디스크립터 프로파일을 사용한 통제어휘 자동색인

Automatic Indexing with Controlled Vocabulary Using a Descriptor Profile

김판준, 연세대학교 {dpbluesea@hanmail.net}

Pan-Jun Kim, Yonsei University

통제어휘를 사용하는 주제색인 작업에서 색인전문가를 효율적으로 지원할 수 있는 자동색인 방법으로 프로파일 방법의 성능과 특성을 검토해 보았다. 자동색인의 성능에 영향을 미치는 주요 요인들을 검토한 다음, 동일한 조건 하에서 프로파일 기반 방법과 다른 방법들(NB, SVM, VPT)의 성능을 비교하였다. 그 결과, 로치오 알고리즘에 기초한 프로파일을 사용하는 방법이 다른 방법들에 비해 저성능이라는 일부 평가를 일반화하기는 어렵다는 사실이 실험을 통해 드러났다. 또한, 후보 디스크립터 리스트의 생성을 통하여 색인전문가의 색인작업을 지원하는 반자동색인의 경우, F_1척도로는 SVM, VPT와 동등한 수준에 있으면서 재현율이 상대적으로 높은 수준인 프로파일 기반 방법을 우선적으로 고려해 볼 수 있을 것이다.

1. 서론

주제색인의 기본적인 목적은 특정 대상(문헌)의 내용 또는 주제를 표현하는 것으로, 이용자가 원할 때 찾을 수 있도록 그 문헌을 적절하게 표현할 수 있는 색인어나 분류기호를 부여한다. 색인어로는 주제명표목, 디스크립터 등의 통제색인어가 사용될 수도 있고, 본문에 출현한 표현(자연언어) 그대로의 비통제색인어를 사용할 수 있다.

통제어휘를 사용하는 주제색인은 일반적으로 인간의 사고작용과 함께 수작업으로 이루어지는 것으로 받아들여져 왔기 때문에, 통제어휘를 사용하는 주제색인을 컴퓨터를 통하여 처리하기 위한 시도는 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 나날이 증가하고 있어, 종전의 수작업 방식 그대로는 적절한 색인작업이 거의 불가능한 상황에 이르고 있다. 또한, 정보의 생산자 및 소비자로서 이용자 측면에서는 찾아야 할 대상이 너무나 많은 것은 말할 것도 없고 이미 찾은 정보(예를 들면, 정보검색시스템의 검색결과)조차 너무 많아서, 그 중에서 자신이 필요로 하는 정보를 적절하게 판정하기조차 어려운 상황에 처해있다. 이러한 상황에서 색인전문가 측면에서는 보다 효율적으로 많은 양의 문헌을 빠른 시간 내에 일관성 있게 색인하고, 이용자 측면에서는 원하는 정보를 주제 또는 개념 측면에서 접근할 수 있도록 하기 위한 통제어휘 자동색인에 대한 필요성이 커지고 있다.

본 연구에서는 지금까지 대부분 인간에 의해 전적으로 수작업으로 수행되어 온 통제어휘를 사용하는 색인작업에서, 색인전문가를 효율적으로 지원할 수 있는 자동색인 방법을 모색하여 보았다. 이를 위한 실제적인 방법 중의 하나로 정보검색 분야에서 많이 사용되고 있는

벡터공간 모형과 로치오 알고리즘을 기초로 하는 통제색인어 프로파일의 생성 및 입력문헌과의 매칭에 의한 자동색인 방법의 가능성을 알아보고자 하였다. 특히, 정보학 분야 학술지 논문을 대상으로 대표적인 통제색인어인 디스크립터를 자동부여하는 실험을 통하여, 프로파일 기반 통제어휘 자동색인 방법의 실제적인 성능 및 특성을 알아보았다.

2. 통제어휘 자동색인

2.1 통제어휘 자동색인과 텍스트 범주화

통제어휘 자동색인은 컴퓨터를 활용하여 주제명표목 또는 디스크립터를 문헌에 색인어로 부여하는 것이다. 이에 반해, 문헌의 일부 또는 전문에 출현한 대부분의 단어들을 일정한 절차를 거쳐 키워드로 추출하는 자동색인은 대부분 비통제어휘 자동색인이라 할 수 있다. 이들 두 가지 유형의 자동색인은 특정 문헌의 주제를 표현하기 위한 색인어의 부여 과정에서 컴퓨터를 활용하는 것은 동일하지만, 색인어의 통제 여부와 인간 전문가의 참여 정도에 따른 차이가 있다. 데이터베이스 환경에서 비통제어휘 자동색인은 색인어를 전혀 또는 거의 통제하지 않으면서 인간의 개입이 없이 전적으로 컴퓨터에 의해 색인어가 부여되는 경우가 많다. 그러나, 통제어휘 자동색인은 색인어가 엄격하게 통제되고 있는 상황에서, 전적으로 컴퓨터에 의해 색인어가 부여되는 경우와 컴퓨터를 활용하여 색인전문가의 효율적인 색인작업을 지원하는 경우로 구분할 수 있다. 특히, 후자의 경우는 인간의 사고작용까지 포함하는 모든 것을 컴퓨터가 완전하게 대체하는 것이 아니라 인간의 최종 판단을 효율적으로 지원하는 것을 목적으로 한다는 측면에서 '컴퓨터 지원 색인(computer assisted indexing)' 또는 '반자동색인(semi-automatic indexing)'이라고도 한다.

문헌의 자동분류는 특정 알고리즘에 의해 유사한 문헌들을 집단화하는 것으로, 사전 분류체계가 없이 문헌을 집단화하는 문헌 클러스터링과 사전 분류체계에 기초하여 범주들을 문헌에 배정하는 텍스트 범주화로 구분할 수 있다. 여기서 유사한 문헌들을 집단화하기 위해 사전분류체계의 범주들을 문헌에 자동부여하는 텍스트 범주화는, 디스크립터를 사전 분류체계의 범주로 간주한다면 통제어휘를 사용하는 자동색인과 거의 차이가 없다. 따라서 통제어휘 자동색인과 텍스트 범주화는 기본적으로 통제어휘(사전 분류체계)를 사용하여 문헌의 내용을 표현하고, 그 결과로 문헌을 집단화한다는 측면에서 거의 동일한 목적과 절차를 공유한다. 또한, 출력 측면에서 통제어휘 자동색인은 문헌의 내용을 표현하기 위한 것으로 결과가 문헌에 대한 통제색인어의 부여로 나타나고, 텍스트 범주화는 유사한 문헌을 집단화하기 위한 것으로 그 결과가 사전 분류체계 내 범주의 배정이 된다.

2.2 프로파일 기반의 통제어휘 자동색인

프로파일 기반의 통제어휘 자동색인 방법은 기본적으로 벡터공간 모형에 기초하는 통제색인어 프로파일을 생성하고, 이러한 프로파일과 문헌 벡터 간의 유사도 매칭을 통하여 문헌에 색인어를 부여하는 방법이다. 이러한 프로파일은 각 디스크립터를 표현하는 것으로 프로토타입, 센트로이드 또는 전형문서(Propotypical document)라고 부르기도 하는데, 특정 디스크립터의 부여 여부에 따라 구분된 문헌집합에서의 용어 출현 정보에 따라 생성된다.

프로파일의 생성은 대부분 적합성 피드백에 의한 질의확장에서 사용되었던 로치오 알고리즘에 기초한다. 따라서 특정 디스크립터가 부여된 문헌들을 긍정예제로, 부여되지 않은 문헌들을 부정예제로 취급하여 기본적으로 다음

과 같은 공식을 사용한다(Ittner et al. 1995).

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

여기서 R_c는 긍정예제의 수, $\overline{R_c}$는 부정예제의 수이다. 이때, 대부분의 디스크립터가 비교적 작은 수의 긍정예제와 상당히 큰 수의 부정예제를 가지는 경우가 많으므로, 긍정예제와 부정예제에 서로 다른 파라미터 값을 곱해주어 균형을 맞추거나 긍정예제만을 사용하여 프로파일을 생성한다. 예를 들면, 전자의 경우에는 β값(16)을 γ값(4)보다 크게 하여 긍정예제의 상대적 영향력을 높였고(Ittner et al. 1995), 후자의 경우에는 β은 1, γ값은 0으로 하여 부정예제를 제외한 긍정예제만을 사용하였다(Hull 1994). 또한, 복수언어(multilingual) 자동색인을 위한 연구에서 Ferber(1997)는 디스크립터와 용어의 동시출현 정보에 기초하여 프로파일을 생성하고 연관 가중치(association weight)를 부여한 다음, 프로파일과 입력문헌간의 내적(inner product)을 산출하여 OECD 시소러스의 디스크립터를 문헌에 부여하였다. 최근 유럽연합의 공식 문헌들에 대하여 복수 언어로 통제색인어를 부여하는 일련의 연구들(Steinberger et al. 2002; Pouliquen et al. 2003)에서도 EUROVOC 시소러스의 디스크립터를 자동부여하기 위하여, 특정 디스크립터와 연관 용어들(associates) 간의 동시출현 정보에 기초한 프로파일을 생성하여 사용하였다.

이와는 달리, 텍스트 범주화를 위한 연구들에서는 로치오 알고리즘의 여러 장점에도 불구하고 성능 측면에서 다소 떨어진다는 평가에 따라, 다른 분류기(classifier)와의 성능 비교를 위한 기본형으로 사용하는 경우가 많았다(Lewis et al. 1996; Rogati and Yang 2002). 그러나, 다른 알고리즘(k-NN)과의 성능 비교에서 자질의 수가 크게 축소되는 경우를 제외하고는 로치오 알고리즘의 성능이 거의 동등하거나 일부 더 높게 나타난 결과가 있었다(Galavotti et al. 2000). 또한, 고빈도 범주(≥300)를 대상으로 한 성능에서 지지벡터기계(SVM), k-최근접이웃(k-NN), 결정트리, 로치오 알고리즘, 나이브 베이즈(NB)가 모두 유사한 성능을 보이는 것으로 보고된 바도 있었다(Joachims 1998). 따라서, 텍스트 범주화 분야에서 로치오 알고리즘에 기초한 프로파일을 사용하는 방법의 성능이 다른 방법과 비교하여 저성능이라는 평가를 그대로 일반화하기에는 문제가 있음을 알 수 있다.

3. 실험 구성

3.1 문헌집단 및 사전처리

실험에 사용된 문헌집단은 Journal of Citation Reports(2001년~2004년)의 데이터에 기초하여 문헌정보학 분야의 핵심 학술지들을 선정하였다. 또한, 실제적인 데이터 수집은 Cambridge Scientific Abstracts(CSA)사의 CSA Illumina 서비스에서 제공하는 LISA 데이터베이스에서, 1994년부터 2004년까지 3개 학술지(IPM, JASIST, JD)에 수록된 논문들을 검색한 결과를 다운로드하여 <표 1>과 같이 실험집단을 구성하였다.

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.

<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

디스크립터는 학습집합 내 문헌빈도가 최소 15이상인 디스크립터 75개 중에서 임의 선정한 33개를 사용하였다. 또한, 문헌에 디스크립터를 부여하기 위한 단서가 되는 자질집합은 각 문헌의 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 Visual FoxPro로 구현된 프로그램과 Excel을 사용하였다.

3.2 디스크립터 프로파일 및 문헌벡터

로치오 알고리즘에 기초한 프로파일은 학술지 논문에 특정 디스크립터를 부여하기 위한 목적으로 생성되는데, 이때 문헌에 부여되는 디스크립터를 특정 주제를 표현하는 주제 범주라고 한다면 문헌의 자동분류에서 사용되는 분류기(classifer)와 유사한 역할을 한다. 본 연구에서 디스크립터 프로파일($\vec{T}$)은 정보검색에서 일반적으로 사용되는 벡터공간 모형을 기반으로 특정 디스크립터가 부여되었는지의 여부에 따라 문헌에 출현한 용어들을 중심으로 생성된다.

$$\vec{T} = (w_1, w_2, w_3, \cdots, w_m)$$

로치오 공식에서 긍정사례만을 사용하여 가중치합을 산출하는 변형 공식은 다음과 같으며 여기서 w_k는 자질 가중치로서 특정 디스크립터가 부여된 문헌들에 출현한 k번째 단어의 가중치이다.

$$w_k = \sum_{d \in D} d_k$$

디스크립터 프로파일 벡터에 기초한 자동색인 실험에서 기본형(baseline)으로 사용되는 가중치합 프로파일(PV_b)의 기본 가중치는 문헌 내 출현빈도(tf)를 사용한다. 여기서 디스크립터 프로파일은 하나의 디스크립터가 여러 문헌에 부여되어 있는 상황에서 생성되는 것이므로 기본형의 가중치는 결과적으로 해당 디스크립터가 부여된 문헌 내 출현빈도의 합이 된다.

$$PV\_b = (wt_1, wt_2, wt_3, \cdots, wt_m),\ wt_i = \sum_i tf$$

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

$$\vec{d} = (w_1, w_2, w_3, \cdots, w_m),\ wt_{w_i} = tf$$

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

3.3 자질선정 및 프로파일 생성방법에 따른 실험

디스크립터 자동부여에서 성능에 영향을 미치는 요인으로 자질선정 기준 및 가중치부여 방법에 따른 성능을 비교하는 실험을 수행하였다. 먼저, 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 비교하였고, 여기서 좋은 성능을 보인 자질선정 기준에 따라 선정된 자질들로 전체 자질집합을 25% 수준(1,611개)으로 축소하였다.

다음으로, 전체 33개 디스크립터와 선정된 자질집합을 대상으로 가중치합, 가중치 평균, 가중치 최대값의 3가지 프로파일 생성방법을 적용하여, 가장 좋은 성능을 보이는 방법을 이후의 실험을 위한 기본적인 방법으로 사용하였다. 여기서 가중치합을 사용한 방법은 앞에서 언급한 디스크립터 프로파일의 기본형(PV_b)과 동일하며, 가중치 평균(PV_avg)을 사용하는 방법은 로치오 알고리즘에서 긍정사례만을 사용하는 센트로이드와 동일하다. 한편, 가중치 최대값(PV_max)은 해당 디스크립터가 부여된 문헌들의 용어빈도(tf) 중에서 가장 큰 값을 사용하는 것이다.

$$PV\_avg=(ut_1, ut_2, ut_3, \cdots, ut_m), ut_i=\frac{1}{N_T}\sum_i tf$$

$$PV\_\max=(ut_1, ut_2, ut_3, \cdots, ut_m),\ ut_i=\max tf$$

또한, 다른 자동색인 방법들과 비교하여 로치오 알고리즘에 기초한 프로파일 방법이 실제로 저성능인 가를 알아보기 위한 실험을 수행하였다. 이를 위해 자카드계수를 자질선정 기준으로 자질집합을 구성하고 용어빈도(tf)를 기본 가중치로 부여한 다음, 가중치합(Sum_tf)을 프로파일 생성방법으로 적용한 프로파일 기본형(PV_b)과 이와 동일한 자질선정 기준 및 가중치를 적용한 다른 자동색인 방법들(나이브 베이즈/NB, 지지벡터기계/SVM, 투표형 퍼셉트론)의 성능을 비교하였다.

3.4 학습집합의 크기에 따른 실험

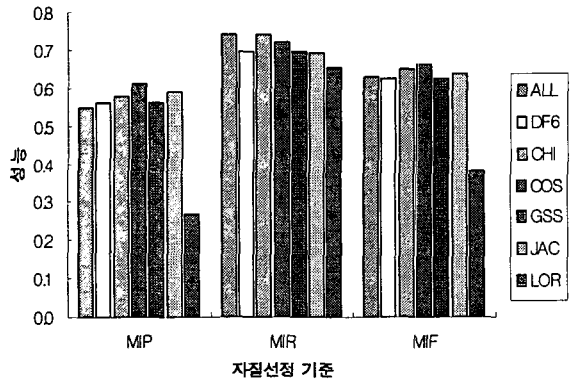
디스크립터 프로파일을 사용한 자동색인의 성능에서 또 다른 주요 영향 요인이 되는 학습집합의 크기에 따른 성능을 비교하였다. 선행연구들에서는 주로 학습집합 내 긍정사례의 수 또는 비율을 일정하게 증가시켜 학습집합의 크기를 조정하였지만, 본 연구에서는 전체 10년 동안의 학습집합(1994년~2003년)을 출판년도에 따라 가장 최근의 연도(2003년)부터 1년씩 추가하여 학습집합의 양을 증가하는 방식으로 학습집합의 크기를 변화시켰다. 그 이유는 대규모 학술 데이터베이스 환경에서 매년 새롭게 입력되는 많은 양의 학술지 논문들에 디스크립터를 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.

4. 실험결과 및 분석

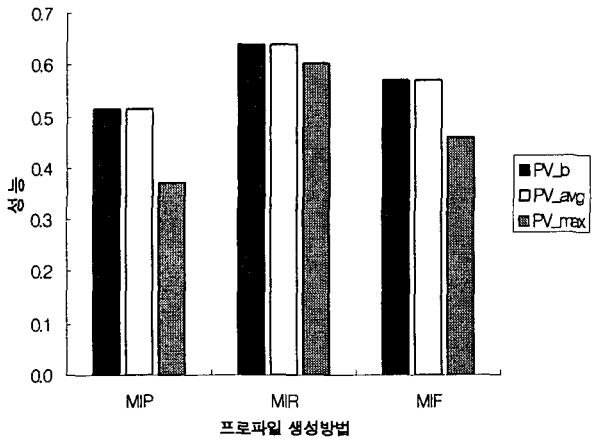
4.1 자질선정 기준 및 프로파일 생성방법에 따른 결과 및 분석

<그림 1>은 여러 자질선정 기준에 따른 디스크립터 자동부여 실험의 결과를 마이크로 평균 정확률(MIP), 마이크로 평균 재현율(MIR), 마이크로 평균 F_1(MIF) 척도로 나타낸 것이다. 전체 자질을 사용한 것(ALL)보다 상위 세 가지 자질선정 기준(CHI, COS, JAC)이 거의 유

사한 수준으로 성능이 향상된 결과를 보였지만, 이후의 실험에서 다른 방법들과의 비교를 위하여 자카드 계수(JAC)를 기본적인 자질선정 기준으로 적용하여 자질집합을 구성하였다(김판준 2006).



<그림 1> 자질선정 기준에 따른 성능

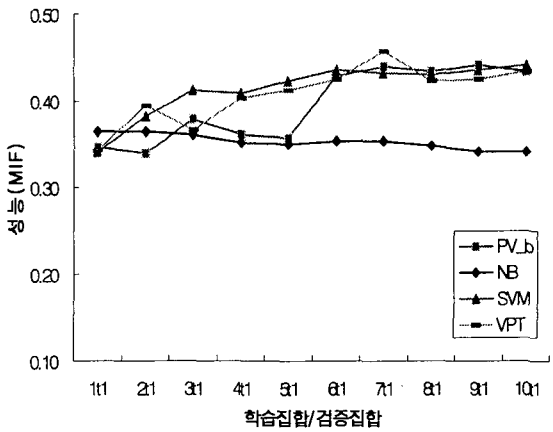


<그림 2> 프로파일 생성방법에 따른 성능

다음으로, 자카드 계수를 자질선정 기준으로 적용하여 25% 수준으로 축소된 자질집합을 사용하면서, 세 가지 프로파일의 생성방법을 적용하여 성능을 산출하였다. <그림 2>는 세 가지 방법을 적용하여 생성한 프로파일을 사용하여 디스크립터를 부여한 결과이다. 그 결과 PV_b과 PV_avg의 성능은 동일한데 반하여, PV_max는 성능은 크게 떨어지는 것으로 나타났다. 따라서 이후의 실험에서는 상대적으로 컴퓨터 처리상의 이점이 있는 가중치합을 적용하여 생성한 프로파일 벡터를 기본형(baseline)으로 사용하였다.

4.2 학습집합의 크기에 따른 결과 및 분석

<그림 3>은 동일한 환경(실험집단, 자질선정 기준, 가중치, 학습집합의 크기)을 적용한 다른 세 가지 방법(나이브 베이즈/NB, 지지벡터기계/SVM, 투표형 퍼셉트론/VPT)의 결과와 디스크립터 프로파일 기본형(PV_b)의 결과를 함께 나타낸 것이다.



<그림 3> 학습집합의 변화에 따른 성능: 프로파일 벡터 기본형(PV_b)과 다른 세 가지 방법 간의 비교(마이크로 평균 F_1/MIF)

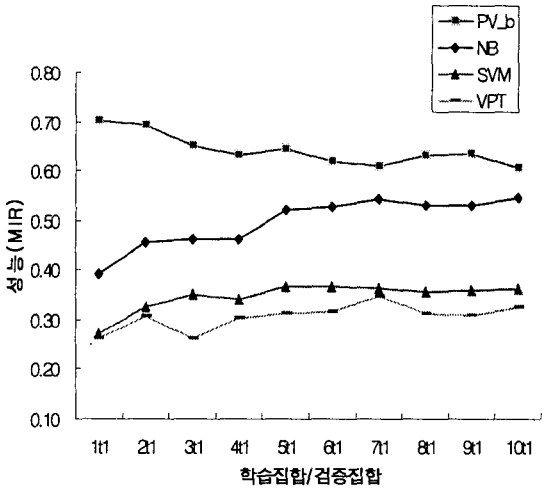
전체 10년의 학습집합(10t1)을 모두 사용하는 경우, 디스크립터 프로파일의 기본형(PV_b)은 선행연구들에서 문헌의 자동분류에서 가장 좋은 성능을 보이는 것으로 알려진 지지벡터기계(SVM)와 거의 유사한 성능(0.441 vs. 0.434)을 보여주었다. 또한, 지지벡터기계와 동등하거

나 약간 떨어지는 성능을 가지는 것으로 알려진 투표형 퍼셉트론(VPT)과는 거의 동일한 성능(0.435 vs. 0.434)이었다. 한편, 학습집합의 크기를 연차적으로 증가시키는 경우(1t1~10t1)에는 지지벡터기계(SVM)는 3년 이후부터 지속적으로 성능이 안정되는 추세를 보이는 반면, 프로파일 기본형(PV_b)은 6년 이상의 학습집합을 사용한 이후부터 안정된 성능을 보여주었다. 이러한 실험 결과는, 자동색인의 성능에 영향을 주는 다른 요인들을 동일한 조건으로 하고 6년 이상의 학습집합을 사용하는 경우, 프로파일 방법(PV_b)은 지금까지 좋은 성능을 보이는 것으로 보고되어온 다른 방법들(SVM, VPT)과 동등한 성능 수준을 유지한다는 것을 보여주고 있다. 따라서 통제어휘 자동색인 또는 텍스트 범주화에서 로치오 알고리즘의 성능이 다른 방법들보다 저성능이라는 평가는 재고되어야 할 것이다.

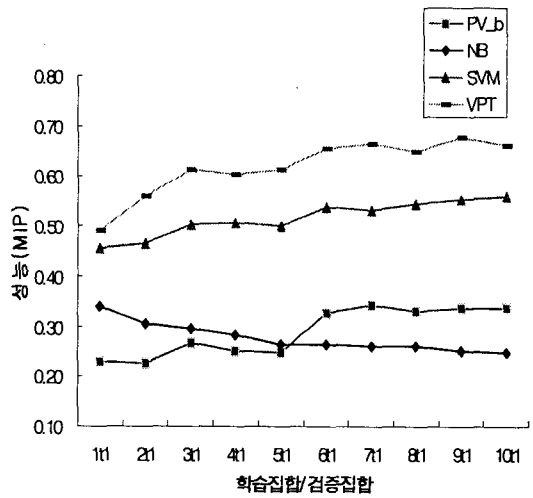
4.3 재현율과 정확률로 본 결과 및 분석

자동색인의 결과에 대한 평가 측면에서, 대개 완전 자동색인은 색인전문가의 개입이 없이 컴퓨터 알고리즘에 의해 색인어가 직접 부여되기 때문에 보다 정확한 결과를 위하여 정확률이 더 중요시되고, 반자동색인은 컴퓨터에 의해 후보 색인어 리스트가 생성되지만 최종 부여 결정은 색인전문가에 의해 이루어지는 관계로 재현율이 더 중요시되는 경향이 있다.

<그림 4>는 프로파일 기본형(PV_b)을 비롯한 여러 자동색인 방법(NB, SVM, VPT)의 성능을 마이크로 평균 재현율(MIR)로 나타낸 것이다. 여기서 학습집합의 크기에 상관없이, PV_b(0.61)의 재현율이 다른 세 가지 방법 보다 상당히 높게 나타난다는 것을 알 수 있다. 이는 일반적으로 다른 방법들보다 비교적 높은 재현율을 나타내는 것으로 알려진 NB(0.55)보다도 높은 수준인 것은 물론, 정확률에서 더 강세를 보이는 것으로 알려진 SVM(0.36)과 VPT(0.32)과는 상당히 큰 차이가 나는 것이다.



<그림 4> 학습집합의 변화에 따른 성능: 프로파일 기본형(PV_b)과 다른 3가지 방법 간의 비교(마이크로 평균 재현율/MIR)



<그림 5> 학습집합의 변화에 따른 성능: 프로파일 기본형(PV_b)과 다른 세 가지 방법 간의 비교(마이크로 평균 정확률/MIP)

그러나, 마이크로 평균 정확률(MIP) 측면에서는 전혀 반대의 결과가 나타났다. <그림 5>

에서와 같이 VPT(0.66)가 가장 높은 성능을 나타냈고, 그 다음은 SVM(0.56), PV_b(0.34), NB(0.25)의 순이었다. 따라서, 색인전문가를 지원하기 위한 목적으로 사용하기 위한 자동색인 방법으로는 F_1척도로는 SVM, VPT와 동등한 수준에 있으면서, 상대적으로 재현율이 가장 높은 프로파일 기반 방법을 사용하는 것을 우선적으로 고려할 수 있을 것이다.

5. 결론

통제어휘를 사용하는 주제색인 작업에서 색인전문가를 효율적으로 지원할 수 있는 자동색인 방법으로 프로파일 기반 방법의 성능과 특성을 검토해 보았다. 특히, 로치오 알고리즘에 기초한 프로파일을 사용하는 방법이 다른 방법들에 비해 저성능이라는 일부의 평가는 재고의 여지가 있다는 사실을 실험을 통해 밝혔다. 또한, 색인전문가의 색인작업을 지원하는 반자동 색인의 경우, F_1척도로는 SVM, VPT와 동등한 수준에 있으면서 재현율이 다른 방법들에 비해서 상당히 높은 수준인 프로파일 기반 방법을 우선적으로 고려해 볼 수 있다.

참고문헌

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

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

Galavotti, L., F. Sebastiani, and M. 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.

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

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

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)*.

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.

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.