 때 문서의 모든 하이퍼링크를 추출해 수집 경로를 형성한다. 수집기는 형성된 수집 경로를 기반으로 문서를 수집하게 된다. 현존해 있는 주제 중심 문서 수집기는 일반적으로 앞에서 언급한 방법을 이용하고 있다. 하지만 (그림 1)에서 보는 것과 같이 중대한 문제점이 있다.

그 문제점들은 (1)관련 없는 문서로 평가되었지만 사용자 요구에 적합한 하이퍼링크를 가질 수 있다는 점과 (2)관련 있는 문서로 평가되었지만 문서 상의 하이퍼링크가 사용자의 요구에 적합하지 않는 문서를 가리키는 링크라 할지라도 링크들을 모두 방문해 평가해야 한다는 점으로 정리할 수 있다.

(그림 1)에서 (a)의 경우는 사용자 요구에 적합한 하이퍼링크를 가지고 있지만 관련 없는 문서로 평가되어 하이퍼링크가 가리키는 문서는 수집되지 않는 경우를 나타낸다. 이 경우 짧은 시간에 관련 있는 문서를 많이 수집하고자 하는 주제 중심 웹 문서 수집기의 목적에 충실하지 못하게 되고 정보의 누수현상이 나타날 수 있게 된다. (b)에서는 관련 문서로 평가되었지만 사용자 요구에 적합한 문서를 가리키는 하이퍼링크가 없는 경우를 나타낸다. 이 경우 문서의 하이퍼링크들을 모두 방문해 평가하게 됨으로써 많은 데이터 처리로 인한 시스템 부하가 생길 수 있게 된다. 이러한 문제점들을 보완하기 위해 문서의 하이퍼링크의 방문 여부를 문



(그림 1) 웹 문서들의 연결 구조

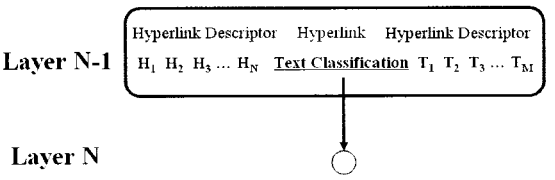
서 평가에 의해 결정되는 것이 아니라 개별적으로 하이퍼링크들을 평가하여 방문 여부를 결정하여 해결할 수 있다.

3. 하이퍼링크 정보를 얻기 위한 문서 분할

문서 분할의 목적은 하이퍼링크 주변의 정보를 이용해 하이퍼링크가 관련 있는 문서를 가리키고 있는지 판단하기 위함이다. 본 논문에서 하이퍼링크는 다른 문서를 연결하기 위한 것으로 문맥상 하이퍼링크 주변에 하이퍼링크에 연관된 글이 기술되어 있다고 가정하고 있다. 따라서 하이퍼링크를 평가하기 위해 하이퍼링크 기술자(Hyperlink Descriptor)를 추출하고 그 정보를 이용함으로써 하이퍼링크의 관련 여부를 판단할 수 있다. 하지만, 하이퍼링크 기술자를 추출한다는 것이 매우 어려운 작업이다.

하이퍼링크 기술자를 추출하기 위해 두 방법이 있다. 첫 번째 방법은 하이퍼링크를 포함하고 있는 문장을 추출하는 방법이 있다. 이 방법은 문장을 분석하는 방법이 아닌 미리 정의된 HTML 태그를 이용해 처리할 수 있다. 두 번째로 하이퍼링크 전, 후의 단어들의 집합을 추출하는 방법이 있다. 이 방법은 추출하는 단어의 수를 명시함으로써 이용 가능하다. 첫 번째 방법과 같이 태그를 이용하는 방법은 HTML 문서를 작성하는 사용자마다 태그의 사용 용도, 표기법, 작성 패턴 및 방식이 틀리기 때문에 명확한 하이퍼링크 기술자를 추출하기 어렵다. 이것은 작성된 HTML 문서의 구성 형태가 모두 틀리기 때문에 태그를 이용한 하이퍼링크 기술자를 추출하는 방법은 매우 위험하며, 경우에 따라서 하이퍼링크 기술자를 추출하지 못할 수도 있다. 그래서 본 논문에서는 하이퍼링크의 전후 10개의 단어들을 하이퍼링크 기술자로 이용하는 두 번째 방법을 이용한다. (그림 2)는 본 논문에서 이용하는 하이퍼링크 기술자를 나타낸 그림이다. Layer N은 Layer N-1인 부모 노드가 가리키고 있으며 Layer N을 설명하기 위해 하이퍼링크 기술자와 "Text Classification"과 같은 하이퍼텍스트를 사용하였다.

본 논문에서 사용하는 방법은 문서상에서 하이퍼링크들을 기준으로 하이퍼링크 기술자 추출 및 하이퍼링크 기반의 문서의 분할을 보장해 준다.



(그림 2) 하이퍼링크 기술자

4. 특징 추출 및 분류기

관련 있는 문서 여부를 평가하는 분류기는 사용자가 초기 제공하는 문서들을 학습하는 사전 작업을 수행하게 된다. 본 논문에서는 가장 일반적으로 많이 사용되고 성능이 입증된 Naïve Bayes 확률 모델을 이용하였다. Naïve Bayes 분류기에서 사용하게 될 특징(feature)의 집합 즉 수집기가 수집하려는 문서들에서 발생하는 유용한 단어집합은 수집기의 성능과 직결되는 요소이다. 본 논문에서는 두 가지 방식으로 특징을 수집한다. 첫 번째 방법은 수집기의 학습단계에 주어지는 학습예제를 사용하여 특징을 추출한다. 이때의 특징추출 여부는 x^2 통계[9]를 사용한다. 두 번째 방법은 구글의 backlink 정보를 이용하여 x^2를 보완한다.

4.1 특징선택 - x^2통계 방법

분류기의 성능을 향상 시키기 위한 특징선택 방법으로 x^2 통계 방법을 이용하였다. x^2 통계 방법은 x^2 분포를 통해 단어(t)와 클래스(c) 사이의 독립성을 측정한다.

$$x^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

A는 클래스 c에서 단어 t가 나타난 문서 수, B는 클래스 c가 아닌 다른 클래스에서 단어 t가 나타난 문서 수, C는 클래스 c에서 단어 t없이 나타난 문서의 수, D는 클래스 c가 아닌 다른 클래스에서 단어 t가 나타나지 않는 문서 수, N는 전체 총 문서 수를 의미한다.

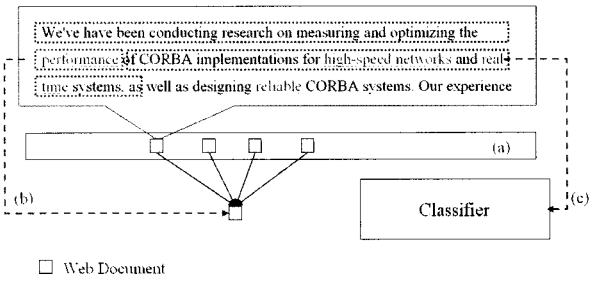
x^2 통계적 방법은 만약 단어 t와 클래스 c가 독립적이라 했을 때, "0"의 값이 나타나게 된다. 우리는 각 클래스 별로 단어 t에 대한 독립성을 계산하여 단어 t가 대변하는 클래스를 찾을 수 있게 하였다.

$$\chi^2_{\max}(t)=\max_{i=1}^{m}\{\chi^2(t,c)\}$$

4.2 구글(Google) 검색엔진을 통한 특징 구성

통계적 방법만으로 주어진 주제에 대한 특징집합의 추출은 문맥정보를 고려하지 않기 때문에 정확도가 떨어지는 경우가 발생하며 다양한 특징의 추출이 어렵다. 이러한 문제점을 해결하기 위해 본 논문에서 하이퍼링크 기술자는 하이퍼링크와 연관된 글이 기술되어 있다고 가정하고 구글 검색엔진의 backlink 정보를 이용하여 제공하는 학습 데이터 URL을 가리키는 부모 웹 문서를 찾아 부모 웹 문서상에서 학습 문서를 가리키는 하이퍼링크의 주변 단어들을 특징으로서 분류기에 반영한다.

(그림 3)에서 보는 것과 같이 (a)학습 문서를 링크하고 있는 웹 문서를 구글 검색엔진을 통해 찾고 (b)찾은 문서에



(그림 3) 구글을 통해 분류기를 개선하는 방식

서 학습 문서를 링크하여 기술하는 정보를 얻는다. (c)그 정보를 분류기에 특징으로 반영한다. 하이퍼링크를 평가하기 위해 하이퍼링크 기술자를 이용함에 있어서 구글 검색엔진을 통해 부모 문서의 하이퍼링크 기술자를 특징으로 이용함으로써 성능을 개선하고자 한다.

4.3 Naïve Bayes 분류기 및 파라미터 추정

Naïve Bayes 분류기는 문서에 대해 가장 적합한 클래스를 예측한다[10]. 확률모델에서 문서 d_j에 따른 클래스 c_i 의 유사도는 다음과 같이 측정된다.

$$P(c_i \mid \vec{d}_j) = \frac{P(c_i)P(\vec{d}_j \mid c_i)}{P(\vec{d}_j)}$$

Naïve Bayes 모델은 임의의 문서가 선택되는 사건과 그 문서에서 임의의 단어가 출현하는 사건이 독립적이라는 것과 k번째 단어가 출현하는 사건과 k' 번째 단어가 출현하는 사건은 서로 독립적이라는 가정이 있다.

$$P(\vec{d}_j \mid c_i) = \prod_{k=1}^{|T|} P(w_k \mid c_j)$$

여기서 w_k 는 이진 독립 변수로 0 또는 1의 값으로써 클래스 c_i 에서 나타났는지의 유무를 나타낸다.

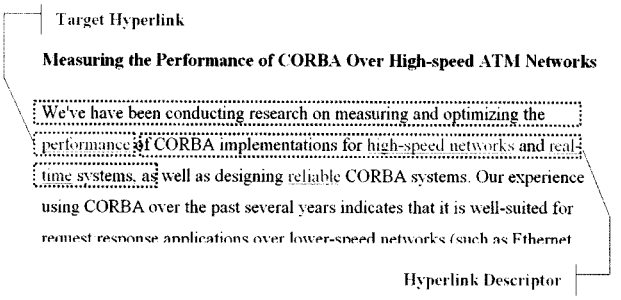
MLE(Maximum Likelihood Estimation)는 파라미터를 추정하는데 일반적으로 사용된다. 하지만 MLE는 학습데이터에 과잉종속(Overfitting)될 수 있으며 학습데이터의 희소성에 취약한 문제점을 가지고 있다. 이러한 문제점을 개선하기 위해 본 논문에서는 m-estimate 방법을 이용하였다[10].

$$\frac{n_c + t_p}{n + t}$$

n은 학습예제에서 출현한 feature의 총 빈도수를 나타내며, n_c는 클래스 c에서 특징이 출현한 빈도수를 나타낸다. t는 샘플의 크기를 나타내고 p는 샘플 크기에 대한 가중치를 나타낸다.

5. 웹 문서 수집기

수집할 주제에 대한 학습이 종료되면 주제 중심 웹 문서 수집기는 seed 페이지부터 문서 수집을 시작한다. 수집기는 seed 페이지를 각각의 주제 스택에 쌓는다. 관련문서를 나타내는 하이퍼링크만을 찾아내기 위해 수집기는 주제 스택에 존재하는 페이지를 방문한다. 수집기는 방문한 페이지를 하이퍼링크와 하이퍼링크 기술자 그리고 하이퍼텍스트로 구성되는 작은 문서의 집합으로 분리한다. 수집기의 문서분류기는 작은 문서 집합을 분류하여 각각의 하이퍼링크가 관련 문서와 연결되어 있는지 판단한다. (그림 4)는 하이퍼링크를 중심으로 하나의 문서를 여러 개의 작은 조각으로 분리하는 방법을 설명한다. 문서에 하이퍼링크가 발생하면 이를 중심



(그림 4) 하이퍼링크 평가 예제

으로 전후 n개의 토큰을 하나의 작은 문서로 분리한다. 본 논문에서는 n의 값으로 10을 사용한다. 이 값은 임의로 설정한 값으로 이 값은 변할 수 있다. 하이퍼링크 전후의 단어들은 문맥상 하이퍼링크와 연관된 글을 기술하고 있기 때문에 작은 단위로 나뉘어진 문서들은 보다 잘 하이퍼링크의 내용을 나타낼 수 있다. 문서의 관련성 여부는 시스템에 미리 설정된 임계치를 사용하여 임계치 이상인 하이퍼링크들만 각 주제에 맞는 주제 스택에 저장한다.

이러한 과정이 각 주제 스택에 더 이상의 URL이 존재하지 않을 때까지 반복적으로 이루어진다.

6. 실험 결과

본 논문에서는 각 주제의 학습 문서를 웹 문서로 제공하고, 수집기는 학습 후 각 주제에 적합한 문서를 찾아 수집하게 된다. 주제 중심 웹 문서 수집기는 크게 분류기 학습 과정과 웹 문서 수집 과정이 있고 그 처리 과정을 <표 1>에서 볼 수 있다.

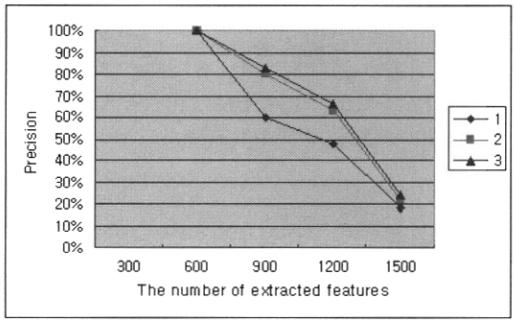
본 논문에서 하이퍼링크 주변에 링크와 관련 있는 글이 있다는 가정이 성립함을 보이기 위해 기존의 주제 중심 웹 문서 수집기와 본 논문에서 제시한 주제 중심 웹 문서 수집기, 구글을 통해 분류기를 개선한 주제 중심 웹 문서 수집기를 실험하였다.

〈표 1〉 주제 중심 웹 문서 수집기의 구동 방식

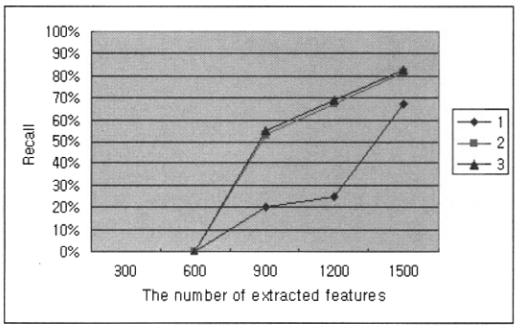
분류기 학습
1. 사용자가 제공한 문서를 기반으로 분류기 학습 1.1 구글을 통해 학습 데이터의 부모 URL 추출 1.2 부모 URL의 웹 문서 상에 학습 문서를 가리키는 하이퍼링크를 기준으로 좌우 10단어의 특징을 분류기에 반영
웹 문서 수집
2. 방문할 URL 목록에 seed URL 추가 3. 방문할 URL 목록에서 첫 번째 URL 획득 4. URL의 하이퍼링크 추출 하이퍼링크를 기준으로 좌,우 10단어와 링크의 단어들을 분류기로 평가 5. 연관성 있는 하이퍼링크로 평가되었을 때 수집기의 방문 할 URL 목록에 평가된 링크 URL을 추가 6. 고려중인 웹 문서가 관련 있는 문서인지 평가 여부에 따라 문서 수집 7. 방문 할 URL의 목록에서 첫 번째 URL 제거 8. 수집기가 방문할 URL의 목록에 URL이 남아있다면 3번으로 이동

〈표 2〉 분류기 학습 데이터

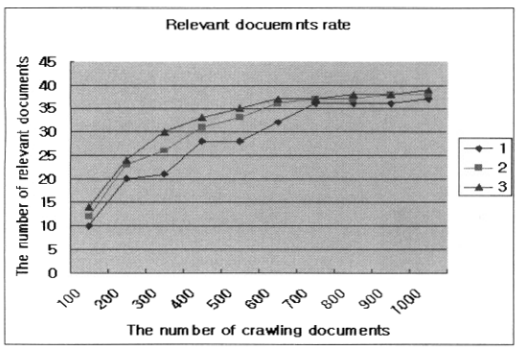
학습 데이터 URL
1. http://www.omg.org/gettingstarted/corbafaq.htm
2. http://www.cs.indiana.edu/~kksiazek/tuto.html
3. http://www2.rad.com/networks/1998/corba/corba_t.htm
4. http://holly.colostate.edu/~mabrake/
5. http://www-mtl.mit.edu/CAPAM/memos/96-7/node7.html
6. http://www-mtl.mit.edu/CAPAM/memos/96-7/node6.html
7. http://www.cs.mtu.edu/~yinma/study/Corba/Corba.html
8. http://www.cs.wustl.edu/~schmidt/corba-overview.html
9. http://www.lnu.edu.cn/corba/c1/c1.html
10. http://corba.ebi.ac.uk/intro.html
11. http://www.iiom.org/corba.htm
12. http://www.omg.org/gettingstarted/history_of_corba.htm



(그림 5) 실험결과: 특징 수에 따른 정확도



(그림 6) 실험결과: 특징 수에 따른 재현률



(그림 7) 수집한 문서에 따른 관련 있는 문서 수

본 논문에서 CORBA에 관련된 12개의 웹 문서를 분류기에 학습시켰다. 수집기의 시작 URL(Seed URL)은 "http://www.omg.org/gettingstarted/corbafaq.htm"으로 설정하였다. 수집결과 8822개의 문서를 CORBA의 관련문서로 수집하였으며 분류기의 기준값은 0.76으로 설정하였다.

(그림 5)와 (그림 6)은 특징의 수에 따른 정확도와 재현율을 나타낸다. 실험 결과에서 (1)은 기존의 주제 중심 웹 문서 수집기이고, (2)는 구글 검색엔진을 통해 분류기에 특징 정보를 추가하지 않은 웹 문서 수집기이다. (3)은 본 논문에서 제시한 수집기에 구글 검색엔진을 통한 부모 문서의 하이퍼링크 기술자를 분류기에 반영한 수집기이다.

(그림 5)에서 가로축은 특징의 수를 나타내며 세로축은 정확도를 나타낸다. 정확도의 성능을 보았을 때 기존 주제 중심 웹 문서 수집기보다 좋은 성능을 보여 주고 있다. 하이퍼링크를 문서 분할 방법을 통해 평가하여 수집 경로를 형성함으로써 연관성 있는 문서를 많이 수집할 수 있었다. 하지만 구글 검색엔진을 통해 분류기를 개선한 수집기와 개선하지 않은 수집기의 성능 차는 크게 나타나지 않았다. (그림 6)에서 가로축은 특징의 수를 나타내며 세로축은 재현율을 나타낸다. 기존의 수집기는 관련 없다고 판단된 문서에 대해서 하이퍼링크를 고려하지 않기 때문에 정보의 누수현상이 나타난다. 하지만 본 논문에서 제시한 주제 중심 수집기 및 구글을 이용해 특징정보를 추가한 수집기는 이러한 문제점을 개선하고 있음을 볼 수 있다.

(그림 7)은 수집기가 문서를 수집하면서 관련 있는 문서를 얼마나 많이 수집해 주는지를 보여준다. 1000개의 문서를 수집하는데, 관련 문서가 얼마나 빨리 수집했는지 보여주는 그래프이다. 구글 검색엔진을 통해 분류기에 특징정보를 반영한 웹 문서 수집기가 다른 웹 문서 수집기보다 관련 문서를 빠르고 많이 수집한 것을 알 수 있다.

실험결과를 통해 하이퍼링크에 관련된 글이 주변에 기술되어 있다는 가정을 실험을 통해 알 수 있었다. 하이퍼링크 기술자를 이용해 하이퍼링크를 평가 함으로써 수집 경로를 관련 있는 문서부터 빠르게 수집함을 보여 주는 예로, CORBA 관련 글 "http://www.omg.org/gettingstarted/specintro.htm" 문서를 기존의 웹 문서 수집기의 경우 7번째에 수집되었다면 구글 검색엔진을 이용하지 않은 수집기에서는 6번째에 수집되었고, 구글 검색엔진을 통해 분류기에 특징정보를 추가한 웹 문서 수집기는 4번째에 수집되었다. 결과적으로 문서 분할법을 통한 하이퍼링크 기술자는 하이퍼링크에 관해 기술하고 있고 개별적인 하이퍼링크 평가로 수집경로를 개선함으로써 웹 문서 수집기의 성능을 향상시킬 수 있었다.

6. 결 론

기존의 주제 중심 문서 수집기는 문서 수집시 (1)사용자 요구에 적합한 하이퍼링크를 가지고 있지만 관련 없는 문서로 평가되어 링크의 문서들이 수집되지 않는 정보의 누수현상과 (2)관련 문서로 평가되었지만 사용자 요구에 적합한 하이퍼링크가 없는 경우로 많은 데이터 처리로 인해 시스템 부하가 생기게 되는 문제점을 가지고 있다.

이러한 문제점을 개선하기 위해 본 논문에서는 문서의 하이퍼링크들을 개별적으로 평가하여 수집 경로를 개선함으로

써 위와 같은 문제점을 해결하였다. 하이퍼링크를 평가하기 위한 방법으로 하이퍼링크 주변의 단어들, 하이퍼링크 기술자를 평가 시 사용하였다.

결과적으로 기존 주제 중심 문서 수집기의 성능을 개선할 수 있었다. 좀더 좋은 주제 중심 문서 수집기를 만들기 위해서 우리는 분류기의 성능을 향상시킬 수 있는 방법을 찾고 하이퍼링크를 평가할 수 있는 다양한 평가함수를 실험을 통해 찾는다면 더욱 향상된 주제 중심 문서 수집기를 개발할 수 있을 것이다.

참 고 문 헌

[1] Matthew K Gray, "Measuring the Growth of the Web, June 1993 to June 1995," http://www.mit.edu/people/mkgray/growth.

[2] S. Chakrabarti, m. Ven den Berg And B.E. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," WWW-8. 1999.

[3] Paul De Bra, "Information Retrieval in Distributed Hypertexts," Proceeding of 4th RIAO Conference, 1994.

[4] M, Hersovici, "The SharkSearch Algorithm-An Application: Tailored Web Site Mapping," Proceeding of 8th Int'l World Wide Web conference, pp.213-225, 1998.

[5] J. Cho, "Efficient Crawling through URL ordering," Computer Networks and ISDN Systems, Vol.30, pp.161-172, 1998.

[6] M. Dologenti, "Focused Crawling Using Context Graphs," Proceeding of 26th Int'l conference, Vwey Large Data Bases, Morgan Kaufmann, pp.527-534, 2000.

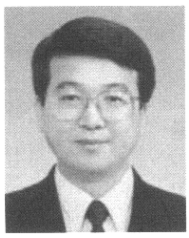
[7] A. McCallum, "Building Domain-Specific Search Engines with Machine Learning Techniques," Proceeding AAAI Symp. Intelligent Agents in Cyberspace, AAAI Press, pp.28-39, 1999.

[8] S. Mukherjea, "WTMS: A System for Collecting and Analyzing Topic-Specific WEB Information," Computer Networks, Vol.33, No.1-6, pp.457-471, 2000.

[9] Yiming Yang, Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proceedings of the Fourteenth International Conference on Machine Learning, pp.412-420, 1997.

[10] Tom Mitchell, Machine Learning, McGraw Hill, pp.154-199, 1998.

조 창 희

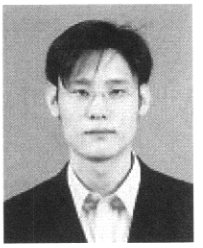


e-mail : lawworld@moleg.go.kr
1994년 홍익대학교 컴퓨터교육학과(석사)
2005년 숭실대학교 컴퓨터학과 박사수료
1985년~1989년 총무처 정부전자계산소 프로그래머 근무
1990년~현재 법제처 법령정보화 총괄담당 사무관
관심분야 : 법령정보데이터베이스, 공공기관 SW발주체계, 소프트웨어공학, 웹서비스, 소프트웨어개발방법론

이 남 용

e-mail : nylee@comp.ssu.ac.k
1979년 숭실대학교 컴퓨터학부(공학사)
1983년 고려대학교 경영정보학과(석사)
1993년 미국 미시시피주립대학 경영정보학과(경영학박사)
1979년~1983년 국군정보사령부 정보처 정보시스템분석 장교
1983년~1999년 한국국방연구원 군수체계 및 정보체계연구부장
1999년 연세대학교 경법대학 겸임교수
2000년 한국전자거래학회 논문편집위원장
2004년~현재 한국정보통신기술사협회 회장
1999년~현재 숭실대학교 컴퓨터학과 교수
관심분야 : 소프트웨어 테스팅, 품질보증, MIS, 정보보호, 시스템 엔지니어링, 소프트웨어 엔지니어링, MIS, CALS/EC, CORBA, Ada, 유비쿼터스, 전자상거래 등

강 진 범

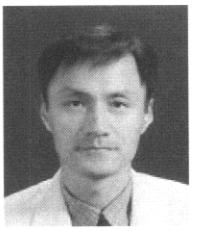


e-mail : jbkang@cse.hanyang.ac.kr
2003년~2004년 (주) BnGRotis
2004년 동명정보대학교 컴퓨터공학과(학사)
2005년~현재 한양대학교 석사과정
관심분야 : 인공지능, 에이전트, 데이터 마이닝

양 재 영

e-mail : isconan@naver.com
1998년 한양대학교 전자계산학과(학사)
2000년 한양대학교 대학원 전자계산학과 (석사)
2003년 한양대학교 대학원 컴퓨터공학과 (박사)
2003년~2005년 ㈜오픈베이스 책임연구원
2005년~현재 동부정보기술주식회사 컨설팅사업팀
관심분야 : 인공지능, 기계학습, 유비쿼터스 컴퓨팅

최 중 민



e-mail : jmchoi@cse.hanyang.ac.kr
1984년 서울대학교 컴퓨터공학과(학사)
1986년 서울대학교 대학원 컴퓨터공학과 (석사)
1993년 뉴욕주립대 (Buffalo) 전산학과 (박사)
1993년~1995년 전자통신연구원(ETRI) 선임연구원
1995년~현재 한양대학교 컴퓨터공학과 교수
관심분야 : 인공지능, 에이전트, 웹 정보처리, 시맨틱 웹