로치오 알고리즘을 이용한 학술지 논문의 디스크립터 자동부여에 관한 연구

A Study on the Automatic Descriptor Assignment for Scientific Journal Articles Using Rocchio Algorithm

김판준(Pan-Jun Kim)

연세대학교 문헌정보학과 시간강사 (dpbluesea@hanmail.net)

초 록

로치오 알고리즘에 기초한 통제어휘 자동색인 또는 텍스트 범주화에서 적용되어 온 여러 성능 요인들을 재검토하였고, 성능 향상을 위한 기본적인 방법을 찾아보았다. 또한, 동등한 조건에서 통제어휘 자동색인을 위한 로치오 알고리즘 기반 방법의 성능을 다른 학습기반 방법들의 성능과 비교하였다. 결과에 따르면, 통제어휘 자동색인을 위한 로치오 기반의 프로파일 방법은 구현의 용이성과 컴퓨터 처리 시간 측면의 경제성이라는 기존의 장점을 그대로 유지하면서도, 다른 학습기반 방법들(SVM, VPT, NB)과 거의 동등하거나 더 나은 성능을 보여주었다. 특히, 색인전문가의 색인작업을 지원하는 반-자동 색인의 목적으로는 비교적 높은 수준의 재현율을 유지하면서 학습 데이터의 증가에 따라 정확률이 크게 향상되는 로치오 알고리즘을 이용한 방법을 우선적으로 고려할 수 있을 것이다.

ABSTRACT

Several performance factors which have applied to the automatic indexing with controlled vocabulary and text categorization based on Rocchio algorithm were examined, and the simple method for performance improvement of them were tried. Also, results of the methods using Rocchio algorithm were compared with those of other learning based methods on the same conditions. As a result, keeping with the strong points which are implementational easiness and computational efficiency, the methods based Rocchio algorithms showed equivalent or better results than other learning based methods(SVM, VPT, NB). Especially, for the semi-automatic indexing(computer-aided indexing), the methods using Rocchio algorithm with a high recall level could be used preferentially.

키워드: 로치오 알고리즘, 자동색인, 통제어휘, 디스크립터, 텍스트 범주화, Rocchio algorithm, automatic indexing, computer assisted indexing, controlled vocabulary, text categorization, descriptors

1. 서론

색인의 기본적인 목적은 특정 대상(문헌)의 주제를 표현하는 것으로, 이용자가 원할 때 찾을 수 있도록 그 대상의 내용을 적절하게 표현할 수 있는 색인어나 분류기호를 부여한다. 색인어로는 주제명표목, 디스크립터 등의 통제색인어가 사용될 수도 있고, 본문에 출현한 표현(자연어) 그대로의 비통제색인어를 사용할 수도 있다. 현재 대부분의 정보검색시스템에서 자동색인은 색인의 주체가 인간이 아닌 컴퓨터로서 주로 텍스트 내 키워드의 추출에 기초하

고 있다. 컴퓨터 프로그램을 통해 텍스트에 출현한 그대로의 표현을 색인어로 추출하는 이러한 형태의 자동색인은 일부 형태소분석 또는 복합어 처리 등 낮은 수준의 어휘통제를 포함하기도 하지만, 대부분 비통제어휘를 사용하여 문헌에 색인어를 부여한다.

인터넷의 발전으로 누구나 대규모의 정보에 언제라도 접근할 수 있는 현재의 환경과는 달리, 비교적 소수의 전문가 또는 학자들 간에 정보가 유통되었던 초기의 온라인 데이터베이스 환경에서는 수록된 정보의 유형에 따라 비통제어휘(자연언어)와 통제어휘를 사용하는 색인은 각기 다른 측면에서 발전되었다. 즉, 컴퓨터를 활용하는 자동색인은 역사적으로 크게 두 가지 발전 노선이 있었다. 첫째로 미국 국립의학도서관(the National Library of Medicine), 국방부(the Department of Defense), 미 항공우주국(the National Aeronautics an Space Administration)과 같은 정부기관이 주도한 서지 데이터베이스(bibliographic database)를 중심으로 한 색인전문가에 의한 통제언어 색인의 발전과, 둘째로 법률분야 등의 민간부문에서 주도한 전문 데이터베이스(full-text database)를 중심으로 컴퓨터에 의해 본문의 단어를 그대로 추출하는 비통제언어 색인의 발전이 그것이다(Lancaster 2003). 1960년대와 1970년대에 걸친 이러한 두 가지 색인언어 간의 우위에 대한 논의는, 1980년대에 와서 양자를 함께 사용하는 복합(hybrid) 시스템이 가장 이상적이라는 합의에 이르러 현재 대부분의 정보검색시스템은 이러한 두 가지 유형의 색인을 상호보완적으로 제공하는 추세이다.

통제어휘를 사용하는 주제색인은 일반적으로 인간의 사고작용을 통하여 수작업으로 이루어지는 것으로 받아들여져 왔기 때문에, 통제어휘를 사용하는 주제색인을 컴퓨터를 통하여 처리하기 위한 시도는 비통제어휘 자동색인보다는 상대적으로 적었다고 할 수 있다. 그러나 과학기술의 발달로 인하여 다양한 유형의 정보가 폭발적으로 증가하면서, 통제어휘 자동색인의 필요성이 여러 측면에서 제기되고 있다. 특히, 정보와 이용자의 매개자로서 색인전문가의 입장에서는 색인하여야 할 문헌의 양이 급속도로 크게 증가하여, 종전의 수작업 방식 그대로는 적절한 색인작업이 거의 불가능한 상황에 이르고 있다. 또한, 이용자 측면에서는 정보검색 대상과 검색결과가 너무 많아 적합한 정보의 판정과 선택이 어려운 상황에 처해있다. 이러한 상황에서 색인전문가 측면에서는 보다 효율적으로 많은 문헌을 빠른 시간 내에 일관성 있게 색인하고, 이용자 측면에서는 원하는 정보를 주제 또는 개념 측면에서 접근할 수 있도록 하기 위한 통제어휘 자동색인시스템에 대한 필요성이 커지고 있다. 한 예로, 유럽연합(EU)의 EUROVOC 시소러스를 사용하여 색인전문가가 수작업으로 디스크립터를 부여하는 경우에는 평균적으로 하루 30개 이하의 문헌을 색인할 수 있었던 반면, 통제어휘 자동색인시스템에서는 한 문헌 당 단지 5초의 처리 시간만을 필요로 하였다(Pouliquen, Steinberger, and Ignat 2003).

본 연구에서는 지금까지 대부분 인간에 의해 전적으로 수행되어 온 통제어휘를 사용하는 주제색인 작업에서 색인전문가를 효율적으로 지원할 수 있는 자동색인 방법을 모색하여 보았다. 특히, 대표적인 통제색인어라고 할 수 있는 디스크립터를 자동부여하는 방법으로, 정보검색 분야에서 많이 사용되고 있는 벡터공간 모형과 로치오 알고리즘을 기반으로 하는 통제어휘 프로파일의 생성 및 입력문헌과의 매칭에 의한 자동색인 방법의 가능성을 알아보고자 하였다. 지금까지 이러한 유형의 자동색인 방법은 비교적 구현하기 쉽고 컴퓨터 처리시간 측면에서 경제적이라는 커다란 장점에도 불구하고, 학습기반의 다른 텍스트 범주화 방법들에 비해 저성능인 것으로 평가되어 다른 방법의 상대적인 성능 우위를 보여주기 위한 기본형(baseline)으로만 사용되는 경우가 많았다(Sebastiani 2002). 그러나, 본 연구에서는 정보

학 분야 학술지 논문을 대상으로 대표적인 통제색인어인 디스크립터를 자동부여하는 실험을 통하여, 로치오 알고리즘에 기초한 방법의 실제적인 성능 및 특성을 알아보았고, 이에 기초하여 이러한 방법의 성능 향상을 위한 방안을 모색하여 보았다.

2. 로치오 알고리즘을 이용한 자동색인과 텍스트 범주화

통제어휘 자동색인과 텍스트 범주화는 기본적으로 컴퓨터 알고리즘과 통제어휘(주제 범주)를 사용하여 문헌의 내용을 표현하고, 그 결과로 문헌을 집단화한다는 측면에서 동일한 목적과 절차를 공유한다. 이를 출력 측면에서 보면, 통제어휘 자동색인은 문헌의 내용을 표현하는 것으로 그 결과가 문헌에 대한 통제색인어 부여로 나타나고, 텍스트 범주화는 유사한 문헌을 집단화하기 위한 것으로 그 결과가 주제범주의 배정으로 나타난다.

2.1 로치오 알고리즘을 이용한 통제어휘 자동색인

로치오 알고리즘을 이용한 통제어휘 자동색인 방법은 기본적으로 벡터공간 모형과 로치오 알고리즘에 기초하는 통제어휘(디스크립터 또는 주제명표목) 프로파일을 생성하고, 이러한 프로파일과 문헌 벡터 간의 유사도 매칭을 통하여 문헌에 색인어를 부여하는 방법이다. 이러한 프로파일은 각 디스크립터를 표현하는 것으로 프로토타입(Ittner, Lewis, and Ahn 1995), 센트로이드(이재윤 2005a; Ruiz and Srinivasan 1999), 또는 전형문서(Prototypica document/Sebastiani 2002)라고 부르기도 하는데, 특정 디스크립터의 부여 여부에 따라 긍정예제와 부정예제로 구분된 문헌집합에서의 단어 출현 정보에 따라 생성된다.

적합성 피드백에 의한 질의확장에서 사용되었던 로치오 알고리즘은 Hull(1994)에 의해 처음으로 통제어휘 자동색인에 적용되었다. 로치오 알고리즘을 이용한 통제어휘 자동색인에서는 특정 디스크립터가 부여된 문헌들을 긍정예제로, 부여되지 않은 문헌들을 부정예제로 취급하여 기본적으로 다음과 같은 공식을 사용한다(Ittner, Lewis, and Ahn 1995).

$$w_{ck} = \beta \frac{1}{|R_c|} \sum_{i \in R_c} w_{ik} - \gamma \frac{1}{|\overline{R_c}|} \sum_{i \in R_c} w_{ik}$$

여기서 R_c는 긍정예제의 수, $\overline{R_c}$는 부정예제의 수이다. 이때, 대부분의 디스크립터가 비교적 작은 수의 긍정예제와 상당히 많은 수의 부정예제를 가지는 경우가 많으므로, 긍정예제와 부정예제에 서로 다른 파라미터 값을 곱해주어 균형을 맞추거나 혹은 긍정예제만을 사용하여 프로파일을 생성한다. 예를 들면, 전자의 경우에는 β값(16)을 γ값(4)보다 크게 하여 긍정예제의 상대적 영향력을 높였고(Ittner, Lewis, and Ahn 1995), 후자의 경우에는 β은 1, γ값은 0으로 하여 부정예제를 제외하고 긍정예제만을 사용하였다(Hull 1994; Schütze, Hull, and Pedersen 1995; Joachims 1998).

통제어휘 자동색인에서 로치오 알고리즘에 기초한 프로파일을 사용하는 방법은 기본적으로 가중치의 평균을 구하는 방식으로 학습하기 때문에 구현이 용이하고 컴퓨터 저장공간 및 처리시간 측면에서 상당한 장점을 가지고 있어 많은 학자들에 의해 사용되었다(Schapire, Singer, and Singhal 1998; Joachims 1996; Ittner, Lewis, and Ahn 1995; Yang 1999). 또한

웹 환경의 기반 기술을 최초로 개발한 바 있는 CERN에서는, 디스크립터 프로파일에 기초한 자동 디스크립터 부여기(the automatic descriptor assigner)를 사용하여 고에너지 물리학(High Energy Physics) 관련 문헌들에 DESY 디스크립터를 할당한 결과로, 재현율과 정확률 양자에서 60%에 가까운 성능을 보고하였다(Montejo-Raez 2002). 최근 미국 국립의학도서관(NLM)은 색인전문가의 색인작업을 지원하기 위한 Medical Text Indexer(MTI) 시스템을 소개하면서, 로치오 알고리즘에 기초한 통제어휘 자동색인에서 문헌의 전문(full-text)을 사용한 경우와 논문의 일부(sections)를 사용한 경우의 성능을 비교하였다(Gay, Kayaalp, and Aronson 2005).

한편, 다언어(multilingual) 자동색인을 위한 연구에서는 영어가 아닌 다른 언어를 사용하여 영어문헌을 검색하기 위한 목적으로 디스크립터 프로파일과 유사한 방법을 사용하였다. 즉, 디스크립터와 단어의 동시출현 정보에 기초한 프로파일을 생성하고 연관 가중치(association weight)를 부여한 다음, 프로파일과 입력문헌간의 내적(inner product) 유사도를 산출하여 복수의 언어로 구성된 OECD 시소러스의 디스크립터를 문헌에 부여하였다.(Ferber 1997). 이외에 유럽연합의 공식 문헌들을 대상으로 통제색인어를 부여하기 위한 일련의 연구들(Pouliquen, Steinberger, and Ignat 2003; Steinberger, Pouliquen, and Hagman 2002; Steinberger 2001; Steinberger, Hagman, and Scheer 2000)에서도, 로치오 알고리즘에 기초한 프로파일을 사용하는 자동색인 방법을 사용하였다. 이들은 유럽연합 국가들의 11개 공식 언어로 구성된 EUROVOC 시소러스의 디스크립터를 문헌에 부여하는 과정에서, 특정 디스크립터와 연관 단어들(associates) 간의 동시출현 정보에 기초한 프로파일을 생성하고 로그 승산비로 축소한 자질집합에 가중치를 부여한 다음, 코사인 유사도를 산출하여 디스크립터를 부여하였다.

2.2 로치오 알고리즘을 이용한 텍스트 범주화

텍스트 범주화를 위한 많은 연구들은 로치오 알고리즘의 여러 장점에도 불구하고 성능 측면에서 다소 떨어진다는 일부의 평가(Cohen & Singer 1999; Yang 1999; Joahchims 1998; Lewis et al. 1996; Schütze, Hull, and Pedersen 1995)에 따라, 로치오 방법을 다른 분류기(분류 알고리즘)와의 성능 비교를 위한 기본형으로 사용하는 경우가 많았다(Rogati and Yang 2002; Moens 2000; Galavotti, Sebastiani, and Simi 2000; Schapire & Singer 2000; Cohen & Singer 1999; Lewis et al. 1996; Schütze, Hull, and Pedersen 1995). 또한, 로치오 알고리즘의 성능이 낮다는 평가를 전제로 이를 통한 텍스트 범주화의 성능 향상을 모색하기 위하여 유사긍정예제(near_positives), 질의 지역화(quey zoning), 동적 피드백 최적화(dynamaic feedback optimization) 등 여러 성능 향상 기법을 적용하는 연구들이 진행되었다(Weigend, Wiener, and Pedersen 1999; Schapire, Singer, and Singhal 1998; Ng, Goh, and Low 1997; Wiener, Pedersen, and Weigend 1995).

그러나, 여러 연구 결과 중에서 로치오 알고리즘의 저성능 평가를 일반화시킬 수 없는 사실들이 발견되었다. 예를 들면, 로치오 알고리즘과 다른 알고리즘(k-NN) 간의 성능 비교에서 자질의 수가 크게 축소되는 경우를 제외하고는 성능이 거의 동등하거나 일부 더 높은 것으로 나타났다(Galavotti, Sebastiani, and Simi 2000). Joachims(1998)의 연구 결과에서도 고빈도 범주(≥300)의 경우에는 지지벡터기계(SVM), k-최근접이웃(k-NN), 결정트리(Decision Tree), 로치오 알고리즘(Rocchio Algorithm), 나이브 베이즈(Naive Bayes)가 모두

유사한 성능을 보이는 것으로 발표되었다. 특히, 학습과정에서 유사긍정예제를 함께 포함시킬 경우에는 오히려 여러 방법들 가운데 로치오 알고리즘이 최고의 성능을 보여주었다(Schapire, Singer, and Singhal 1998). 따라서, 로치오 알고리즘이 자동색인이나 자동분류를 위한 목적에 이용되었을 때, 다른 방법들에 비하여 항상 낮은 성능을 나타내는 것으로 단정하는 것은 문제가 있다.

로치오 알고리즘을 이용한 통제어휘 자동색인이나 텍스트 범주화에서, 로치오 알고리즘의 성능에 영향을 주었던 주요 요인들은 프로파일 생성 방법, 자질선정 기준, 가중치부여 방법, 학습집합의 크기(긍정, 부정예제의 수 또는 비율), 유사도 산출 공식, 기준치 적용방법 등이었다. 여기서 기본형으로 쓰인 로치오 알고리즘의 구현을 위한 프로파일 생성 방법으로는 문헌에 출현한 용어의 가중치 평균을 주로 사용하고 있어, 이렇게 생성된 벡터는 곧 문헌들의 센트로이드가 된다. 또한, 센트로이드를 구성하는 각 자질의 가중치부여 방법으로는 단어빈도(*tf*), 또는 이러한 단어빈도에 역문헌빈도를 곱한 값(*tfidf*)을 사용하고, 입력문헌과의 유사도 산출에는 내적(inner product)을 사용하는 경우가 많았다. 따라서 대부분의 연구 결과가 이러한 기본형 로치오 알고리즘에 여러 자질선정 기준을 사용하거나 학습집합의 크기를 변화시켜 여러 알고리즘(또는 분류기)간에 또는 복합(hybrid) 알고리즘과의 성능을 비교하는 것이었다.

본 연구에서는 로치오 알고리즘의 성능에 영향을 주는 요인으로서 자질선정 기준과 프로파일 생성방법을 검토한 결과에 따라 생성한 로치오 알고리즘 기본형을 대상으로, 여러 가중치 부여방법을 적용하여 성능 향상을 모색하여 보았다. 또한, 동일한 조건하에서 로치오 알고리즘 기본형과 다른 학습기반 방법들의 성능을 비교하는 실험을 통하여 각 방법의 실제적인 성능 및 특성을 알아보았다.

3. 실험 설계

먼저, 로치오 알고리즘을 사용하는 통제어휘 자동색인에서 여러 자질선정 기준과 프로파일 생성방법을 적용하였고, 그 결과에 따라 기본형 프로파일을 생성하였다. 다음으로, 기본형 프로파일에 세 가지 가중치 부여방법을 적용하여 가장 좋은 성능을 보였던 가중치를 부여한 다음, 최근 1년부터 10년까지 연차적으로 증가시킨 학습집합의 크기에 따른 성능을 알아보기 위한 실험을 수행하였다. 또한, 여러 성능 요인을 동일하게 적용한 환경에서 로치오 알고리즘에 기초한 기본형 프로파일 방법(PV_b)과 다른 학습기반 방법들(지지벡터기계/SVM, 투표형 퍼셉트론/VPT, 나이브 베이즈/NB)의 성능을 비교하였다.

3.1 문헌집단 및 사전처리

실험에 사용된 문헌집단은 Journal of Citation Reports(2001년~2004년)의 데이터에 기초하여 정보학 분야의 핵심 학술지들을 선정하여 구성하였다. 특히, 학술지의 영향력을 Impact Factor(IF), Immedicacy Index(II), Cited Half-Life(CH)의 세 가지 요소를 중심으로 학술지를 선정한 결과, IPM(Information Processing & Management), JASIST(Journal of the American Society for Information Science & Technology), JD(Journal of Documentation)의 3개 학술지를 정보학 분야의 핵심 학술지로 선정하였다.

실제 데이터 수집을 위해서는 Cambridge Scientific Abstracts(CSA)사의 CSA Illumina에

서 제공하는 여러 데이터베이스 중에서 LISA 데이터베이스를 대상으로, 1994년부터 2004년까지 3개 학술지에 실린 논문들을 검색한 결과를 다운로드하여 실험 문헌집단을 구성하였다(김판준 2006). 이러한 실험 문헌집단의 구체적인 내용은 <표 1>와 같다.

<표 1> 실험 문헌집단

항 목	내 역
전체 문헌 수(학습/검증집합의 문헌 수)	1,858(1,666/192)
디스크립터 종수	33
디스크립터 당 학습문헌 수(최대/최소/평균)	409/16/68.3
학습문헌 당 디스크립터 수(최대/최소/평균)	12/1/4
학습집합의 용어 종수	6,433
학습문헌 당 용어 종수(최대/최소/평균)	119/11/46.2
학습문헌 당 용어 수(최대/최소/평균)	184/18/69.2

텍스트 범주화에서와 마찬가지로 프로파일을 사용하는 디스크립터 자동부여를 위한 세 가지 기본적인 요소는 (1) 문헌집단(collection), (2) 디스크립터(descriptors), (3) 자질(features)이다. 여기서 문헌집단은 크게 학습집합과 검증집합으로 구분된다. 실험에서 학습집합은 1994년부터 2003년까지 이전 10년 동안 발행된 3개 학술지에 수록된 논문들로, 검증집합은 가장 최근인 2004년 한 해 동안 발행된 3개 학술지에 수록된 논문들로 구성하였다.

LISA 데이터베이스의 디스크립터 필드에 출현한 디스크립터들 가운데 학습집합에 속한 논문들에 한 번 이상 부여된 디스크립터는 총 1,565개이지만, 본 연구에서는 학습집합 내 문헌빈도가 최소 15이상인 디스크립터 75개 중에서 임의 선정한 33개의 디스크립터를 사용하였다. 또한, 문헌에 디스크립터를 부여하기 위한 단서가 되는 자질집합은 각 논문의 제목과 초록 필드에서 다운로드한 단어를 대상으로 Porter Stemmer를 사용한 형태소분석(stemming)과 전치사, 조사, 숫자 등의 불용어 제거 절차를 거친 다음, 여러 자질 선정 기준을 적용한 결과에 따라 구성하였다.

실험집단에 대한 사전처리와 자질선정 기준의 적용은 Python으로 구현된 프로그램과 Access를 사용하였고, 디스크립터의 자동부여 및 성능 비교를 위한 실험은 Python 프로그램과 Excel을 사용하였다.

3.2 자질선정, 프로파일 생성방법, 가중치 부여방법에 따른 실험

학습집합의 문헌들에 부여된 디스크립터와 함께 출현한 단어 정보에 기초하여 디스크립터 프로파일을 생성하였다. 로치오 알고리즘에 기초한 디스크립터 프로파일은 학술지 논문에 특정 디스크립터를 자동부여하기 위한 것으로, 이때 문헌에 부여되는 디스크립터를 특정 주제를 표현하는 주제 범주로 보면 텍스트 범주화에서 사용되는 분류기(classifier)와 동일한 역할을 한다. 본 연구에서 디스크립터 프로파일은 정보검색에서 일반적으로 사용되는 벡터공간 모형과 로치오 알고리즘을 기반으로 특정 디스크립터가 부여되었는지의 여부에 따라 문헌에 출현한 단어들을 중심으로 생성된다.

로치오 알고리즘을 이용한 디스크립터 자동부여에서 자질선정과 가중치부여 방법은 성능에 상당한 영향을 주는 요인으로 작용할 수 있다. 먼저, 로치오 알고리즘을 이용한 디스크립

터 자동부여에서 자질선정 기준에 따른 성능을 비교하는 실험을 수행하였다. 즉, 33개 디스크립터 중 1/3에 해당하는 11개 디스크립터를 대상으로 전체(ALL), 문헌빈도 6이상(DF6), 카이제곱 통계량(CHI), 코사인 계수(COS), 자카드 계수(JAC), GSS 계수(GSS), 로그 승산비(LOR)의 7개 자질선정 기준을 적용한 결과를 비교하였다. 여기서 가장 좋은 성능을 보인 자질선정 기준에 따라 선정된 자질들로 전체 자질집합을 약 25% 수준(1,611개)으로 축소하였다.

이러한 자질집합으로 구성된 벡터에 가중치로 용어빈도(*tf*)를 사용하였고, 다음으로 가중치합, 가중치 평균, 가중치 최대값의 세 가지 생성방법(Sum_ *tf*, Avg_ *tf*, Max_*tf*)을 적용하여 산출한 결과 중에서, 가장 좋은 성능을 보이는 방법을 이후의 실험을 위한 기본적인 프로파일 생성방법으로 사용하였다. 여기서 가중치합을 사용한 방법은 긍정예제에 출현한 단어들의 가중치합(PV_b)을 사용하여 디스크립터 프로파일을 생성하는 것이며, 가중치 평균(PV_avg)을 사용하는 방법은 긍정예제만을 사용하여 센트로이드 벡터를 생성하는 것과 동일하다. 한편, 가중치 최대값(PV_max)은 해당 디스크립터가 부여된 문헌들의 용어빈도(*tf*) 중에서 가장 큰 값을 사용하는 것이다.

$$\vec{T} = (wt_1, wt_2, wt_3, \ldots, wt_i), \quad wt_i = tf$$

$$\text{PV\_b(가중치합 프로파일)}: \; wt_i = \sum_i tf$$

$$\text{PV\_avg(가중치 평균 프로파일)}: \; wt_i = \frac{1}{N_T}\sum_i tf$$

$$\text{PV\_max(가중치 최대값 프로파일)}: \; wt_i = \max tf$$

프로파일를 구성하는 자질의 가중치는 기본적으로 문헌 내 출현빈도(tf)를 사용하였는데, 출현빈도 가중치를 사용한 결과는 흔히 검색 결과의 평가에서 단어빈도 요소의 유용성을 측정하는 기준으로 사용될 수 있다(정영미 2006). 이러한 디스크립터 프로파일은 하나의 디스크립터가 여러 문헌에 부여되어 있는 상황에서 생성되는 것이므로, 기본형의 가중치는 결과적으로 해당 디스크립터가 부여된 문헌 내 출현빈도의 합이 된다. 사전실험 결과, 위의 세 가지 프로파일 생성방법 중에서 좋은 성능을 보인 가중치합 프로파일(PV_b)을 이후의 실험에서 기본형으로 사용하였다.

다음으로, 전체 33개 디스크립터를 대상으로 디스크립터 프로파일 기본형(PV_b)의 성능 향상을 위한 방법으로서, 자질집합에 대하여 여러 가중치부여 방법을 적용하는 실험을 수행하였다. 자질집합을 구성하는 용어들에 대한 가중치부여 방법으로는 빈도, 확률, 동시출현정보 등을 사용할 수 있는데, 본 연구에서는 출현빈도에 기초한 기본적인 가중치부여 방법 세 가지를 사용하였다. 즉, 기본형으로 문헌 내 출현빈도(PV_b: *tf*)를 사용하는 방법과 역문헌빈도를 적용하여 정규화하는 방법(PV_i: *tfidf*), 그리고 용어빈도에 로그를 취하고 역문헌빈도로 정규화하는 방법(PV_L: *Ltfidf*)의 세 가지 가중치부여 방법을 사용하였다.

$$wt_i = \sum_{T \in D} tf,$$

$$wt_i = \sum_{T \in D} tfidf,$$

$$wt_i = \sum_{T \in D} \log(tf) idf$$

디스크립터가 부여될 검증집합의 각 문헌은 문헌에 출현한 단어들로 구성된 자질벡터로 표현하였고, 각 자질의 가중치는 해당 문헌 내 빈도(*tf*)를 가중치로 부여하였다.

$$\vec{d} = (wd_1, wd_2, wd_3, ..., wd_m), \quad wd_i = tf$$

특정 디스크립터의 부여 결정은 디스크립터 프로파일 벡터와 이러한 문헌벡터 간의 유사도를 계산하여 결정한다. 본 실험에서는 정보검색에서 많이 사용되고 있는 코사인 유사계수를 이용하여 두 개의 가중치 벡터 간의 유사도를 산출하였다.

$$Cosine(T,d) = \frac{\sum_T wt_i \cdot \sum_d wd_i}{\sqrt{\sum_T wt_i^2} \cdot \sqrt{\sum_d wd_i^2}}$$

3.3 학습집합의 크기에 따른 실험

로치오 알고리즘을 사용한 통제어휘의 성능에서 또 다른 주요 영향 요인이 되는 학습집합의 크기에 따른 성능을 비교하였다. 선행 연구들에서는 주로 학습집합 내 긍정예제의 수 또는 비율을 점진적으로 증가시켜 학습집합의 크기를 조정하였지만, 본 연구에서는 전체 10년 동안의 학습집합(1994년~2003년)을 출판년도에 따라 가장 최근의 연도(2003년)부터 1년씩 추가하여 학습집합의 양을 증가하는 방식으로 학습집합의 크기를 변화시켰다. 그 이유는 대규모 학술 데이터베이스 환경에서 매년 새롭게 입력되는 많은 양의 학술지 논문들에 디스크립터를 부여하기 위한 목적으로 프로파일 기반의 자동색인을 적용하는 경우, 이전 몇 년 동안의 학습집합을 사용하는 것이 가장 좋은 결과를 산출하는 지를 검토해 보는 것이 의미가 있을 것으로 생각하였기 때문이다.

먼저, 사전 실험 결과에 기초하여 모든 33개 디스크립터에 대하여 로치오 알고리즘에 기초한 프로파일의 생성방법을 가중치합으로, 자질선정 기준으로는 자카드 계수를 적용하여 전체 자질집합의 25%(1,611개) 수준으로 자질집합을 축소하였다. 다음으로, 용어빈도에 기초한 세 가지 방법을 사용하여 자질 가중치를 부여한 3개 유형의 디스크립터 프로파일 벡터(PV_b, PV_i, PV_L)를 생성한 다음, 각 프로파일을 최근부터 1년씩 추가되는 10개(1t1~10t1)의 학습집합을 사용하여 학습하였다. 또한, 유사도 순으로 정렬된 후보 디스크립터 리스트에서 특정 디스크립터의 부여 여부를 결정하는 과정에서는, 학습문헌의 코사인 유사도 값에 대하여 일정 구간의 기준치(0~0.95)를 0.05씩 증가시켜 적용하고, 이 중에서 최대 성능을 산출하는 기준치를 검증문헌에 적용함으로써 학습에 따른 각 프로파일의 성능을 비교하였다.

3.5 다른 학습기반 방법과의 성능 비교

다른 학습기반 방법들과 비교하여 로치오 알고리즘에 기초한 프로파일 기본형의 성능이

실제로 저성능인가를 알아보았다. 이를 위해 모든 33개 디스크립터를 대상으로 프로파일 기본형(PV_b)과 다른 세 가지 학습기반 방법(NB, SVM, VPT)에 동일한 성능요인을 적용한 결과를 비교하였다. 즉, 자카드 계수(JAC)를 자질선정 기준으로 약 25% 수준으로 축소된 자질집합을 구성하고, 가중치합을 사용하여 프로파일을 생성한 다음, 단어빈도(*tf*)를 가중치로 부여하는 것으로 성능 요인들을 동일하게 적용하였다.

또한, 다른 세 가지 학습기반 방법들(NB, SVM, VPT)과 비교하여 로치오 알고리즘에 기반한 프로파일 방법(PV_b)이 여러 성능 척도(마이크로 평균 F_1/MIF, 마이크로 평균 재현율/MIR, 마이크로 평균 정확률/MIP) 측면에서 어떠한 특성을 가지고 있는 가를 알아보았다.

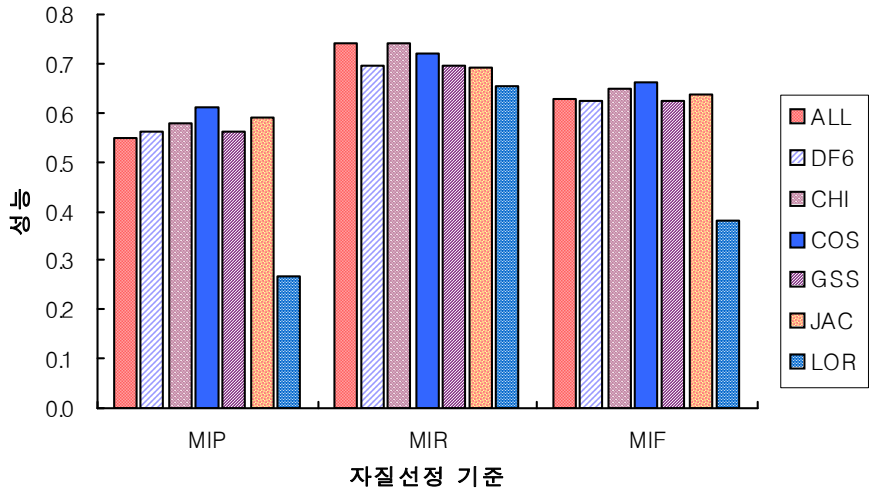
4. 실험 결과 및 분석

본 연구에서 수행된 모든 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 함께 사용하였다. 또한, 이들 성능 척도의 평균을 구하는 방법으로는 문헌 중심의 마이크로와 범주 중심의 매크로 방법이 있는데, 이 중에서 최근 텍스트 범주화에서 많이 사용되고 있는 마이크로 평균을 사용하였다(김판준 2006).

4.1 자질선정 및 프로파일 생성방법에 따른 결과 및 분석

로치오 알고리즘에 기초한 디스크립터 자동부여에서 여러 자질선정 기준과 프로파일 생성 방법을 적용해 보았다. 자질선정 기준으로는 전체 자질(ALL: 6,433개)을 사용한 결과와 함께 다양한 자질선정 기준(DF6, CHI, COS, GSS, JAC, LOR)을 적용하여 전체 자질을 축소(25%: 1,611개)하였을 때의 성능을 알아보았다. 또한, 디스크립터 프로파일 벡터를 생성하는 방법으로 가중치합(PV_b), 가중치 평균(PV_avg), 가중치 최대값(PV_max)의 세 가지 방법을 적용하고 그 성능을 알아보았다.

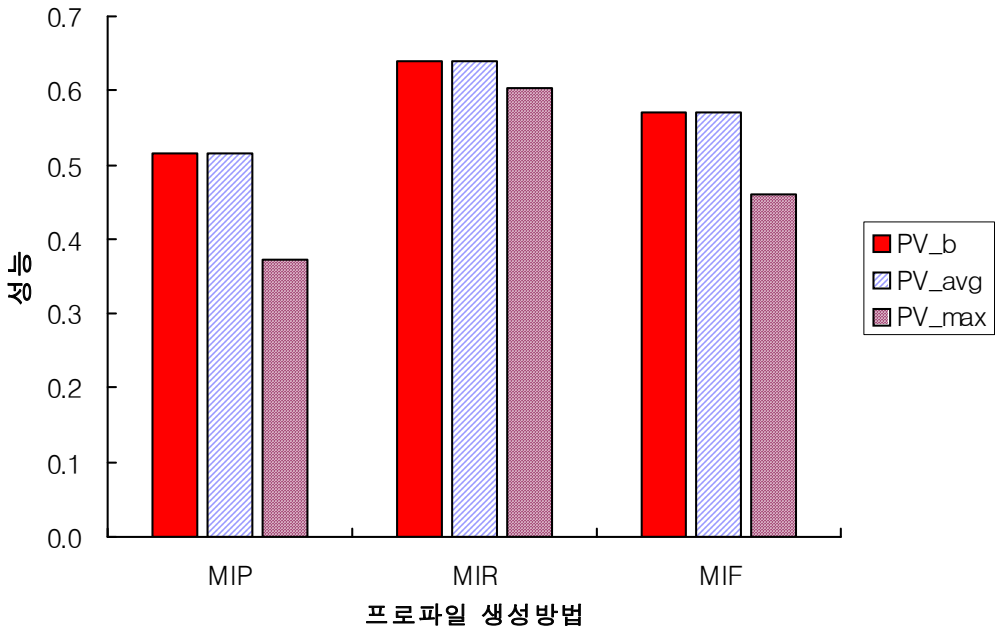
<그림 1>은 여러 자질선정 기준에 따른 로치오 알고리즘을 이용한 디스크립터 자동부여 실험의 결과를 각각 마이크로 평균 정확률(MIP), 마이크로 평균 재현율(MIR), 마이크로 평균 F_1(MIF) 척도로 나타낸 것이다. 여기서 ALL은 전체 자질을 모두 사용한 결과이고, DF6는 문헌빈도가 6이상($DF \geq 6$)인 자질집합, CHI는 카이제곱 통계량을 적용한 자질집합, COS는 코사인 계수를 적용한 자질집합, GSS는 카이제곱 통계량의 변형인 GSS 계수를 적용한 자질집합, JAC는 자카드 계수를 적용한 자질집합, LOR은 로그 승산비를 적용한 자질집합을 사용하여 산출한 결과이다.



<그림 1> 자질선정 기준에 따른 성능

마이크로 평균 F_1값으로 본 성능 측면에서 모든 자질을 사용하였을 경우(ALL/0.63)와 비교하면, 코사인 계수(COS/0.66), 카이제곱 통계량(CHI/0.65), 자카드 계수(JAC/0.64)의 세 가지 자질선정 기준을 적용하여 자질집합을 축소한 경우에는 성능이 더 향상된 결과를 보였고, GSS 계수(GSS/0.62)와 문헌빈도(DF6/0.62)를 적용한 결과는 거의 동등한 성능을 나타냈다. 대조적으로, 로그 승산비(LOR/0.38)를 적용한 경우에는 상당한 성능 감소를 보였는데, 이러한 결과는 저빈도어 선호 경향이 있는 로그 승산비의 특성에 따른 것(이재윤 2005b)으로 이전 연구결과와 유사한 경향을 보여주었다(김판준 2006). 실험 결과, 전체 자질을 사용한 것(ALL)보다 상위 세 가지 자질선정 기준(CHI, COS, JAC)이 거의 유사한 수준으로 성능이 향상된 결과를 보였지만, 이후의 실험에서는 다른 학습기반 방법들과의 비교를 위하여 자카드 계수(JAC)를 기본적인 자질선정 기준으로 적용하여 자질집합을 구성하였다.

다음으로, 자카드 계수를 적용하여 축소된 자질집합을 사용하면서 로치오 알고리즘에 기초한 프로파일 벡터의 생성방법으로 가중치를 합하는 방법(PV_b)과, 평균값을 취하는 방법(PV_avg), 최대값을 취하는 방법(PV_max)을 각각 적용하고 그 성능을 산출하였다. <그림 2>는 이러한 세 가지 방법을 적용하여 디스크립터를 부여한 결과를 나타낸 것이다.



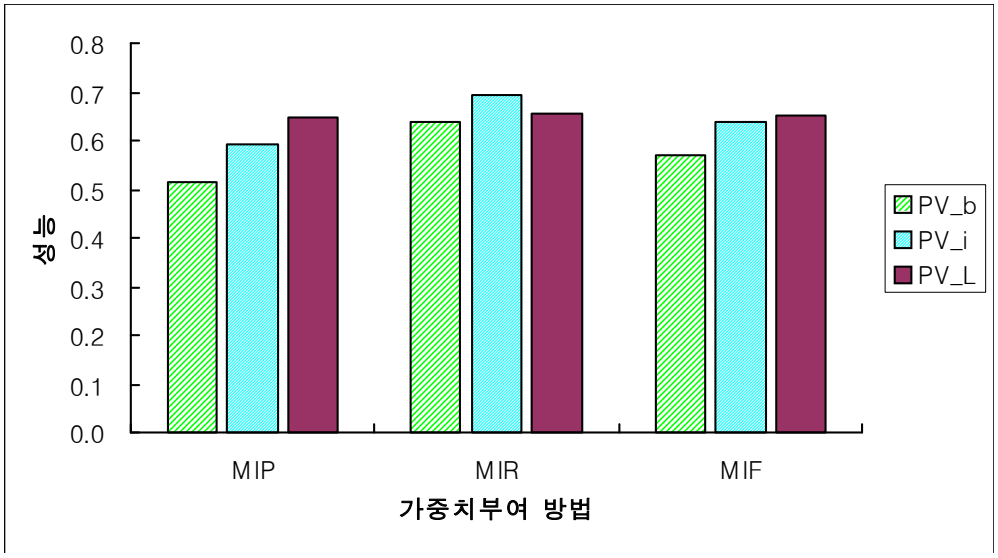
<그림 2> 프로파일 생성방법에 따른 성능

<그림 2>에서 보는 바와 같이 가중치합(PV_b/0.57)과 가중치 평균(PV_avg/0.57)의 성능은 동일한데 반하여, 가중치 최대값(PV_max.0.46)의 성능은 크게 떨어지는 것으로 나타났다. 여기서 가중치합(PV_b)과 가중치 평균(PV_avg)의 결과가 동일하게 나타난 것은 디스크립터 부여 단계에서 코사인 유사도를 사용하였기 때문이다. 따라서 이후의 실험에서는 상대적으로 단순하여 컴퓨터 처리상의 이점이 있는 가중치합을 적용하여 생성한 프로파일 벡터(PV_b)를 기본형으로 사용하였다.

4.2 가중치부여 방법에 따른 결과 및 분석

이전 실험에서의 결과에 따라 자질선정 기준으로 자카드계수(JAC)를 적용하여 자질집합을 구성하고 가중치합을 생성방법으로 사용한 프로파일 기본형(PV_b)을 생성한 다음, 기본형의 성능 향상을 위한 한 방법으로 여러 가중치부여 방법의 적용에 따른 성능을 알아보았다. 그 결과, 가중치부여 방법 중에서 용어의 출현빈도에 기초한 기본적인 가중치부여 방법 세 가지(PV_b/*tf*, PV_i/*tfidf*, PV_L/*Ltfidf*)를 사용한 결과는 <그림 3>과 같다. 여기서, 각 자질에 부여된 가중치는 본 연구에서 사용하는 기본적인 프로파일 벡터 구성방법의 특성에 따라 디스크립터가 부여된 각 문헌에 출현한 용어의 가중치를 모두 합산한 값이 된다.

가중치부여 방법 중에서 단순히 용어빈도만을 사용한 기본형(PV_b: 0.57)보다 다른 두 가지 가중치부여 방법이 뚜렷하게 더 좋은 성능을 보였다. 가장 좋은 성능은 용어빈도에 로그를 취하여 역문헌빈도로 정규화한 것(PV_L: 0.65)이었지만, 로그를 취하지 않고 용어빈도를 역문헌빈도로 정규화한 방법(PV_i: 0.64)도 이와 거의 동등한 수준이었다. 따라서 전체 디스크립터(33개)를 대상으로 하는 학습집합의 크기에 따른 실험에서, 이들 세 가지 가중치부여 방법의 성능을 여러 측면에서 다시 검토하였다.



<그림 3> 자질 가중치부여 방법에 따른 성능

4.3 학습집합의 크기에 따른 결과 및 분석

전체 33개 디스크립터를 대상으로 세 가지 가중치부여 방법을 적용한 3개 유형의 프로파일(PV_b, PV_i, PV_L)을 생성하고, 전체 10년 동안의 학습집합(1994년 ~2003년)을 출판년도에 따라 가장 최근의 연도(2003년)부터 1년씩 추가하여, 학습집합의 양을 증가하는 방식으로 학습집합의 크기를 변화시킨 성능을 산출하였다.

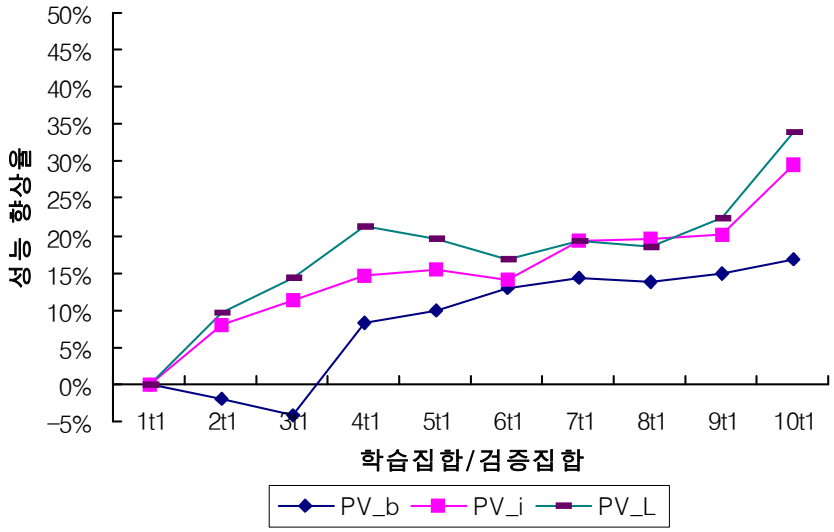
<표 2>는 세 가지 가중치부여 방법을 적용한 각 디스크립터 프로파일에 대하여 학습집합의 크기를 연차적으로 변화시켜 얻은 결과를 마이크로 평균 F_1(MIF) 척도로 나타낸 것이다. 이에 따르면 전체 10년의 학습집합(10t1)을 모두 사용한 경우의 성능에서, 이전 실험의 결과와 마찬가지로 기본형(PV_b)보다 다른 두 가지 가중치부여 방법(PV_i, PV_L)을 사용한 경우에 상당히 향상되었음을 알 수 있다(PV_i: 0.447 vs. 0.542, +21.3%; PV_L: 0.447 vs. 0.546, +22.1%). 즉, 정보검색에서와 마찬가지로 디스크립터의 자동부여에서도 자질 가중치로서 단순 빈도(*tf*)만을 사용하는 것보다는 역문헌빈도로 정규화(*tfidf*)하거나 단어빈도에 로그를 취하여 역문헌빈도로 정규화 (*Ltfidf*)하는 경우에 성능이 향상되었다.

<표 2> 학습집합의 크기에 따른 성능: 세 가지 가중치부여 방법에 의한 로치오 알고리즘 기반 프로파일의 비교

	1t1	2t1	3t1	4t1	5t1	6t1	7t1	8t1	9t1	10t1
PV_b	0.383	0.375	0.367	0.415	0.421	0.433	0.438	0.436	0.440	**0.447**
PV_i	0.418	0.451	0.466	0.479	0.483	0.477	0.499	0.501	0.503	**0.542**
PV_L	0.407	0.447	0.465	0.494	0.487	0.475	0.486	0.482	0.98	**0.546**

학습집합의 크기를 변화시킨 성능 측면에서, PV_i(*tfidf*)와 PV_L(*Ltfidf*)은 2년 이상의 학습집합(≥3t1)을 사용하는 경우에는 PV_b(*tf*)에서 전체 10년의 학습데이터(10t1)를 사용하는 것보다 항상 더 높은 성능을 보여주었다. 이러한 결과는 PV_i(*tfidf*)와 PV_L(*Ltfidf*)이 더 작

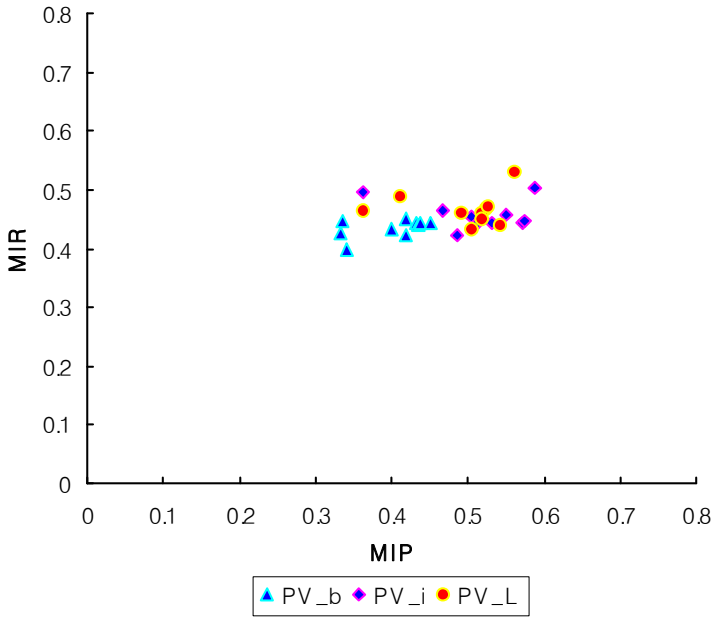
은 규모의 학습집합(≥2t1)으로도 PV_b(*tf*)가 전체 학습집합(10t1)을 사용하는 것과 동등하거나 더 좋은 결과를 가져올 수 있다는 것을 의미한다.



<그림 4> 학습집합의 크기에 따른 성능: 세 가지 가중치부여 방법에 의한 로치오 알고리즘 기반 프로파일의 비교(성능 향상율)

학습집합의 크기에 따른 성능을 다른 측면에서 보면, <그림 4>는 1년(1t1)을 기준으로 하여 학습집합의 증가에 따른 세 가지 가중치부여 방법의 성능 향상율을 나타낸 것이다. 전체 학습집합(10t1)을 사용하였을 때, 최근 1년(1t1)의 결과와 비교하여 PV_b는 향상율이 16.7%(0.383 vs. 0.447)에 불과하지만 PV_i과 PV_L은 각각 30%(0.418 vs. 0.542)와 34%(0.407 vs. 0.546)로 더 큰 성능 향상율을 보여주고 있다. 또한, 학습집합의 크기를 변화시킨 경우에 PV_b가 3년까지(3t1)는 성능 향상이 없거나 오히려 떨어지는 반면, PV_i와 PV_L은 학습 데이터의 증가에 따라 지속적으로 성능이 향상되었다.

<그림 5>는 학습집합의 크기 변화에 따른 세 가지 가중치부여 방법에 의한 로치오 알고리즘 기반 3개 프로파일의 결과를 정확률과 재현율의 산포도로 작성한 것이다. 여기서 로치오 알고리즘 기반의 3개 프로파일이 재현율 측면에서는 비교적 높은 수준을 유지하고 있지만, 정확률은 상대적으로 이보다 낮은 수준에 있다는 것을 알 수 있다. 여기서, 특히 PV_i(*tfidf*)와 PV_L(*Ltfidf*)은 동일한 재현율 수준에서 PV_b(*tf*)보다 나은 정확률을 보이고 있어, 전반적인 성능(MIF)에서 우위에 있다는 것을 알 수 있다.



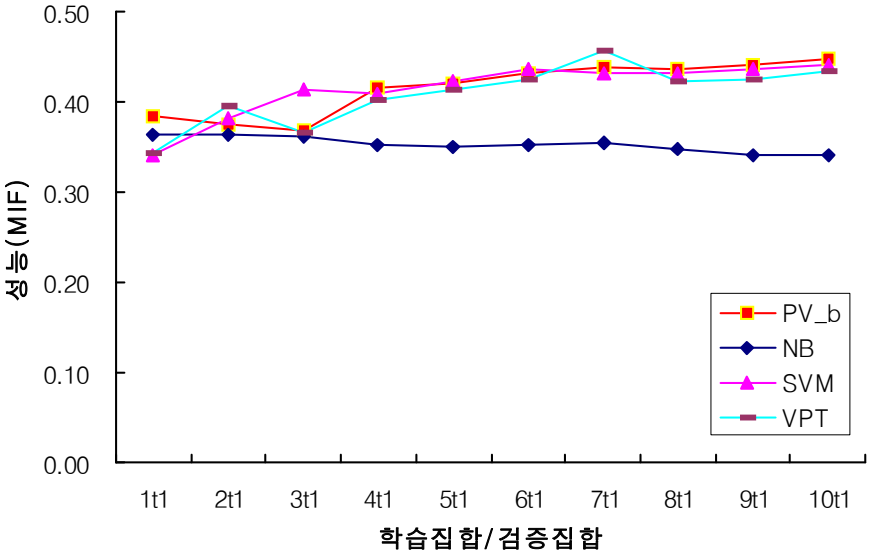
<그림 5> 학습집합의 크기에 따른 성능: 세 가지 가중치부여 방법에 의한 로치오 알고리즘 기반 프로파일의 정확률(MIP)과 재현율(MIR) 산포도

4.4 다른 학습기반 방법과의 비교 분석

동일한 환경(실험집단, 자질선정 기준, 가중치부여 방법, 학습집합의 크기)에서 실험한 다른 3가지 학습기반 방법(나이브 베이즈/NB, 지지벡터기계/SVM, 투표형 퍼셉트론/VPT)의 결과와 로치오 알고리즘 기본형(PV_b)의 결과를 다음의 <그림 6>으로 함께 제시하였다.

전체 10년의 학습집합(10t1)을 모두 사용하였을 경우, 로치오 알고리즘 기본형(PV_b)은 선행연구들에서 텍스트 범주화에서 일반적으로 가장 좋은 성능을 보이는 것으로 알려진 지지벡터기계(SVM)과 거의 동등한 성능(0.447 vs. 0.441: +1.4%)을 보여주었다. 또한, 투표형 퍼셉트론(VPT)이나 나이브 베이즈(NB)와는 이보다 더 큰 성능 차이(0.447 vs. 0.435: +2.8%, 0.447 vs. 0.341: +31.1%)를 보여주고 있다. 한편, 학습집합의 크기를 연차적으로 증가시키는 경우에는 나이브 베이즈(NB)를 제외한 모든 방법(SVM, VPT, PV_b)에서 4년 이상의 학습집합(≥4t1)을 사용한 이후에는 지속적으로 안정된 성능을 나타내었다.

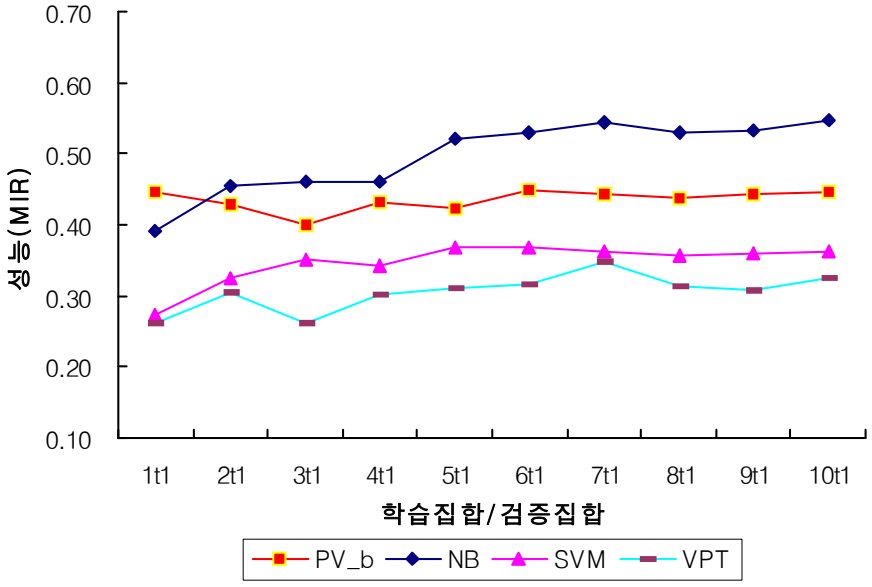
이러한 실험 결과는 자동색인 방법의 성능에 영향을 주는 주요 요인들을 동일한 조건으로 하였을 때, 로치오 알고리즘에 기초한 프로파일 방법(PV_b)은 지금까지 좋은 성능을 보이는 것으로 보고되어온 다른 방법들(SVM, VPT)과 거의 동등하거나 더 나은 성능 수준을 유지한다는 것을 실제적으로 보여주고 있다. 따라서 학술지 논문을 대상으로 하는 경우, 통제어휘 자동색인 또는 텍스트 범주화를 위한 로치오 알고리즘의 성능이 다른 방법들보다 저성능이라는 평가는 재고되어야 한다.



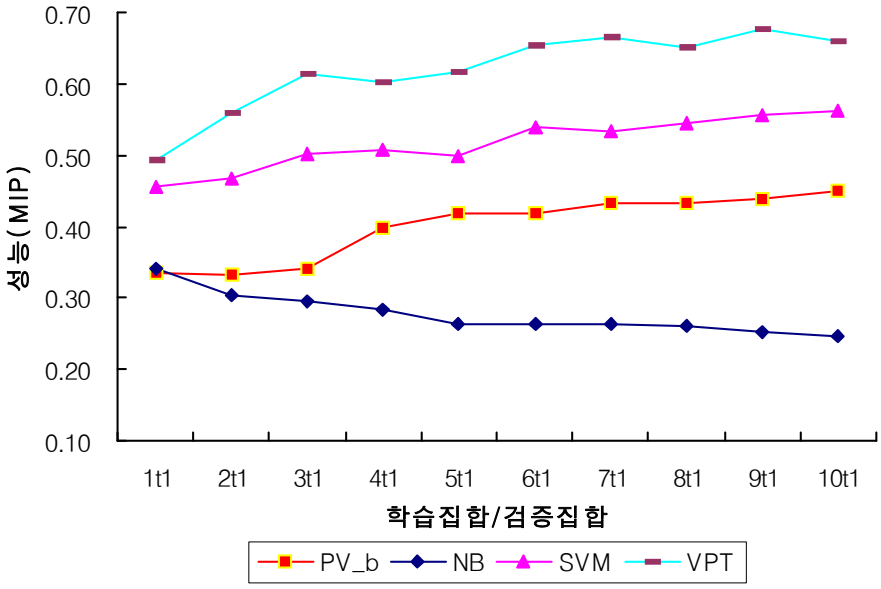
<그림 6> 학습집합의 변화에 따른 성능: 프로파일 기본형(PV_b)과 다른 3개 학습기반 방법 간의 비교(마이크로 평균 F_1)

한편, 통제어휘 자동색인의 결과에 대한 평가에서 일반적으로 완전 자동색인은 색인전문가의 개입이 없고 컴퓨터 알고리즘에 의해 색인어가 직접 부여되기 때문에 보다 정확한 결과를 위하여 정확률을 더 중요시하고, 반자동색인은 컴퓨터에 의해 후보 색인어 리스트가 생성되고 최종 색인어 부여 결정은 색인전문가에 의해 이루어지는 관계로 재현율을 더 중요시하는 경향이 있다. 본 연구에서 사용하는 통제어휘 자동색인은 기본적으로 벡터공간 모형과 로치오 알고리즘에 기초한 방법으로서 알고리즘의 출력이 후보 디스크립터의 순위화된 리스트인 관계로 엄밀하게는 반자동색인에 포함된다. 따라서 로치오 알고리즘 기반 통제어휘 자동색인 방법의 세부적인 성능 평가는 정확률보다는 재현율에 초점을 맞추어야 할 것이다.

<그림 7>은 프로파일 기본형(PV_b)을 비롯한 여러 학습기반 방법(NB, SVM, VPT)의 성능을 마이크로 평균 재현율(MIR)로 나타낸 것이다. 여기서 전체 10년의 학습집합(10t1)을 모두 사용하는 경우와 학습집합의 크기를 달리 하는 경우에 모두, PV_b가 비교적 높은 재현율을 나타내고 있다. 이는 결과에서 재현율이 높은 특성을 나타내는 것으로 알려진 NB보다는 낮은 수준이지만, 상대적으로 정확률에서 강세를 보이는 SVM과 VPT 보다는 상당히 높은 재현율 수준을 보여주는 것이다.



<그림 7> 학습집합의 변화에 따른 성능: 프로파일 기본형(PV_b)과 다른 3개 학습기반 방법 간의 비교(마이크로 평균 재현율/MIR)



<그림 8> 학습집합의 변화에 따른 성능: 프로파일 기본형(PV_b)과 다른 3개 학습기반 방법 간의 비교(마이크로 평균 정확률/MIP)

반면, 마이크로 평균 정확률(MIP) 측면에서는 대조적인 결과가 나타났다. <그림 8>에서 보는 바와 같이 VPT가 가장 높은 성능을 보였고, 그 다음이 SVM, PV_b, NB의 순이었다. 여기서 학습집합의 크기를 연차적으로 증가시키는 경우, PV_b와 함께 재현율에서는 강세를 보이지만 정확률에서 다소 떨어지는 유사한 경향을 보이는 NB는 정확률 향상이 없는 반면, PV_b는 꾸준하게 정확률이 향상되는 것으로 나타났다. 따라서 반-자동색인 즉, 색인전문가를 지원하기 위한 목적으로 사용하기 위해서는, 재현율이 비교적 높은 수준에 있으면서 학습 데이터의 증가에 따라 정확률이 지속적으로 개선되는 로치오 알고리즘을 이용한 방법을 우선적으로 고려할 수 있을 것이다.

5. 결론

지금까지의 로치오 알고리즘에 기초한 통제어휘 자동색인 또는 텍스트 범주화에서 일반적으로 적용되어 온 요소들을 성능 측면에서 재검토해 보고, 성능 향상을 위한 방법을 모색해 보았다. 또한, 동등한 환경에서 통제어휘 자동색인을 위한 로치오 알고리즘 기반 방법의 성능을 다른 학습기반 방법들의 성능과 비교하였다. 먼저, 로치오 알고리즘 기반의 통제어휘 자동색인을 위하여 여러 자질선정 기준과 프로파일을 생성하는 방법을 재검토하였고, 그 결과 생성한 기본형의 성능을 향상하기 위하여 빈도에 기초한 세 가지 가중치부여 방법을 적용하였다. 다음으로, 학습집합을 단순하게 문헌 수 또는 비율로 구분하지 않고 출판년도에 따른 학습집합의 적용에 따른 성능 변화를 다양한 측면에서 살펴보았다. 또한, 로치오 알고리즘에 기초한 프로파일을 사용하는 방법이 다른 학습기반 방법들에 비해 저성능인 것으로 평가되어 온 측면에서, 이를 일반화하기에는 문제가 있다는 것을 실험을 통해 밝혔다.

로치오 알고리즘에 기초한 통제어휘 자동색인 방법은 구현의 용이성과 컴퓨터 저장 공간 및 처리시간 측면의 경제성이라는 기존의 장점을 그대로 유지하면서도, 좋은 성능을 보이는 것으로 알려진 다른 학습기반 방법들(NB, SVM, VPT)보다 거의 동등하거나 더 나은 성능을 보여주었다. 또한, 색인작성의 주체가 색인전문가인 반-자동색인의 경우, 재현율이 비교적 높은 수준에 있으면서 학습 데이터의 증가에 따라 정확률이 지속적으로 개선되는 로치오 알고리즘을 이용한 방법을 우선적으로 고려할 수 있을 것이다. 실험 및 분석을 통해 밝혀진 구체적인 사실은 다음과 같다.

첫째, 로치오 알고리즘 기반의 통제어휘 자동색인에서 여러 자질선정 기준을 적용한 결과는 전체 자질집합을 25%(16,11개) 수준으로 축소하였을 때, 코사인계수(COS/0.66), 카이제곱통계량(CHI/0.65), 자카드계수(JAC/0.64)가 거의 동등하게 높은 수준으로 나타났다.

둘째, 자질선정 기준을 자카드계수로 적용하고 세 가지 프로파일 생성방법을 적용하였을 경우에, 가중치합(PV_b/0.57)와 가중치 평균(PV_avg/0.57)은 동일한 결과를 산출하였고, 가중치 최대값(PV_max/0.46)은 이에 비해 낮은 성능을 보였다.

셋째, 사전 실험 결과에 따라 생성한 로치오 알고리즘 기본형(PV_b)의 성능 향상을 모색하기 위하여 가중치부여 방법과 학습집합의 크기를 달리한 실험에서는, PV_i(*tfidf*)와 PV_L(*tfidf*)이 모두 학습집합의 크기에 상관없이 더 좋은 결과(PV_i: 0.447 vs. 0.542, +21.3%; PV_L: 0.447 vs. 0.546, +22.1%)를 보여주었다.

넷째, 동일한 환경(실험집단, 자질선정 기준, 가중치부여 방법, 학습집합의 크기)에서, 프로파일 생성방법을 가중치합으로 하는 프로파일 기본형(PV_b)과 다른 세 가지 자동색인 방

법(NB, SVM, VPT)의 성능을 비교하였다. 그 결과, PV_b(0.447)는 일반적으로 최고의 성능을 보이는 것으로 알려진 SVM(0.441: +1.4%)이나 VPT(0.435: +2.8%)와 거의 동등한 성능을 나타내었고, NB(0.341: +31.1%)와는 상당히 큰 성능 차이를 보였다.

다섯째, 재현율과 정확률로 본 성능 측면에서, 색인전문가의 색인작업을 지원하기 위한 목적으로는 다른 방법들(SVM, VPT)에 비하여 비교적 높은 수준의 재현율을 유지하면서, 학습 데이터의 증가에 따라 정확률이 상당히 개선(1t1/0.335 vs. 10t1/0.450, +34.3%)되는 로치오 알고리즘을 이용한 방법을 우선적으로 고려할 수 있을 것이다.

<참고문헌>

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

정영미. 2005. 『정보검색연구』. 서울: 구미무역(주) 출판부.

이재윤. 2005a. 문헌간 유사도를 이용한 SVM 분류기의 문헌분류성능 향상에 관한 연구. 『정보관리학회지』, 22(3): 261-287.

_____. 2005b. 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 관한 연구. 『문헌정보학회지』, 39(2): 123-146.

Cohen, W. W., and Y. Singer. 1999. "Context-sensitive learning methods for text categorization." *ACM Transactions on Information Systems*, 17(2): 141-173.

Dattola, R. T. 1969. "A fast algorithm for automatic classification". *Journal of Library Automation*, 2(1): 31-48.

Ferber, Reginald. 1997. "Automated indexing with thesaurus descriptors: a co-occurrence based approach to multilingual retrieval." In: Peters, Carol, and Costantino Thanos eds. In: *Lecture Notes in Computer Science 1324, Research and Advanced Technology for Digital Libraries, First European Conference*, Springer: 233-251.

Galavotti, Luigi, Fabrizio Sebastiani, and Maria Simi. 2000. "Experiments on the use of feature selection and negative evidence in automated text categorization." In: *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries*: 59-68.

Gay, Clifford W., Mehmet Kayaalp, and Alan R. Aronson. 2005. "Semi-automatic indexing of full text biomedical articles". *In: Proceedings of AMIA 2005 Symposium*: 271-275.

Hull, D. A. 1994. "Improving text retrieval for the routing problem using latent semantic indexing." In: *Proceedings of SIGIR 94*: 282-289.

Ittner, D. J., D. D. Lewis, and D. D. Ahn. 1995. "Text categorization of low quality images." In: *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*: 301-315.

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.

Lancaster, F. W. 2003. *Indexing and Abstracting in Theory and Practice*. Third ed. London: facet publishing.

Lewis, D. D. et al. 1996. "Training algorithms for linear text classifiers." In: *Proceedings of SIGIR '96*: 298-306.

Moens, Marie-Francine. 2000. *Automatic Indexing and Abstracting of Document Texts. The Kluwer International Series on Information Retrieval*. Boston: Kluwer Academic Publishers.

Montejo-Raez, Arturo. 2002. "Toward conceptual indexing using automatic assignment of descriptors". In: *Proceedings of the AM 2002 Workshop on Personalization Techniques in Electronig Publishing*. Malaga, Spain, May 2002. [cited 2006. 5. 11.].
<http://www.dimi.uniud.it/mizzaro/AH2002/proceedings/pdfs/Bmontejo.pdf >

Ng, H. T., W. B. Goh, and K. L. Low. 1997. "Feature selection, perceptron learning, and a usability case study for text categorization." In: *Proceedings of SIGIR '97*: 67-73.

Pouliquen, Bruno, Ralf Steinberger, and Camelia Ignat. 2003. "Automatic annotation of multilingual text collections with a conceptual thesaurus". In: *Proceedings of the workshop 'Ontologies and Information Extraction' at the Summer School 'The Semantic Web and Language Technology' (EUROLAN 2003), Bucharest*.

Rogati, M., and Y. Yang. 2002. "High-performing feature selection for text classification." In: *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management*: 659-661.

Ruiz, Miguel E., and Padmini Srinivasan. 2002. "Hierarchical text categorization using neural networks." *Information Retrieval,* 5(10): 87-118.

Schapire, R. E., Y. Singer, and A. Singhal. 1998. "Boosting and rocchio applied to text filtering." In: *Proceedings of SIGIR '98*: 215-223.

Schapire, R. E., and Y. Singer. 2000. "BoosTexter: A boosting-based system for text categorization." *Machine Learning*, 39(2/3): 135-168.

Schütze, H., D. A. Hull, and J. O. Pedersen. 1995. "A comparison of classifiers and document representations for the routing problem." In: *Proceedings of the SIGIR '95*: 229-237.

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

Steinberger, Ralf, Bruno Pouliquen, and Johan Hagman. 2002. "Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC." In: A. Gelbukh ed. *Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing 2002*: 415-424.

Steinberger, Ralf, Johan Hagman, and Stefan Scheer. 2000. "Using thesauri for automatic

indexing and for the visualisation of multilingual document collections." In: *Proceedings of the workshop on Ontologies and lexical knowledge bases (Ontolex, 2000)*, Sozopol, Bulgaria, pp. 130-141.

Steinberger, Ralf. 2001. "Cross-lingual keyword assignment." In: *Proceedings of the SEPLN 2001*: 273-280.

Weigend, A. S., Wiener, E. D., and Pedersen, J. O. 1999. "Exploiting hierarchy in text categorization." *Information Retrieval*, 1(3): 193-216.

Wiener, E. D., Pedersen, J. O. and Weigend, A. S. 1995. "A neural network approach to topic spotting." In: *Proceedings of SDAIR 1995, 4th Annual Symposium on*

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