연관도를 계산하는 자동화된 주제 기반 웹 수집기☆

An Automated Topic Specific Web Crawler Calculating Degree of Relevance

서 혜 성* 최 영 수** 노 상 욱*** 최 경 희**** 정 기 현*****
Haesung Seo Youngsoo Choi Sanguk Noh Kyunghee Choi Gihyun Jung

요 약

인터넷을 사용하는 사람들에게 그들의 관심사와 부합하는 웹 페이지를 제공하는 것은 매우 중요하다. 이러한 관점에서 본 논문은 각 웹 페이지의 주제와 연관된 정도를 계산하여 웹 페이지 군(cluster)을 형성하며, 단어빈도/문서빈도, 엔트로피(entropy) 및 컴파일된 규칙을 이용하여 수집된 웹 페이지를 정제하는 주제 기반 웹 수집기를 제안한다. 실험을 통하여 주제 기반 웹 수집기에 대한 분류의 정확성, 수집의 효율성 및 수집의 일관성을 평가하였다. 첫째, C4.5, 역전파(back propagation) 및 CN2 기계학습 알고리즘으로 컴파일한 규칙을 이용하여 실험한 웹 수집기의 분류 성능은 CN2를 사용한 분류 성능이 가장 우수 하였으며, 둘째, 수집의 효율성을 측정하여 각 범주별로 최적의 주제 연관 정도에 대한 임계값을 도출할 수 있었다. 마지막으로, 제안한 수집기의 수집정도에 대한 일관성을 평가하기 위하여 서로 다른 시작 URL을 사용하여 수집된 웹 페이지들의 중첩정도를 측정하였다. 실험 결과에서 제안한 주제 기반 웹 수집기가 시작 URL에 큰 영향을 받지 않고 상당히 일관적인 수집을 수행함을 알 수 있었다.

Abstract

It is desirable if users surfing on the Internet could find Web pages related to their interests as closely as possible. Toward this ends, this paper presents a topic specific Web crawler computing the degree of relevance, collecting a cluster of pages given a specific topic, and refining the preliminary set of related web pages using term frequency/document frequency, entropy, and compiled rules. In the experiments, we tested our topic specific crawler in terms of the accuracy of its classification, crawling efficiency, and crawling consistency. First, the classification accuracy using the set of rules compiled by CN2 was the best, among those of C4.5 and back propagation learning algorithms. Second, we measured the classification efficiency to determine the best threshold value affecting the degree of relevance. In the third experiment, the consistency of our topic specific crawler was measured in terms of the number of the resulting URLs overlapped with different starting URLs. The experimental results imply that our topic specific crawler was fairly consistent, regardless of the starting URLs randomly chosen.

☞ Keyword : topic specific Web crawler (focused crawler), degree of relevance, Web page classification, machine learning, compiled rules

1. 서 론

인터넷의 사용이 증가함에 따라 웹 콘텐츠의 종류와 웹 페이지의 개수는 계속하여 증가하고 있으며, 방대한 양의 웹 정보에서 특정한 주제와 관련된 웹 페이지를 분류 또는 정제할 수 있는 기능은 인터넷에서 자신의 관심과 관련한 웹 페이지를 찾고자 하는 인터넷 사용자들을 위하여 필요한 요소임에 틀림없다. 일반적인 수집기와는 다르게 주제기반 웹 수집기(topic specific Web crawler or focused crawler) [1-5]는 전체 웹 정보에서 주어진 주제 또는 사용자의 관심과 관련한 일정한 세그먼트를 유지하고 관리한다. 따라서, 전체 웹의

* 준 회 원 : 아주대학교 정보통신전문대학원 정보통신공학과 retry@ajou.ac.kr
** 준 회 원 : 아주대학교 정보통신전문대학원 정보통신공학과 drabble@ajou.ac.kr
*** 정 회 원 : 가톨릭대학교 컴퓨터정보공학부 교수 sunoh@catholic.ac.kr(교신저자)
**** 정 회 원 : 아주대학교 정보통신전문대학원 정보통신공학과 교수 khchoi@ajou.ac.kr
***** 정 회 원 : 아주대학교 전자공학부 교수 khchung@ajou.ac.kr
[2005/12/19 투고 - 2006/01/09 심사 - 2006/04/12 심사완료]
☆ 본 연구는 2005년도 가톨릭대학교 교비연구비와 2006년도 가톨릭대학교 전공특성화 사업의 지원으로 이루어졌음.

정보를 수시로 갱신하고 관리해야 하는 부담을 줄일 수 있으며, 주제와 관련한 상대적으로 적은 양의 정보를 효율적으로 갱신하고 유지할 수 있을 것이다. 본 논문은 이러한 주제기반 웹 수집기를 제안하며, 기존의 주제기반 웹 수집기[1]와 다른 점은 다음과 같다: 제안하는 주제기반 웹 수집기는 구체적으로 (1) 특정한 주제와 연관된 정도(degree of relevance)를 정의하여 이러한 모델을 바탕으로 대략적인 웹 페이지 집단(cluster)을 형성하며, (2) 기계학습 알고리즘을 사용하여 자동적으로 생성한 웹 페이지 분류 규칙을 이용하여 웹 페이지 집단을 주제와의 관련성이 높도록 정제한다.

주어진 주제와 관련된 웹 페이지를 수집하기 위하여 주제와의 연관된 정도(이후에는 '연관도'라고 함)를 사용한다. 연관도는 웹 페이지가 주어진 주제와 얼마나 관련이 있는지를 나타내는 정도이다. 연관도의 계산을 위하여 미리 주제와 관련된 단어들의 집합을 정의한다. 이 단어들은 사전이나 주제와 관련된 문서 등을 참고로 선택된다. 연관도 값은 첫째, 미리 정의된 단어들 중에서 해당 웹 페이지에 나타난 단어들의 비율과, 둘째, 해당 웹 페이지를 가리키는 다른 웹 페이지의 연관도 값을 재귀적으로 고려하여 결정한다. 연관도에 의하여 수집된 웹 페이지 집단은 웹 페이지 분류기에 의하여 정교하게 분류될 수 있다. 웹 페이지의 내용을 뚜렷하게 식별하는 단어를 선택하기 위하여 웹 페이지에 나타난 단어들의 단어빈도/문서빈도 [6-8]의 값과 엔트로피 [9] 값을 계산한다. 그리고 선택된 단어 집합과 분류하고자 하는 웹 페이지에 포함된 단어들을 비교하여 주어진 웹 페이지를 적절한 범주로 분류한다. 웹 페이지와 미리 정의된 범주간의 관계를 식별하기 위하여 본 논문에서 제안하는 접근 방법은 기계학습에 의하여 컴파일된 단어기반 분류 규칙을 이용한다. 따라서, 제안하는 주제 기반 수집기는 연관도를 이용하여 특정 주제와 관련된 웹 페이지 군(cluster)을 수집하고, 컴파일된 분류규칙을 이용하여 수집된 웹 페이지 집단을 관련된 범주로 정제하여, 웹 페이지에 대한 분류의 정확성을 높일 수 있는 장점을 가진다.

본 논문의 구성은 다음과 같다. 2장은 주제 기반 수집기와 웹 페이지 분류에 관련된 연구들에 대하여 정리하며, 3장은 주어진 주제와 웹 페이지간의 연관도를 계산하는 방법과 웹 페이지를 식별하는 단어를 선택하기 위한 방법을 설명한다. 4장은 벤치마크 데이터를 이용하여 제안한 주제기반 웹 수집기를 실험한 결과를 보여준다. 결론에서, 본 논문의 내용을 정리하며, 앞으로의 연구 방향을 제시한다.

2. 관련연구

웹 수집기는 검색엔진(search engine)이 빠르게 WWW(World Wide Web) 상의 특정한 페이지를 방문할 수 있도록 하기 위하여 웹 페이지 내용에 대한 목록(indexing)을 구축한다. 웹 수집기에 의한 웹 페이지의 수집 과정은 모든 대상 웹 페이지 중에서 중요한 페이지라 판단할 수 있는 기준을 정의하고 기준을 만족하는 페이지들을 우선적으로 수집한다. 이때, 중요한 페이지들의 우선 순위는 넓이 우선(Breath-First), 역링크 계수(Backlink Count), 페이지 랭크(PageRank) [10] 등의 방법으로 결정되며, 임의의 도메인의 특성에 따라 각각의 방법에 의한 수집의 효율성이 다르게 나타난다. 이러한 기법들은 웹 수집기들이 웹 전체를 다운로드 받을 수 없다면, 인터넷 상의 임의의 집단을 수집하는 것보다 주어진 질의와 가장 관련 있는 웹 페이지들을 우선적으로 수집하는 것이 바람직하다는 사실을 제공한다.

위에서 설명한 일반적인 웹 수집기와는 달리 특정 주제와 관련된 웹 페이지만을 수집하는 주제 기반 수집기(topic-specific crawler or focused

1) 현재 운영중인 웹 수집기(웹 로봇 또는 스파이더 등으로 불림)의 이름, 유형 및 관련정보는 http://www.robotstxt.org/wc/active/html/index.htm 를 참조할 수 있다.

crawler) [1-5]에 대한 연구도 진행되어 왔다. Chakrabarti, van den Berg와 Dom [1] 등이 제안한 주제 중심적 수집기는 특정 주제와 관련된 단어를 이용하지 않고 표본이 되는 문서들을 기초로 하여 수집한 웹 페이지 그룹을 구성한다. 주제 기반 웹 수집기는 분류기(classifier), 증류기(distiller), 수집기(crawler)의 세가지 중요한 모듈로 동작한다. 분류기는 링크를 확장하기 위하여 수집된 페이지들에 대한 관련성을 판정하며, 증류기는 방문의 우선순위를 결정하기 위하여 수집된 페이지들이 주어진 주제와 얼마나 근접하는가를 측정한다. 또한, 수집기는 분류기 모듈과 증류기 모듈에 의하여 결정된 우선순위에 따라 페이지들을 수집한다. 주제 기반 웹 수집기에 대한 연구에서 가장 중요한 문제는 특정한 웹 페이지를 실질적으로 다운로드 하기 전에 주제와 관련된 정도를 예측할 수 있어야 한다는 것이다. 기존의 주제 기반 웹 수집기는 URL을 수집하여 사전에 구축된 분류 트리 중 URL이 속하는 노드를 계산한 후, 주제와 관련이 있는 노드일 경우에만 URL에서 링크를 추출한다. 이 때, URL 간의 링크 정보는 이용하지 않는다. 반면에 본 논문에서 제안하는 연관도의 계산기법은 URL 간의 링크 정보가 어느 정도 연관성이 있는가를 상대적으로 비교하여 이용한다는 점에 그 차이가 있다.

Diligenti 등은 참고문헌 [2] 에서 문맥 그래프(Context Graphs)를 이용한 주제 중심적 수집기를 제안하였다. 수집기가 수집을 하면서 주제와의 관련성을 측정하고, 아직 방문하지 않은 페이지를 수집하기 위한 경로를 정하는데 도움을 주기 위하여 문맥 그래프를 이용하는 것이다. 문맥 그래프는 중심에 목표 웹 페이지들이 위치하고 중심 페이지로부터의 링크 거리에 따라서 바깥쪽으로 층 (layer)이 나누어져 있다. 문맥 그래프는 시작 문서로부터 역 방향 수집(backward crawling)을 하여 층을 구축한 후, 각 층에 속한 페이지의 단어들의 단어빈도/역문서빈도(TF-IDF)를 수정하여 색인하는 방식으로 분류기를 학습시킨다. 수집된 페이지들은 분류기에 의하여 여러 층 중의 하나에 속하게 되며, 수집기는 중심에서 가까운 페이지들을 선택함으로써 주제와 관련이 높은 페이지들을 우선적으로 검색한다.

기존의 주제 기반 수집기에 대한 연구에서 주목할 점은 아직 방문하지 않은 웹 페이지들이 특정한 주제와 어느 정도 연관이 있는가를 식별하기 위하여 해당 웹 페이지의 내용만을 이용하여 분류를 수행한다는 것이며, 웹 페이지간의 링크 정보는 사용하지 않았다는 것이다. 반면에, 본 논문은 주어진 주제와의 연관도(degree of relevance)를 해당 웹 페이지와 그 페이지를 링크하고 있는 페이지의 링크 정보를 함께 사용하여 정의함으로써 특정 주제와 관련되었을 가능성이 높은 웹 페이지들을 우선적으로 수집할 수 있도록 하였다. 또한, 기존의 주제 기반 수집기는 수집과 분류가 동시에 이루어지는 반면, 본 논문에서 제안하는 주제 기반 수집기의 구조는 수집단계와 분류단계를 분리하여 설계하였다. 이와 같은 구조는 분류단계에서 수행하는 단어빈도/문서빈도 및 엔트로피 계산 이전에 수집단계에서 실시간으로 주제와 관련성이 있는 웹 페이지의 수집을 가능하게 한다. 다시 말하면, 네트워크 자원에 의존적인 수집과정과 이후에 집중적인 계산을 필요로 하는 분류과정을 분리 함으로써 각각의 구조적 특성에 적합한 효율적인 수집이 가능하도록 하였다.

웹 페이지의 분류를 위하여 대부분의 접근 방법들은 기계학습 알고리즘을 이용하였다 [7,11-14]. Sebastiani [12] 는 문서 범주화(또는 문서 분류)를 위한 기계학습 알고리즘의 역할 및 기계학습 알고리즘에 입력되는 특성들의 차원을 감소시킬 수 있는 전략의 필요성에 대하여 언급하였다. 동일한 시각에서 Noh [7] 등은 경험적으로 선택된 특성들이 웹 페이지를 정확하게 분류할 수 있을 뿐만 아니라 각각의 범주를 특징적으로 표현할 수 있어야 함을 강조하였다. 특성들의 크기를 합리적으로 결정하기 위하여, Ruger와 Gauch [11]

는 특정한 단어들의 문서 빈도수에 대한 최소 한계와 최대 한계 값을 설정하여 구간 내에 포함되는 단어들만을 선택하였다. 그러나, 단어들의 문서 빈도수에 의존하는 방법론은 제한된 수의 문서를 유지하는 경우에 선택된 단어들이 특정한 범주를 대표하지 못하는 단점을 가진다. 반면에, 본 논문에서 제안하는 방법론은 단어빈도/문서빈도 뿐만 아니라 엔트로피까지 고려하여 특정한 범주를 나타내는 단어들을 선택하였다. 지금까지 수행한 기초 연구 [7] 를 하나의 모듈로 구축하여 본 논문의 주제 기반 수집기의 분류단계로 확장하였다.

다양한 방식으로 제안된 주제 기반 수집기의 성능을 평가하기 위하여 다음과 같은 요소들이 측정되었다: (1) 수집한 웹 페이지 군에서 주제와 연관된 페이지의 비율(수확비율: harvest ratio [1, 3]), (2) 서로 다른 초기 URL 에서 출발하여 일정 갯수의 웹 페이지를 수집했을 때 URL 들이 중첩된 비율(수집의 일관성: URL overlapping or crawling robustness [1]), (3) 지정된 목표 웹 페이지를 모두 찾기 위해 수집한 웹 페이지의 갯수(탐색 길이: search length [1,4]), 및 (4) 수집된 페이지들의 평판 정도(수집의 정확성: hubs and authorities [15,16]) 와 같은 요소들에 대한 측정이 이루어졌다. 본 논문에서 제안한 주제 기반 수집기의 성능을 평가하기 위하여 위와 같은 측정 요소들을 다음과 같이 세가지로 정리하였다. 실험 및 평가에서 첫째, 주제의 분류 정확도, 둘째, 주제와 연관된 웹 페이지의 비율, 셋째, 서로 다른 초기 URL 로 시작하여 수집된 웹 페이지의 중첩 비율 정도를 측정하여 제안한 기법의 성능을 검증하고자 한다.

3. 지능적인 주제 기반 웹 수집기

특정한 웹 페이지가 주어진 주제와 어느 정도 연관성이 있는가를 나타내는 연관도(degree of relevance)를 정의하고, 이를 웹 페이지의 수집 전략으로 사용하는 알고리즘을 소개한다. 또한, 분류단계에서의 단어빈도/문서빈도와 엔트로피를 결합한 분류전략을 설명한다.

3.1 수집 전략

주어진 주제와 관련된 웹 페이지를 수집하기 위하여 웹 페이지가 주제와 얼마나 관련이 있는지를 나타내는 주제와의 연관된 정도, 즉, 연관도를 정의한다. 연관도의 계산을 위하여 주어진 웹 페이지를 링크하고 있는 다른 웹 페이지의 연관도 값과 주어진 웹 페이지의 단어 중에서 사전에 정의한 단어 집합과 일치하는 단어 개수의 비율을 고려한다. 웹 페이지 i 의 연관도 R_i 는 다음과 같이 계산된다.

$$R_i = \frac{(1-\rho)\,\lambda_i}{|K| + \rho R_j} \qquad (1)$$

이때,

- R_j 는 웹 페이지 j 의 관련도를 나타내며, 웹 페이지 j 는 수집기가 이미 수집한 웹 페이지로써 웹 페이지 i 의 URL 을 포함한다;
- ρ 는 R_i 가 R_j 에 의하여 얼마나 영향을 받는가를 나타내는 상수 값 ($0 < \rho < 1$);
- λ_i 는 웹 페이지 i 에 나타난 단어 중 미리 정의된 단어와 일치하는 단어의 개수;
- K 는 주어진 주제와 부합하는 미리 정의된 단어의 집합. $|K|$ 는 집합 K의 원소의 개수.

주제 기반 수집기는 주어진 주제와 연관될 확률이 많은 URL을 선택하기 위하여 삽입(enqueue)과 삭제(dequeue)의 두 동작을 지원하는 우선순위 큐(priority queue)를 제공한다. 그림 1은 주제 기반 수집기의 알고리즘을 나타낸다.

주어진 주제를 대표적으로 나타내는 웹 페이지를 가리키는 시작 URL(seed URL 또는 starting URL)이 있다고 가정하자. 이러한 URL들은 수집

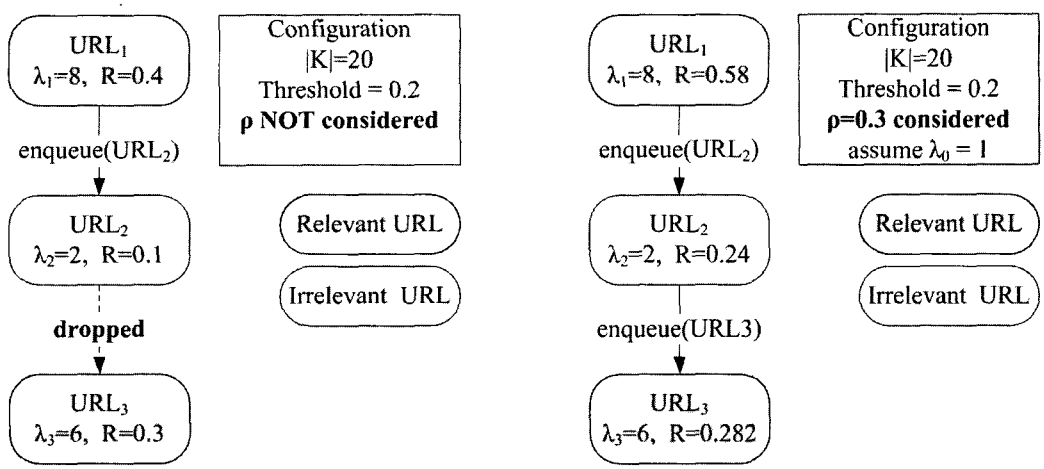
```
ITSC(topic, starting_urls, pre-defined keywords) {
    for each url in starting_urls {
        put (url, MAX_VALUE_OF_R) into todo_q;
    }
    while (! is_empty(todo_q) || visited < MAX_PAGES) {
        get url from todo_q;

        doc = fetch(url);
        R = degree_of_relevance(doc);

        for each hyper_link in doc {
            if(hostname_of(hyper_link) != hostname_of(url)) {
            put (hyper_link, R) into todo_q;
        }
        reorder(todo_q);
        put url into done_q;
    }

    for each url in done_q {
        classify(url);
    }
}
```

〈그림 1〉 주제 기반 수집기의 수집 알고리즘.

기의 출발 점으로, R_i 값은 1로 설정되어 큐[2](queue)에 삽입된다. 수집기는 큐로부터 이 URL을 꺼내서, URL에 해당하는 웹 페이지를 서버로부터 가져온다(fetch(url)). 가져온 웹 페이지의 주제와의 연관도 R_i를 계산하고, 웹 페이지가 포함하는 URL중 웹 페이지와 다른 호스트 이름을 가지는 URL만을 선택하여 웹 페이지의 R_i 값을 부여하여 다시 큐(todo_q)에 입력한다. 큐는 URL들의 R_i 값으로 정렬되어 있으며, 수집기는 사용자가 정한 개수의 URL을 가져오거나, 큐에 있는 모든 웹 페이지들이 처리될 때까지 위 작업을 반복한다.

R_i 의 값이 해당 웹 페이지와 주제와의 연관된 정도를 잘 반영하기 위해서는 K, ρ 의 선택이 중요하다. 본 논문에서 K에 속하는 단어들은 사전(dictionary)과 출발점으로 주어지는 웹 페이지를 분석하여 결정하며, ρ 값의 결정은 경험적인 분석을 통하여 이루어진다. 그림 2 는 ρ 값을 고려하지 않고 R_i 값을 계산한 경우와 ρ 값을 고려하여 R_i 를 계산한 경우 – 즉, 자신을 가리키던 웹 페이지의 연관도를 포함하였을 경우 – 의 차이를 보여준다.

그림 2의 (a)는 R_i를 계산할 때 ρ 를 고려하지 않은 경우이고, (b)는 ρ 를 0.3으로 고려하여 웹 페이지를 수집하는 경우이다. URL_i는 각각 웹 페이지 i 의 URL을 나타낸다. 먼저 (a)에서는 웹 페이지 1에서 나타난 단어들 중 미리 정의된 단어 집합 K에 포함된 단어의 개수는 8개 이며, ρ 를 고려하지 않았으므로 R 은 0.4 이다. 또한, 웹 페이지 2에서도 ρ 를 고려하지 않았으므로 웹 페이지 1의 R 값과 관계없이 웹 페이지 2는 자신의 페이지 안에서 나타난 단어의 개수만을 가지고 R



(a) ρ 를 고려하지 않은 경우 (b) ρ=0.3 인 경우

〈그림 2〉 연관도 식에서 ρ 값의 역할.

2) 그림 1의 알고리즘에서 'todo_q' 에 해당한다.

이 0.1로 결정된다. 따라서 웹 페이지 2의 R 이 임계치인 0.2를 넘지 못하므로 웹 페이지 2가 가리키고 있던 URL_3은 실제로 주제와 부합되는 페이지이지만 수집되지 않는다.

반면에 그림 2의 (b)에서는 웹 페이지 1에서 나타난 단어들 중에서 K에 포함된 단어의 개수는 왼쪽 설정 (a)에서와 같이 8개 이지만, ρ 를 포함하여 웹 페이지를 수집하므로 식 (1)에 의하여 R 은 0.58이 된다. 웹 페이지 2에서도 (a)에서와 같은 개수의 단어가 나타났지만, ρ 를 고려한 식 (1)에 의하여 R 은 0.24의 값을 갖게 된다. 따라서 웹 페이지 2의 R 은 임계치인 0.2를 넘게 되어 웹 페이지 2가 가리키고 있던 주제와 부합하는 URL_3이 수집된다.

3.2 웹 페이지 분류 전략

주제와의 연관된 정도를 계산하여 생성한 웹 페이지 군(cluster)을 주제와의 관련성을 중심으로 정제하는 자동화된 웹 페이지 분류 알고리즘 [7] 을 제안한다. 웹 페이지를 분류하는 분류 규칙은 각 범주를 특징적으로 식별(또는 분류)할 수 있는 단어를 기계학습 알고리즘에 의하여 컴파일함으로써 얻어진다. 이때, 각 범주를 특징적으로 나타내는 단어는 구체적으로 각각의 범주 특성을 대표하는 단어이며, 다른 범주와 뚜렷이 구분하는 단어를 의미한다.

특정 범주에 해당하는 웹 페이지들의 집합을 J, $J = \{1, 2, ..., n\}$, 로 정의하고, 각 범주에 해당하는 웹 페이지들의 집합에 나오는 단어의 집합을 I, $I = \{1, 2, ..., m\}$, 로 가정하자. 각 문서에 나타나는 단어의 가중치는 정보 검색에서 사용되는 단어 빈도(term frequency)와 문서 빈도(document frequency)의 개념을 이용하여 아래와 같이 결정된다.

$$W_{i,j} = \frac{TF_{i,j}}{\max_{k \in I} TF_{k,j}} \times \frac{DF_i}{n} \qquad (2)$$

이때,

- $W_{i,j}$ 는 웹 페이지 $j \in J$ 에 속하는 단어 $i \in I$ 의 가중치;
- $TF_{i,j}$ 는 웹 페이지 $j \in J$ 에 속하는 단어 $i \in I$ 의 문서빈도;
- DF_i 는 단어 $i \in I$ 가 나타난 웹 페이지의 개수;
- n 은 특정 범주에 속하는 웹 페이지의 전체 개수.

위 식 (2)는 특정 단어의 가중치는 각각의 페이지에서 나타나는 비율과 비례하고 또한 그 단어가 나타나는 웹 페이지의 개수의 비율에 비례 한다는 것을 의미한다. 그러므로 특정 범주의 각 페이지에서 많이 언급되고 또한 다수의 웹 페이지들에서 자주 출현 할수록 그 단어는 특정 범주를 대표하는 단어가 된다는 직관을 표현한다. 이렇게 선택된 단어들은 범주의 계층을 대표적으로 표현하는 단어가 될 것이다.

범주를 대표할 만한 단어들을 선택한 후, 웹 페이지를 주어진 분류 범주 중 하나에 대입하기 위하여 하나의 범주와 다른 범주를 뚜렷이 구별할 수 있는 단어를 선택해야 한다. 특징적인 단어의 선택은 분류하여야 할 범주들이 서로 유사한 경우에 더욱 중요하다. 왜냐하면, 유사 범주의 경우에는 사용되는 용어나 표현 양식이 비슷하기 때문에 식 (2)에 의하여 선택된 단어들이 중복될 가능성이 많기 때문이다. 따라서, 본 논문은 각 단어의 엔트로피(entropy)를 이용하여 각 단어가 주어진 범주를 얼마나 잘 구별할 수 있는가를 계산한다. 단어의 엔트로피는 정확한 분류를 위하여 필요한 정보량의 기대 값을 제공한다. 엔트로피가 낮을수록 주어진 웹 페이지를 임의의 범주로 대입하는데 필요한 정보량이 작음을 뜻한다. 그러므로 효율적인 분류를 위하여 엔트로피가 작은 단어들을 우선적으로 이용한다. 단어의 엔트로피[3]는 다음 식을 이용하여 결정하였다.

$$E_S = -\sum_{i=1}^{n} p_i \ln p_i \quad (3)$$

이때,

- S는 n개의 범주로 분류될 웹 페이지의 집합이다;
- p_i는 S의 임의의 웹 페이지가 범주 i로 분류될 확률을 의미하며, 집합 S의 웹 페이지 중에서 범주 i에 속하는 웹 페이지의 비율로 계산된다;
- E_S는 집합 S의 엔트로피를 나타내며, S의 모든 웹 페이지를 분류하는데 필요한 정보량의 기대값을 의미한다.

예를 들어, 집합 S의 웹 페이지가 긍정과 부정의 두 범주로 구성된다고 하자. 그러면 p_1은 집합 S의 전체 웹 페이지의 개수에 대한 긍정 범주에 속하는 웹 페이지의 개수의 비율이 되며, p_2 또한 같은 방법으로 계산된다. 그러므로 집합 S의 엔트로피 값은 $-(p_1 \ln p_1 + p_2 \ln p_2)$ 와 같다.

위의 식 (3)으로 정리한 엔트로피는 집합의 속성 정보를 전혀 이용하지 않은 값이며, 속성 정보를 이용하면 집합 S를 특정한 범주들로 분류하는데 필요한 정보량을 줄일 수 있다. 이 경우 각 속성 정보가 줄여주는 엔트로피 양을 정보 이득(Information Gain)이라고 하며 아래와 같이 계산한다.

$$Gain(\alpha) = E_S - \sum_{j=1}^{m} \left(\frac{|S_{\alpha_j}|}{|S|} \times E_{S_{\alpha_j}} \right) \quad (4)$$

이때,

- $S_{\alpha_j}, S_{\alpha_j} \subset S$, 는 속성 α의 값이 $j \in \{1, \ldots, m\}$인 웹 페이지의 집합을 나타낸다. $|S_{\alpha_j}|$는 집합 S_{α_j}에 속하는 웹 페이지의 총 개수이다.

본 논문에서 엔트로피의 속성정보는 단어를 나타내며, 각각의 속성이 가지는 정보 이득 값은 분류에 사용되는 각 단어들의 우선순위를 결정한다. 따라서, 각 단어들의 정보 이득 값이 높으면 우선적으로 선택되어진다.

웹 페이지를 분류하기 위하여 여러 개의 속성을 하나의 튜플로 조합시킬 필요가 있다. A를 속성들의 집합이라고 한다면, 속성의 선택은 다음과 같이 식 (5)를 적용하여 수행한다.

$$H_\mu = \{\alpha \mid \alpha \in A, Gain(\alpha) \geq \mu\} \quad (5)$$

임계값 μ는 각각의 범주에 있는 단어들 중 상대적으로 중요하지 않은 단어들을 걸러주는 역할을 한다. 위의 식을 수행한 결과로 만들어진 집합 H_μ는 μ가 높을수록 해당 범주와 관련이 깊은 속성들의 집합이 된다.

지금까지 정의한 용어를 바탕으로 전체적인 웹 페이지 정제과정은 다음과 같다. 첫째로, 단어빈도/문서빈도를 계산하여 해당 범주를 대표할 수 있는 단어들을 선정한다. 둘째로, 첫번째 과정을 통하여 선정된 단어들의 정보이득 값을 계산하여 속성들의 우선순위를 정하고, 하나의 튜플로 구성한다. 세째로, 구성된 튜플을 다양한 기계학습 알고리즘 [17-20] 에 입력하여, 컴파일된 분류 규칙 집합을 생성한다. 이러한 과정을 거쳐 생성된 규칙은 임의의 웹 페이지를 주어진 범주 중의 하나로 분류할 수 있도록 한다.

4. 실험 및 평가

관련연구에서도 설명한 바와 같이 주제 기반 웹 수집기의 성능을 평가하기 위한 다양한 측정 요소들이 제안되어 왔다. 본 논문에서는 이와 같은 측정 요소 중에서 공통적으로 사용된 (1) 분류

3) 참고문헌 [9] 에서 제안한 엔트로피를 웹 페이지의 분류에 적용하기 위한 형태로 작성한 것이다.

의 정확도, (2) 수집의 효율성, (3) 수집의 일관성이라는 세가지 관점에서 제안하는 주제 기반 웹 수집기의 성능을 평가하고자 한다. 첫번째, 분류의 정확도는 도메인 전문가에 의하여 사전에 준비된 데이터 집합을 기계학습 알고리즘에 의하여 컴파일된 규칙으로 분류하는 경우, 분류 규칙이 얼마나 정확하게 데이터를 분류하는지를 나타낸다. 두번째, 수집의 효율성(또는 집중도)은 주제 기반 웹 수집기가 수집한 페이지를 분류하여 주제와 관련한 범주와 주제와 관련하지 않은 범주로 나눌 때, 전체 수집된 문서 중에서 주제와 관련한 범주에 속하는 웹 페이지의 비율이 어느 정도 인지를 나타낸다. 세번째, 수집의 일관성은 다양한 시작 URL(seed URL)을 사용하여 수집기를 실행한 후, 결과적으로 분류된 웹 페이지들이 중복되는 비율을 나타낸다. 중복되는 비율이 높을수록 수집기는 시작 URL에 적게 영향을 받음을 알 수 있으며, 수집 결과에 대한 일관성 또한 보장된다.

주제 기반 웹 수집기의 검증에 사용된 평가 문서 집합은 Sinka [21] 등이 제공하는 데이터 집합을 이용하였다. Sinka의 데이터 집합은 'Banking and Finance', 'Programming Language', 'Science'와 'Sport'의 네 개의 상위 범주로 구성된다. 'Banking and Finance'는 'Commercial Bank', 'Building Society'와 'Insurance Agency'의 세 개의 하위 범주로 구성되며, 'Programming Language'는 'Java', 'C/C++'와 'Visual Basic'의 하위 범주로, 'Science'는 'Astronomy'와 'Biology'의 하위 범주로, 'Sport'는 'Soccer', 'Motor Sport'와 'Sport'의 하위 범주로 구성된다.

4.1 분류의 정확도

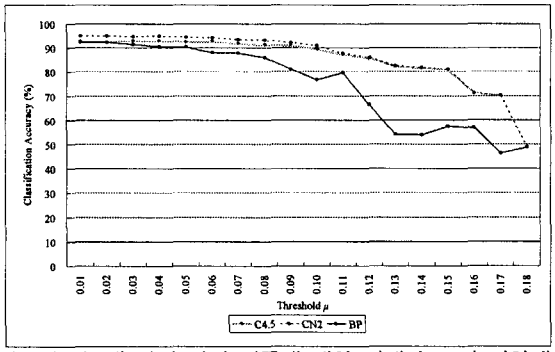
첫 번째 실험인 분류의 정확성과 관련하여 제안된 분류 방식이 다양한 수준의 범주들에 대하여 어느 정도의 정확도를 나타내는가를 측정하였다. 이를 위하여 네 개의 상위 범주에 해당하는 문서들을 이용하여 분류 정확성을 측정하였으며, 또한 각각의 유사범주에 속하는 문서들을 이용하여 하위범주에 대한 분류 정확성을 측정하였다.

네 개의 상위 범주의 분류 정확성을 측정하기 위하여 각 범주에서 300, 300, 200, 300개의 문서가 임의로 선택되었다. 선택된 문서에 대하여 식 (2)와 (4)를 적용하여, 범주의 대표성을 가진 단어들의 집합과 각각의 범주를 특징적으로 구별하는 단어들의 집합을 결정하였다. 또한, 실험에서 선택된 웹 페이지들은 평균 277개의 단어를 가지고 있었다.

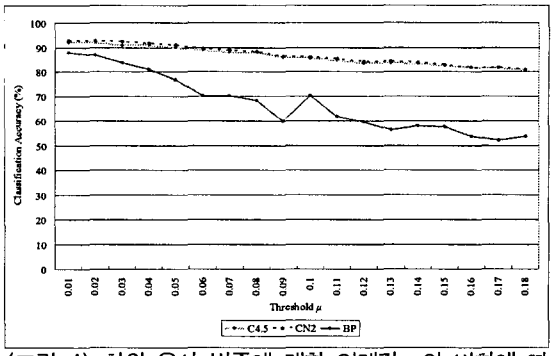
기계학습 알고리즘의 입력으로 사용할 단어를 선택하기 위하여 식 (5)를 선택된 단어 집합에 적용하여 임계값 μ보다 큰 단어들을 선택하였다. 실험에서 임계값을 0.01부터 0.18까지 0.01 간격으로 변환하면서 선택한 단어들을 검증하기 위하여 아직 사용하지 않은 9,900개의 문서에 적용하여 분류 정확성을 측정하였다. 각각의 임계값 μ으로 선택된 튜플의 집합은 C4.5 [19], CN2 [17] 및 역전파(back propagation) [20] 기계학습 알고리즘에 의하여 컴파일되어 분류 규칙을 생성하였다. 세가지 기계학습 알고리즘에 의하여 컴파일된 분류 규칙들의 성능은 그림 3에 나타내었다.

그림 3에서 임계값 μ가 감소하면, 사용된 단어의 개수가 늘어나지만 분류 성능은 좋아짐을 알 수 있다. 이러한 결과의 관찰을 통하여 사용된 단어의 개수와 분류 성능 사이에 최적화된 조합이 필요함을 알 수 있다. 왜냐하면, 사용된 단어가 증가할수록 분류를 위한 계산량이 증가하며, 증가하는 계산량에 비하여 분류 성능의 향상 정도가 일정 수준 이하의 임계값에서는 크게 향상되지 않기 때문이다. 따라서, 주제 기반 웹 수집기의 분류 성능이 90.2% 일 때 사용된 77개의 단어를 선택하여 범주의 계층 구조를 구성하였다.

둘째로, 각 범주 내의 하위 범주들에 대한 분류 정확성을 측정하였다. 실험환경은 위의 실험과 동일한 방식으로 진행되었다. 각 범주 별로 100개의 문서를 임의로 선택하여 단어를 선별하고



〈그림 3〉 네 가지 상위 범주에 대한 임계값 μ 의 변화에 따른 분류 성능 비교.



〈그림 4〉 하위 유사 범주에 대한 임계값 μ의 변화에 따른 평균 분류 성능 비교

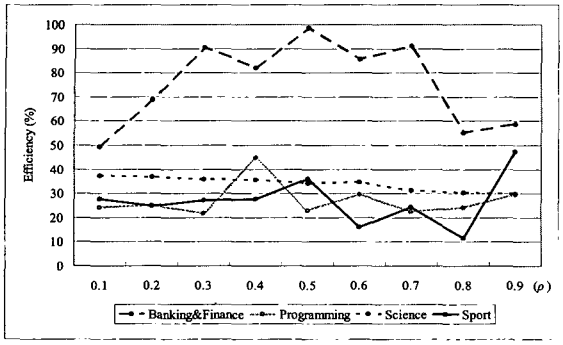
범주 분류 규칙을 생성한 후, 분류 규칙이 나머지 문서들을 각각의 범주로 올바르게 분류한 비율을 측정하였으며, 실험결과는 그림 4에 나타내었다.

그림 4에 나타난 바와 같이 하위 유사 범주에 대한 분류 결과는 C4.5와 CN2 알고리즘으로 컴파일한 분류 규칙을 사용한 경우, 모든 임계값 범위에서 80% 이상의 분류 성능을 보였다. 역전파 알고리즘으로 컴파일한 분류 규칙을 사용한 경우에는 임계값이 적으면 – 분류를 위하여 상대적으로 많은 수의 단어 사용 – 분류 성능이 좋았지만, 임계값이 커질수록 분류의 정확도가 결정트리 방식(C4.5와 CN2)보다 낮게 나타남을 알 수 있었다. 결과적으로 그림 3과 그림 4에 나타난 분류의 정확도를 종합하면, 제안된 주제 기반 웹 수집기가 다양한 레벨의 범주(상위 개념 및 하위 유사 개념)들에 대하여 일관적인 분류 성능을 보장함을 알 수 있었다.

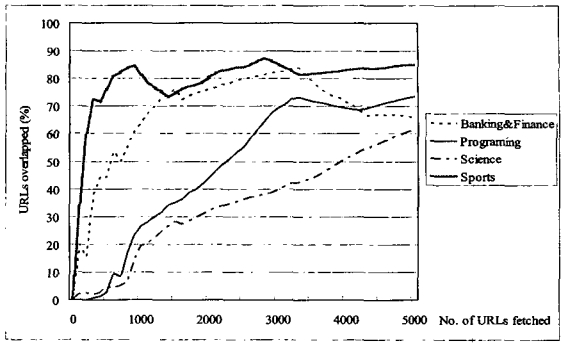
4.2 수집의 효율성

본 논문에서 제안한 주제기반 웹 수집기의 수집에 대한 효율성은 (주어진 주제와 관련된 URL의 개수 ÷ 수집한 URL의 총 개수) 로 정의한다. 수집에 대한 효율성은 각 URL의 주제와의 관련도 R_i값에 영향을 주는 ρ의 최적값을 구하기 위하여 측정된다. 분류의 정확성 실험에서 사용한 단어 중 임의의 20개 단어를 미리 정의된 단어 집합 K(식 (1) 참조)로 선택하였으며, 각 범주에서 임의의 URL을 10개씩 추출하여 시작 URL로 선택하였다. 위의 4.1 절에서 수행한 분류 정확도 실험 결과(그림 3 참조)에서 CN2의 분류 정확도가 가장 좋았기 때문에, CN2를 이용하여 범주 분류 규칙을 컴파일하였다. 생성된 분류 규칙을 이용하여 ρ값을 0.1부터 0.9까지 0.1 간격으로 변화시키면서 10,000개의 웹 페이지를 수집하였으며, 이에 대한 수집의 효율성을 측정하였다. 제안한 주제 기반 웹 수집기의 수집의 효율성에 대한 실험 결과는 그림 5와 같다.

주제와의 연관도(식 (1) 참조)를 계산할 때, 웹 페이지 상호간의 영향력을 결정하는 ρ가 각각 0.1과 0.9일 경우의 수집의 효율성은 네 가지 범주 공통적으로 비교적 좋지 않았다. ρ가 0.1일 때, 웹 페이지 i의 연관도 R_i는 이 웹 페이지를 포함하는 웹 페이지 j에 영향을 거의 받지 않아서 실제로 웹 페이지 i가 주제와 관련이 높다 하더라도 다른 웹 페이지보다 우선적으로 수집되지 않을 수 있다. 또한, ρ가 0.9일 때는 웹 페이지 i가 이 웹 페이지를 포함하는 웹 페이지 j의 영향을 절대적으로 받기 때문에 실제로 웹 페이지 i가 주제와 관련이 적다 하더라도 다른 웹 페이지 보다 우선적으로 수집될 수 있는 가능성이 존재한다. 그림 5에서 'Banking and Finance'에 대한 최적의 ρ값



〈그림 5〉 네 가지 상위 범주에 대한 ρ값의 변화에 따른 수집 효율성 비교



〈그림 6〉 네 개의 상위 범주에 대한 수집의 일관성 비교. 각각의 수집기는 5,000개의 URL을 수집한 후 수집을 중단 함.

은 0.5이었으며, 수집의 효율성은 98% 로 측정되었다. 또한 'Programming Language', 'Science' 및 'Sport'에 대한 최적의 ρ값은 각각 0.4, 0.1, 0.5로 나타났으며, 이때 수집의 효율성은 각각 44%, 37%, 36% 로 측정되었다.

4.3 수집의 일관성

주제 기반 웹 수집기의 수집의 일관성은 URL 중첩정도 [1] 를 측정하여 평가하였다. 이 방법은 서로 다른 URL을 시작 URL로 하여 두 개의 수집기를 실행하여 동일한 개수의 URL을 수집한 후, 수집된 두 URL집합을 비교하여 중복되는 비율을 측정하는 것이다. 위의 4.1절과 4.2절의 실험결과를 바탕으로, 주제 기반 웹 수집기는 분류 정확도(그림 3 참조) 실험에서 가장 좋은 성능을 보여준 CN2 알고리즘에 의하여 컴파일된 분류 규칙을 내장하였으며, 수집 효율성(그림 5 참조) 실험에서 가장 좋은 결과를 보여준 ρ값을 설정하였다. 주제 기반 수집기의 일관성에 대한 실험은 총 2회 실시하였으며, 2회 실험의 평균값에 대한 결과를 그림 6에 정리하였다.

네 개의 상위 범주에 대하여 수집된 URL의 중복 비율은 수집된 URL 의 개수가 증가함에 따라 전체적으로 증가함을 알 수 있었으며, 최종적으로 수집된 URL 의 개수가 5,000개에 가까워 짐에 따라 70%에서 85% 사이의 비율을 유지하였다. 이러한 실험결과는 주제 기반 수집기가 시작 URL에 영향을 받지 않고 일관적인 수집을 수행한다는 것을 보여준다.

5. 결 론

특정 주제와 관련된 웹 페이지를 수집하기 위하여 본 논문은 (1) 웹 페이지와 해당 웹 페이지를 링크하는 웹 페이지간의 관계를 이용한 주제와의 연관도 및 (2) 웹 페이지에 나타난 미리 정의된 단어의 개수 정보를 이용하였다. 또한, 수집된 웹 페이지를 정제하여 주제와 부합하는 웹 페이지만을 분류하기 위하여 단어빈도/문서빈도와 엔트로피를 이용하여 핵심적인 단어를 선택하였다. 그리고, 귀납학습 알고리즘과 신경망 알고리즘을 이용하여 선택된 단어를 기반으로 분류 규칙을 생성하였다. 이와 같이 제안한 주제 기반 웹 수집기를 검증하기 위한 실험에서 분류 정확도를 비교한 결과, 77개의 대표적인 단어를 사용하는 경우에 CN2의 분류 성능이 93.2%로 가장 좋았다. 두 번째 수행한 실험에서 수집의 효율성을 측정하여 각 범주별로 연관도에 미치는 영향을 경험적으로 도출할 수 있었다. 마지막으로, 제안한

수집기의 수집정도에 대한 일관성을 평가하기 위하여 서로 다른 시작 URL을 사용하여 수집된 웹 페이지들의 중첩정도를 측정하였다. 실험 결과에서 서술한 바와 같이 제안한 주제 기반 웹 수집기가 시작 URL에 큰 영향을 받지 않고 상당히 일관적인 수집을 수행함을 알 수 있었다.

본 논문에서 제안한 주제 기반 수집 전략의 타당성을 분류의 정확성, 수집의 효율성 및 수집의 일관성 측면에서 검증하였으며, 향후 계속적인 연구를 통하여 다른 방법론과 비교 및 검증하고자 한다. 구체적으로, 연관도에 기반한 본 연구의 수집 전략의 타당성을 웹 페이지들이 하이퍼 링크(hyperlink)로 서로 연결된 웹 그래프(web graph) 구조를 이용하여 웹의 구조적 차원에서 분석할 것이다. 또한, 웹 그래프 시뮬레이션을 통하여 본 연구에서 제안한 전략과 다른 웹 페이지 수집 기법들을 비교 평가할 수 있을 것으로 기대한다.

참 고 문 헌

[1] S. Chakrabarti, M. van den Berg and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Resource Discovery," in Online Proceedings of 8th International World Wide Web Conference, Toronto, Canada, 1999.

[2] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori, "Focused Crawling Using Context Graphs," in Proceedings of 26th International Conference on VLDB, Cairo, Egypt, 2000, pp. 527-534.

[3] M. Ehrig and A. Maedche, "Ontology-Focused Crawling of Web Documents," in Proceedings of the 2003 ACM symposium on Applied Computing, Melbourne, Florida, 2003.

[4] A. Grigoriadis and G. Paliouras, "Focused Crawling Using Temporal Difference-Learning," Lecture Notes in Computer Science, vol. 3025, pp. 142-153, 2004.

[5] F. Menczer, G. Pant and P. Srinivasan, "Topic-Driven Crawlers: Machine Learning Issues," ACM TOIT, Submitted for publication, 2002.

[6] R. Baeza-Yates and B. Ribeiro-Neto , "Modern Information Retrieval," ACM Press/Addison-Wesley, 1999.

[7] S. Noh, H. Seo, J. Choi, K. Choi and G. Jung, "Classifying Web Pages Using Adaptive Ontology," in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Washington, D.C., U.S.A., Oct. 2003, pp. 2144-2149.

[8] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing Management, vol. 24, no. 5, pp. 513-523, 1988.

[9] C. E. Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, Jul. and Oct. 1948.

[10] J. Cho, H. Garcia-Molina and L. Page, "Efficient Crawling Through URL Ordering," Computer Networks and ISDN Systems, vol. 30, pp. 161-172, 1998.

[11] S. M. Ruger and S. E. Gauch, "Feature Reduction for Document Clustering and Classification," Technical Report, Department of Computing, Imperial College, London, England, 2000.

[12] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp.

1-47, 2002.

[13] M. Shaoping and H. Ke, "Chinese Web Page Classification based on Statistical Word Segmentation," in Proceedings of 2002 IEEE International Conference on Systems, Man and Cybernetics, vol. 1, 2002, pp. 636-640.

[14] A. Sun, E. Lim and W. Ng, "Web Classification Using Support Vector Machine," in Proceedings of the Fourth International Workshop on Web Information and Data Management, 2002, pp. 96-99.

[15] F. Menczer, G. Pant, P. Srinivasan and M. E. Ruiz, "Evaluating Topic-Driven Web Crawlers," in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 241-249.

[16] A. Passerini, P. Frasconi and G. Soda, "Evaluation Methods for Focused Crawling," Lecture Notes in Computer Science, vol. 2175, pp. 33-39, 2001.

[17] P. Clark and T. Niblett, "The CN2 Induction algorithm," Machine Learning Journal, vol. 3, no.4, pp. 261-283, 1989.

[18] L. Holder, ML v2.0, available on-line: http://www-cse.uta.edu/~holder/ftp/ml2.0.tar.gz.

[19] J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann, 1993.

[20] D. R. Tveter, Backprop Package, Available on-line: http://www.dontveter.com/nnsoft/bp042796.zip, 1996.

[21] M. P. Sinka and D. W. Corne, "A Large Benchmark Dataset for Web Document Clustering," Soft Computing Systems: Design, Management and Applications, vol. 87, pp. 881-890, 2002.

◐ 저 자 소 개 ◑

서 혜 성 (Haesung Seo)
2003년 아주대학교 정보통신 대학 정보 및 컴퓨터 공학부 (학사)
2003년 ~ 현 재 아주대학교 정보통신전문대학원 정보통신공학과
관심분야 : 네트워크 보안, 분산 시스템, 인공지능 등
E-mail: retry@ajou.ac.kr

최 영 수 (Youngsoo Choi)
2003년 아주대학교 정보통신 대학 정보 및 컴퓨터 공학부 (학사)
2003년 ~ 현 재 아주대학교 정보통신전문대학원 정보통신공학과
관심분야 : 네트워크 보안, 실시간 시스템, 인공지능 등
E-mail: drabble@ajou.ac.kr

노 상 욱 (Sanguk Noh)
1987년 서강대학교 생명과학 (학사)
1989년 서강대학교 컴퓨터공학 (공학석사)
1999년 텍사스 주립대 (Arlington) 컴퓨터공학 (공학박사)
1989년 ~ 1995년 국방과학연구소 연구원
2000년 ~ 2002년 미조리 주립대 (Rolla) 컴퓨터학과 조교수
2002년 ~ 현 재 가톨릭대학교 컴퓨터정보공학부 부교수
관심분야: 지식관리, 기계학습, 지능형 에이전트, 멀티 에이전트 시스템, 인공지능, 분산 실시간 시스템 등
E-mail: sunoh@catholic.ac.kr

최 경 희 (Kyunghee Choi)
1976년 서울대학교 사범대학 수학교육과 졸업 (학사)
1979년 프랑스 그랑데꼴 Enseeiht 정보공학과 졸업 (공학석사)
1982년 프랑스 Paul Sabatier 정보공학과 졸업 (공학박사)
1982년 ~ 현 재 아주대학교 정보통신전문대학원 교수
관심분야: 운영체제, 분산시스템, 실시간 및 멀티미디어 시스템 등
E-mail: khchoi@madang.ajou.ac.kr

정 기 현 (Gihyun Jung)
1984년 서강대학교 공과대학 전자공학과 졸업 (학사)
1988년 미국 Illinois 주립대 EECS 졸업 (공학석사)
1990년 미국 Perdue 전기전자공학부 졸업 (공학박사)
1991년 ~ 1992년 현대전자 반도체 연구소
1993년 ~ 현 재 아주대학교 전자공학부 교수
관심분야: 컴퓨터구조, VLSI설계, 멀티미디어 및 실시간 시스템 등
E mail: khchung@madang.ajou.ac.kr