웹 문서 클러스터링에서의 자질 필터링 방법

박 흠[†] · 권 혁 철[††]

요 약

색인전문가에 의해 분류된 웹문서들을 통계적 자질 선택방법으로 자질을 추출하여 클러스터링을 해 보면, 자질 선택에 사용된 데이터셋에 따라 성능과 결과가 다르게 나타난다. 그 이유는 많은 웹 문서에서 문서의 내용과 관계없는 단어들을 많이 포함하고 있어 문서의 특징을 나타내는 단어들이 상대적으로 잘 두드러지지 않기 때문이다. 따라서 클러스터링 성능을 향상시키기 위해 이런 부적절한 자질들을 제거해 주어야 한다. 따라서 본 논문에서는 자질 선택에서 자질의 문서군별 자질값뿐만 아니라, 문서군별 자질값의 분포와 정도, 자질의 출현여부와 빈도를 고려한 자질 필터링 알고리즘을 제시한다. 알고리즘에는 (1)단위 문서 내 자질 필터링 알고리즘(FFID : feature filtering algorithm in a document), (2)전체 데이터셋 내 자질 필터링 알고리즘(FFIM : feature filtering algorithm in a document matrix), (3)FFID와 FFIM을 결합한 방법(HFF : a hybrid method combining both FFID and FFIM)을 제시한다. 실험은 단어빈도를 이용한 자질선택 방법, 문서간 동시-링크 정보의 자질확장, 그리고 위에서 제시한 3가지 자질 필터링 방법을 사용하여 클러스터링 했다. 실험 결과는 데이터셋에 따라 조금씩 차이가 나지만, FFID보다 FFIM의 성능이 좋았고, 또 FFID와 FFIM을 결합한 HFF 결과가 더 나은 성능을 보였다.

키워드 : 자질선택, 자질필터링, 클러스터링, 웹문서

Feature Filtering Methods for Web Documents Clustering

Heum Park[†] · Hyuk-Chul Kwon[††]

ABSTRACT

Clustering results differ according to the datasets and the performance worsens even while using web documents which are manually processed by an indexer, because although representative clusters for a feature can be obtained by statistical feature selection methods, irrelevant features(i.e., non-obvious features and those appearing in general documents) are not eliminated. Those irrelevant features should be eliminated for improving clustering performance. Therefore, this paper proposes three feature-filtering algorithms which consider feature values per document set, together with distribution, frequency, and weights of features per document set : (1) features filtering algorithm in a document (FFID), (2) features filtering algorithm in a document matrix (FFIM), and (3) a hybrid method combining both FFID and FFIM (HFF). We have tested the clustering performance by feature selection using term frequency and expand co link information, and by feature filtering using the above methods FFID, FFIM, HFF methods. According to the results of our experiments, HFF had the best performance, whereas FFIM performed better than FFID.

Key Words : Feature Selection, Feature Filtering, Clustering, Web Document

1. 서 론

웹 문서 검색의 성능 향상 기법 중 하나로 검색 후 클러스터링 기법을 많이 사용한다. 이는 검색어에 의해 1차 검색된 웹사이트나 웹 문서들을 다시 문서의 특성에 따라 클러스터링 하여 보여줌으로써, 검색자에게 보다 정확한 문서를 검색하도록 해 준다.

이번 웹 문서 자질선택과 클러스터링 실험을 위해 색인전문가에 의해 수작업으로 분류하여 디렉토리 서비스를 하고 있는 국내 포털 사이트 3곳의 웹 문서를 수집하여 자질 선택과 선택된 자질로 클러스터링 해 보았다.

그 결과 단어빈도에 의한 웹 문서 클러스터링 결과에 비해 통계적 방법에 의해 추출한 자질로 클러스터링 한 결과가 더 나쁘거나, 자질선택에 사용된 데이터셋에 따라 성능에 많은 차이가 나타났다.

그래서 디렉토리 서비스의 웹 문서들의 특징을 조사해 보니, 많은 문서에서 문서를 대표하는 단어들뿐만 아니라, 문서의 주제와 관계없는 클러스터링에 부적절한 단어들, 즉 편집자 소개, 메뉴, 공지사항, 안내, 광고, 관련 사이트, 뉴스, 이벤트, 멀티미디어 자료 등을 포함하고 있었다. 따라서 이런

※ 논문은 교육인적자원부 지방연구중심대학육성사업(차세대물류IT기술연구사업단)의 지원에 의하여 연구되었음.
† 정 회 원 : 유비텍㈜ 이사
†† 정 회 원 : 부산대학교 전자전기정보컴퓨터공학부 교수
논문접수 : 2006년 3월 9일, 심사완료 : 2006년 7월 7일

부적절한 단어들은 자질선택 과정에서 제거해야 한다[7, 8].

자질이란 문서 내 단어 중 문서의 내용과 특징을 잘 나타내고 있는 단어를 말하는데, 이 자질들은 문서 내 단어 중 얼마나 중요한 지에 대한 가중치로 평가되어 문서 클러스터링의 분류자질로 사용된다[2]. 자질에 대한 가중치 평가방법은 X^2, Mutual information, Information gain, Odds ratio 등이 있다. 하지만 이런 방법으로 얻은 자질값만을 이용한 클러스터링으로는 불필요한 자질을 모두 제거해 주지 못한다. 또 너무 많은 여분의 자질들도 적절하게 제거해주어야 한다[6, 8, 10].

그리고 실험을 위해 웹 문서 내 하이퍼링크 정보와 동시-링크 정보를 이용했다. 하이퍼링크 정보는 문서간에 연결시켜 주므로 한 단계 정도는 자질로 확장시킬 수 있다. 그리고 대문페이지로 구성된 일부 웹 페이지의 경우는 추출할 자질이 없기 때문에 하이퍼링크 정보를 이용해 한 단계 혹은 두 단계까지 자질을 확장할 수 있다[5, 16]. 또 단어 자질뿐만 아니라 웹 문서간 동시-링크 정보도 문서간 관련성이 아주 높기 때문에 자질로서 확장할 수 있다[16]. 그래서 자질이 부족한 문서에는 하이퍼링크 정보를 이용해 자질을 확장하였다. 또 일부 데이터셋에는 실험을 위해 동시-링크 정보를 결합하여 확장하였다.

본 논문에서는 웹문서의 특성상 기존의 자질선택 방법으로는 제거하지 못한 불용어들을 제거하고 대표 자질은 더욱 두드러지게 하기 위한 자질 필터링 방법으로, (1)단위 문서 내 자질 필터링 알고리즘 (FFID : feature filtering algorithm in a document)과 (2)전체 데이터셋 내 자질 필터링 알고리즘 (FFIM : feature filtering algorithm in a document matrix) 5가지, 또 (3)FFID와 FFIM을 결합한 방법(HFF : a hybrid method combining both FFID and FFIM) 5가지를 제시한다.

먼저 자질의 단어빈도를 이용한 클러스터링 결과와 기존 자질선택방법으로 추출한 자질과 자질값을 이용한 클러스터링 결과를 비교하고, 또 문서간 동시-링크 정보를 자질로 확장한 클러스터링 실험 결과를 비교하겠다. 그리고 이 결과와 본 논문에서 제시한 3가지 방법(FFID, FFIM, HFF)에 의한 클러스터링 결과를 비교하겠다.

논문의 구성은 2장에 관련연구, 3장에는 본론으로 자질 추출 방법과 자질 필터링 방법, 클러스터링 실험 환경 등에 대해 설명하고, 4장은 실험 및 결과 분석, 5장은 결론, 그리고 6장은 향후 연구 과제에 대해 기술하겠다.

2. 관련 연구

전통적으로 문서 클러스터링을 위한 전 처리과정인 자질선택을 위한 방법은 통계학적 기법인 X^2, Mutual Information(MI), Information Gain(IG), Odds Ratio(OR) 등이 있으며[3, 7], 이와 같은 자질선택 방법으로 자질 매트릭스를 축소하여 클러스터링 성능을 비교했다[15]. 그리고 추출된 자질을 Filtering과 Wrapper 기법을 이용해 자질을 축소하고 여분의 자질을 제거해 줌으로써 클러스터링 성능을 향상시켰고[9], Filter기법과 Wrapper 기법의 장점을 이용한 Tow-phase 자질선택 방법으로 부적절한 자질을 제거하는 방법을 개선했다[6]. 그리고 대량의 데이터베이스에 대한 실험에서 Filter가 Wrapper보다 처리 속도가 빠른 반면 이산 분류에 문제가 있는데, 이런 이산과 연속 분류에 대한 문제점을 correlation-based 자질 선택 방법으로 해결하는 연구가 있었다[8]. 또 자질 선택에 사용된 데이터셋에 따라 성능이 달라지는 문제점을 보완해, 문서 내 자질의 문서군에 따라 문서의 대표 문서군을 정해 주어 클러스터링 이전에 문서의 특징을 미리 정해 자질을 축소하는 Max Feature Selection 기법을 사용하여 클러스터링 성능을 향상시켰다[7]. 또 문서 내 대표자질을 선정하는 방법으로 자질값투표 기법을 사용한 연구도 있었다[1]. 웹 문서의 링크를 확장해 문서를 문서의 자질로 포함시키거나, 웹 문서간 동시-링크 빈도를 계산해 문서의 자질을 확장해 단어 자질로만 사용했을 때와 클러스터링 성능을 비교하였다[7, 16]. 자질 선택 방법에 의해 얻은 자질을 SVM (support vector machine)기법으로 training시켜 성능을 비교하는 연구도 있었다[14].

3. 본 론

기존의 자질선택 방법인 X^2, Mutual Information(MI), Information Gain(IG), Odds Ratio(OR) 등에 의해 나온 자질과 자질값을 다른 데이터셋에 적용해 클러스터링 했을 때, 자질선택에 사용한 데이터셋에 따라 클러스터링 성능에 많은 영향을 받는다[7]. 이유는 자질선택에서 추출된 자질들 중에 여전히 문서의 주제와 관계없는 불필요한 단어가 많이 존재하기 때문이다. 이렇게 기존의 자질선택 방법으로 제거하지 못한 웹 문서 내 자질들의 특징을 보면 다음과 같다.

1) 자질이 서로 관계없는 문서군에서 동시에 출현한다.
2) 자질의 전체 문헌빈도가 높으며, 특정 문서군에서도 문헌빈도가 높다
3) 자질 선택 방법에 의해 나온 자질값이 너무 낮거나 잘 두드러지지 않는다.
4) 자질 선택 방법에 의해 나온 자질값이 서로 관계없는 문서군에서도 높다.

따라서 본 논문에서는 클러스터링 전단계인 자질선택 단계에서 기존 자질선택방법에 의한 실험과 문서간 동시-링크 정보를 추가한 자질확장으로 클러스터링 실험을 했다. 그리고 위 방법으로 제거하지 못한 문서 내 불필요한 자질들을 제거하기 위해 기존 자질선택 방법으로 추출된 자질과 자질값을 이용해, (1)단위 문서 내 자질 필터링 알고리즘(FFID)과 (2)전체 데이터셋 내 자질 필터링 알고리즘(FFIM), (3)HFF 등을 이용해 클러스터링 실험을 했다.

3.1 기존 자질 추출 방법

단어빈도로 구성된 문서*단어 매트릭스를 기존 자질선택

(그림 1) 단어 자질과 동시-링크 정보를 결합한 문서 매트릭스

알고리즘 X^2, Mutual Information(MI), Information Gain(IG), Odd Ratio(OR) 등을 이용해 자질과 문서군별 자질값을 추출하였다. 자질 선택 알고리즘 수식은 다음과 같다[3, 7].

A : 범주(문서군) c에 속해 있는 문서 중 단어 t를 포함하고 있는 문서 수
B : 범주(문서군) c에 속하지 않은 문서 중 단어 t를 포함하고 있는 문서 수
C : 범주(문서군) c에 속해 있는 문서 중 단어 t를 포함하고 있지 않은 문서 수
D : 범주(문서군) c에 속하지 않은 문서 중 단어 t를 포함하고 있지 않은 문서 수
N : 전체 문서 수

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

$$MI(t,c)=\log\frac{\Pr(t\wedge c)}{\Pr(t)\times\Pr(c)} \quad MI(t,c)\approx\log\frac{A\times N}{(A+C)\times(A+B)}$$

$$IG(t,c)=-\sum_{i=1}^{m}\Pr(c_i)\log\Pr(c_i)+\Pr(t)\sum_{i=1}^{m}\Pr(c_i|t)\log\Pr(c_i|t)+\Pr(\bar{t})\sum_{i=1}^{m}\Pr(c_i|\bar{t})\log\Pr(c_i|\bar{t})$$

$$IG(t,c)\approx\frac{A}{(A+B)}\frac{A\times N}{(A+C)\times(A+B)}+\frac{C}{(C+D)}\frac{A\times N}{(A+C)\times(C+D)}$$

$$OR(t,c_i)=\frac{\sum_{i=1}^{m}\Pr(t\mid c_i)\sum_{i=1}^{m}\Pr(\bar{t}\mid\bar{c}_i)}{\sum_{i=1}^{m}\Pr(\bar{t}\mid c_i)\sum_{i=1}^{m}\Pr(t\mid\bar{c}_i)} \qquad OR(t,c)\approx\frac{A\times D}{B\times C}$$

여기서 얻은 자질과 문서군별 자질값을 데이터셋 매트릭스 파일에 적용시켰다.

3.2 문서간 동시-링크 빈도를 자질로 확장

2개의 문서가 동시에 링크 한 문서 수를 구해 문서*문서 매트릭스를 만들었다. 동시-링크 문서 수는 Alltheweb(http://www.alltheweb.co.kr)의"AdvancedSearch" 옵션에서 검색식 "LINK : url_A AND LINK : url_B"를 사용했다[5].

〈표 1〉 자질 선택 알고리즘 X^2를 적용한 자질의 문서군별 자질값과 대표문서군.

번호	자질	문서군1 X2값	문서군2 X2값	문서군3 X2값	문서군4 X2값	문서군5 X2값	문서군6 X2값	대표 문서군
1	연구	4	24	6	4	4	14	2
2	수학	0	4	0	0	0	4	2, 6
3	공학	3	1	0	48	3	13	4
4	전공	5	6	14	2	5	7	3
5	전기	0	27	0	0	15	0	2
6	물리	23	0	2	0	0	3	1
7	암컷	0	0	54	0	0	0	3
8	핵	14	0	8	0	0	11	1
9	하늘	1	4	5	0	20	20	5, 6
10	세상	4	2	1	1	8	5	5

이렇게 얻은 동시-링크 문서 매트릭스와 문서*자질 매트릭스를 결합하면 (그림 1)과 같다.

그리고 이 매트릭스를 이용해 클러스터링 한 실험 결과를 CL()로 부르겠다.

3.3 단위 문서 내 자질 필터링 방법

앞에서 추출한 자질과 문서군별 자질값을 이용하여 자질의 대표문서군을 정할 수 있다. 즉 대표문서군이란 자질이 데이터셋 내 전체 문서군 중 어떤 문서군에 속하는 지를 나타낸다. 자질 t의 대표 문서군 $C_f(t)$는 다음과 같다.

$$C_f(t)=\max_{i=1}^{k}\{FV_i(t)\}$$

(f : 대표문서군 번호, k : 문서군 수, FV : Feature Value $FV_i(t)$: 문서군 i에서 자질 t의 자질값)

그리고 이 자질과 자질값을 적용한 데이터셋 매트릭스 내 각 문서는 자질과 자질값으로 문서의 대표문서군을 정할 수 있다. 즉 문서 d내 자질들의 대표문서군 수가 가장 많은 문서군이 문서 d의 대표문서군이 된다. 문서 d의 대표문서군은

$$C_f(d)=\max_{i=1}^{k}\{Count_i(C_i(t_1:t_n))\}$$

(f : 대표문서군 번호, k : 문서군 수, $C_i(t_1:t_n)$: 문서 d에 있는 자질 t_1에서 t_n까지의 대표문서군, $Count_i(C_i(t_1:t_n))$: 대표문서군 i에 속한 자질 수)

따라서 데이터셋 매트릭스 내 각 문서의 자질들 중 문서의 대표문서군에 속한 자질만 남기고 모두 제거하여 문서가 속할 문서군을 미리 정한다. 대표문서군 $C_f(d)$에 속한 자질 외 다른 자질을 제거한 문서 d는 ($FV_1^{C_f}$, $FV_2^{C_f}$, ... , $FV_n^{C_f}$)로 표현할 수 있다.

예를 들어 자질선택 방법 X^2를 사용하여 얻은 자질과 자질값, 대표문서군을 보면 <표 1>과 같다.

그리고 <표 1>을 이용해 문서군별 대표 자질들을 보면 <표 2>과 같다.

〈표 2〉 문서군별 대표 자질

<표 1>에서 추출 *는 2개 이상 문서군에 속함

	문서군1	문서군2	문서군3	문서군4	문서군5	문서군6
대표 자질	물리 핵	연구 수학* 전기	전공 암컷	공학	하늘* 세상	수학* 하늘

따라서 문서 d의 자질과 자질값이 다음과 같다면,

d : 물리 3 연구 1 공학 1 핵 4

<표 2>에 의해 문서 d의 대표 문서군은 1이 된다. 즉 자질 '물리'와 '핵'의 대표문서군은 1이고, '연구'의 대표문서군은 2이고, '공학'의 대표문서군은 4이기 때문에 대표 문서군 1에 속한 자질은 2개므로 문서 d의 대표문서군은 1이 된다. 만약 대표 문서군 수가 같으면 자질값의 합이 큰 문서군을 대표 문서군으로 한다.

같은 방법으로 데이터셋 내 모든 문서의 대표 문서군을 정할 수 있다. 그리고 각 문서는 대표 문서군에 속한 자질만 남기고 다른 자질을 제거한 문서들을 문서*자질 매트릭스 파일로 만들어 클러스터링 한다. 또 MI, IG, OR 에 의해 추출된 자질과 자질값에 대해서도 같은 방법으로 자질을 필터링 하여 클러스터링 실험을 한다.

이 실험을 Max(X^2), Max(MI), Max(IG), Max(OR) 등으로 부르고, 기존 자질선택 방법 X^2, Mutual Information, Information Gain, Odd Ratio 등을 이용한 클러스터링 실험을 Org(X^2), Org(MI), Org(IG), Org(OR) 등으로 부르겠다.

3.4 전체 데이터셋 내 자질 필터링 방법

전체 데이터셋 내 자질 필터링 방법으로는 자질의 문서군별 자질값과 문서군별 문헌빈도, 자질의 출현 문서군 수 등이 기준치를 넘을 때 자질을 제거하거나 선택하는 방법 5가지를 제시한다.

3.4.1 데이터셋 내 자질이 출현하는 문서군의 수가 기준치 이상인 자질 제거

자질이 데이터셋 내 몇 군데의 문서군에서 출현하는 가에 따라 기준치를 넘는 자질을 데이터셋 매트릭스에서 제거한다. 기준치는 80% 혹은 60%, 40% 등으로 한다. 예를 들어 전체 문서군수가 6이고, 기준치가 80%이면 5개 문서군이상, 60%이면 4 개 문서군이상, 40%이면 3개 문서군이상에서 출현하는 자질을 데이터셋 매트릭스에서 모두 제거한다. 자질 t가 출현한 문서군 수

$$\underset{i=1}{\overset{k}{count}}(C_i(t))$$

(k : 문서군 수, $C_i(t)$: 자질 t_i의 대표문서군)

가 기준치 이상인 자질은 제거한다.

이렇게 축소된 자질로 구성한 데이터셋 매트릭스 파일을 클러스터링 한다. 이 실험을 F1(X^2), F1(MI), F1(IG), F1(OR) 등으로 부르겠다. 그리고 3.3단위 문서 내 자질 필터링 방법을 사용하여 만든 매트릭스 파일을 다시 이 방법을 이용해 자질을 필터링 한 실험을 F1(Max(X^2)), F1(Max(MI)), F1((Max(IG)), F1(Max(OR)) 등으로 부르겠다.

3.4.2 특정 문서군에서의 자질값이 전체 문서군의 자질값 합과 비교

전체 문서군의 자질값 합과 비교해 기준치보다 큰 자질값을 갖는 자질만 선택하고 나머지는 제거. 기준치는 50%로 했다.

자질 t의 특정 문서군에서의 자질값 $FV(t)$가 전체 문서군의 자질값의 합의 1/2보다 클 때만 자질로서 선택한다.

$$\frac{1}{2}\sum_{i=1}^{k} FV_i(t) \le FV(t)$$

(k : 문서군 수, $FV(t)$: 특정 문서군에서의 자질값)

이렇게 선택된 자질로만 구성된 매트릭스 파일로 클러스터링 한다. 이 실험을 F2(X^2), F2(MI), F2(IG), F2(OR) 등이라 하고, 3.3단위 문서 내 자질 필터링 방법을 사용하여 만든 매트릭스 파일을 다시 이 방법을 이용해 자질을 필터링 한 실험을 F2(Max(X^2)), F2(Max(MI)), F2((Max(IG)), F2(Max(OR)) 등으로 부르겠다.

3.4.3 특정 문서군에서의 자질의 문헌빈도가 전체 문서군의 문헌빈도 합과 비교해 기준치보다 큰 값을 갖는 자질만 선택하고 나머지는 제거. 기준치는 50%.

특정 문서군 내 자질 t 문헌빈도 $DF_f(t)$가 전체 문서군의 문헌빈도의 합의 1/2보다 큰 자질만 선택한다.

$$\frac{1}{2}\sum_{i=1}^{k} DF_i(t) \le DF_f(t)$$

(k : 문서군 수, $DF_f(t)$: 특정 문서군에서의 문헌빈도)

이렇게 선택된 자질로만 구성된 매트릭스 파일로 클러스터링 한다. 이 실험을 F3(X^2), F3(MI), F3(IG), F3(OR) 등이라 하고, 3.3단위 문서 내 자질 필터링 방법을 사용하여 만든 매트릭스 파일을 다시 이 방법을 이용해 자질을 필터링 한 실험을 F3(Max(X^2)), F3(Max(MI)), F3((Max(IG)), F3(Max(OR)) 등으로 부르겠다.

3.4.4 자질의 출현 문서군수에 의해 제한하면서(3.4.1), 특정 문서군에서의 자질의 문헌빈도와 전체 문서군의 문헌빈도 합을 비교해 기준치보다 큰 자질만 선택(3.4.3).

이 실험은 3.4.3 실험 기준치

$$\frac{1}{2}\sum_{i=1}^{k} DF_i(t) \le DF_f(t)$$

(k : 문서군 수, $DF_f(t)$: 특정 문서군에서의 문헌빈도)

〈표 3〉 자질 필터링 실험 방법

		tf	X2	MI	IG	OR	단위 문서 내 자질 필터링(FFID)			
							X2	MI	IG	OR
기존 자질 선택 방법		Org(tf)	Org(X2)	Org(MI)	Org(IG)	Org(OR)	Max(X2)	Max(MI)	Max(IG)	Max(OR)
데이터셋 내 자질 필터링 (FFIM)	3.4.1		F1 (X2)	F1 (MI)	F1 (IG)	F1 (OR)	FFID와 FFIM을 결합한 자질 필터링(HFF)			
							F1(Max X2)	F1(Max MI)	F1(Max IG)	F1(Max OR)
	3.4.2		F2 (X2)	F2 (MI)	F2 (IG)	F2 (OR)	F2(Max X2)	F2(Max MI)	F2(Max IG)	F2(Max OR)
	3.4.3		F3 (X2)	F3 (MI)	F3 (IG)	F3 (OR)	F3(Max X2)	F3(Max MI)	F3(Max IG)	F3(Max OR)
	3.4.4		F4 (X2)	F4 (MI)	F4 (IG)	F4 (OR)	F4(Max X2)	F4(Max MI)	F4(Max IG)	F4(Max OR)
	3.4.5		F5 (X2)	F5 (MI)	F5 (IG)	F5 (OR)	F5(Max X2)	F5(Max MI)	F5(Max IG)	F5(Max OR)

에 의해 선택된 자질들 중 3.4.1 실험으로

$$\underset{i=1}{\overset{k}{count}}(C_i(t))$$

(k: 문서군 수, $C_i(t)$: 자질 t_i의 대표문서군)

의 기준치를 넘지 않는 자질만 선택한다. 즉 자질 t는

$$\underset{i=1}{\overset{k}{count}}(C_i(\frac{1}{2}\sum_{i=1}^{k} DF_i(t) \le DF_f(t)))$$

가 기준치 미만인 자질만 선택하고 나머지는 제거한다. 기준치는 80% 혹은 60%, 40% 등으로 한다.

이렇게 선택된 자질로만 구성된 매트릭스 파일을 클러스터링 한다. 이 실험을 F4(X^2), F4(MI), F4(IG), F4(OR) 등이라 하고, 3.3단위 문서 내 자질 필터링 방법을 사용하여 만든 매트릭스 파일을 다시 이 방법을 이용해 자질을 필터링 한 실험을 F4(Max(X^2)), F4(Max(MI)), F4((Max(IG)), F4(Max(OR)) 등으로 부르겠다.

3.4.5 자질을 출현 문서군수에 의해 제한하면서(3.4.1), 특정문서군에서의 자질값이 전체 문서군의 자질값 합이 기준치보다 큰 자질값을 갖는 자질만 선택(3.4.2).

이 실험은 3.4.2 실험 기준치

$$\frac{1}{2}\sum_{i=1}^{k} FV_i(t) \le FV(t)$$

(k: 문서군 수, $FV(t)$: 특정 문서군에서의 자질값)

에 의해 선택된 자질들 중 3.4.1 실험으로

$$\underset{i=1}{\overset{k}{count}}(C_i(t))$$

. (k: 문서군 수, $C_i(t)$: 자질 t_i 의 대표문서군)

의 기준치를 넘지 않는 자질만 선택한다. 즉 자질 t는

$$\underset{i=1}{\overset{k}{count}}(C_i(\frac{1}{2}\sum_{i=1}^{k} FV_i(t) \le FV(t))$$

가 기준치 미만인 자질만 선택하고 나머지는 제거한다. 기준치는 80% 혹은 60%, 40% 등으로 한다.

이렇게 선택된 자질로만 구성된 매트릭스 파일을 클러스터링 한다. 이 실험을 F5(X^2), F5(MI), F5(IG), F5(OR) 등이라 하고, 3.3단위 문서 내 자질 필터링 방법을 사용하여 만든 매트릭스 파일을 다시 이 방법을 이용해 자질을 필터링 한 실험을 F5(Max(X^2)), F5(Max(MI)), F5((Max(IG)), F5(Max(OR)) 등으로 부르겠다.

기존 자질선택 방법과 단위 문서 내 자질 필터링 방법, 전체 데이터셋 내 자질 필터링 방법을 모두 열거하면 <표 3>과 같다.

그리고 본 논문에서의 실험방법은 (그림 2)와 같다.

3.5 클러스터링 실험 환경

클러스터링은 k-means clustering 방법을 사용하였고, 클러스터링 수는 12개로 했다. 그리고 criterion function은 다음 2개의 식을 사용했다.

$$\max \sum_{r=1}^{k} \sum_{d_i \in S_r} \cos(d_i, C_r)$$

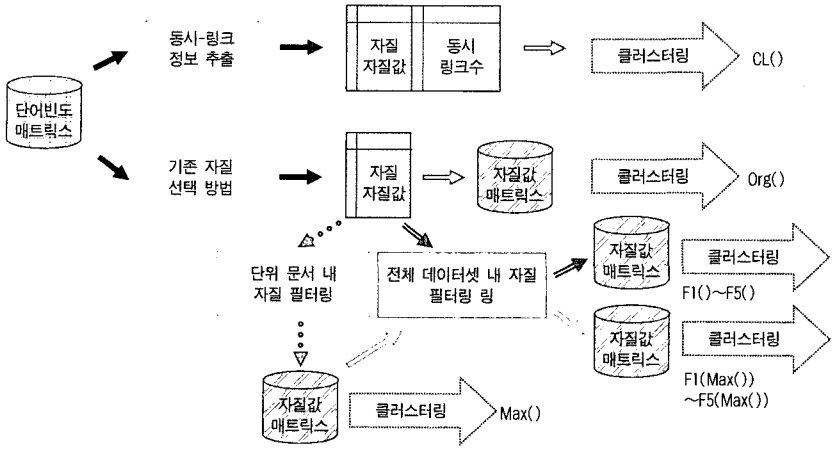
(d_i: 문서, C_r: 문서군, S_r: 문서군 r에 속한 문서)[11]

$$\min \sum_{r=1}^{k} {n_r}^2 \frac{\sum_{d_i \in S_r, d_j \in S-S_r} \cos(d_i, d_i)}{\sum_{d_i, d_j \in S_r} \cos(d_i, d_i)}$$

(n_r: 문서군 r에 속한 문서 수)[11]

유사도 계수 cosine함수 $\cos(d_i, d_j) = \frac{d_i^t d_j}{\|d_i\| \|d_j\|}$ 를 사용했다[10].

그리고 클러스터링에 사용된 데이터셋 매트릭스 파일의 column은 자질값 혹은 자질값에 대한 *idf(inverse-document-frequency)* 값을 사용했다. 즉 매트릭스 내 각 문서는 자질값을 적용했을 경우 $d_{tf} = (FV_1, FV_2, \ldots, FV_n)$, 자질값의 *idf* 값을 적용하면 $d_{tfidf} = FV_1 \log(N/FV_1), FV_2 \log(N/FV_2), \ldots, FV_n \log(N/FV_n)$[10]과 같다.



(그림 2) 자질 추출과 클러스터링 구현 방법

〈표 4〉 데이터셋의 세분류 별 문서 수와 문서 번호

세분류	Yahoo(문서번호)	Naver(문서번호)	Empas(문서번호)	합계
물리학	65(1~65)	85 (1~85)	143(1~143)	293
생물학	365(66~430)	327(86~412)	560(144~703)	1,252
지구과학	134(431~564)	114(413~526)	163(704~866)	411
화학	61(565~625)	91 (527~617)	117(867~983)	269
수학	125(626~750)	83 (618~700)	76 (984~1,059)	284
천문학	79(751~829)	221(701~921)	114(1,060~1,173)	414
합계	829	921	1,173	2,923

4. 실험 및 결과 분석

실험에 사용된 데이터셋과 실험 도구에 대해 간략하게 설명하고, 실험방법에 따른 클러스터링 결과를 비교 분석하겠다.

4.1 실험데이터

실험 데이터는 포탈사이트Yahoo, Naver, Empas의 디렉토리 서비스 중 자연과학 분류 중 '물리학', '생물학', '지구과학', '화학', '수학', 천문학' 등 6개 분야에서 추출했다. 이 데이터들은 색인 전문가에 의해 수작업으로 분류된 문서들이다. 각 데이터셋의 세분류 별 문서 수와 문서번호는 <표 4>와 같다.

본 논문에서는 각 데이터셋의 이름을 익명으로 데이터셋 A, 데이터셋B, 데이터셋C 등으로 부르겠다. 그리고 각 데이터셋은 6개의 문서군으로 이루어져 있는데, 각각 '문서군1', '문서군2', … , '문서군6' 등으로 부르겠다.

데이터셋은 용도에 따라 자질선택용과 클러스터링용으로 나뉜다. 자질선택용 데이터셋은 자질선택 알고리즘을 이용해 자질과 자질값을 얻는데 사용되고, 클러스터링용 데이터셋은 자질선택 알고리즘에서 얻은 자질값을 적용시켜 클러스터링 실험을 하는데 사용된다.

실험에 사용되는 데이터셋은 색인 전문가에 의해 수작업으로 잘 분류된 문서들이지만 각 데이터셋을 단어빈도 클러스터링으로 실험한 결과, 데이터셋 A는 클러스터링 결과가 아주 좋지 않았고, 데이터셋 B는 대체로 양호, 데이터셋 C는 가장 양호하였다. 각 데이터셋을 이용한 단어빈도 클러스터링 결과는 <표 5>와 같다.

〈표 5〉 데이터셋 별 단어빈도 클러스터링 결과 Entropy와 Purity

	데이터셋 A	데이터셋 B	데이터셋 C
Entropy	0.592	0.400	0.335
Purity	0.593	0.759	0.795

이 데이터셋들은 자질선택과 클러스터링 실험에 서로 역할을 바꾸어가며 실험에 사용했다. 그리고 문서간 동시-링크 정보를 이용한 실험은 데이터셋 B만 사용했다.

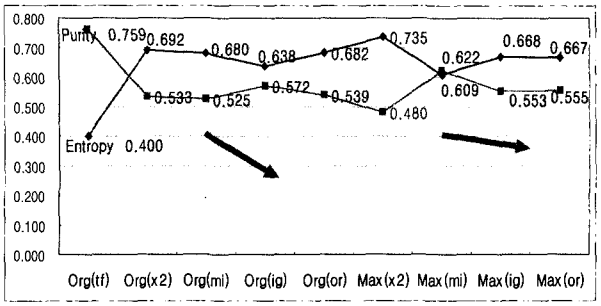
4.2 실험 도구

웹사이트로부터 문서 수집은 부산대학교 한국어정보처리연구실에서 개발한 로봇을 이용했고, 수집된 문서에서 색인어를 추출하는 프로그램 역시 부산대학교 한국어정보처리연구실에서 개발한 문서색인시스템을 사용했다. 그리고 클러스터링 실험과 평가는 미국 미네소타대학교에서 개발한 clustering toolkit Cluto2.1을 사용했다[12].

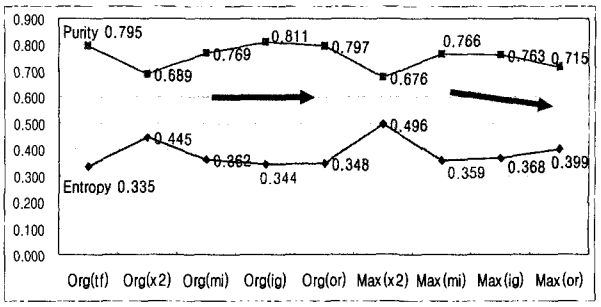
4.3 실험 결과 분석

클러스터링 성능 평가는 Entropy와 Purity를 사용했다[11]. 먼저 단어빈도 매트릭스 파일의 클러스터링 결과 Org(tf)와 기존 자질선택 방법(X^2, MI, IG, OR)에 의한 클러스터링 결과 Org(X^2), Org(MI), Org(IG), Org(OR), 동시-링크 정보를 이용한 클러스터링 결과 CL(), 단위 문서 내 자질 필터링 방법으로 만든 매트릭스 파일의 클러스터링 결과 Max(X^2), Max(MI), Max(IG), Max(OR))를 비교하겠다.

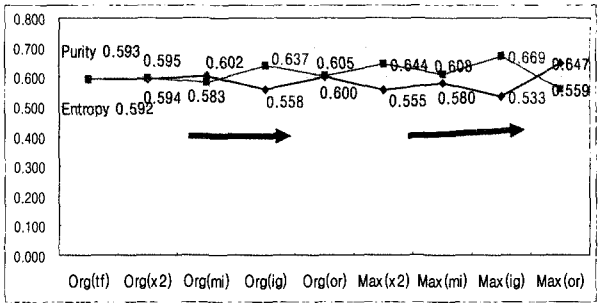
그리고 전체 데이터셋 내 자질 필터링 방법으로 실험한 결과 5가지 F1(), F2(), F3(), F4(), F5() 은 실험에 사용한 자질선택방법 X^2, MI, IG, OR 을 이용한 결과를 평균하여 나타냈고, 또 단위 문서 내 자질 필터링 방법과 전체 데이터셋 내 자질 필터링 방법을 결합한 실험 결과 역시 평균값으로 F1(Max), F2(Max), F3(Max), F4(Max), F5(Max)를 비교하였다.



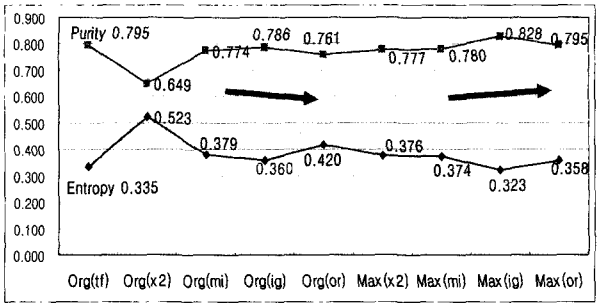
(그림 3) 자질선택 데이터셋 A, 클러스터링 데이터셋 B



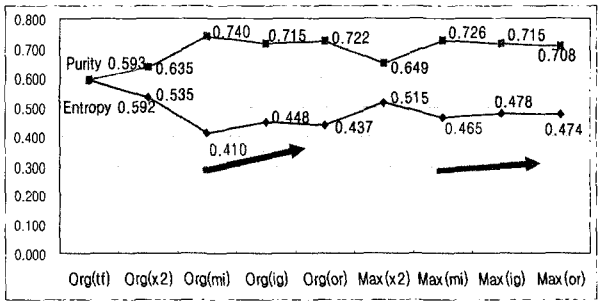
(그림 4) 자질선택 데이터셋 A, 클러스터링 데이터셋 C



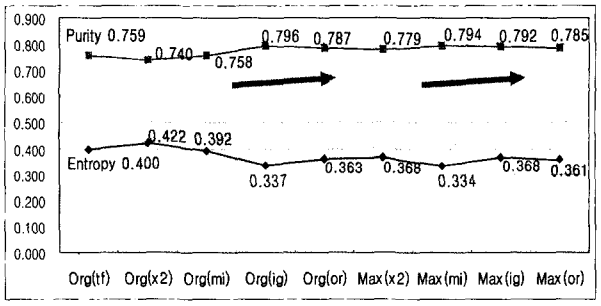
(그림 5) 자질선택 데이터셋 B, 클러스터링 데이터셋 A



(그림 6) 자질선택 데이터셋 B, 클러스터링 데이터셋 C



(그림 7) 자질선택 데이터셋 C, 클러스터링 데이터셋 A



(그림 8) 자질선택 데이터셋 B, 클러스터링 데이터셋 C

〈표 6〉 단어빈도 클러스터링과 동시-링크 정보를 이용한 클러스터링 결과

	Org(tf)	CL(tf)	CL(tf+co)
Entropy	0.400	0.477	0.450
Purity	0.759	0.635	0.660

4.3.1 기존 자질선택 방법과 단위문서 내 자질필터링방법에 의한 클러스터링 결과

Org(tf)는 단어빈도 클러스터링 결과이고, Org(X^2)부터 Org(OR)까지는 기존 자질선택 방법에 의한 클러스터링 결과다. 그리고 Max(X^2)부터Max(OR)까지는 단위 문서 내 자질 필터링 방법에 의한 클러스터링 결과다. 자질선택에 사용된 데이터셋 별 클러스터링 결과 Entropy와 Purity를 보면 다음과 같다.

(그림 3)와 (그림 4)은 데이터셋 A를 자질선택에 사용하여, 데이터셋 B, C에 적용한 클러스터링 결과인데, Org() Max() 모두 좋지 않다.

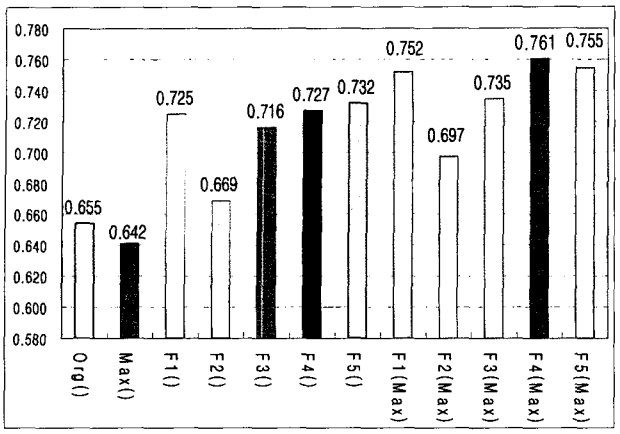
(그림 5)와 (그림 6)는 데이터셋 B를 자질선택에 사용한 결과인데, 클러스터링 결과는 Org()는 좋지 않고 Max()에서는 약간 좋은 것을 볼 수 있다.

하지만 (그림 7)과 (그림 8)은 데이터셋 C를 자질선택에 사용한 결과인데, 다소 양호한 것을 볼 수 있다. 이유는 데이터셋 C가 단어빈도 클러스터링 실험 결과가 좋은 데이터셋이기 때문이다. 따라서 기존 자질선택 방법을 사용한 클러스터링 결과는 자질선택에 사용한 데이터셋에 따라 많은 영향을 받는 걸 볼 수 있다.

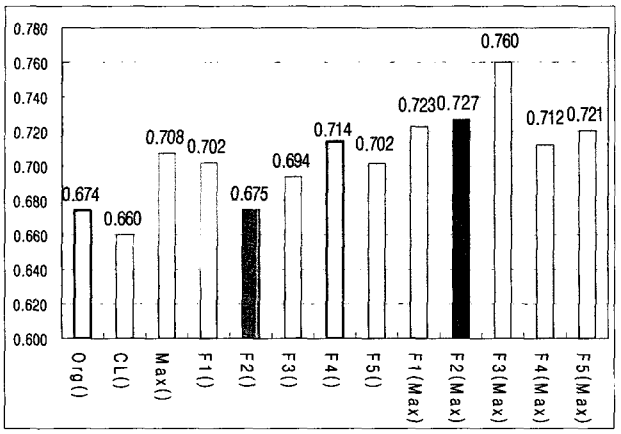
4.3.2 동시-링크 정보를 결합한 실험 결과 비교

데이터셋 B에 대해 단어빈도를 이용한 클러스터링 결과 Org(tf)와 문서간 동시-링크 정보만 갖는 매트릭스를 이용한 클러스터링 실험 CL(tf), 단어빈도와 동시-링크 정보를 결합한 매트릭스의 클러스터링 실험 CL(tf+co) 결과를 비교하면 <표 6>과 같다.

Org(tf)가 동시-링크 정보를 이용한 클러스터링 결과 CL(tf)가 단어빈도를 이용한 클러스터링 결과 Org(tf) 보다 성능이 좋지 않다.



(그림 10) 자질선택 데이터셋 A. 클러스터링 데이터셋 B와 C의 Purity 평균.



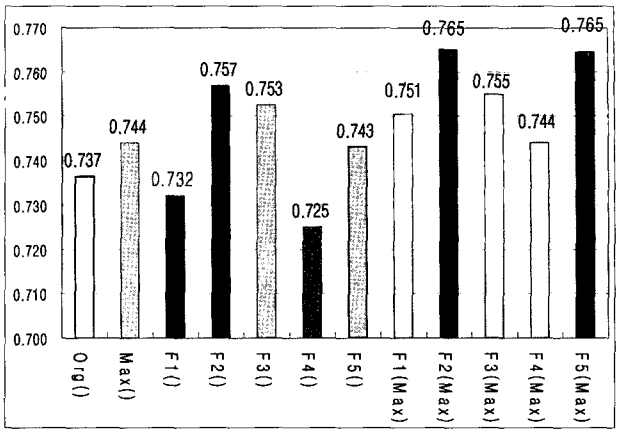
(그림 11) 자질선택 데이터셋 B. 클러스터링 데이터셋 A와 C의 Purity 평균

그 이유는 데이터셋에 따라 차이는 있지만, 단어빈도로 선택한 자질을 갖는 문서군과 동시-링크 정보로 선택한 자질을 갖는 문서군은 자질의 특성이 서로 다르기 때문에 성능 비교가 어렵다. 그래서 이 두 자질 정보를 결합하여 데이터셋 매트릭스에 적용해(그림 1) 클러스터링 했으나, 그 결과 CL(tf+co) 역시 좋지 않다.

4.3.3 전체 데이터셋 내에서의 자질 필터링에 의한 실험 결과와 비교

위에 나타난 그래프는 자질선택방법 X^2, MI, IG, OR 을 이용해 자질을 필터링 한 결과로 Purity의 평균값이다. 즉 F1()은 F1(X^2), F1(MI), F1(IG), F1(OR) 실험 결과 Purity의 평균값이다.

기존 자질선택 방법에 의한 실험결과 평균값Org()와 단위 문서 내 자질 필터링 방법에 의한 실험 결과 평균값Max(), 그리고 데이터셋 내 자질 필터링 방법 5가지 실험 결과 평균값 F1(), F2(), F3(), F4(), F5(), 단위 문서 내 자질 필터링 방법과 5가지 필터링 방법을 혼합한 실험 결과 평균값을 F1(Max), F2(Max), F3(Max), F4(Max), F5(Max)라 하자. 각 클러스터링 결과를 보면 다음과 같다.



(그림 12) 자질선택 데이터셋 C. 클러스터링 데이터셋 A와 B의 Purity 평균

(그림 10) 데이터셋 A을 이용한 자질 선택 실험에서는, Org()와 Max()는 아주 좋지 않다. 하지만 F1(), F3(), F4(), F5()와 F1(Max), F3(Max), F4(Max), F5(Max)에서 좋은 결과를 보여준다. 그 중에도 F1(Max), F4(Max), F5(Max)에서 아주 좋은 결과를 볼 수 있다. 따라서 데이터셋 A를 자질선택에 사용했을 경우는 자질 필터링 방법 F1+F3+Max 기법을 혼합한 F4(Max)가 가장 성능이 우수했다.

(그림 11) 데이터셋 B를 이용한 자질 선택 실험에서는 링크정보를 이용한 자질 확장에 의한 결과는 아주 좋지 않았고, 단위 문서 내 자질 필터링 방법 Max()는 기존 자질 선택 방법 Org()보다 좋았다. 그리고 F2() 외 모두 좋은 결과를 보였다. 그 중에도 F3(Max)에서 가장 우수한 성능을 보였다. 그리고 F1(Max)와 F2(Max), F5(Max)에서도 성능이 아주 좋다. 따라서 데이터셋 B를 자질선택에 사용했을 경우는 자질 필터링 방법 F3+Max 기법을 혼합한 F3(Max)가 가장 성능이 우수했다.

(그림 12) 데이터셋 C를 이용한 자질 선택 실험에서는, Org()와 Max() 성능 모두 좋다. 하지만 F2() F3()에서는 성능이 좋지만, F1()과 F4()에서는 성능이 떨어진 것을 볼 수 있다. 그리고 F1(Max), F2(Max), F3(Max), F5(Max) 에서 아주 좋은 결과를 보였다. 따라서 데이터셋 C를 자질선택에 사용했을 경우는 자질 필터링 방법 F2+Max 기법을 혼합한 F2(Max)과 F2+F3+Max 기법을 혼합한 F5(Max)가 가장 성능이 우수했다.

5. 결 론

색인 전문가에 의해 수작업으로 분류된 웹 문서의 경우 기존 자질선택방법에 의해 클러스터링 했을 때 데이터셋에 따라 다소 차이는 있지만 좋은 성능을 얻지 못했다. 그 이유는 대부분의 웹 문서는 문서 내 대표하는 자질 외 클러스터링에 불필요한 단어들을 포함하고 있기 때문이다. 그래서 기존 자질선택방법으로는 제거하지 못한 불필요한 자질들을 자질들의 문서군별 출현여부와 문헌빈도 등의 분포와 정도

를 기준치로 정해 필터링 해 보았다.

그 결과 단위 문서 내 자질 필터링 방법(FFID)도 좋은 성능을 보였고, 또 전체 데이터셋 내 자질 필터링 방법(FFIM)을 사용했을 때 더 좋은 성능을 보였다. 그리고 이 두 가지를 결합한 자질 필터링 방법(HFF)을 사용했을 때 가장 좋은 성능을 보였다.

특히 단어빈도를 이용한 클러스터링 결과 성능이 나쁜 데이터셋 A로 추출한 자질을 다른 두 데이터셋에 적용해 클러스터링 한 결과(그림 10) 아주 좋은 성능을 보였다.

즉 단어빈도를 이용한 자질선택방법에서는 본 논문에서 제시한 자질 필터링 방법을 사용하면 기존 자질선택방법에 의한 클러스터링 성능보다 아주 좋은 결과를 얻을 수 있다는 증명한 것이다.

6. 향후 연구 과제

웹 문서 클러스터링에서 기존의 자질선택방법을 사용하면 데이터셋에 따라 성능에 차이를 보인다. 그래서 본 논문에서 제시한 자질 필터링 방법을 사용했을 때 좋은 성능을 보였다. 하지만 자질 필터링 방법은 자질을 제거하는 방법이므로 문서 내 단어 수가 적을 경우 클러스터링 성능을 좋지 않는 영향을 미친다.

따라서 자질 필터링방법을 적용할 때, 데이터셋에 따라 기준치에 대한 학습이 필요하다. 또 동시-링크 정보는 웹 문서간의 의미 있는 정보지만 결과가 좋지 않았다. 그래서 동시-링크 정보를 확장한 클러스터링 연구가 계속해서 필요하다.

그리고 포탈사이트에서 전문가에 의해 수작업으로 분류된 웹 문서가 통계적 방법론에 의한 문서 클러스터링에서는 성능이 좋지 않기 때문에 웹 문서의 특징과 자질 선택에 많은 연구가 있어야 한다.

참 고 문 헌

[1] 이재윤, "자질값투표 기법과 문서측 자질 선정을 이용한 고속 문서 분류기", 12회 정보관리학회지, pp.71-78, 2005.

[2] 정영미, 이재윤, "지식 분류의 자동화를 위한 클러스터링 모형 연구", 정보관리학회지, Vol.18권, No.2, pp.203-230, 2001.

[3] 고영중, 서정연, "문서 관리를 위한 자동 문서 범주화에 대한 이론 및 기법", 정보관리연구논문지, Vol.33, No.2, pp.16-32, June, 2002.

[4] 국민상, 정영미, "자질선정에 따른 Naïve Bayesian 분류기의 성능 비교", 7회 정보관리학회 제7회 학술대회 논문집, pp.33-36, 2000.

[5] 이원희, 이교운, 박흠, 김영기, 권혁철, "웹 문서의 단어정보와 링크정보 결합을 이용한 클러스터링 기법", 15회 한국정보과학회지, pp.101-107, 2003.

[6] H.Yaun, S.S.Tseng, W.Gangshan, and Z.Fuyan. "A two-phase feature selection method using both filter and wrapper", In IEEE International conference on Systems, Man, and Cybernetics, Vol.2, pp.132-136, 1999.

[7] Heum Park, "A Feature Selection for Korean Web Document Clustering", The 30th Annual Conference of IEEE Industrial Electronics Society, 2004.

[8] Hall, M. "Correlation-based feature selection of discrete and numeric class machine learning", In Proceedings of the International Conference on Machine Learning, pp.359-366, San Francisco, CA. Morgan Kaufmann Publishers, 2000.

[9] A.Y. Ng, "On feature selection : learning with exponentially many irrelevant features as training examples". In Proc. 15th Intl. Conf. on Machine Learning, pp.404-412, 1998.

[10] Zhao, Ying and Karypis, George, "Criterion functions for document clustering - experiment and analysis", Technical Report TR #01-40, Department of Computer Science, University of Minnesota, 2001.

[11] Zhao, Ying and Karypis, George, "Evaluation of hierarchical clustering algorithms for document datasets", Technical Report TR #02-22, Department of Computer Science, University of Minnesota, 2002.

[12] Karypis, George, "CLUTO : A Clustering Toolkit", Technical Report TR #02-017, Department of Computer Science, University of Minnesota, 2002.

[13] Zhi-Hong Deng, Shi-Wei Tang, Dong-Qing Yang, Ming Zhang, Xiao-Bin Wu and Meng Yang, "Two Odds-Radio-Based Text Classification Algorithms", Proceedings of Web Information Systems Engineering(Workshops) pp.223-231, 2002.

[14] Brank, J., Grobelnik, M., Mili´c-Frayling, N. & Mladeni´c, D., "Interaction of feature selection methods and linear classification models", Proceedings of the ICML-02 Workshop on Text Learning, Sydney, AU, 2002.

[15] Y. Yang and J. P. Pedersen, "A comparative study on feature selection in text categorization", In Proceedings of the International Conference on Machine Learning, pp.412-420, 1997.

[16] Kyo-Woon Lee, Young-Gi Kim, Hyuk-Chul Kwon, "Clustering of Web Documents with the Use of Term Frequency and Co-link in Hypertext", Proceedings of the International Conference on APIS2003, 2003.

박 흠

e-mail : parkheum2@empal.com

1988년 부산대학교 자연과학대학 계산통계학과(학사)

1998년 부산대학교 일반대학원 인지과학협동과정(이학석사)

2005년 부산대학교 일반대학원 정보통신협동과정 박사과정 수료

1988년~1990년 코닉시스템㈜

1990년~1998년 부산일보

2005년~현재 유비텍㈜ 이사

관심분야 : 한국어정보처리, 정보검색, 유비쿼터스, 텔레메틱스

권 혁 철

e-mail : hckwon@pusan.ac.kr

1982년 서울대학교 공과대학 전산학 학사

1984년 서울대학교 공과대학 전산학 석사

1987년 서울대학교 공과대학 전산학 박사

1988년~현재 부산대학교 전자전기정보컴퓨터공학부 교수

1988년~현재 한국정보과학회 프로그래밍언어 연구회 운영위원

1990년~현재 한국정보과학회 한국어정보처리 연구회 운영위원

1992년~1993년 미국 Stanford 대학 CSLI연구소 연구원

1992년~1993년 Xerox Palo Alto Research Center 자문위원

2003년~현재 BK21 산업자동화 및 정보통신분야 인력양성사업단 단장

2004년~현재 한국정보과학회 이사

2006년~현재 한국인지과학회 이사

관심분야 : 한국어정보처리, 정보검색, 프로그래밍언어, 인공지능, 시맨틱웹