장르와 주제 범주간 용어 편차정보를 이용한 디지털 문서의 장르기반 분류

(A Genre-based Classification of Digital Documents by using Deviation Statistic of Genre-revealing Term and Subject-revealing Term)

이 용 배 † 맹 성 현 ††
(Yong-Bae Lee) (Sung Hyon Myaeng)

요 약 장르기반 분류는 문서를 내용이나 주제가 아닌 문서의 형식 또는 스타일에 의해 분류하는 것을 의미한다. 현재 장르분류 방법은 기존의 주제기반 분류방법에 사용되었던 알고리즘을 그대로 이용하거나 자질선택 방법에 있어서도 효과적이지 못하고 비교적 단순하여 분류 정확률 또한 상대적으로 낮았다. 본 연구에서는 장르기반으로 문서를 자동 분류할 수 있는 새로운 방법론을 제시한다. 장르분류 방법은 크게 두 가지 정보를 이용하여 학습과 분류를 하는데 장르 간 용어의 편차정보와 장르 내에 분포되어 있는 주제 범주 간 용어의 편차정보를 이용한다. 제안된 방법의 성능을 측정하기 위해 인터넷상에서 정제되지 않은 문서를 수집하였으며 이를 대상으로 실험한 결과 기존의 카이제곱 자질선택 방법 및 베이지안 분류 알고리즘과 비교하여 약 30% 정도 우수한 정확도를 나타내었다.

키워드 : 장르, 자동분류, 자질선택

Abstract A genre-based classification means classifying documents by the purpose for which they were written, not by the semantics or subject areas. Most genre classifying methods in the past were based on the existing documents categorization algorithms and ineffective for feature selections, resulting in low quality classification results. In this research, we propose a new method for automatic classification of digital documents by genre. The genre classifier we developed uses the deviation statistic between the genre-revealing term frequencies and between the subject-revealing term frequencies within a genre. We collected Web documents to evaluate the proposed genre classification method. The experimental results show that the proposed method outperforms a direct application of a kai-square feature selection and bayesian classifier often used for subject classification by proving an excellent accuracy of about 30 percent.

Key words : Genre, Automatic Classification, Feature Selection

1. 서 론

주제별 분류(content-based classification)란 문서를 사회, 과학, 문화, 스포츠 등과 같이 문서의 주제 또는 의미에 따라 분류하는 것을 뜻하며 장르별 분류(genre-based classification)란 사건기사, 보고서, 논문, 판결문 등과 같이 문서의 형식이나 문체에 따라 분류하는 것을 의미한다.

현재 정보의 보고라고 불리는 인터넷에서 문서의 양이 무한정 증가하므로 문서관리나 검색효율 측면에서 두가지 중요한 문제가 발생하고 있다. 첫째는 웹 상에서 임의의 문서를 검색하기 위해 질의를 하면 검색결과가 수천에서 수백만 건에 달하므로 사용자들은 검색결과 내에서도 자신이 원하는 장르의 문서를 찾기조차 어렵다는 것이며 둘째는 인터넷 상에서 이처럼 기하급수적으로 증가하는 문서들의 효율적인 관리는 단순히 하드웨어 용량을 늘린다고 해결되지 않는다는 것이다.

위에서 제기되고 있는 문서의 검색성능이나 문서관리

· 본 연구는 한국과학재단지정 소프트웨어 연구센터의 기본 프로그램 연구비지원에 의해서 수행된 연구결과임
† 비 회 원 : 전주교육대학교 컴퓨터교육과 교수
yblee98@hotmail.com
†† 종신회원 : 한국정보통신대학교 교수
shmyaeng@cs.cnu.ac.kr
논문접수 : 2003년 5월 9일
심사완료 : 2003년 8월 12일

측면에서의 문제들은 장르기반 분류기를 이용하여 쉽게 해결할 수 있으며 현재 국외의 소수 연구팀에서 90년대 중반부터 미온적인 연구[1-9]가 진행되어 왔으나 실제 대량의 디지털 문서로 확장하여 적용하기에는 학습방법이나 분류방법에 있어 풀어야 할 과제가 많이 남아있다.

본 논문에서 장르기반으로 디지털 문서를 자동학습하고 분류할 수 있는 방법론을 제시한다. 제안하는 장르기반 자동분류 방법은 다음과 같은 가정하에 시작된다. "특정장르를 대표하는 용어들은 그 장르 내에서의 빈도수가 높으며 다른 장르에서는 상대적으로 빈도수가 낮을 것이다. 또한 장르 내에 주제별 범주가 존재한다면 특정 장르를 대표하는 용어들은 그 장르 안의 여러 주제별 범주에 걸쳐 골고루 높은 빈도로 분포되어 있을 것이다." 라는 가정하에 장르간 용어의 분포정보와 장르내에서 주제 범주간 용어의 분포정보를 함께 이용한다.

본 논문의 구성은 다음과 같다.

2장에서는 현재까지 장르기반 분류와 관련된 연구 결과들을 분석해 보고 이에 대한 문제점이나 한계에 대하여 분석하여 본 다음, 3장에서는 본 논문에서 장르기반 분류를 위해 제안하고 있는 방법론에 대하여 상세히 기술한다. 4장에서 실험환경 구축에 대하여 설명하고 5장에서는 수집된 문서들을 대상으로 논문에서 제시한 방법을 기반으로 다각적인 측면에서 실험한 결과를 분석하여 본다. 마지막 6장에서는 본 논문에서 시도한 연구가 갖는 의의와 실험결과에 대해 정리하고 연구과정에서 미진하여 앞으로 좀 더 연구의 필요성이 있는 부분에 대하여 기술한다.

2. 관련 연구

장르기반 분류는 기존의 주제별 문서분류와는 분류기준이 다르다. 주제별 분류가 문서의 내용과 토픽을 기준으로 분류하는 것이라면 장르기반 분류는 문서의 스타일이나 문체에 의해 분류하는 것을 의미한다. 분류기준이 다르므로 학습방법이나 분류방법 또한 판이할 수 있다. 따라서 장르기반의 새로운 분석방법 및 학습방법과 분류 알고리즘은 필수적이고 당면한 과제이다.

장르기반 분류방법에 대하여 비교적 소수의 연구지만 현재까지의 연구결과를 다음의 각 절에서 분석해 본다.

2.1 디지털 장르의 조사 및 분석

인터넷의 확산으로 온라인상의 문서가 팽창하면서 문서 장르 체계는 오프라인상의 장르 체계와는 다를 뿐 아니라 훨씬 더 다양해졌다. 이에 따라 웹상에서의 장르 체계 정립에 대한 연구뿐만 아니라 새로운 장르분석 방법에 대한 연구가 함께 병행되어 왔다.

Dillon의 연구[1]는 새로운 디지털 장르로서 개인 홈페이지(personal home page)를 정의하고 100개 정도의 문서를 수집하여 개인 홈페이지 내부의 구성 요소들과 내용을 세부적으로 분석해 놓았으며 Haas의 연구[2,3]에서는 75개의 웹 문서에서 나타나는 링크형태와 앵커의 역할 등을 분석하고 웹 장르를 Organizational, Documentation, Text, Home page, Multimedia, Tool, Database entry의 7가지로 분류하였다. 그러나 위의 연구들은 비교적 적은 문서를 대상으로 분석하였으며 문서수집 방법이 체계화되지 못하였다는 단점이 있었다.

비교적 체계적이고 정돈된 웹 장르의 정의[4]는 DropJaw 프로젝트의 연구결과로서 여러 인터넷 사용자들 648명(스톡홀름 대학의 학생, 연구원, 교수들)을 대상으로 웹상에서의 장르라고 판단되는 문서 종류가 어떤 것들이 있는지 전자우편으로 설문조사 하고 그 중 67명으로부터 답변을 받아 종합하고 정리한 결과 표 1과 같이 11가지의 장르로 정의하였다. 이는 웹상의 장르를 정의하는 과정에서 사용자의 피드백을 기반으로 하여 장르표를 정립하였으므로 장르 정의 결과에 대한 객관적인 판단기준을 수립할 수 있었다는 의의를 갖는다.

표 1 웹 장르 정의 결과

	Genre	Sample
1	Informal, Private (Persona home pages)	128
2	Public, commercial (Home pages for the general public)	197
3	Searchable indices (Pages with feed-back : customer dialogue; searchable indexes)	73
4	Journalistic materials (Press: news, reportage, editorials, reviews, popular reporting, e-zines)	94
5	Reports (Scientific, legal, and public materials; formal text)	113
6	Other running text	160
7	FAQs	12
8	Link Collections	148
9	Other listings and tables	225
10	Asynchronous multi-party correspondence (Contributions to discussions, requests, comments; Usenet News materials)	24
11	Error Messages	184
Total		1,358

2.2 장르분류 방법과 한계

• Karlgren의 연구

Biber의 텍스트의 다변량 통계분석 방법[5]을 확장한 Karlgren의 연구[6,7]에서는 인칭대명사, 강조표현, 부드러운 표현 및 숫자, 단어의 평균길이 등을 분류를 위한 자질(feature)로서 이용하며 여러 개의 자질들을 조합하여 장르 분류함수를 만들어 냈다. 이 연구에서는 말뭉치가 장르별로 비교적 정리가 잘 되어있는 브라운 말뭉치(Brown Corpus)[10]를 4개 장르로 나누어 실험했을 때 정확률은 0.73을 나타내었으며 15개 장르로 나누어 실험했을 때의 정확률은 0.52를 나타내었다. 또한 표 1의 웹 문서 11개 장르로 성능실험을 했을 때에는 0.67~0.75의 정확률을 가졌다.

```
if
    - there are more because than average,
    - longer words than averge,
    - type-token ratio is above average,
then
    - the object is of class Textual with a certainty of 90%
```

그림 1 Karlgren의 장르 분류함수

Karlgren의 연구에서 학습으로 이용한 자질들은 많은 반면 실제 분류에서 사용한 자질들은 그림 1과 같이 적었으며 분류함수도 자질의 통계치를 확률함수가 아닌 규칙함수로 사용하였으므로 분류정확률도 상대적으로 낮았다.

• Stamatatos의 연구

장르기반 분류라기보다는 하나의 장르 내에서 여러 사람들이 각각 어떠한 문체로 글을 쓰는지를 분석한 Stamatatos의 연구[8]에서는 문서처리 도구를 이용하여 문서에서 문장개수, 토큰개수, 특수기호개수, 구개수, 단어개수 등의 자질을 추출한 후, 자질 빈도수를 선형 조합하여 분류함수를 만들었다.

분류실험을 위해 그리스의 주간신문 TO BHMA[9]의 기자 10명을 대상으로 기자별로 20개씩의 기사를 발췌하여 절반은 학습을 하고 나머지 절반을 이용하여 실험하였다. 실험결과 정확률은 최고 0.7을 넘지 못하였지만 저자의 글을 쓰는 스타일에 따라 문서를 분류하는 최초의 시도였다는 데 의의가 있다.

• Kessler의 연구

Kessler의 연구[10]에서는 장르 분류를 위한 자질 종류를 표 2와 같이 문서의 구조, 단어, 문자로 구분하고 언어처리를 통해 자질을 추출한 후 문서분석을 시도했다.

Kessler의 장르분류 함수는 표 2의 자질 통계치를 이용하여 이진결정 방법과 유사한 비교적 간단한 방법을 사용하여 구성하였다. 장르분류 실험시에는 브라운 말뭉치[11]를 대상으로 Reportage, Editorial, Scitech, Regal, Nonfiction, Fiction의 6개 장르에 대하여 분류를 시도하였고 분류정확도는 0.65이하로 보고되었다.

표 2 분류에 이용한 자질의 종류

종 류	자 질
구조정보	주제문장, POS태그개수
단어정보	연설문 등에 많이 쓰이는 단어 Mr/Ms, 라틴접사, 과학·학술용어, 뉴스의 날씨 표현에 사용된 단어
문자정보	물음표, 느낌표, '-'포함어, 약어

• 기존연구의 한계

현재까지 진행된 장르 분석 연구에서는 비교적 적은 문서들을 대상으로 실험하였으며 임시적(ad hoc)인 분류 방법을 사용하였으므로 실제 대량의 디지털 문서로 확장하여 적용하기 위해서는 자질선택 방법이나 분류 방법에서 좀 더 세심한 연구가 필요하다.

본 연구에서는 실제 한국 인터넷 사용자가 필요로 하는 장르의 문서들을 수집하여 장르간 용어의 빈도와 장르내 주제 범주별 용어의 빈도를 이용한 자질선택 방법을 제시하며 학습결과인 지식베이스를 활용하여 디지털 문서의 자동분류를 시도하고자 한다.

3. 디지털 문서의 장르기반 분류 방법

3.1 자질의 빈도수를 이용한 장르학습

장르학습 과정은 내용기반 분류의 학습방법과 마찬가지로 크게 전처리 단계, 자질선택(feature selection) 단계, 지식베이스 구축단계로 구성된다. 먼저, 전처리 단계에서는 학습환경에 맞도록 학습문서를 전처리하고 문서에서 중요한 용어들을 추출하는 색인기능을 수행한다. 그 다음 학습과정의 핵심단계인 자질선택 과정을 거치며 그 결과 생성된 용어와 가중치를 빠르게 검색하여 분류에 이용할 수 있는 지식베이스 구축으로 학습과정을 종결한다.

다음에는 장르학습의 핵심과정인 자질선택과정을 단계별로 상세히 기술한다.

• 1단계 : 빈도수 높은 용어추출

자질선택의 첫번째 단계에서는 색인결과의 용어들에서 장르별로 문서출현빈도(document frequency)가 높은 용어들을 추출하고 장르 내에서도 주제별 범주가 존재하면 주제별 범주별로 문서출현빈도가 높은 용어들을 추출하여 빈도수의 내림차순으로 정렬시킨다. 빈도수를 문서출현빈도로 이용하는 이유는 본 논문에서 제안하는 방법이자 연구과정에서 분석한 결과에 의존한다.

표 3은 본 논문의 연구과정에서 수집한 문서 중에서 개인 홈페이지 장르의 한글문서 절반을 대상으로 추출한 용어의 문서빈도 분포를 보여준다. 표 3의 장르와 주

제별 범주에서 괄호 안의 숫자는 전체 문서개수를 의미하고 각 열의 용어와 대응되는 숫자는 용어를 포함한 문서출현빈도를 의미한다. 예를 들어, 표 3에서 학생범주 전체 문서는 120건이며 이 중에서 '취미'이라는 용어를 포함하는 문서는 63건이다.

• 2단계 : 장르별 대표용어 계산

장르별 대표용어 계산 단계에서는 우선, 앞 단계에서 각 장르별로 문서 빈도수가 높은 순으로 추출된 용어들에 대하여 빈도수가 너무 작은 용어들은 일정 비율의 한계값(threshold)을 두어 잘라낸다. 이렇게 잘라낸 각 장르별 문서출현빈도가 높은 상위 N개의 용어들에 대하여 용어가 주제별로 어떻게 분포되어 있는지 편차를 계산한다.

주제별로 빈도수가 높으면서 편차가 작은 용어는 장르를 대표할 수 있는 용어일 확률이 높은 반면 특정 주제에서만 빈도가 높거나 편차가 큰 용어는 상대적으로 장르를 대표할 수 있는 용어가 될 확률이 낮다. 이러한 용어의 문서출현빈도와 주제별 편차가 나타내는 특성을 이용하여 장르를 대표할 수 있는 용어의 가중치를 계산해 낸다. 예를 들어, 표 3의 '학교'라는 용어는 개인 홈페이지 장르에서 빈도수가 가장 높을 뿐 아니라 주제별 범주에서도 높은 빈도로 출현하고 있는 것을 볼 수 있으며 '대학교'라는 용어는 장르 전체에서는 높은 빈도수를 기록하지만 학생이나 연예인의 범주에서는 순위에 들지 못하는 것을 알 수 있다. 이는 '학교'라는 용어가 장르 전체와 주제별로 높은 빈도수를 나타내므로 개인 홈페이지 장르를 대표할 수 있는 확률이 높다는 것을 뜻하며 '대학교'라는 용어는 장르 전체 문서출현빈도는 높지만 장르를 대표하는 용어가 될 확률은 작다는 것을 의미한다.

$$R\_Val_m(t_k) = (1 - \sqrt{\frac{\sum_{i=1}^{n_c} (DF_m(t_k) - DF_m(t_k^i))^2}{n_c}}) \quad (1)$$

(1) 장르별 대표용어 확률값

- $DF_m(t_k)$: 장르 m내에서 용어 t_k의 문서 빈도비율. 예를 들어, 장르 m의 문서 개수는 500이고 용어 t_k의 문서출현빈도수는 100이라면 $DF_m(t_k)$는 0.2(100/500)가 된다. 즉, 문서출현빈도를 전체 장르문서 개수로 표준화(normalization)한 값임
- $DF_m(t_k^i)$: 장르 m안의 주제별 범주 i 내에서 용어 t_k의 문서 빈도비율
- n_c : 장르 m안의 주제별 범주의 개수

(1)에서 $DF_m(t_k)$과 $DF_m(t_k^i)$은 장르별로 또는 주제 범주별 문서의 개수에 따라 문서출현빈도가 상대적으로 많고 적을 수 있으므로 장르별 또는 주제 범주별 전체문서 개수로 나눈 값으로 표준화하여 장르별 대표용어 확률값 $R_Valm(t_k)$계산에 사용한다. 계산된 확률값에 장르 내에서 용어의 가중치 만큼을 곱해준 값이 용어가 특정 장르 m을 대표할 수 있는 최종값인 (2)

표 3 문서출현빈도가 높은 용어

장르		주제별범주							
개인홈페이지(453)		학생(120)		교사/교수(125)		회사원(103)		연예인(105)	
학교	181	이름	78	대학교	98	졸업	46	가족	60
이름	164	취미	63	대학	71	고등학교	39	cm	60
대학교	158	학교	62	현재	60	대학교	36	생년월일	60
생년월일	157	가족	54	대학원	60	현재	34	영화	52
가족	153	생년월일	48	연구	57	서울	34	kg	50
취미	149	소개	46	졸업	56	이름	33	사랑	47
졸업	136	음식	45	한국	50	사람	30	혈액형	45
고등학교	128	특기	44	교육	49	학교	30	학교	45
서울	128	음악	40	mail	48	취미	29	관계	43
현재	126	사람	39	교수	47	생년월일	28	취미	43
대학	123	친구	39	석사	45	중학교	28	데뷔	43
사람	119	홈페이지	37	학교	44	홈페이지	28	출연	42
중학교	107	학년	36	분야	43	디자인	25	드라마	41
mail	106	나이	36	서울	43	대학	25	서울	40
cm	104	중학교	35	고등학교	42	mail	24	사람	37
영화	104	성격	35	박사	42	생활	24	별명	35
여자	102	초등	33	논문	37	소개	24	이름	34
초등	102	여자	32	경과	37	사랑	23	연기	34
홈페이지	101	저의	29	학과	33	가족	22	음식	34
특기	98	혈액형	28	교사	32	근무	22	kbs	31
...		...		...		...		...	
...		...		...		...		...	

$WR_Val_m(t_k)$가 된다. 이 용어들의 값은 학습 과정의 다음 단계에서 장르간을 구분할 수 있는 변별치 계산에 이용된다.

$$WR\_Val_m(t_k) = R\_Val_m(t_k) \times DF_m(t_k) \quad (2)$$

장르 대표용어의 최종값

• 3단계 : 장르간 변별치 계산

여러 장르에 걸쳐 문서출현빈도가 높거나 혹은 동시에 문서출현빈도가 낮은 용어로는 어떤 장르의 문서인지를 판단하기가 어렵다. 특정 장르에만 문서출현빈도가 높거나 낮은 용어만이 장르간을 구분할 수 있는 자질이 될 수 있기 때문이다.

(3)에서는 용어 tk의 장르간 구분 변별치 $D_Val_m(t_k)$을 장르 대표용어 최종값의 장르간 편차로 계산한다. 용어의 편차가 크다는 것은 문서를 판별할 때 이 용어를 이용하면 다른 장르와 구분할 수 있는 확률이 높게 된다는 의미이고 편차가 작다는 것은 이 용어로는 타 장르와 구분할 수 있는 확률이 낮게 된다는 것을 뜻한다.

$$D\_Val_m(t_k) = \sqrt{\frac{\sum_{i=1}^{n_g}(WR\_Val_m(t_k) - WR\_Val_i(t_k))^2}{n_g}} \quad (3)$$

용어의 장르간 변별치 계산

- n_g : 전체 장르의 개수

3.2 장르 분류 모델

장르 학습이 끝나고 새로운 문서에 대한 장르 분류 요구가 있을 경우에는 지식베이스를 이용하여 문서의 장르를 판단하게 된다. 이때 판단하는 방법은 여러 가지가 있을 수 있지만 본 논문에서는 장르 대표벡터와 문서벡터와의 유사도(similarity)를 이용하여 문서 분류를 시도한다.

$$gID = \max_m[sim(G_m, D)] \quad (4)$$

장르 분류 모델

- gID : 장르 식별자
- G_m : 지식베이스에 있는 장르 m의 대표벡터. 즉, 용어-변별치의 결합 테이블
- D : 장르 분류를 하기 위한 문서의 용어벡터

문서 장르 분류 과정은 먼저 지식베이스의 용어-변별치 결합테이블을 장르의 개수만큼 장르 대표벡터로 구성하고 분류하고자 하는 문서에서도 용어를 추출하여 문서 벡터를 구성한다. 다음으로 각 장르의 대표벡터와 문서 벡터와의 유사도를 계산하여 가장 유사한 장르에 문서를 할당하게 된다.

(4)에서 이용하는 유사도란 정보검색의 벡터모델에서 이용하는 질의어벡터와 문서벡터 간의 유사도 계산과 같이 장르대표 벡터 G_m와 분류대상 문서 벡터 D의 내적으로 계산된다.

4. 실험환경 구축

지금도 계속해서 늘어나고 있는 웹 문서들의 장르를 정의하기란 쉬운 작업이 아니다. 웹 장르의 정확한 분류체계와 기준을 마련하기는 어렵지만 [1,2,4]의 연구에서는 비교적 객관적인 근거를 가지고 웹 장르 정의에 접근을 시도하였다. 본 연구에서는 제안된 장르기반 분류 방법에 기인하여 실험하고 향후 다른 목적의 장르기반 분류 또는 웹 문서 기반 실험을 위해 웹 문서 장르를 선정하고 수집하였다.

수집된 장르는 한글문서에 표준화된 질의[12,13]를 이용하여 수집된 샘플 문서의 통계치와 기존의 연구결과 [1,4] 및 사용자들이 웹 검색결과에서 상대적으로 많이 필요로 하는 장르만을 선정하여 1차로 구성하였다. 이 말뭉치(corpus)를 이용하여 다음 5장에서는 제안된 방법으로 분류실험을 하고 그 결과를 산출하였다.

수집된 문서 장르는 신문의 사건기사와 사설, 개인 홈페이지, 리뷰, 논문, Q&A, 상품의 스펙으로 총 7가지로 구성된다. 각 문서들은 장르내의 주제 범주별 통계정보를 이용하기 위해 장르와 장르내의 주제 범주에 대한 태깅을 해 놓았다.

표 4는 수집된 문서 정보를 보여주고 있다.

표 4의 첫 번째와 두 번째 열은 장르종류와 그에 속하는 주제별 범주들을 보여주고 있으며 세 번째와 네 번째 열은 각 장르별로 수집된 한글문서와 영어문서의

표 4 장르 및 장르 내의 주제별 범주

장르 및 장르 내의 주제범주		한글	영어
사건 · 사고	강도, 교통, 사기, 살해, 폭력, 약물, 유괴, 자살, 절도, 폭력, 화재	929	815
사설 · 칼럼	경제, 교육, 국제, 문화, 북한, 사회, 스포츠, 정보통신, 정치	750	849
논문	가정예술, 공학, 기초과학, 농수임, 생명의학, 인문사회, 전기전자	1,051	1,200
리뷰	교육, 금융, 문화, 쇼핑, 스포츠, 유아, 음식, 의류, 자동차, 컴퓨터, 화장품	2,362	1,490
개인홈	초중고대학생, 교사교수, 회사원, 연예인	906	1,067
Q&A	법률, 소비자, 영어, 요리, 유학, 의학, 청소년, 컴퓨터	960	1,020
Spec	귀금속, 스포츠, 비디오, 자동차, 전자, 컴퓨터, 화장품, 휴대폰	870	1,174
합계		7,828	7,615

개수를 나타낸다. 표 4에서 보는 바와 같이 각각의 장르는 여러 개의 주제별 범주로 구분해 놓았는데 이는 본 논문에서 제안하는 장르기반 분류방법이 장르별 특성과 더불어 장르 안의 주제별 범주정보를 함께 이용하기 때문이다.

문서수집 과정에서 한글문서는 언어학 전공자 1명과 컴퓨터과학 전공자 2명이 수집하였고 영어문서는 영어학 전공자 3명과 언어학 전공자 1명이 함께 수집하였다. 수집된 문서는 서로 교환하여 적합한 장르의 문서인지를 확인하고 필터링하는 작업을 두 번 거쳤다. 또한 장르별로 다수의 대표적인 포털 사이트에서 문서를 수집하여 문서의 형식이나 스타일이 한 방향으로 편중되는 것을 막았다.

5. 실험 및 결과분석

본 연구에서의 실험은 제안한 장르기반 문서분류 방법의 적합성을 평가해 보는 데 목적이 있다. 따라서, 여러 가지 비교 실험을 통하여 제안된 자질추출 방법이나 분류 알고리즘의 타당성을 분석하고 그 결과를 기술한다.

실험에서 사용한 실험대상 문서로는 제4장에서의 수집문서를 대상으로 장르기반 분류실험을 하였다. 즉, 한글문서 7,828건, 영어문서 7,615건을 대상으로 절반(한글문서 3,914건 영어문서 3,808건)은 장르 학습을 위해 사용하고 나머지 절반(한글문서 3,914건 영어문서 3,807건)을 이용하여 장르 분류에 사용하였다.

5.1 자질선택 방법 비교

기존의 장르분류를 위한 자질선택은 여러 종류의 자질에서 높은 빈도의 자질만을 선택하는 단순한 방법이었으며 이는 분류시에 문서의 자질이 학습된 자질의 평균치보다 높은지 낮은지만을 판단하는데 이용했었다. 이러한 방법은 분류 정확도가 떨어질뿐 아니라 기존의 주제기반 분류시의 자질선택 방법보다도 성능이 떨어진다.

표 5는 기존의 장르분류에 이용했던 자질과 추가적으로 대명사, 종결어미, 인명을 각각 자질로 이용했을때 분류결과를 나타낸다. 이 중에서 명사 자질이 가장 높은 정확도를 나타내었으며 기존의 방법에서 제시한 감탄사, 특수기호 및 본 논문에서 추가적으로 실험한 대명사, 종결어미, 인명 등은 실제로 장르를 분류하는데 자질로서 역할을 수행하지 못하고 있음을 알 수 있었다.

본 절에서는 위의 실험에서 가장 높은 정확도를 나타내는 자질인 명사만을 대상으로 기존의 문서 분류를 위한 학습 과정에서 이용하던 자질선택 방법과 논문에서 제안하는 자질선택 방법으로 자질을 추출한 후 분류실험을 통해 정확도를 비교해 본다. 이 실험결과를 통해서는 제안하는 장르 학습에서의 자질선택 방법이 어느 정도 성능을 발휘하는지 효과를 알 수 있다.

기존의 자질선택 방법은 여러 가지가 있지만 그 중에서 카이제곱(5) 방법[14-17]은 용어와 범주 사이의 독립성을 계산하는 방식으로 카이제곱으로 자질 추출을 한 후 문서 분류시에 비교적 높은 정확도[18,19]를 나타내는 것으로 알려져 있다. 따라서, 본 절에서는 카이제곱 방법으로 문서 학습을 시키고 그 분류결과를 기술한다.

$$G_m(t_k) = \chi^2(t_k, m) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (5)$$

카이제곱 계산

	m	$\sim m$
t_k	A	B
$\sim t_k$	C	D

그림 2 카이제곱 변수

- A : t_k를 포함하고 장르 m에 속하는 문서의 수
- B : t_k를 포함하고 장르 m에 속하지 않는 문서의 수
- C : t_k를 포함하지 않고 장르 m에 속하는 문서의 수
- D : t_k를 포함하지 않고 장르 m에도 속하지 않는 문서의 수
- N : 장르 학습 문서 전체의 개수

그림 2는 카이제곱 (5)의 각 변수에 대한 설명이다.

본 연구에서는 유형별로 실험문서의 개수가 상대적으로 많고 적으므로 정확도를 계산할 때 micro average precision/recall 을 사용하였으며 정확도의 계산 방법은 아래와 같다.

정확도 = (맞게 분류된 문서의 개수 / 전체 실험 문서의 개수)

카이제곱 방법으로 자질을 추출한 후 (4)를 이용하여 장르 분류를 한 결과와 제안된 방법으로 자질을 선택

표 6 자질추출 결과비교

정확도		정확도비교(%)
카이제곱	제안방법	
0.87	0.90	+3.4

표 5 자질 종류별 분류 정확도

자질종류	명사	대명사	감탄사	종결어미	인명	특수기호
정확도	0.90	0.36	0.41	0.45	0.18	0.11

(1~3)한 후 (4)를 이용하여 장르 분류를 한 결과는 표 6과 같다.

위의 실험 표 6에서는 기존의 자질선택 방법보다 제안된 자질선택 방법으로 장르학습을 하는것이 문서 분류시 정확도를 더 높여준다는 사실을 알 수 있다. 즉, 장르기반 분류를 위해서는 기존의 분류에 이용한 학습방법 보다는 새로운 학습방법이 필요하다.

5.2 분류결과의 장르별 세부 분석

본 절에서는 제안된 방법으로 자질선택과 장르분류를 한 후 장르별 세부분류 결과를 분석하여 오류 원인을 찾아보도록 한다.

표 7에서 첫 번째 열에서 괄호 안의 숫자는 실험문서의 개수를 나타내고 두 번째 열부터는 분류된 문서의 개수를 의미한다. 예를 들어, 논문으로 분류테스트를 했을 경우, 실험문서의 개수는 526건이며 이 중에서 사설로 분류된 것은 3건, 사건으로 분류된 것은 2건, 리뷰로 분류된 것은 7건, 논문으로 정확히 분류된 것은 497건, 프로필로 분류된 것은 6건, Q&A로 분류된 것은 9건, 스펙으로 분류된 것은 2건이라는 것을 뜻한다.

표 7은 대체적으로 정확히 분류되었지만 사설이나 사건 장르의 문서로 분류 실험을 했을 경우에, 오 분류가 많이 나오는 것을 알 수 있다. 즉, 사설이 사건 장르로 분류(100)되고 사건이 사설 장르로 분류(110)된 것이 오류의 대부분이다. 이는 사설이나 사건기사 장르가 모두 같은 신문기사 장르에서 수집되었기 때문에 서로 글이 쓰여진 문체가 비슷하므로 오류가 발생한 것으로 보이며, 표 8에서 사설과 사건장르를 신문기사로 합쳐서 실험했을때 정확도가 상승하는 것으로 이 사실을 입증해 준다.

또한 표 7, 표 8에서는 영어문서에 적용하였을 경우 분류 정확도가 각각 0.87, 0.90으로 한글문서의 정확도와 비교할 정도가 됨으로 본 논문에서 제안하는 자질선택 및 장르 분류 알고리즘이 언어에 종속적이지 않고 독립적인 방법임을 실험으로 증명하였다.

5.3 분류 모델과의 비교

본 논문에서 제안하는 장르분류 방법은 자질 추출 후 생성된 지식베이스의 장르 대표벡터와 문서 벡터와의 유사도(similarity)를 기반으로 하며 유사도는 두 벡터 사이의 내적(inner product)으로 계산된다. 본 절에서는 제안된 분류 방법의 효용성을 측정하기 위해 기존의 문서 분류 모델과 비교 실험을 하여 평가를 한다.

비교를 위한 다양한 분류 모델 중에서 분류 알고리즘이 비교적 간단하고 다양한 분류 응용[20]에서 많이 사용되고 있는 베이지안(Bayesian) 모델[14,21-23]을 선정하였다. 베이지안 모델은 분류대상 문서가 각 범주에 속할 확률을 구해 가장 큰 확률값을 갖는 범주에 그 문서를 할당하는 방법이다.

$$P(d \mid g_i) = P(g_i)\prod_{k=1}^{T} P(t_k \mid g_i) \quad (6)$$

베이지안 모델

- T : 전체 문서 집합내의 용어의 수
- $P(g_i)$: 전체 문서 집합에서 범주 i의 문서가 나올 확률
- $P(t_k|g_i)$: 장르범주 i에서 용어 t_k가 나올 확률

표 7 한글문서 장르별 분류결과

장르구분	사설	사건	리뷰	논문	프로필	Q&A	스펙
사설(375)	238	100	9	1	18	5	4
사건(464)	110	353	0	0	0	1	0
리뷰(1,181)	0	0	1,169	0	6	1	5
논문(526)	3	2	7	497	6	9	2
프로필(453)	3	5	15	22	404	4	0
Q&A(480)	0	0	15	0	3	437	25
스펙(435)	0	0	14	0	0	0	421
정확도 = 0.90 (영어문서 정확도 = 0.87)							

표 8 한글문서 6개 장르로 분류

장르구분	신문기사	리뷰	논문	프로필	Q&A	스펙
신문기사(840)	808	6	1	16	5	4
리뷰(1,181)	1	1,168	0	5	2	5
논문(525)	10	6	490	7	10	2
프로필(453)	9	9	17	415	3	0
Q&A(480)	1	13	0	3	438	25
스펙(435)	1	13	0	0	0	421
정확도 = 0.96 (영어문서 정확도 = 0.90)						

표 9는 제안된 자질선택 방법으로 장르학습 후, 유사도를 기반으로 분류한 것과 베이지안 확률값 기반으로 분류한 결과이다. 또한 자질선택 제2단계(장르 내에서 주제 범주간의 용어의 빈도를 이용하여 장르내 대표용어를 계산)의 효과를 알아보기 위해 2단계를 포함시킨 실험과 2단계를 포함시키지 않은 실험을 함께 병행하였다.

표 9 베이지안과의 분류 정확도 비교

분류방법 / 장르학습방법	제안된방법 (유사도기반)	베이지안 방법	비교(%)
1단계,3단계만을 이용	0.87	0.74	+17.6
1,2,3단계 모두이용	0.90	0.70	+28.6

※ 참고 : 제안된 방법
1단계 : 빈도수 높은 용어 추출
2단계 : 장르내 대표용어 계산
3단계 : 장르간 변별치 계산

분류 모델만을 비교해 볼 때 1단계와 3단계를 이용한 장르내 주제별 범주 정보를 이용하지 않은 결과를 보면 유사도를 기반으로 분류한 결과(0.87)가 베이지안 방법으로 분류한 결과(0.74)보다 정확도가 약 20% 높게 나왔고, 1단계, 2단계, 3단계를 모두 이용한 결과(0.90)도 유사도를 이용한 방법이 베이지안 방법(0.70)보다 정확도가 약 30% 높게 나왔다.

주제 범주간의 용어빈도 정보 이용의 효과를 알아보기 위해 표 9의 유사도를 기반으로 분류된 결과만을 볼 때, 장르내의 주제별 범주정보를 이용했을 경우(1,2,3단계를 모두 이용)가 주제별 범주정보를 이용하지 않았을 때(1단계와 3단계만을 이용)보다 정확도가 약 3.5% 높아짐을 볼 수 있다. 그러나 베이지안 방법으로 분류된 결과에서는 주제 범주간의 정보를 이용한 분류결과(0.70)가 주제 범주간의 정보를 이용하지 않은 분류결과(0.74)보다 정확도 약 5.3% 정도 떨어졌다.

위의 표 9를 보면서 두 가지 사실을 알 수 있다.

- 제안된 유사도 기반의 분류 모델(4)로 장르 분류를 할 경우에는 기존의 분류 모델(확률값 기반, 예(6)) 보다 분류 정확도가 우수하다.
- 제안된 자질선택 방법에서 장르내 주제 범주간의 용어의 분포를 이용하는 것이 주제 범주간의 용어 분포를 이용하지 않는 것 보다 장르 분류시에 정확도를 향상시킬 수 있다.

6. 결론 및 향후연구

본 연구에서는 문서를 장르기반으로 분류하기 위해 새로운 학습 방법과 분류 모델을 제시하였다. 제시된 분류 방법은 다음과 같은 가정을 전제로 만들어졌다. "특정 장르를 대표하는 용어들은 그 장르 내에서의 빈도수가 높으며 다른 장르에서는 상대적으로 빈도수가 낮을 것이다. 또한 장르 내에 주제별 범주가 존재한다면 특정 장르를 대표하는 용어들은 그 장르 안의 여러 주제별 범주에 걸쳐 골고루 높은 빈도로 분포되어 있을 것이다."

위와 같은 가정으로 시작하여 개발된 자질선택 방법(편차이용) 및 분류 알고리즘(유사도기반)은 실험결과 분류 정확도가 기존의 자질선택 방법인 카이제곱(x^2) 및 베이지안(Bayesian) 분류 모델과 비교하여 높게 평가되었다.

본 연구에서 제안한 장르기반 자질선택 방법은 장르간의 용어 빈도수와 장르 내부의 주제별 범주간의 용어 빈도수를 함께 이용하며 분류 알고리즘은 장르 대표벡터와 대상문서 벡터와의 유사도(similarity)를 기반으로 분류하는 방법이다.

본 논문에서 연구한 결과를 종합하여 보면 다음과 같이 요약될 수 있다.

• 장르별 문서수집

장르기반 분류의 테스트베드를 마련하기 위해 본 연구과정에서는 언어학 전공자 1, 컴퓨터공학 전공자 2, 영문학 전공자 3명으로 하여금 디지털 웹 문서를 수집하도록 하였다. 수집된 문서는 사건·사고, 사설·칼럼, 개인 홈페이지, 논문, 리뷰, Q&A, 상품 스펙의 7가지 장르로 한글 문서는 7,828건, 영어 문서는 7,615건으로 구성된다. 이 말뭉치는 차후 장르기반 연구나 개발의 테스트베드로 이용될 수 있을 것이다.

• 새로운 장르학습 및 분류방법

본 연구에서 제안하는 장르기반 학습 방법은 용어의 장르간 빈도수 및 장르내의 주제 범주간 용어의 빈도수를 이용한다. 이를 기반으로 분류 자질을 추출하며 정형화된 수식으로 도출해 내었으며 카이제곱 방법과 비교하여 상대적으로 높은 정확도를 나타내었다. 또한 장르 분류 방법으로 유사도를 기반으로 하는 알고리즘을 제시하였으며 기존의 분류 방법인 베이지안과의 정확도 비교에서 약 28%나 높게 나와 본 연구에서 제안한 분류 방법이 정확도 측면에서 우수함을 입증하였다.

• 언어의 독립성 입증

본 연구에서 제안하는 자질 추출 방법 및 분류 알고리즘이 한글문서에만 제한적으로 동작하는지 알아보기 위해 수집된 영어문서로 장르분류 테스트를 시도하였다. 그 결과, 높은 분류 정확도(약 87%)를 나타내어 본 연구에서 제안한 자질선택 방법과 분류 모델이 특정언어에 종속적이지 않고 외국어 영역으로 무난하게 확장할 수 있는 언어적으로 독립적인 방법임을 증명하였다.

본 논문에서는 장르기반 문서분류 방법을 제안하고

이 방법의 타당성을 실험하고 평가하는 쪽에 연구 초점을 맞추었다. 따라서, 아직까지 확장하여 적용해 볼 몇 가지 과제가 남아있으며 그 내용은 다음과 같다.

• 내용기반 분류에의 적용

본 연구에서 제안한 분류방법이 시발점부터 장르기반 분류에 초첨을 맞추어 개발되었지만 기존의 내용기반 분류에도 적용해 볼 필요가 있다. 개발된 자질선택 방법과 분류 방법으로 문서를 내용에 의해 분류해 보고 기존의 자질선택 방법 및 분류 알고리즘과의 성능을 비교해 보면 본 논문의 연구과정에서 제안한 분류 방법이 내용기반 분류에도 적용 가능한지 아니면 장르기반 분류에만 적합한 방법인지를 알 수 있을 것이다.

• 주제별 범주 개수 및 역할 분석

본 논문에서 제안된 방법은 장르 내에서 주제별 범주정보를 이용하는 것이 특징이다. 따라서 주제 범주의 개수가 분류 정확도에 어떠한 영향을 주는지에 대한 분석이 필요하다. 즉, 장르내 주제별 범주의 개수를 다변화 시키면서 분류 정확도를 측정해 볼 필요가 있다. 또한 베이지안 분류방법으로 분류했을 경우 주제별 범주정보를 이용한 분류결과가 주제별 범주정보를 이용하지 않은 분류결과보다 떨어졌는데 이 원인에 대한 분석이 필요하다.

참 고 문 헌

[1] Andrew Dillon, Barbara Gushrowski, "Genre and the Web: Is the Personal Home Page the First Uniquely Digital Genre?," JASIS, 51(2), 2000.

[2] Stephanie Haas, Erika Grams, "Readers, Authors, and Page Structure: A Discussion of Four Questions Arising from a Content Analysis of Web Pages," JASIS, 51(2), 2000.

[3] Stephanie Haas, Erika Grams, "Page and Link Classifications: Connecting Diverse Resources," Digital Libraries 98, Pittsburgh USA, 1998.

[4] Johan Dewe, Jussi Karlgren, Ivan Bretan, "Assembling a Balanced Corpus from the Internet," 11th Nordic Conference of Computational Linguistics, Copenhagen, 1998.

[5] Douglas Biber, A Typology of English Texts, Linguistics, 27:3-43, 1989.

[6] Jussi Karlgren, Ivan Brettan, Johan Dewe, Anders Hallberg, Niklas Wolkert, "Iterative Information Retrieval Using Fast Clustering and Usage-Specific Genres," 8th DELOS Workshop on User Interfaces in Digital Libraries, 1998.

[7] Jussi Karlgren, Douglass Cutting, "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis," Proc. of COLING94, Kyoto, 1994.

[8] Efstathios Stamatatos, Nikos Fakotais, George Kokkinakis, "Automatic Authorship Attribution," Proc. of the 9th Conference on EACL'99, Norway, 1999.

[9] TO BHMA, http://tovima.dolnet.gr

[10] Brett Kessler, Geoffrey Nunberg, Hinrich Schtze, "Automatic Detection of Text Genre", ACL'97, July 1997.

[11] Brown Corpus Manual, http://www.hit.uib.no/icame/brown/bcm.html

[12] 이석훈, 맹성현, 김지영, 장동현, 서정현, 김현, "정보검색 평가체제 구축을 위한 HANTEC 테스트 컬렉션의 패키징", 제5회 한국과학기술 정보인프라 워크샵(KOSTI) 학술발표논문집.

[13] 맹성현, 이석훈, 이준호, 이응봉, 송사광, "정보검색 시스템 평가를 위한 균형 테스트 컬렉션 구축", 한국정보관리학회지, 제16권, 제2호 1999.

[14] David Lewis, Marc Ringuette, "A Comparison of Two Learning Algorithm for Text Categorization," Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[15] David Lewis, Robert Schapire, James Callan, Ron Papka, "Training Algorithms for Linear Text Classifier," Proc. of the 19th ACM SIGIR'96, 1996.

[16] Yiming Yang, Jan Peterson, "A comparative study on feature selection in text categorization," Proc. of 14th Int. Conf. On Machine Learning, 1997.

[17] Robert Gaizauskas et. al., "Comparison between a Method based on the Chi-square Test and a Support Vector Machine for Document Classification," Proc. of ACM SIGIR 2001.

[18] 이상준, 김유원, "계층구조를 고려한 텍스트데이터 분류의 성능향상 방안 연구", 한국데이타마이닝학회 2002 추계학술대회 논문집, 2002.

[19] 염기종, 권영식, "Suffix Tree를 이용한 웹문서 클러스터의 제목 생성 방법 성능 비교", 한국데이타마이닝학회 2002 추계학술대회 논문집, 2002.

[20] Hyo-Jung Oh, Sung Hyon Myaeng, Mann-Ho Lee, "A Practical Hypertext Categorization Method using Links and Incrementally Available Class-Information," Proc. of the 23rd ACM SIGIR Conference, Athenes, Greece, 2000.

[21] Mehran Sahami, "Learning Limited Dependence Bayesian Classifiers," Proc. of the 2nd International Conference on KDD'96, 1996.

[22] Yiming Yang, Xin Liu, "A Re-examination of Text Categorization Methods," Proc. of the 22nd ACM SIGIR'99, 1999.

[23] Andrew McCallum, Kamal Nigram, "A Comparison of Event Models for Nave Bayes Text Classification," AAAI '98 Workshop on Learning for Text Categorization, 1998.

이 용 배

1996년 충남대학교 컴퓨터과학과 학사과정. 1998년 충남대학교 컴퓨터과학과 대학원 석사. 2003년 충남대학교 컴퓨터과학과 대학원 박사. 현재는 전주교육대학교 컴퓨터교육과 교수. 관심분야는 정보검색, 자연어처리, 디지털도서관, 장르분류, 지식관리시스템, 하이퍼미디어시스템

맹 성 현

미국 Southern Methodist University에서 전산학 석사 및 박사(1985년, 1987년). 미국 Temple 대학교 조교수, Syracuse 대학교 부교수. 충남대학교 교수 역임. 현재 한국정보통신대학교(ICU) 교수. 관심분야는 정보검색, 자연어처리, 텍스트마이닝, 디지탈도서관, 시맨틱웹