# 용어 가중치부여 기법을 이용한 로치오 분류기의 성능 향상에 관한 연구

## A Study on the Performance Improvement of Rocchio Classifier with Term Weighting Methods

김판준(Pan-Jun Kim)*


### 초 록

로치오 알고리즘에 기반한 자동분류의 성능 향상을 위하여 두 개의 실험집단(LISA, Reuters-21578)을 대상으로 여러 가중치부여 기법들을 검토하였다. 먼저, 가중치 산출에 사용되는 요소를 크게 문헌요소(document factor), 문헌집합 요소(document set factor), 범주 요소(category factor)의 세 가지로 구분하여 각 요소별 단일 가중치부여 기법의 분류 성능을 살펴보았고, 다음으로 이들 가중치 요소들 간의 조합 가중치부여 기법에 따른 성능을 알아보았다. 그 결과, 각 요소별로는 범주 요소가 가장 좋은 성능을 보였고, 그 다음이 문헌집합 요소, 그리고 문헌 요소가 가장 낮은 성능을 나타냈다. 가중치 요소 간의 조합에서는 일반적으로 사용되는 문헌 요소와 문헌집합 요소의 조합 가중치(*tfidf or ltfidf*)와 함께 문헌 요소를 포함하는 조합(*tf**cat or *ltf**cat) 보다는, 오히려 문헌 요소를 배제하고 문헌 집합 요소를 범주 요소와 결합한 조합 가중치 기법(*idf**cat)이 가장 좋은 성능을 보였다. 그러나 실험집단 측면에서 단일 가중치와 조합 가중치를 서로 비교한 결과에 따르면, LISA에서 범주 요소만을 사용한 단일 가중치(cat only)가 가장 좋은 성능을 보인 반면, Reuters-21578에서는 문헌집합 요소와 범주 요소간의 조합 가중치(*idf**cat)의 성능이 가장 우수한 것으로 나타났다. 따라서 가중치부여 기법에 대한 실제 적용에서는, 분류 대상이 되는 문헌집단 내 범주들의 특성을 신중하게 고려할 필요가 있다.

### ABSTRACT

This study examines various weighting methods for improving the performance of automatic classification based on Rocchio algorithm on two collections(LISA, Reuters-21578). First, three factors for weighting are identified as document factor, document factor, category factor for each weighting schemes, the performance of each was investigated. Second, the performance of combined weighting methods between the single schemes were examined. As a result, for the single schemes based on each factor, category-factor-based schemes showed the best performance, document set-factor-based schemes the second, and document-factor-based schemes the worst. For the combined weighting schemes, the schemes(*idf**cat) which combine document set factor with category factor show better performance than the combined schemes(*tf**cat or *ltf**cat) which combine document factor with category factor as well as the common schemes(*tfidf* or *ltfidf*) that combining document factor with document set factor. However, according to the results of comparing the single weighting schemes with combined weighting schemes in the view of the collections, while category-factor-based schemes(cat only) perform best on LISA, the combined schemes(*idf**cat) which combine document set factor with category factor showed best performance on the Reuters-21578. Therefore for the practical application of the weighting methods, it needs careful consideration of the categories in a collection for automatic classification.


키워드: 가중치, 가중치부여 기법, 로치오 알고리즘, 문헌분류, 텍스트 범주화, 분류기
weighting, Rocchio algorithm, classifier, automatic classification, text categorization, feature selection principles

* 연세대학교 문헌정보학과 강사(long4azure@paran.com)

■ 논문접수일자 : 2008년 2월 19일 ■ 게재확정일자 : 2008년 3월 10일
■ 情報管理學會誌, 25(1): 211-233, 2008. [DOI:10.3743/KOSIM.2008.25.1.211]

# 1. 서 론

문헌의 자동분류에서 분류 성능의 향상을 위해 자질로서 사용되는 용어에 가중치를 부여하는 여러 방법들을 용어 가중치부여 기법(term weighting schemes)이라 한다. 이러한 가중치 부여 기법은 문헌을 표현하는 색인어에 적절한 가중치를 부여하여 정보검색의 성능을 높이는 것과 마찬가지로, 문헌의 분류를 위한 자질로서 사용되는 용어에 적절한 가중치를 부여함으로써 분류 성능을 향상시키기 위한 것이다.

지금까지 자동분류에서 많이 사용되어 온 가중치부여 기법은 정보검색 분야에서 사용되어 온 것으로 용어의 문헌 또는 문헌집합 내 출현정보에 기반하는 것이었다. 그러나 대부분의 정보검색 환경에서 활용가능한 정보가 용어의 출현정보만으로 제한되는 것과는 달리, 문헌의 자동분류에서는 이러한 출현정보와 함께 분류문헌에 부여되어 있는 범주정보를 추가적인 단서로 활용할 수 있다는 특징이 있다. 다시 말해서 자동분류는 가중치 산출을 위한 세 가지 요소로서 문헌, 문헌집합, 범주 요소를 모두 고려할 수 있으므로, 정보검색보다 가중치 부여를 위한 가용 정보가 더 많다고 할 수 있다. 또한, 일반적으로 정보검색의 목적이 이용자의 질의에 대한 적합문헌을 제시하는 것인 반면, 자동분류의 목적은 입력문헌에 대한 적합범주를 결과물로서 제시하는 것이다. 따라서 자동분류에서는 이러한 목적의 달성을 위해 개별 문헌과 전체 문헌집합 단위의 정보는 물론 범주 단위의 정보가 보다 중요한 자원으로서 활용되어야 한다.

본 연구는 활용 가능한 정보나 목적 측면에서 정보검색과는 다소 다른 특징을 갖는 자동분류의 성능 향상을 위하여 효과적인 가중치부여 기법을 모색하여 보고자 하는 것이다. 특히, 구현이 쉽고 컴퓨터 저장 공간 및 처리시간 측면에서 상당한 장점을 갖고 있는 로치오 분류기에 기반한 자동분류의 성능 측면에서, 여러 가중치부여 기법들의 성능을 비교 및 분석함으로써 가장 효율적인 방안을 모색하여 보고자 한다.

# 2. 자동분류를 위한 용어 가중치부여 기법

자동분류에 적용된 용어 가중치부여 기법들은 분류 성능을 향상시키기 위해 분류 자질로 사용되는 용어에 적절한 가중치를 부여하기 위한 것이다. 이러한 용어 가중치부여 기법은 정보검색 분야에서와 마찬가지로 특정 용어의 중요성을 반영할 수 있도록 적절한 가중치를 부여하는데 사용되어 왔다.

정보검색에서는 전통적으로 가중치부여 기법들을 한 문헌 내 용어의 출현에 기반한 지역적 가중치 기법(Local weighting schemes)과 전체 문헌집합 내 용어의 출현에 기반하는 전역적 가중치 기법(Global weighting schemes)으로 구분하였다. 여기서 지역적 가중치 기법은 단순히 특정 문헌 내 용어의 출현 여부(1 또는 0)만을 고려하는 이진 값이나 단순 출현빈도(tf), 또는 이러한 단순 출현빈도의 다양한 이형(tf's variants)들을 포함한다. 반면, 전역적 가중치 기법은 전체 문헌집합 내에서 특정 용어의 출현 정보를 고려하는 것으로 주로 역문헌 빈도

(idf)에 기반한 것이다. 많은 자동분류 연구에서 사용되어 온 것으로, 이러한 출현빈도와 역문헌빈도의 조합에 따른 가중치부여 기법(tfidf)은 특정 문헌 내 출현빈도(tf)가 높으면서, 전체 문헌집단 내 출현빈도(df)는 낮은 용어가 더 중요하다는 가정에서 비롯된 것이다(Salton and McGill 1983). 한편 최근의 기계학습 기반의 자동분류 연구들에서는 학습문헌에 부여된 범주 정보의 사용 여부에 따라 가중치부여 기법을 구분하는 경향이 나타났다. 즉 문헌 또는 문헌집합 내 출현 정보만을 사용하는 경우를 비지도 가중치부여 기법(unsupervised weighting methods)이라 하고, 용어가 출현한 문헌에 부여된 범주 정보를 사용하는 경우는 지도 가중치부여 기법(supervised weighting methods)으로 구분하는 것이다(Debole and Sebastiani 2003; Deng et al. 2004; Soucy and Mineau 2005; Lan et. al. 2006; 김판준 2007).

이와 같이 자동분류를 위한 용어 가중치부여 기법은 초기에는 정보검색 분야의 대표적인 가중치부여 기법으로서 문헌과 문헌집합 내 출현 정보를 조합한 가중치(*tfidf*)를 주로 사용해 왔으나, 최근에 와서 범주 정보의 사용을 적극적으로 검토하는 다양한 시도가 진행되고 있다. 즉, 자동분류를 목적으로 적용된 가중치 산출의 기반이 되는 단서로 앞에서 언급한 기존의 두 가지 요소 이외에, 가중치 산출을 위한 새로운 정보원으로서 범주 요소를 사용하는 연구가 활발히 이루어지고 있는 것이다.

자동분류 측면에서 범주 정보를 사용하는 지도 가중치 부여 기법과 유사한 것으로 정보검색 분야에서 질의어에 부여하기 위한 목적으로 제안된 적합성 가중치(relevance weight)가 있다(Robertson & Sparck Jones 1976; Robertson 2007). 이러한 적합성 가중치는 이용자의 질의어를 대상으로 문헌의 적합성 정보에 기초한 가중치를 부여하는 방법이다. 검색 측면에서 특정 문헌의 적합성 판정을 분류 측면에서 해당 문헌에 대한 범주의 부여로 간주한다면, 정보검색에서의 적합성 가중치와 자동분류에서 범주 정보에 기초한 일부 가중치부여 기법들은 거의 차이가 없는 것이다.

또한 자동분류를 위한 주요 단계 중의 하나로서 자질선정 기준에 관한 연구들은 용어의 중요성에 기초하여 자동분류를 위한 자질을 선정한다는 측면에서, 여러 가중치부여 기법들과 상당히 유사한 목적과 절차를 공유하고 있다(Yang and Pedersen 1997; Yang 1999; Brank, J. et al. 2002; Forman 2003; Rogati and Yang 2003). 이에 따라 최근 자동분류 분야에서는 이전의 자질선정 관련 연구들에서 좋은 성능을 보인 여러 자질선정 기준에 의해 산출된 자질 값을 자동분류를 위한 가중치로서 그대로 활용하는 단순하면서도 효율적인 접근법들이 제안되고 있다(Deng et al. 2004; Liu et al. 2007).

지금까지 자동분류에 적용된 가중치부여 기법들은 분류 자질이 되는 용어에 부여할 가중치의 산출에 사용된 요소 측면에서 다음과 같이 크게 세 가지로 구분할 수 있다.

첫째, 특정 문헌 내 용어의 출현 정보에 따른 문헌 요소 기반의 가중치부여 기법(document factor: *bin*, *tf* and its variants)

둘째, 전체 문헌집합 내 용어의 출현 정보에 따른 문헌집합 요소 기반의 가중치부여 기법

(document set factor: *idf*)

셋째, 특정 범주의 부여 여부에 따른 범주 요소 기반의 가중치부여 기법(category factor: relevance weights or feature selection weights)

아래의 〈표 1〉은 이러한 세 가지 요소 측면에서, 본 연구에서 사용된 여러 가중치부여 기법들을 제시한 것이다. 여기서 적합성 가중치(RW: Relevance Weight) 또는 자질선정 기준에 따른 가중치(FSW: Feature Selection Weight)로 제시된 여러 가중치부여 기법들은 적용 분야나 연구자에 따라 부르는 명칭만 다를 뿐, 서로 동일한 것이거나 일부 변형이 이루어진 것으로 모두가 범주 요소에 기반하고 있다는 특징을 갖는다.

## 2.1 단일 가중치부여 기법

### 2.1.1 문헌 요소

문헌 요소는 용어의 특정 문헌 내 출현정보에 따른 것이다. 가장 기본적인 문헌 요소 기반의 가중치부여 기법은 특정 문헌 내 개별 용어의 출현 여부에 따른 이진값 또는 단순 출현빈도를 그대로 가중치로 사용하는 것이다. 또한 단순 출현빈도의 이형으로서 이를 정규화하거나 일부 변형한 것으로 문헌 요소에 기반한 다양한 가중치 부여 기법들이 제안되었다. 본 연구에서 사용된 것으로 문헌 요소에 기반한 가중치부여 기법들은 〈표 2〉와 같다.

### 2.1.2 문헌집합 요소

문헌집합 요소는 용어의 전체 문헌집합 내 출현정보에 기초한 것이다. 문헌집합 요소에 기반한 대표적인 가중치부여기법으로서 역문헌빈도(*idf*)는 1972년에 Sparck Jones의 논문에서 용어 특정성(term specificity)의 척도로 처음 제안되었다(Sparck Jones 1972). 이는 특정성 측면에서 용어의 중요성은 많은 수의 문헌에 출현한 용어(고빈도어: frequent term)보다 적은 수의 문헌에 출현한 용어(저빈도어: rare term)가 더 크기 때문에, 전체 문헌집합 내에서 더 적은 수의 문헌에 출현한 용어에 높은 가중치를 부여해야 한다는 직관에서 비롯된 것이다. 자동분류 측면에서 이러한 역문헌빈도는 범주 정보(적합성 정보)를 사전에 알 수 없

〈표 1〉 가중치 산출을 위한 세 가지 요소 측면에서 구분한 가중치부여 기법들의 분류

| 구 분 | | 약어 표현 | 설 명 | 비 고 |
|---|---|---|---|---|
| 문헌 요소 | | *bin, tf, variants of tf(atf, itf, ltf, otf, stf)* | 이진값(1 또는 0), 문헌 내 출현빈도, *tf*의 이형들 | 지역적 또는 비지도 |
| 문헌집합 요소 | | *idf* | 역문헌빈도<br>범주 정보를 사용하지 않는 적합성 가중치 | 전역적 또는 비지도 |
| 범주 요소 | 적합성 가중치 (RW) | *rw1(=mi), rw2, rw3, rw4(=lor), simplified rws(srw1 ~ srw7)* | 적합성 가중치1(=상보정보량), 적합성 가중치2(로그 승산비의 이형), 적합성 가중치3(로그 승산비의 이형), 적합성 가중치4(=로그 승산비), 적합성 가중치의 단순형1~7 | 지역적 또는 지도 |
| | 자질선정 가중치 (FSW) | *CC(cos, dice, jac), chi , pi, rmij, mi(=rw1), lor(=rw4), yule's y* | 상관계수(코사인계수, 다이스계수, 자카드계수), 카이제곱 통계량($\chi^2$), 상호정보량의 이형(proximity index), 상대적 상호정보량 *j*, 상호정보량, 로그 승산비, 율의 *y* | |

* 가중치부여 기법을 지역적/전역적 또는 지도/비지도로 분류한 것은(김판준 2007)을 참조.

〈표 2〉 문헌 요소에 기반한 단일 가중치부여 기법: 이진값(*bin*), 단순 출현빈도(*tf*), 단순 출현빈도의 이형들(variants of *tf*)

| 약 어 | 공 식 | 설명/출처 |
|---|---|---|
| *bin* | 1 or 0 | $tf$ = 1, if $tf$ 〉 0 |
| *tf* | 단순 출현빈도 | 문헌 내 용어 출현빈도 |
| *atf* | $0.5+(0.5\times(\frac{tf}{\max tf}))$ | 보정 *tf*(augmented *tf*) |
| *itf* | $1-(\frac{1}{(1+tf)})$ | 역용어빈도(Inverse Term Frequency) |
| *ltf* | $\log(1+tf)$ | log *tf* |
| *otf* | $\frac{tf}{(2+tf)}$ | okapi *tf* |
| *stf* | $\sqrt{tf}$ | square *tf* |

을 경우에 적용가능한 적합성 가중치로 볼 수도 있다. 즉, 범주 정보를 전혀 사용할 수 없는 경우에 단순히 특정 용어가 전체 문헌집합 내에서 얼마나 많이 출현했는가에 따라 용어의 중요성을 산출할 수 있는 가중치부여 기법으로서 기본 공식은 다음과 같다.

$$idf(t_i)=\log\frac{N}{n_i}$$

2.1.3 범주 요소

범주 요소는 용어가 출현한 문헌에 부여된 범주 정보에 기초한 것으로, 최근 이러한 범주 정보를 이용하여 용어의 중요성을 나타내는 가중치 값을 산출하는 여러 가중치부여 기법들이 제안되었다(Debole and Sebastiani 2003; Deng et al. 2004; Lan et al. 2006; Liu et al. 2007). 이들은 정보검색 분야의 적합성 가중치 공식 또는 자동분류를 위한 자질선정 기준에 따른 가중치를 산출하여 용어에 부여하는 접근법을 사용하고 있는데, 적용 분야 또는 연구자에 따라 동일한 척도를 다른 이름으로 부르기도 하고 일부 변형하여 적용하기도 하였다. 한편 데이터 마이닝 분야에서도 특정 패턴(용어)의 중요성을 흥미 척도(interestingness measures)라는 명칭으로 산출하고 있는데, 사실상 위에서 언급한 여러 가중치부여 기법과 동일하거나 유사한 척도를 사용하고 있다(Geng et al. 2006).

본 연구에서 사용되는 범주 요소 기반 가중치부여 기법들은 기본적으로 벡터-공간 모형과 확률 모형에 기반하고 있으며, 벡터를 구성하는 각 요소에 대한 가중치의 산출을 위해서는 〈표 3〉과 같은 분할표가 필요하다.

〈표 3〉 2×2 분할표

| | $c_i$ | $\overline{c_i}$ |
|---|---|---|
| $t_k$ | $a$ | $b$ |
| $\overline{t_k}$ | $c$ | $d$ |

a: 용어 $t_k$가 적어도 한번 출현한 범주 $c_i$에 속한 문헌 수

b: 용어 $t_k$가 적어도 한번 출현한 범주 $c_i$에 속하지

않은 문헌 수

c: 용어 $t_k$가 출현하지 않은 범주 $c_i$에 속한 문헌 수

d: 용어 $t_k$가 출현하지 않은 범주 $c_i$에 속하지 않은 문헌 수

또한 이러한 분할표 상의 각 셀의 값에 기반하여 〈표 4〉의 범주 요소에 포함되는 여러 가중치부여 기법들을 적용할 수 있다.

## 2.2 조합 가중치부여 기법

자동분류 분야에서 1990년대 후반까지 가장 많이 사용되어 온 가중치는 본질적으로 조합 가중치부여 기법에 속한다. 즉 자동분류에서 일반적으로 사용되어 온 것으로 *tfidf* 또는 *ltfidf*는 문헌 요소와 문헌집합 요소를 조합한 가중치부여 기법이라 할 수 있다(Joachims 1996; Joachims 1998; Yang 1999). 그러나 1970년대 정보검색 분야에서는 질의어에 대한 가중치로서 문헌 요소와 적합성 정보에 기초한 범주 요소의 조합 가중치를 부여한 결과, 검색 성능 측면에서 상당한 효과를 보여주었고(Salton et al. 1981), 최근 몇 년간 자동분류에서도 이와 유사한 접근법들이 활발히 제안되고 있다. 이들은 대부분 문헌과 문헌집합 요소의 조합 가중치 기본형(*tfidf* or *ltfidf*)에서 문헌요소(*idf*)를 적합성 가중치(RW: Relevance Weight) 또는 자질선정 기준에 의한 가중치(FSW: Feature Selection Weight)로 대체하는 형식을 취하고 있다.

Debole & Sebastiani(2003)는 SVM 분류기를 사용하는 경우에 이러한 형식의 조합 가중치($tf\chi^2$, *tfig*)가 모두 낮은 성능을 보이는 것으로 보고하였지만, Deng et al.(2004)의 SVM 기반한 연구에서는 문헌 요소와 범주 요소의 조합($tf\chi^2$)이 조합 가중치 기본형(*tfidf*)보다 더 효과적이라고 주장하였다. 또한 Lan 등(2006)이 다양한 가중치부여 기법들을 두 개의 문헌집단과 SVM, kNN 알고리즘에 대하여 적용해 본 결과, 대부분의 조합 가중치($tf\chi^2$, *tfig* and *tfor*)들이 단일 가중치들보다 오히려 낮은 성능을 보였지만 자신들이 제안한 조합 가중치(*tfrf*)는 상대적으로 좋은 성능을 보였다고 주장하였다. 그러나 이들의 연구에서 제안된 조합 가중치들은 기본형으로서 *tfidf* 조합 가중치 또는 여러 단일 가중치들보다 일관성 있게 나은 성능을 보여주지는 못하였다. 한편 Liu 등(2007)은 다양한 조합 가중치부여 기법들의 성능을 비교하는 연구에서, 실험에 사용된 16개 기법 중에서 정규화된 문헌 요소(Normalized *tf*)와 적합성 가중치의 단순형(*srw*: simplified relevance weight)을 조합한 가중치들이 일관성 있게 좋은 성능을 보였다고 보고하였다. 이처럼 최근 자동분류 연구에 적용되고 있는 여러 조합 가중치부여 기법은 단일 가중치부여 기법은 물론 다른 조합 가중치부여 기법들에 비하여 일관성 있게 우월한 성능을 보여주지는 못하고 있으며, 특히 조합을 위한 구성요소로서 문헌집합과 범주 요소 이외에 문헌집합 요소를 전혀 고려하지 않고 있다.

본 연구에서는 최근 여러 연구에서 제안된 접근법으로 문헌 요소와 범주 요소에 기반한 조합 가중치부여 기법은 물론, 새로운 접근법으로서 문헌 요소를 배제하고 문헌집합 요소(*idf*)와 범주 요소(category: RW or FSW)로 구성된 조합 가중치부여 기법들을 추가적으로

〈표 4〉 범주 요소에 기반한 단일 가중치부여 기법

| 번호 | 약어 | 공 식 |
|---|---|---|
| 1 | *cos* | $\frac{a}{\sqrt{((a+b)(a+c))}}$ |
| 2 | *dice* | $\frac{2a}{(2a+b+c)}$ |
| 3 | *jac* | $\frac{a}{(a+b+c)}$ |
| 4 | *chi* | $\frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$ |
| 5 | *pi* | $\frac{Na}{(a+b)(a+c)}$ |
| 6 | *rmij* | $\frac{\log(\frac{Na}{(a+b)(a+c)})}{\log(\frac{N}{a+b})+\log(\frac{N}{a+c})-\log(\frac{Na}{(a+b)(a+c)})}$ |
| 7 | *yule's y* | $\frac{(\sqrt{ad}-\sqrt{bc})}{(\sqrt{ad}+\sqrt{bc})}$ |
| 8 | *rw1*<br>*(mi)* | $\log(\frac{Na}{(a+b)(a+c)})$ |
| 9 | *rw2* | $\log(\frac{ab+ad}{ab+bc})$ |
| 10 | *rw3* | $\log(\frac{ac+ad}{ac+bc})$ |
| 11 | *rw4*<br>*(lor)* | $\log(\frac{ad}{bc})$ |
| 12 | *srw1* | $\log(1+\frac{a}{c})$ |
| 13 | *srw2* | $\log(1+\frac{a}{b})$ |
| 14 | *srw3* | $\log(1+\frac{a}{b}\frac{a}{c})$ |
| 15 | *srw4* | $\log(1+\frac{a}{b})\log(1+\frac{a}{c})$ |
| 16 | *srw5* | $\log(1+\frac{(a+c)}{(b+d)})$ |
| 17 | *srw6* | $\log(1+(\frac{a}{(a+b)}))$ |
| 18 | *srw7* | $\log(1+(\frac{a}{(a+c)}))$ |

* *cos, dice, jac, chi, pi, rmij, yule's y, mi, lor*: 자동분류를 위한 자질선정 기준

** *rw1*~*rw4*: 정보검색 분야의 적합성 가중치(relevance weight)

*** *srw1*~*srw4*: 적합성 가중치의 단순형(simplified relevance weight)

검토하였다. 문헌의 자동분류를 위한 가중치부여 기법으로서 이러한 유형의 조합 방식은 지금까지는 보고된 바가 없는 것이다.

# 3. 실 험

## 3.1 실험집단

실험집단은 LISA database의 학술지 논문 집단과 Reuters-21578의 신문 기사 집단의 2가지를 사용하였다. 먼저, Journal of Citation Reports(2001년~2004년)의 데이터에 기초하여 정보학 분야의 3개 핵심 학술지를 선정하고, LISA 데이터베이스에서 1994년부터 2004년까지 이들 학술지에 수록된 논문을 다운로드하여 실험집단을 구성하였다(김판준 2006a). 다음으로, 지금까지 텍스트 범주화 연구에서 가장 많이 사용된 신문 기사 문헌집단으로서 Reuters-21578의 ModApte split을 사용하였다. 이들 두 가지 문헌집단에 대한 세부 내용은 〈표 5〉와 같다.

자동분류 실험을 위해 LISA 데이터베이스의 디스크립터 필드에 있는 디스크립터들을 논문에 부여된 범주로 간주하여 사용하였고, Reuter-21578에서는 신문기사에 부여된 범주를 그대로 사용하였다. 또한 분류 성능 측면에서 여러 가중치부여 기법의 비교를 위해, 두 실험집단에서 소속 학습문헌 수를 기준으로 가장 큰 15개 범주를 대상으로 자동분류 실험을 수행하였다.

문헌에 특정 범주를 부여하기 위한 단서가 되는 자질집합의 구성을 위해 LISA 데이터베이스에서는 검색한 각 논문의 제목과 초록(title & abstract) 필드의 단어를 사용하였고, Reuters-21578에서는 신문기사 제목과 본문(title & body) 필드의 단어를 사용하였다. 이러한 자질집합은 가중치부여 기법의 효과를 최대한 반영하기 위해 자질선정 단계를 수행하지 않고 Porter Stemmer를 사용한 형태소분석(stemming)과 불용어 제거 이후에 남아있는 모든 단어를 대상으로 구성하였다.

## 3.2 로치오 분류기

자동분류를 위한 분류기로서 사용되는 로치오 분류기는 다른 분류기들에 비하여 구현이 용이하고 컴퓨터 저장 공간 및 처리 시간 측면에서 효율적인 것으로 알려져 있다. 또한 성능 측면에서도 다른 기계학습 기반 분류기(SVM 등)와 비교하여 동등하거나 오히려 우수하다는 사

〈표 5〉 실험 문헌집단: LISA & Reuters-21578

| | LISA | Reuters-21578 |
|---|---|---|
| 전체 문헌수 | 1,888 | 21,578 |
| 학습문헌수/검증문헌수 | 1,666/192 | 7,058/2,742 |
| 범주 당 최대/평균 학습문헌수 | 409/15.5 | 2877/106.5 |
| 실험 범주수 | 15 | 15 |
| 용어 종수(unique terms) | 6,433 | 23,543 |

실을 이전의 연구에서 밝힌 바 있으므로, 이를 기본 분류기로 사용하여 여러 가중치부여 기법을 적용하였다(김판준 2006b).

본 연구에서 로치오 분류기로서 기능하는 범주 프로파일은 벡터공간 모형과 로치오 알고리즘에 기반하는 것으로, 특정 범주의 부여 여부에 따라 문헌에 출현한 단어들을 중심으로 생성하였다. 즉, 본 연구에서 사용된 로치오 분류기는 긍정예제(POS: Positive examples)과 부정예제(NEG: Negative examples)를 함께 사용하는 일반적인 알고리즘과는 달리, 긍정예제에 출현한 단어들만으로 구성되며 부정예제는 고려하지 않았다. 따라서 벡터-공간 모형 측면에서 각 문헌은 해당 문헌에 출현한 용어들로 구성된 문헌벡터로서 하나의 문헌 프로파일이 되는 것이며, 이들 문헌에 부여된 범주 정보에 따라 각 범주가 부여된 문헌벡터의 요소들만으로 하나의 범주 프로파일이 생성되는 것이다.

로치오 분류기로서 범주 프로파일의 생성과 및 가중치부여를 위한 절차는 다음과 같다. 먼저, 문헌 프로파일로서 각 문헌벡터는 학습문헌에 출현한 용어들에 기반하여 생성하였다. 다음으로, 범주 프로파일로서 각 범주벡터는 해당 범주가 부여된 문헌벡터의 용어들로 구성하였고, 벡터를 구성하는 각 요소의 값은 범주에 속한 문헌 내 용어에 부여된 가중치들의 합으로 부여하였다.

$$\vec{d}_{train} = (t_1, t_2, t_3, ..., t_k),\ t_k = wt_k$$

$$\vec{c}_i = (wt_1, wt_2, wt_3, ..., wt_k),\ wt_k = \sum_{d \in POS} wt_k$$

디스크립터가 부여될 입력문헌으로서 검증집단의 각 문헌은 문헌에 출현한 단어들로 구성된 벡터로 생성하였고, 벡터를 구성하는 각 단어의 가중치는 단순 출현빈도($tf$)를 사용하였다.

$$\vec{d}_{test} = (wd_1, wd_2, wd_3, ..., wd_k),\quad wd_k = tf_k$$

범주의 부여 결정은 범주 프로파일($\vec{c}$)과 이러한 입력 문헌($\vec{d}$) 간의 유사도를 산출하여 결정하였다. 본 연구에서는 정보검색에서 많이 사용되고 있는 코사인 상관계수를 이용하여 두 가중치 벡터 간의 유사도를 기준으로 범주를 부여하였다. 또한, 범주의 최종 판정을 위한 기준치는 학습집단에서 가장 좋은 성능을 나타낸 기준치를 검증집단에 동일하게 적용하는 방법을 채택하였다.

$$Cosine(c, d) = \frac{\sum_c wt_k \cdot \sum_d wd_k}{\sqrt{\sum_c wt_k^2} \cdot \sqrt{\sum_d wd_k^2}}$$

### 3.3 실험 구성

로치오 분류기 기반의 자동분류에서 여러 가중치부여 기법의 적용에 따른 성능을 알아보았다. 먼저 용어 가중치부여 기법을 가중치 계산에 사용되는 세 가지 요소에 따라 구분하고, 각 요소별 단일 가중치부여 기법의 성능과 이들 요소의 결합에 따른 조합 가중치부여 기법의 성능을 알아보기 위한 실험을 수행하였다. 다음으로, 이전 실험에서 가장 좋은 성능을 보인 단일 가중치부여 기법과 조합 가중치부여 기법

들의 성능을 최종적으로 비교 및 분석하였다. 여기서 다양한 가중치부여 기법들의 성능을 비교하기 위한 기본형으로는 자동분류 연구에서 많이 사용되어 온 조합 가중치(*tfidf* or *ltfidf*)의 성능을 기준으로 삼았다.

먼저 세 가지 요소별 단일 가중치부여 기법들의 성능을 비교하는 실험을 수행하였다. 〈표 6〉은 본 연구에서 실험에 사용된 3개 요소별 단일 가중치부여 기법들이다.

다음으로 가중치부여 기법을 구성하는 세 가지 요소의 조합에 따른 성능을 알아보기 위한 실험을 수행하였다. 가중치 조합의 방법으로는 각 요소 간의 곱(product)을 사용하였으며, 조합 가중치 기본형의 성능을 기본형으로 삼아 크게 두 가지 유형의 조합 가중치 기법의 성능을 비교하였다. 여기서 첫 번째 유형은 최근 여러 연구에서 제안된 것으로 문헌 요소를 포함하는 범주 요소와의 조합 가중치이고, 두 번째 유형은 문헌 요소를 배제하고 대신 문헌집합 요소와 범주 요소를 조합한 것이다. 〈표 7〉은 본 연구의 실험에 사용된 가중치 요소들의 조합에 따른 여러 조합 가중치부여 기법들이다. 여기서 조합에 사용된 각 요소별 단일 가중치부여 기법들은 사전 실험에서의 결과에 따라 선정한 것이다. 즉 문헌 요소에 속한 단일 가중치부여 기법들 간에는 이진 가중치부여 기법을 제외하고는 서로 간에 성능 차이가 거의 없었으므로, 가장 많이 사용되는 *tf*와 *ltf*를 사용하였다. 그리고 문헌집합 요소로는 역문헌빈도(*idf*)를 사용하였고, 범주 요소로는 사전 실험에서 가장 좋은 성능을 보인 것으로 상위 집단에 속한 10개 기법을 사용하였다.

전체 실험의 결과 및 분석에 사용된 성능 척도는 정보검색에서 일반적으로 사용되는 재현율과 정확률, 그리고 이들 두 가지 성능 척도를 하나의 지표로 표현할 수 있는 $F_1$ 척도를 사용

〈표 6〉 실험에 사용된 각 요소별 단일 가중치 부여기법

| 구 분 | 단일 가중치부여 기법들(갯수) |
|---|---|
| 문헌 요소 | *bin, tf, atf, itf, ltf, otf, stf(7)* |
| 문헌집합 요소 | *idf(1)* |
| 범주 요소 | *cos, dice, jac, chi, gss, rw1(mi), pi, rw2, rw3, rw4(lor), yule's y, srw1~srw7(18)* |

〈표 7〉 실험에 사용된 요소 간의 조합 가중치부여 기법의 유형

| 유 형 | 조합 방법 | 약어 표현 |
|---|---|---|
| 기본형 | 문헌 요소 × 문헌집합 요소 | *tfidf, ltfidf* |
| 문헌 요소 포함 | 문헌 요소(*tf*) × 범주 요소 | *tfchi, tfpi, tfrmij, tfyule's y, tfrw1(=tfmi), tfrw2, tfrw3, tfrw4(=tflor), tfsrw2, tfsrw6* |
| | 문헌 요소(*ltf*) × 범주 요소 | *ltfchi, ltfpi, ltfrmij, ltfyulel, ltfrw1(=ltfmi), ltfrw2, ltfrw3, ltfrw4(=ltflor), ltfsrw2, ltfsrw6* |
| 문헌 요소 배제 | 문헌집합 요소(*idf*) × 범주 요소 | *idfchi, idfrw1, idfrw2, idfrw3, idfrw4, idfpi, idfsrw2, idfsrw6, idfrmij, idfyule* |

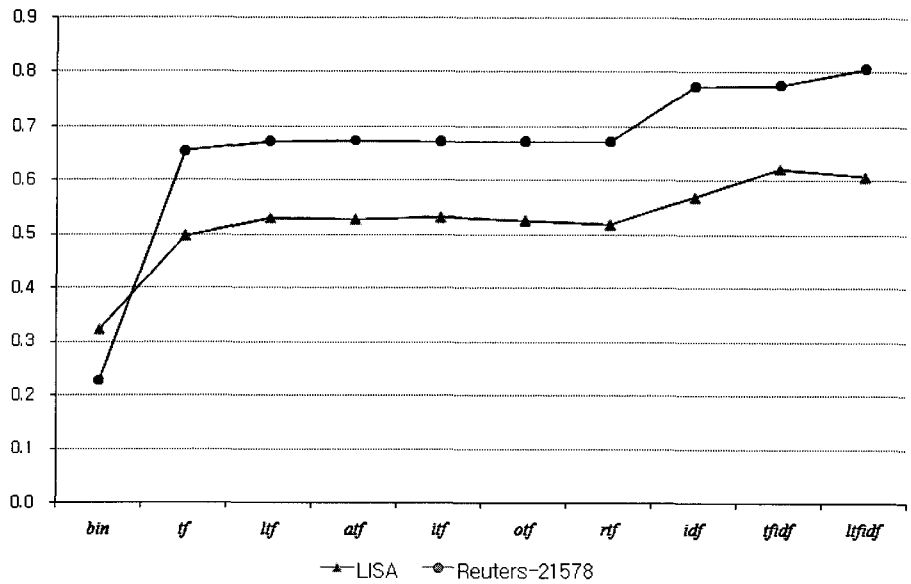
하였다. 이러한 성능 척도의 평균을 구하는 방법으로는 문헌 중심의 마이크로와 범주 중심의 매크로 방법이 있는데, 본 연구에서는 전자인 마이크로 평균 $F_1$ 척도를 사용하여 분류 성능을 평가하였다.

## 4. 실험 결과 및 분석

### 4.1 단일 가중치부여 기법의 실험 결과 및 분석

크게 세 가지 가중치 요소별로 단일 가중치 기법들의 성능을 알아보았다. 먼저, 범주 정보를 사용하지 않고 문헌 내 용어의 출현 정보와 문헌집합 내 용어의 출현정보만을 사용한 것으로 문헌 요소(*tf* and its variations)와 문헌집합 요소(*idf*)에 기반한 각 단일 가중치부여 기법의 성능은 〈그림 1〉과 같다. 여기서 자동분류에서 일반적으로 많이 사용되어 온 조합 가중치부여 기법인 *tfidf*와 *ltfidf*는 성능 비교를 위한 기본형(baseline)으로 제시하였다.

두 실험집단 모두에서 이진 가중치(*bin*)를 제외한 문헌 요소 기반 가중치부여 기법(*tf* and its variations)들은 문헌집합 요소 기반 가중치부여 기법(*idf*)과 조합 가중치 기본형(baseline: *tfidf*, *ltfidf*)에 비해서는 상당히 낮은 성능 수준에 있으면서 서로 간에는 성능 차이가 크지 않은 것으로 나타났다. 즉, 문헌 요소 기반의 단일 가중치부여 기법들은 서로 간에 최대 성능과 최저 성능의 차이가 LISA와 Reuters-21578에서 각각 0.034와 0.019로서 거의 차이가 없었다. LISA에서는 *itf*(0.531), 그리고 Reuters-21578에서는 *atf*(0.673)가 가장 좋은 성능을 나타내



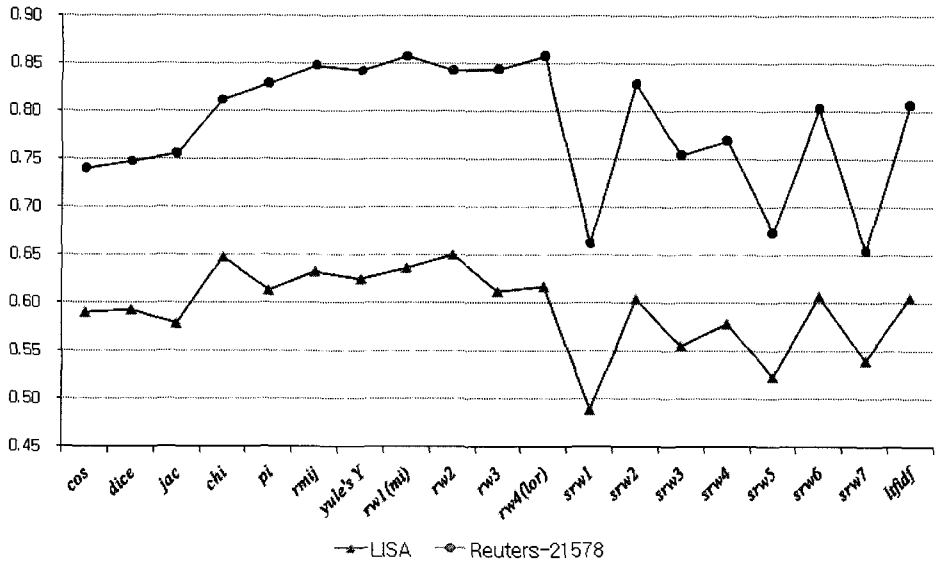

〈그림 1〉 LISA와 Reuters-21578에 대한 단일 가중치부여 기법의 성능: 문헌 요소와 문헌집합 요소

기는 하였지만, 이러한 성능은 이진 가중치이외에 가장 낮은 성능을 보인 *tf*(LISA: 0.497, Reuters-21578: 0.654)와 비교하여도 큰 차이가 없는 것이었다. 반면, 문헌집합 요소 기반의 단일 가중치부여 기법과 비교하는 경우에는 상당한 성능 차이가 발견되었다. 두 실험집단 모두 문헌집합 요소 기반의 단일 가중치부여 기법(*idf*)이 어떤 문헌 요소 기반의 단일 가중치 기법보다도 상당히 좋은 성능을 나타냈다(LISA: 0.568, Reuters-21578: 0.773). 그러나 문헌집합 요소 기반의 단일 가중치부여 기법은 baseline으로 제시한 두 개의 조합 가중치부여 기법보다는 오히려 성능이 다소 낮은 것으로 나타났다(LISA: *tfidf*/0.621, *ltfidf*/0.606 ; Reuters-21578: *tfidf*/0.776, *ltfidf*/0.806).

두 실험집단에 대한 문헌 요소 및 문헌집합 요소에 기반한 단일 가중치부여 기법들의 성능을 요약하면 다음과 같다. 첫째, 문헌 요소에 기반한 단일 가중치부여 기법들은 이진 가중치부여 기법(*bin*)을 제외하면 서로 간에 큰 성능차이가 없었다. 둘째, 문헌 요소와 문헌집합 요소 간의 비교에서는 문헌집합 요소 기반의 단일 가중치부여 기법의 성능이 더 높은 수준에 있는 것으로 나타났다. 셋째, 문헌 요소와 문헌집합 요소 기반의 모든 단일 가중치부여 기법들은 baseline으로서 조합 가중치 기본형(*tfidf*, *ltfidf*)보다는 전반적으로 낮은 성능을 보이는 것으로 나타났다. 따라서 자동분류에서 범주 정보를 전혀 사용할 수 없는 경우에는 지금까지 많은 연구에서 사용된 것으로, 문헌 요소와 문헌집합 요소의 조합에 따른 조합 가중치의 기본형이 좋은 대안이 될 수 있을 것이다.

다음으로 LISA와 Reuters-21578 실험집단에서 분류문헌에 부여된 범주 정보를 사용하는 단일 가중치부여 기법들의 성능은 〈그림 2〉와 같다. 범주 요소 기반의 전체 18개 단일 가중치부여 기법들은 두 개의 실험집단에 대하여 대체로 유사한 성능 패턴을 보여주고 있다. 즉, 순위에서 약간의 변동이 있기는 하지만 동일한 10개의 기법이 대체로 상위 집단에 속하고(*chi*, *pi*, *rmij*, *yule*, *rw1*(*mi*), *rw2*, *rw3*, *rw4*(*lor*), *srw2*, *srw6*: LISA/0.604~0.651 : Reuters-21578/0.802~0.859), 상관계수에 기초한 단일 가중치부여 기법들과 일부 적합성 가중치의 기본형들이 중간 집단(*cos*, *dice*, *jac*, *srw3*, *srw4*: LISA/0.555~0.592 ; Reuters-21578/0.74~0.770), 그리고 나머지 적합성 가중치의 기본형들이 가장 하위 집단(*srw1*, *srw5*, *srw7*: LISA/0.489~0.540 ; Reuters-21578/0.662~0.673)에 포함되었다. 또한 baseline으로서 *ltfidf*(LISA: 0.606, Reuters-21578: 0.806)와 비교할 때는 상위 집단에 속한 대부분의 기법들은 이보다 높은 성능을 나타내었고, 나머지 기법들은 상대적으로 이보다 낮은 성능을 보여주었다.

이러한 결과를 요약하면 두 실험집단 모두 단일 가중치부여 기법들에서는 범주 요소 기반의 기법들의 성능이 가장 높은 수준에 있고, 그 다음이 문헌집합 요소 기반의 기법, 그리고 문헌 요소 기반의 기법들이 가장 낮은 수준인 것으로 나타났다. 조합 가중치 기본형의 성능과 비교하면 문헌 요소와 문헌집합 요소에 기반한 단일 가중치들은 모두 기본형 보다 낮은 성능을 나타낸 반면, 범주 요소 기반의 단일 가중치들은 절반 이상이 더 높은 성능을 보였다. 이러한 결과는 단일 가중치부여 기법 측면에서 볼 때, 문헌 또는 문헌집합 내 출현정보에 의존하는




〈그림 2〉 LISA와 Reuters-21578에 대한 단일 가중치부여 기법의 성능: 범주 요소

것보다 범주 정보를 사용하는 것이 더 효과적이라는 것을 의미한다. 다시 말해서, 문헌과 문헌집합 내 출현정보에만 의존하는 조합 가중치 기본형(*tfidf* or *ltfidf*)을 사용하는 것보다 범주 정보에 기초한 단일 가중치들 중에서 상위 집단에 속한 가중치부여 기법들이 분류 목적으로는 더 효과적일 수 있다는 것이다.

## 4.2 조합 가중치부여 기법의 실험 결과 및 분석

가중치 산출을 위한 요소들의 조합에 의한 가중치부여 기법들의 성능을 알아보기 위한 실험을 수행하였다. 이전 실험과 마찬가지로 조합 가중치 기본형(*tfidf, ltfidf*)의 성능을 기준으로 각 요소들 간의 조합에 따른 여러 가중치부여 기법의 성능을 비교하였다. 단일 가중치부여 기법과는 달리 조합 가중치부여 기법들의 성능 측면에서 두 실험집단의 양상이 다소 다르게 나타났기 때문에, 먼저 실험집단에 따라 LISA와 Reuters-21578의 결과를 구분하여 기술하였고, 다음으로 각 실험집단에 대한 결과에서 문헌 요소를 포함하는 유형과 문헌 요소를 배제하는 유형으로 구분하여 결과를 제시하였다.

LISA 문헌집단을 대상으로 단순 출현빈도(*tf*)와 상위 10개 범주 가중치의 조합에 따른 조합 가중치부여 기법들의 성능을 마이크로 평균 정확률, 마이크로 평균 재현율, 마이크로 평균 $F_1$ 척도로 제시한 것이 〈표 8〉이다. 단일 척도인 마이크로 평균 $F_1$ 척도 측면에서 가장 좋은 성능을 보인 것은 단순 용어빈도와 카이제곱 통계량의 조합(*tfchi*)이었다. 이를 제외한 나머지 대부분의 조합 가중치부여 기법들은 baseline(*tfidf*)과 동등하거나 오히려 낮은 성능을 보였다. 또한, 로그를 취한 출현빈도

(*ltf*)와 10개 범주 가중치의 조합에 따른 조합 가중치부여 기법들의 마이크로 평균 $F_1$ 성능인 〈표 9〉에서도, 이와 유사한 경향이 나타났다. 따라서, 최근의 여러 연구들에서 제안된 접근법으로서 조합 가중치 기본형(baseline: *tfidf* or *ltfidf*)에서 문헌 요소는 그대로 두고 문헌집합 요소를 범주 요소로 대체한 대부분의 가중치 기법들은 성능 향상 효과가 크지 않은 것으로 나타났다.

LISA 실험집단에 대하여 위의 접근법과는 달리 문헌 요소를 배제하고, 대신 문헌집합 요소(*idf*)와 범주 요소를 조합하는 새로운 유형의 조합 가중치부여 기법들의 성능은 〈표 10〉과 같다.

여기서 주목할 것은 전반적으로 문헌 요소를 배제하는 기법들이 기존의 문헌 요소를 포함하는 조합 가중치부여 기법들보다 나은 성능 수준에 있는 것은 물론, 이들의 성능이 기본형(*ltfidf*)보다 크게 향상된 점이다. 즉, 지금까지의 연구들에서 채택하여 온 접근법으로서 단순히 기존의 조합 가중치 기본형에서 문헌 요소는 그대로 두고 문헌집합 요소만을 범주 요소로 대체하는 방식보다는 문헌 요소를 배제하고 대신 문헌집합 요소를 범주 요소와 조합하는 형식이

〈표 8〉 LISA에 대한 문헌×범주(*tf*∗top 10 cat) 조합 가중치부여 기법들의 성능

| 성 능 | *tfchi* | *tfpi* | *tfrmij* | *tfyule's y* | *tfrw1 (tfmi)* | *tfrw2* | *tfrw3* | *tfrw4 (tflor)* | *tfsrw2* | *tfsrw6* | *tfidf* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MIP* | 0.642 | 0.589 | 0.615 | 0.627 | 0.622 | 0.604 | 0.574 | 0.591 | 0.595 | 0.585 | 0.643 |
| MIR** | 0.637 | 0.582 | 0.619 | 0.623 | 0.579 | 0.637 | 0.557 | 0.571 | 0.586 | 0.582 | 0.601 |
| MIF*** | 0.640 | 0.586 | 0.617 | 0.625 | 0.600 | 0.620 | 0.565 | 0.581 | 0.59 | 0.583 | 0.621 |

* MIP: 마이크로 평균 정확률, ** MIR: 마이크로 평균 재현율, *** MIF: 마이크로 평균 $F_1$

〈표 9〉 LISA에 대한 문헌×범주(*ltf*∗top 10 cat) 조합 가중치부여 기법들의 성능

| 성 능 | *ltfchi* | *ltfpi* | *ltfrmij* | *ltfyule's y* | *ltfrw1 (tfmi)* | *ltfrw2* | *ltfrw3* | *ltfrw4 (tflor)* | *ltfsrw2* | *ltfsrw6* | *ltfidf* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MIP* | 0.643 | 0.586 | 0.626 | 0.592 | 0.620 | 0.629 | 0.578 | 0.589 | 0.599 | 0.581 | 0.627 |
| MIR** | 0.641 | 0.586 | 0.590 | 0.568 | 0.586 | 0.590 | 0.542 | 0.571 | 0.586 | 0.590 | 0.586 |
| MIF*** | 0.642 | 0.586 | 0.608 | 0.579 | 0.603 | 0.609 | 0.560 | 0.580 | 0.593 | 0.585 | 0.606 |

* MIP: 마이크로 평균 정확률, ** MIR: 마이크로 평균 재현율, *** MIF: 마이크로 평균 $F_1$

〈표 10〉 LISA에 대한 문헌집합×범주(*idf*∗top 10 cat) 조합 가중치부여 기법들의 성능

| 성 능 | *idfchi* | *idfpi* | *idfrmij* | *idfyule's y* | *idfrw1 (idfmi)* | *idfrw2* | *idfrw3* | *idfrw4 (idflor)* | *idfsrw2* | *idfsrw6* | *ltfidf* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MIP* | 0.641 | 0.652 | 0.632 | 0.659 | 0.676 | 0.667 | 0.648 | 0.640 | 0.66 | 0.644 | 0.627 |
| MIR** | 0.641 | 0.637 | 0.659 | 0.630 | 0.648 | 0.615 | 0.608 | 0.626 | 0.626 | 0.648 | 0.586 |
| MIF*** | 0.641 | 0.644 | 0.645 | 0.644 | 0.662 | 0.640 | 0.628 | 0.633 | 0.643 | 0.646 | 0.606 |

* MIP: 마이크로 평균 정확률, ** MIR: 마이크로 평균 재현율, *** MIF: 마이크로 평균 $F_1$

더 큰 성능 향상을 가져올 수 있는 것으로 나타났다.

Reuters-21578 실험집단을 대상으로 한 자동분류 실험에서 문헌 요소와 범주 요소의 조합에 따른 조합 가중치부여 기법들의 성능은 〈표 11〉, 〈표 12〉와 같다.

Reuters-21578의 경우에는 문헌 요소와 범주 요소를 조합한 모든 가중치부여 기법들이 기본형(*tfidf, ltfidf*)보다 낮은 성능을 보였다. 즉, Reuters-21578을 대상으로 조합가중치 기본형에서 문헌 요소는 그대로 두고 문헌집합 요소를 범주 요소로 대체하는 형식의 조합 가중치부여 기법들을 적용하는 경우에는 성능 향상 효과가 없는 것으로 나타났다.

그러나 조합가중치 기본형에서 문헌 요소를 배제하고 대신 문헌집합 요소와 범주 요소를 조합한 결과인 〈표 13〉에서는 이와는 다른 양상을 보였다. 대부분의 문헌집합 요소와 범주 요소 간의 조합에 따른 가중치부여 기법들이 기본형(baseline)보다 나은 성능을 보이는 것은 물론, 〈표 11〉과 〈표 12〉에서 제시한 문헌 요소와 범주 요소를 조합한 유형의 가중치부여 기법들에 비해서는 상당히 크게 성능이 향상되었다. 결과적으로 LISA와 Reuters-21578의

〈표 11〉 Reuters-21578에 대한 문헌×범주(*tf**top 10 cat) 조합 가중치부여 기법들의 성능

| 성 능 | *tfchi* | *tfpi* | *tfrmij* | *tfyule's y* | *tfrw1 (tfmi)* | *tfrw2* | *tfrw3* | *tfrw4 (tflor)* | *tfsrw2* | *tfsrw6* | *tfidf* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MIP* | 0.684 | 0.633 | 0.657 | 0.657 | 0.672 | 0.690 | 0.652 | 0.671 | 0.683 | 0.639 | 0.696 |
| MIR** | 0.862 | 0.846 | 0.882 | 0.869 | 0.878 | 0.873 | 0.875 | 0.873 | 0.843 | 0.833 | 0.875 |
| MIF*** | 0.763 | 0.724 | 0.753 | 0.748 | 0.761 | 0.771 | 0.747 | 0.759 | 0.754 | 0.723 | 0.776 |

* MIP: 마이크로 평균 정확률, ** MIR: 마이크로 평균 재현율, *** MIF: 마이크로 평균 $F_1$

〈표 12〉 Reuters-21578에 대한 문헌×범주(*ltf**top 10 cat) 조합 가중치부여 기법들의 성능

| 성 능 | *ltfchi* | *ltfpi* | *ltfrmij* | *ltfyule's y* | *ltfrw1 (tfmi)* | *ltfrw2* | *ltfrw3* | *ltfrw4 (tflor)* | *ltfsrw2* | *ltfsrw6* | *ltfidf* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MIP* | 0.696 | 0.641 | 0.673 | 0.656 | 0.757 | 0.720 | 0.656 | 0.684 | 0.713 | 0.619 | 0.779 |
| MIR** | 0.862 | 0.861 | 0.876 | 0.880 | 0.854 | 0.860 | 0.885 | 0.874 | 0.842 | 0.868 | 0.835 |
| MIF*** | 0.770 | 0.735 | 0.761 | 0.752 | 0.803 | 0.784 | 0.753 | 0.768 | 0.772 | 0.723 | 0.806 |

* MIP: 마이크로 평균 정확률, ** MIR: 마이크로 평균 재현율, *** MIF: 마이크로 평균 $F_1$

〈표 13〉 Reuters-21578에 대한 문헌집합×범주(*idf**top 10 cat) 조합 가중치부여 기법들의 성능

| 성 능 | *idfchi* | *idfrmij* | *idfpi* | *idfyule's y* | *idf rw1* | *idfrw2* | *idfrw3* | *idfrw4* | *idfsrw2* | *idfsrw6* | *ltfidf* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MIP* | 0.736 | 0.719 | 0.7908 | 0.790 | 0.811 | 0.843 | 0.788 | 0.816 | 0.818 | 0.791 | 0.779 |
| MIR** | 0.872 | 0.900 | 0.874 | 0.872 | 0.878 | 0.829 | 0.876 | 0.852 | 0.859 | 0.868 | 0.835 |
| MIF*** | 0.799 | 0.799 | 0.830 | 0.829 | 0.843 | 0.835 | 0.830 | 0.833 | 0.838 | 0.828 | 0.806 |

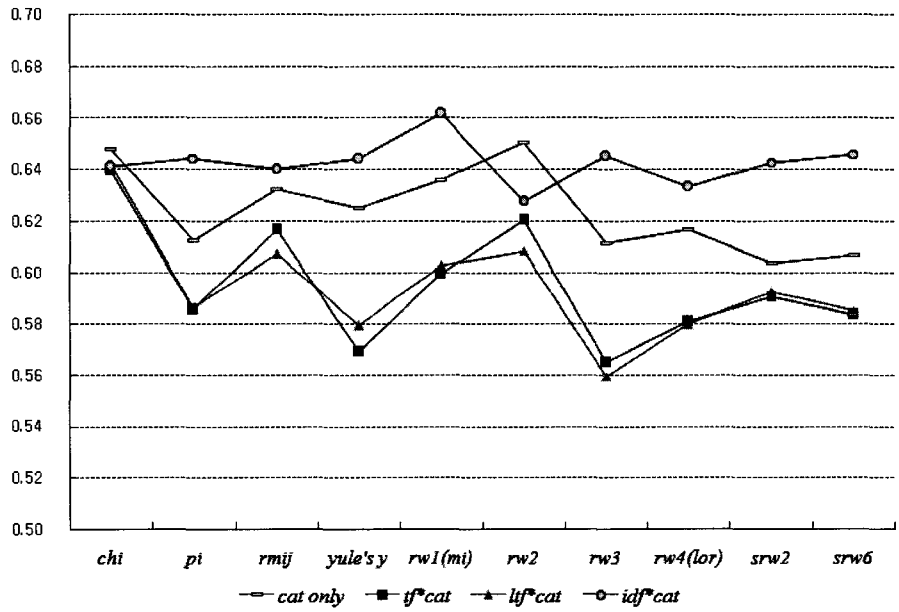
* MIP: 마이크로 평균 정확률, ** MIR: 마이크로 평균 재현율, *** MIF: 마이크로 평균 $F_1$

두 실험집단 모두에서 조합 가중치 기본형에서 문헌 요소와 범주 요소를 조합하는 것보다는 문헌집합 요소와 범주 요소를 조합하는 기법들이 더 높은 성능을 보이는 경향이 나타났다.

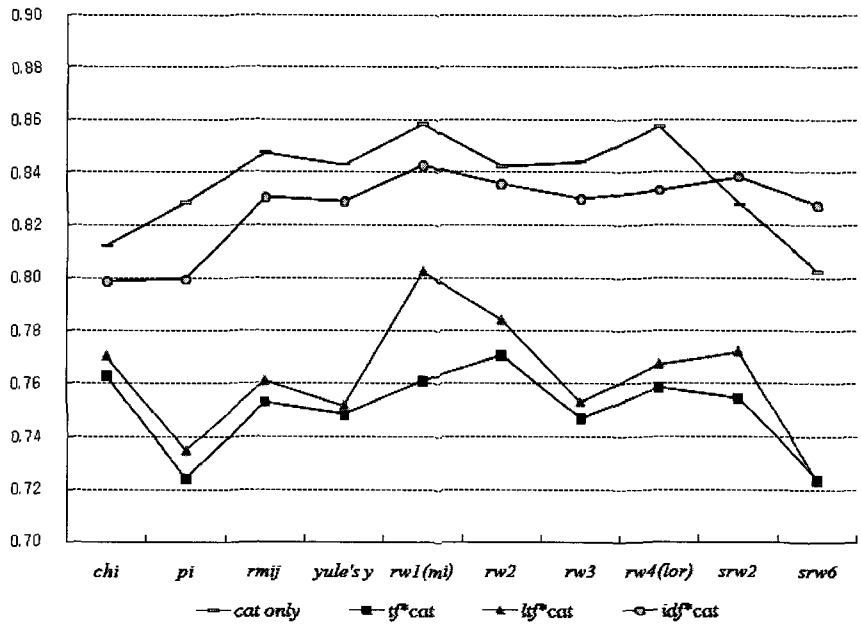
### 4.3 단일 가중치부여 기법과 조합 가중치부여 기법 간의 비교 및 분석

실험에 사용된 두 실험집단에 대하여 단일 가중치부여 기법 중에서 가장 좋은 성능을 보인 범주 요소 기반의 상위 10개 단일 가중치부여 기법(top 10 cat)과 나머지 두 가지 요소인 문헌 요소(*tf, ltf*) 또는 문헌집합 요소(*idf*) 간의 조합 가중치부여 기법들에 대한 성능을 비교 및 분석해 보았다. 먼저, LISA 실험집단에 대한 결과에서는 〈그림 3〉과 같이 문헌집합 요소와 범주 요소를 결합한 조합 가중치부여 기법(*idf**cat)이 대부분의 단일 가중치부여 기법(cat only)은 물론, 다른 조합 가중치부여 기법들(*tf**cat or *ltf**cat)보다는 상당히 높은 성능을 보였다. 특히 범주 요소에 속한 자질선정 기준 및 적합성 가중치들(*chi, rw1(mi), rw4(lor), srw2, srw6*)의 성능이 좋다는 것은 이전의 연구들에서도 보고된 바가 있으나(Debole and Sebastani 2003; Deng et al. 2004; Lan et al. 2006; Liu et al. 2007), 본 연구에서 새롭게 적용한 적합성 가중치들(*rmij, rw2, rw3, pi*)의 성능도 상당히 높은 수준인 것으로 나타났다.

다음으로 〈그림 4〉에서 Reuters-21578 실험집단에 대한 단일 가중치부여 기법과 조합 가중치부여 기법의 성능을 비교한 결과는, LISA에서와는 달리 범주 요소 기반의 단일 가중치




〈그림 3〉 LISA에 대한 단일 가중치와 조합 가중치 간의 성능 비교




〈그림 4〉 Reuters-21578에 대한 단일 가중치와 조합 가중치 간의 성능 비교

부여 기법(cat only)들이 가장 좋은 성능을 나타냈다. 즉 대부분의 단일 가중치부여 기법들이 LISA에서 가장 좋은 성능을 보인 문헌집합 요소와 범주 요소를 결합한 조합 가중치부여 기법들(*idf**cat)은 물론, 문헌 요소와 범주 요소 간의 조합 가중치부여 기법들에 비하여 상당히 높은 성능을 보였다. 그러나 이러한 단일 가중치부여 기법들 중에서도 *srw2*와 *srw6*는 예외적으로 문헌집합 요소와의 조합에서 더 높은 성능을 보였다. 한편, *chi*는 양 실험집단 모두에서 조합 가중치보다는 단일 가중치로 적용하는 경우에 가장 성능이 좋은 것으로 나타났다.

두 실험집단에 대하여 단일 가중치부여 기법과 조합 가중치부여 기법의 성능을 종합적으로 비교해 본 결과, 다음과 같은 사실들을 발견하였다.

첫째, 두 실험집단 모두에서 문헌 요소가 없는 범주 요소 기반의 단일 가중치부여 기법과 문헌집합 요소와 범주 요소의 조합 가중치부여 기법이, 문헌 요소를 포함하는 조합 가중치부여 기법들보다 나은 성능을 보이는 것으로 나타났다(only cat, *idf**cat 〉 *tf**cat, *ltf**cat).

둘째, 실험집단의 특성에 따라 문헌 요소를 배제하는 경우에 가장 좋은 성능을 산출하는 기법의 유형이 서로 다른 것으로 나타났다. 즉, LISA 실험집단에서는 문헌집합 요소와 범주 요소 간의 조합 가중치부여 기법(*idf**cat)이 가장 좋은 성능을 나타낸 반면, Reuters-21578 실험집단에서는 대부분의 단일 가중치부여 기법(only cat)이 가장 우위에 있는 것으로 나타났다.

셋째, 상대적으로 성능이 낮은 수준에 있는 문헌 요소를 포함하는 두 가지 조합 가중치부

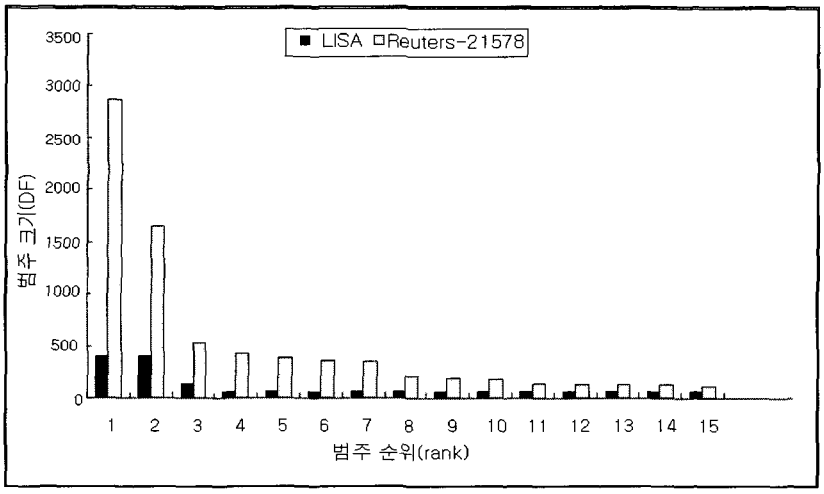
여 기법들 간에는 성능 차이가 그다지 크지 않았다($tf*\text{cat} \rightleftharpoons ltf*\text{cat}$)

전반적으로 문헌 요소를 배제하는 것이 포함하는 경우보다 더 나은 성능을 나타낸 결과에서, 자동분류 목적으로 가중치부여 기법을 적용할 경우에는 문헌 요소가 긍정적인 영향을 주지 못한다는 사실을 추정할 수 있다. 그 이유는 특정 문헌 단위의 검색이 목적인 정보검색에서와는 달리, 특정 범주의 부여가 목적인 자동분류에서는 개별 문헌 단위가 아닌 범주 단위 또는 전체 문헌집합 단위에서 추출된 정보가 더 유용할 수 있기 때문이다.

두 실험집단 간에 가장 좋은 성능을 보인 가중치부여 기법의 유형이 서로 다른 것은 두 실험집단 내 범주의 특성 측면에서 그 원인을 찾을 수 있을 것이다.

〈그림 5〉는 범주의 크기 측면에서 두 실험집단에서 실험에 사용된 상위 15개 범주를 나타낸 것이다. 정보학 분야의 학술지 논문들로 구성된 LISA는 Reuters-21578에 비하여 상대적으로 적은 수의 레코드로 이루어져 있고, 이에 따라 특정 논문들에 부여된 디스크립터로서 범주가 부여된 문헌 수가 상대적으로 작은 편이다. 그림에서 보는 바와 같이 두 실험집단 모두 상위 15개의 범주를 대상으로 실험하였지만 양 집단의 범주 크기는 상당한 규모의 차이가 있다(LISA의 평균 범주 크기 90.93; Reuters-21578의 평균 범주 크기: 521.67). 특히, LISA는 최상위 3개 범주를 제외한 나머지 12개 범주가 모두 53 이하의 작은 범주에 속하고, Reuters-21578은 상대적으로 15개 범주가 모두 124 이상의 큰 범주라 할 수 있다. 따라서 LISA의 경우에는 소규모 범주의 분류에서 상대적으로 유리한 저빈도어 선호의 문헌집합 요소를 결합한 조합 가중치가 더 좋은 성능을 나타낸 것이라 할 수 있다.

또한, 〈표 14〉에서 보는 바와 같이 범주의 내용 측면에서 LISA의 범주들은 색인 전문가에 의해 정보학 분야의 학술지 논문에 부여된 디스크립터들로서 대부분이 특정한 소주제들이라 할 수 있다. 반면 Reuters-21578의 범주




〈그림 5〉 LISA와 Reuters-21578의 상위 15개 범주의 크기

〈표 14〉 LISA와 Reuters-21578의 상위 15개 범주의 내용

| 순위 | LISA | Reuters-21578 |
|---|---|---|
| 1 | online information retrieval | earn |
| 2 | searching | acq |
| 3 | world wide web | money-fx |
| 4 | research | grain |
| 5 | citation analysis | crude |
| 6 | automatic text analysis | trade |
| 7 | information seeking behaviour | interest |
| 8 | evaluation | wheat |
| 9 | information science | ship |
| 10 | internet | corn |
| 11 | models | money-supply |
| 12 | bibliometrics | dlr |
| 13 | information communication | sugar |
| 14 | computerized information storage and retrieval | oilseed |
| 15 | information work | coffee |

들은 신문기사에 부여된 것으로 상대적으로 일반적인 대주제들이다. 따라서, 특정한 범주가 많은 LISA를 대상으로 한 텍스트 범주화에서는 문헌집합 요소를 포함하는 조합 가중치가 더 유리하지만, 보다 일반적인 범주로 구성된 Reuters-21578에서는 범주 요소만으로 이루어진 단일 가중치가 더 좋은 결과를 가져오는 것으로 볼 수 있다.

## 5. 결 론

로치오 분류기 기반의 자동분류에서 두 개의 문헌집단을 대상으로 여러 가중치부여 기법들을 적용하여 성능 향상을 위한 방법을 모색해 보았다. 이를 위해 가중치부여 기법에서 가중치 산출에 사용되는 요소를 크게 문헌요소(document factor), 문헌집합 요소(document set factor), 범주 요소(category factor)의 세 가지로 구분하였다. 그리고 각 세 가지 요소별 단일 가중치부여 기법에 따른 성능과 이들 요소간의 조합 가중치부여 기법에 따른 성능을 알아보기 위한 실험을 수행하였다. 실험 결과에 따르면, 각 요소별 단일 가중치 측면에서는 범주 요소에 기반한 기법(RW, FSW)들이 가장 좋은 성능을 보였고 다음으로 문헌집합 요소(*idf*), 그리고 문헌 요소(*bin*, *tf*, and variants of *tf*)가 가장 낮은 성능이었다. 특히, 대부분의 범주 요소 기반의 기법(cat only)들은 다른 요소들에 기반한 단일 가중치부여 기법들뿐만 아니라 본 연구에서 기본형으로 사용한 조합 가중치 기본형(*tfidf* or *ltfidf*)보다도 상당히 높은 성능을 보여 주었다. 따라서 세 가지 요소로 구분하여 적용한 단일 가중치부여 기법들 중에서는 문헌 요소나 문헌집합 요소 보다는 범주 정보에 기반한 기법들이 가장 좋은 결과를 가져올

수 있었다.

조합 가중치부여 기법 측면에서는 많은 정보검색 및 자동분류 연구에서 사용되어 온 문헌요소와 문헌집합 요소의 조합 가중치를 기본형(*tfidf* or *ltfidf*)으로 하고, 다른 요소와의 조합에 의한 두 가지 유형의 기법 간의 성능을 비교하였다. 즉, 최근의 연구들에서 제안된 것으로 조합 가중치 기본형에서 문헌 요소를 그대로 두고 범주 요소를 조합하는 유형(*tf**cat or *ltf**cat)과 본 연구에서 새롭게 제안하는 것으로 문헌 요소를 배제하고 문헌 집합 요소와 범주 요소를 조합하는 유형(*idf**cat)의 성능을 비교하였다. 그 결과, 범주 요소를 포함하지 않는 조합 가중치 기본형과 비교하여 범주 요소를 포함하는 조합 가중치들의 성능이 더 높은 경향을 보여주었고, 특히 범주 요소를 기반으로 하는 조합 가중치 중에서도 문헌 요소를 포함하는 유형보다 문헌 요소를 배제하는 유형이 더 큰 성능 향상을 가져오는 것으로 나타났다.

LISA와 Reuters-21578의 두 실험집단에 대하여 좋은 성능을 보인 범주 요소 기반의 단일 가중치부여 기법과 조합 가중치부여 기법을 종합적으로 비교 및 분석하여 다음과 같은 사실들을 발견하였다.

첫째, 두 실험집단 모두에서 문헌 요소가 없는 단일 가중치부여 기법(only cat)과 조합 가중치부여 기법이(*idf**cat) 문헌 요소를 포함하는 모든 가중치부여 기법들보다 나은 성능을 보였다. 이는 분류 목적의 가중치부여 기법에서는 문헌 요소보다 문헌집합 요소와 범주 요소가 더 중요한 역할을 한다는 것을 의미한다.

둘째, 문헌 요소가 없는 두 가지 유형의 가중치부여 기법들의 경우에도 실험집단의 특성에 따라 가장 좋은 성능을 산출하는 유형이 다른 것으로 나타났다. 즉, LISA 실험집단에서는 문헌집합 요소와 범주 요소 간의 조합 가중치부여 기법(*idf**caty)이 가장 좋은 성능을 나타낸 반면, Reuters-21578 실험집단에서는 대부분의 단일 가중치부여 기법(only cate)이 가장 우위에 있는 것으로 나타났다. 따라서 가중치부여 기법의 적용을 통한 분류 성능의 향상을 위해서는 먼저 문헌집단 내 범주집합의 특성(크기 및 내용)을 파악하고, 이에 따라 적절한 기법을 적용하는 경우에 가장 좋은 결과를 가져올 수 있다. 예를 들면, LISA와 같이 보다 특정한 소규모 범주들로 구성된 경우에는 범주 요소에 문헌집합 요소를 추가한 조합 가중치를 사용하는 것이 좋고, Reuters-21578처럼 보다 일반적인 대규모 범주들의 경우에는 범주 요소 기반의 단일 가중치를 사용하는 것이 좋은 결과를 가져올 수 있을 것이다.

결과적으로 실험을 통하여 자동분류를 위한 가중치부여 기법에서는 범주 정보가 가중치 산출의 가장 중요한 요소가 되며, 또한 다른 요소와의 조합에서는 문헌 요소를 배제하는 것이 오히려 더 좋은 결과를 가져올 수 있다는 새로운 사실을 발견하였다. 특히, 이러한 범주 정보 기반 가중치들의 실제 적용에서는 분류 대상이 되는 문헌집단의 범주 특성을 먼저 파악하여야 하고, 이에 따라 적절한 가중치부여 기법을 선택하여 적용하는 것이 분류 성능의 최적화를 가져올 수 있을 것이다.

본 연구의 제한점으로는 두 실험집단의 범주집합 내 모든 범주가 아닌 상위의 범주만을 대상으로 하였기 때문에, 상대적으로 하위의 범주들에 대하여 이러한 결론을 그대로 적용하기

는 어려울 것이다. 따라서 실험집단의 모든 범주집합을 대상으로 집단 내 전체 범주들에 대한 이러한 결과의 일반화를 검토하여야 할 것이다. 이에 더하여 다양한 다른 문헌집단과 분류 알고리즘에 대한 적용 및 검증도 필요할 것이다.

## 참 고 문 헌


김판준. 2007. 로치오 알고리즘을 이용한 자동분류에서 용어 가중치 기법. 『문헌정보학논집』, 명지대학교, 문헌정보학회, 제9호: 157-185.

김판준. 2006a. 기계학습을 통한 디스크립터 자동부여에 관한 연구. 『정보관리학회지』, 23(1): 279-299.

김판준. 2006b. 로치오 알고리즘을 이용한 학술지 논문의 디스크립터 자동부여에 관한 연구. 『정보관리학회지』, 23(3): 69-89.

이재윤. 2005. 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구 『한국문헌정보학회지』, 39(2): 123-146.

이재윤, 최보영, 정영미. 2000. 문헌 자동분류에서 용어가중치 기법에 대한 연구. 제7회 한국정보관리학회 학술대회 논문집, 2000년 8월 17일 이화여자대학교, pp.41-44.

정영미. 1993. 『정보검색론』. 서울: 구미무역(주) 출판부.

Brank, J., M. Grobelnik, N. Milic-Frayling & D. Mladenic. 2002. "Interaction of feature selection methods and linear classification models." In: *Proceedings of the ICML-02 Workshop on Text Learning, Sydney*. [cited 2007. 5. 3]. 〈http://citeseer.ist.psu.edu/brank02interaction.html〉.

Castillo M. D. and Serrano J. I. 2004. "A multistrategy approach for digital text categorization from imbalanced documents." *ACM SIGKDD Explorations Newsletter: Special Issue on Leaning from Imbalanced Datasets*, 6(1): 70-79.

Debole, Franca and F. Sebastiani. 2003. "Supervised term weighting for automated text categorization." In: *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, New York: ACM, 784-788.

Deng, Zhi-Hong et al. 2004. A Comparative study on feature weight in text categorization. In *Proceedings of The Sixth Asia Pacific Web Conference(APWEB 2004), Hangzhou, China, April 14-17, LNCS 3007*, 588-597.

Forman G. 2003. "An extensive empirical study of feature selection metrics for text classification." *The Journal of Machine Learning Research, Special*

*Issue on Variable and Feature Selection*, 3: 1289-1305.

Geng, L. and Howard J. Hamilton. 2006. "Choosing the right lens: finding what is interesting in data mining." eds. by Guillet, Fabrice, Howard J. Hamilton. *Quality Measures in Data Mining*. Springer, pp. 3-24.

How, Bong Chih and Narayanan K. 2004. An empirical study of feature selection for text categorization based on term weightage. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pp.592-602.

Joachims, Thorsten. 1996. "A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization." *Proceedings of ICML-97, 14th International Conference on Machine Learning, Nashville, TN*: 143-151.

Joachims, Thorsten. 1998. "Text categorization with support vector machines: learning with many relevant features." In: *Proceedings of the 10th European Conference on Machine Learning*: 137-142.

Lan, Man, Chew-Lim Tan, and Hwee-Boon LOW. 2006. "Proposing a new term weighting scheme for text categorization." In *21st National Conference on Artificial Intelligence, AAAI-2006, 16-20 July, 2006, Boston, Massachusetts, USA*. [cited 2007. 3. 7.]. 〈http://www.comp.nus.edu.sg/~tancl/Papers/AAAI06/AAAI2006-LanMan-final.pdf〉

Liu, Ying, Han Tong Loh, Kamal Youcef-Toumi, and Shu Beng Tor. 2007. "Handling of imbalanced data in text classification: category-based term weights." In: Kao, Anne and Stephen R. Poteet eds. *Natural Language Processing and Text Mining*. Springer., pp.171-192.

Papineni, K. 2001. "Why inverse document frequency?" *Proceedings of the North American Association for Computational Linguistics, NAACI, New York*, pp. 25-32.

Prabowo, Ruby and Mike Thelwall. 2006. "A comparison of feature selection methods for an evloving RSS feed corpus." *Information Processing and Management*, 42: 1491-1512.

Robertson S. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5): 503-520.

Robertson, S. E. and K. Sparck Jones. 1976. "Relevance weighting of search terms." *JASIS*, 27(3): 129-146.

Rogati, M. and Y. Yang. 2002. High-Performing Feature Selection for Text Classification." In: *Proceedings of the eleventh international conference on Information and knowledge management, CIKM*

'02. [cited 2007. 8. 23.]. <citeseer.ist.psu.edu/rogati02highperforming.html>.

Salton, G., H. Wu, and C. T. Yu. 1981. "The Measurement of term importance in automatic indexing." *JASIS*, 32(3): 175-186.

Salton, G. and M. J. McGill. 1983. *Introduction to Modern Information Retrieval.* N. Y.: McGraw-Hill.

Sebastiani, Fabrizio. 2002. "Machine learning in automated text categorization." *ACM Computing Surveys*, 34(1): 1-47.

Soucy P. and Guy W. Mineau. 2005. "Beyond TFIDF weighting for text categorization in the vector space model." *IJCAI-05 proceedings*, 1130-1135. [cited 2007. 2. 20]. <http://dli.iiit.ac.in/ijcai/IJCAI-05/PDF/0304.pdf>

Yang, Y and Liu X. 1999. "A re-examination of text categorization methods." In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Yang, Y. 1999. "Evaluation of statistical approaches to text categorization." Information Retrieval, 1: 69-90.

Yang, Y. and Pedersen J. O. 1997. "A comparative Study on Feature Selection in Text Categorization." In: *Proceedings of ICML-97*, 14th International Conference on Machine Learning: 412-420.

Yu, C. T. and G. Salton 1976. "Precision weighting-an effective automatic indexing method." *Journal of Association for Computing Machinery*, 23(1): 76-88.

Yu, C. T., K. Lam. and G. Salton. 1982. "Term weighting in information retrieval using the term precision model." *Journal of Association for Computing Machinery*, 29(1): 152-170.

Zheng Z., X. Wu and R. Srihari. 2004. "Feature selection for text categorization on imbalanced data." *ACM SIGKDD Explorations Newsletter: Special Issue on Leaning from Imbalanced Datasets*, 6(1): 80-89.