스타일에 따른 웹 문서의 자동 분류

이 공 주[†] · 임 철 수[††] · 김 재 훈[†††]

요 약

스타일 또는 장르는 문서의 주제와는 다른 문서를 보는 또 하나의 관점이 될 수 있다. 그렇기 때문에 문서의 스타일은 문서 분류의 기준으로 사용될 수 있다. 문서의 스타일에 따른 자동 분류 시스템에 대한 여러 연구들이 수행되어 왔다. 그러나 이런 연구들의 대부분이 일반 문서를 대상으로 수행하였으며, 몇몇 일부의 연구만이 웹 문서를 대상으로 스타일 분류에 대한 연구를 수행하였다. 웹 문서는 일반 문서와는 달리 URL과 HTML을 갖고 있다. 본 연구에서는 이와 같은 URL과 HTML로부터 추출한 자질들을 웹 문서의 스타일 분류에 사용해 보고자 한다. 실험을 통해서 이와 같은 자질들이 웹 문서의 스타일 분류에 어떤 영향을 미치는지를 밝혀보고자 한다.

Automatic Classification of Web documents According to their Styles

Kong Joo Lee[†] · Chul Su Lim[††] · Jae-Hoon Kim[†††]

ABSTRACT

A genre or a style is another view of documents different from a subject or a topic. The style is also a criterion to classify the documents. There have been several studies on detecting a style of textual documents. However, only a few of them dealt with web documents. In this paper we suggest sets of features to detect styles of web documents. Web documents are different from textual documents in that they contain URL and HTML tags within the pages. We introduce the features specific to web documents, which are extracted from URL and HTML tags. Experimental results enable us to evaluate their characteristics and performances.

키워드 : 웹 문서(Web Document), 장르(Genre), 스타일(Style), 자동 분류(Automatic Classification), URL, HTML

1. 서 론

인터넷 상의 문서가 기하급수적으로 증가함에 따라 정보 검색 시스템을 통해서조차 원하는 정보를 찾는 것이 그리 쉬운 일만은 아니게 되었다. 정보검색 시스템들은 이에 대한 해결책으로써 추출한 문서를 다양한 기준에 따라서 분류하여 제시하고자 하는 시도들을 진행하고 있다. 이와 같은 대부분의 문서 분류 시스템들은 주로 문서의 주제에 따른 분류들을 다루고 있다. 문서의 스타일 또는 문서의 장르는 문서의 주제와는 다른 또 하나의 문서에 대한 분류 기준이 될 수 있다. '스포츠', '경제', '게임'과 같은 것이 문서의 주제별 분류라고 한다면, '신문기사', '홈페이지', '상품 스펙' 등은 문서의 스타일에 대한 분류라고 할 수 있다[1].

문서의 스타일 분류에 대한 연구는 오래 전부터 연구자들의 관심 분야 중 하나였다[2-9]. 이런 연구들의 대부분은 일반 문서를 대상으로 스타일 분류를 시도하였다. 그렇기 때문에 실험에 사용한 코퍼스와 문서 스타일의 종류, 사용한 문서 자질들이 모두 일반 문서에서만 적합한 것들이었다. 최근에 몇몇 연구에서만이 웹 문서를 대상으로 문서의 스타일 분류를 시도하고자 하였다[1, 10].

자동 문서 분류 시스템에서 각각의 문서는 그 문서의 특질을 나타내줄 수 있는 자질값들로 표현된다. 자동 분류기는 문서의 자질값에 따라 문서를 분류하기 때문에, 그 문서의 특질을 잘 나타낼 수 있는 자질을 추출하는 것 또한 매우 중요하다. 기존의 문서 스타일 분류 시스템에서는 주로 문서의 내용(content-text) 측면에서 추출할 수 있는 자질들을 이용하여 문서를 분류하였다. 문서 내의 글자개수, 단어개수, 문장수, 각 품사별 개수, 또는 특정 단어의 사용 빈도수 등이 대표적인 자질들로 사용되었다. 몇몇 시스템만이 문서 내용이 아닌 문서의 메타 정보인 HREF의 링크 개수나 문서 내의 이미지 개수 등을 사용하기도 하였다. 그러나 이러한 문서의 메타 정보 자질들에 대한 자동 분류 시스템에서의 성능에 대한 자세한 평가는 보고되지 않았다.

본 연구에서는 웹 문서를 스타일에 따라 분류하는데 있어서 중요하게 작용할 수 있는 자질들을 살펴보고자 한다. 웹 문서는 일반 문서와는 달리 URL과 HTML정보를 내포하고 있다. 이와 같은 정보는 웹 문서의 스타일을 예측할 수 있도록 해줄 수 있다. 예를 들어, URL정보에 'index.html'과 같은 스트링이 있을 경우, 그 웹 문서는 홈페이지일 가능성이

† 정 회 원 : 경인여자대학 전산정보과 교수
†† 준 회 원 : 한국과학기술원 대학원 전산학과, (주)시멘틱케스트 연구원
††† 정 회 원 : 한국해양대학교 컴퓨터공학과 교수
논문접수 : 2004년 5월 15일, 심사완료 : 2004년 7월 29일

높아진다. 본 연구에서는 이와 같은 웹 문서 고유의 정보를 문서 분류 자질로써 제시해 보고자 한다. 우선, 웹 문서에서 추출해 볼 수 있는 가능한 많은 자질들을 사용하여 자동 분류 시스템을 구축하고 실험을 통하여 가장 좋은 성능을 발휘하는 자질들을 제시하고자 한다. 이와 같은 연구는 정보검색 시스템의 사용자 인터페이스 개발에 스타일에 따른 문서 분류를 가능하게 할 수 있을 것으로 예상된다.

본 논문은 다음과 같이 구성되었다. 2장에서는 기존의 연구들에서 다룬 문서의 스타일 부류와 자질 정보에 대해서 살펴보고 3장에서는 본 연구가 기반으로 하고 있는 문서의 스타일 부류와 스타일별 코퍼스에 대해서 살펴본다. 4장에서는 문서의 스타일 분류를 위한 자질들을 소개한다. 5장에서는 각 자질별 실험 결과를 살펴보고 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

2.1 웹 문서의 스타일 부류(class)

자동 스타일 분류 시스템을 구축해야 하기 때문에 어떤 말뭉치를 학습 데이터로 사용하는지가 중요하다. 초기의 연구[4, 7]에서는 주로 Brown 말뭉치(Corpus)가 사용되었다. Brown 말뭉치는 약 500개의 문서로 구성되어 있으며 15개의 부류로 구분되어 있다. 이 구분은 스타일로 간주되기는 하지만 스타일 분류를 목적으로 만들어진 구분이 아니기 때문에, 같은 스타일에 속하는 문서들의 형식이 동일하지 않은 경우가 많았다. 또한 'general fiction'과 같은 스타일은 너무 넓은 범위의 일반적인 분류인 반면, 'belles letters'나 'religion' 같은 스타일은 실제적인 응용에 적합하지 않은 부류였다[2].

[1]에서는 'Reportage', 'Editorial', 'Research articles', 'Reviews', 'Homepage', 'Q&A', 'Specification'의 7가지 스타일 부류의 말뭉치를 직접 구축하여 사용하였다. [11] 연구에서는 직접 웹 문서의 스타일 부류를 정의하고 이를 기반으로 말뭉치를 구축하였다. 이 말뭉치는 사용자 조사(user survey)를 기초로 사용자의 기대에 부합하는 부류를 정의하였다. 정의된 부류의 개수는 11개이며, 그 중에서 홈페이지, 링크 모음, 대화형(interactive) 부류가 웹 문서를 위한 부류이다.

[6]에서는 'public affairs', 'scientific', 'journalistic', 'everyday communication', 'literary'의 5개를 스타일 부류로 정의하였다. [9]는 Brown말뭉치 대신에 대부분이 뉴스 장르인 Wall Street Journal(WSJ) 말뭉치를 사용했다. 이 연구에서는 WSJ 말뭉치를 헤드라인(headline)의 설명을 이용하여 휴리스틱하게 4가지의 스타일 부류로 구분했다. 그 4가지 부류는 'editorial', 'letters to the Editor', 'Reportage', 'Spot news'이다. [8]에서는 'Editorial, Reportage', 'Academic prose', 'Official document', 'Literature', 'Recipes', 'Curricula vitae', 'Interviews', 'Planned speeches', 'Broadcast news'의 10개의 부류를 가진 현대 그리스어 말뭉치를 사용했다. 말뭉치는 웹 사이트(web site)로부터 수집하여 직접 구축했다.

2.2 자동 분류를 위해 사용한 자질들

[4]에서는 20개의 자질을 채택하였다. 어휘 출현 횟수(lexical counts) 6개('Therefore', 'me', 'I', 'it', 'That', 'Which'), 품사 출현 횟수(part-of-speech counts) 7개(명사, 1인칭 대명사, 2인칭 대명사, 부사, 전치사, 현재분사, 현재형 동사), 통계정보(statistical information) 7개(문자 수, 긴 단어(long word) 수, 문장당 평균 단어 수, 문장당 문자 수, 문장 수, 단어당 문자 수 평균)의 자질이다.

[6]에서는 하나의 문서를 'formality', 'elegance', 'syntactic complexity', 'verbal complexity'의 4가지의 주요 자질로 표현하였다. 그리고 주요 자질은 문장당 단어 수, 동사-명사 비율, 관용적 표현, 딱딱한 표현(formal word)과 같은 여러 개의 스타일 마커(style markers)로 표현되었다.

[7]에서는 세 가지 수준의 자질을 사용하여 문서를 분류했다. 첫 번째 수준의 자질은 어휘 수준으로 'Mr'나 'Ms'와 같은 어휘를 사용하였다. 이러한 어휘는 신문기사와 같은 스타일에 빈번하게 사용된다. 두 번째는 문자 수준으로 구두점, 느낌표, 하이픈으로 연결한 단어의 출현횟수를 사용한다. 세 번째는 단어당 문자 수, 문장당 단어 수와 같이 전체 문서를 읽어서 얻어 낼 수 있는 자질들이었다.

[8]에서는 스타일 분류를 위한 자질들을 추출하기 위해 구문 분석기와 같은 자연언어 처리 도구를 사용하였다. 이전의 연구와는 달리 구(phrase)수준과 분석(analysis) 수준의 자질들을 사용했다. 이 연구에서는 전체 구절당 명사구(NP)의 비율, 명사구에 포함된 단어의 평균 개수, 형태소나 구문의 중의성 정도 등을 새로운 자질로 제안했으며 이 자질들이 다른 자질들보다 좋은 성능을 보였다고 보고하였다.

[9]에서는 자주 사용되는 단어 발생 횟수만을 사용하여 문서의 스타일 분류를 시도하였다. 이 연구에서는 문어체와 구어체를 모두 가진 BNC(British National Corpus) 말뭉치에서 50개의 최빈도(가장 발생빈도가 높은) 단어를 추출하였으며 WSJ 말뭉치를 이용하여 결과를 평가하였다. 또한 최빈도 구두점 자질 정보가 일반 문서의 스타일을 분별하는 데 매우 중요한 역할을 담당함을 보였다.

[10]은 [11]의 연구에서 정의한 스타일 부류와 말뭉치를 사용하여 문서를 분류했다. 이 연구에서 사용한 자질 중에서 어휘 수준 자질, 품사 수준 자질, 일반적인 문서 통계정보는 이전의 연구와 유사하고, 웹 문서를 위한 자질로는 이미지의 개수와 문서 내의 HREF 링크의 수를 사용하였다.

최근의 [1] 연구에서는 단어의 통계 자료를 이용하여 스타일을 분류하였다. 이들은 문서 분류의 정확도를 높이기 위해 스타일보다 주제(topic)에 더 의존하는 단어들을 여러 방법을 이용하여 제거하였다.

3. 스타일에 따른 웹 문서 코퍼스 구축

3.1 웹 문서의 스타일 부류(class)

웹 문서에 대한 잘 구성된 스타일 부류(class)는 찾기 어렵다. [11]의 연구가 유일한 웹 문서에 대한 스타일 부류에 관

한 연구로 알려져 있다. 이 연구에서는 웹 부류를 크게 문장위주(textual)인 것과 비문장위주(non-textual)인 것으로 나누고 이를 다시 세분화하여 11개의 부류로 나누었다. 비문장위주의 부류는 'personal homepage', 'public/commercial homepage', 'interactive pages', 'link collections', 'other listing/tables', 'error message'의 6가지로 세분화하였고 문장위주의 부류는 'journalistic materials', 'reports', 'other running text', 'FAQs', 그리고 'discussions'의 5가지로 세분화하였다. 본 연구에서는 [11]의 웹 스타일 부류를 기초로 사용하고, 웹 사용자의 관점에서 이 스타일 부류를 재구성하고자 한다. 즉, 사용자가 원하는 정보를 찾는데 어떤 스타일을 포함시키는 것이 유용한가에 대해 간단한 사용자 조사(user survey)를 통해서 얻은 <표 1>과 같은 웹 문서 부류를 사용코자 한다.

〈표 1〉 본 연구에서 사용한 웹 문서의 스타일 부류

	웹문서의 스타일 부류	예 제	[11]의 장르
비문장 위주 (non-text)	(A) Personal homepage(개인, 비형식적 홈페이지)	이력서	personal homepages
	(B) Public homepages(공공기관 홈페이지)	정부, 병원, 기관, 학교 등의 홈페이지	public/commercial homepages
	(C) Commercial homepages(상업용 홈페이지)	각종 쇼핑몰 사이트	public/commercial homepages
	(D) Bulletin collection(게시판)	BBS, 게시판	N/A
	(E) Link collection (링크모음)	링크 모음 페이지	link collection
	(F) Image collection(이미지모음)	이미지나 사진 모음 페이지	N/A
	(G) Simple table/list(테이블/리스트)	간단한 테이블이나 리스트	other listing and tables
	(H) Input page(입력페이지)	검색 페이지, log-in 페이지	interactive pages
문장 위주 (text)	(I) Journalistic materials(신문기사)	신문기사, 사설, 리뷰 등	journalist materials
	(J) Research report(연구보고서)	전자화된 각종 형식의 논문들	reports
	(K) Official materials(형식문서)	회사 정보, 계약서, 법률 문서, 연락 정보 (contact info)	reports
	(L) Informative materials(정보문서)	강의 노트, 요리법, 백과 사전식 정보들…	reports
	(M) FAQs	faq	FAQs
	(N) Discussions(의견페이지)	개인 의견 게시 페이지, 질의 응답 페이지, 뉴스 그룹의 개별 페이지들	discussion
	(O) Product Specification(상품소개)	상품 소개	N/A
	(P) Others (informal texts)(기타)	시, 인터넷 소설, 개인 일기…	other running text

<표 1>에서는 비교를 위해 4번째 열에 [11] 연구에서 사용한 장르를 대응시켰다. 부류 (A)~부류 (C)는 홈페이지 스타일이다. [11]에서 하나의 장르로 구분했던 홈페이지를 공공 홈페이지와 상업적 홈페이지로 나누었다. 공공 홈페이지와 상업적 홈페이지는 유사한 점이 많아 자동으로 구분할 경우, 뚜렷이 구분해 내기가 어렵다. 그러나 온라인 쇼핑몰과 같은 상업적인 홈페이지들이 급격히 증가하고 있고 사용자는 검색 결과로부터 상업적인 홈페이지를 제외시키기를 원할 수 있기 때문에 두 부류를 따로 구분했다. 부류 (D)~부류 (F)는 주로 다른 대상들을 모아놓은 페이지들이다. 부류 (D)는 개인의 의견이나 질문, 답변을 모아 놓은 게시판의 목록 페이지들이다. (G)는 간단한 테이블과 리스트를 가진 페이지를 위한 부류이다. (H)는 사용자의 입력을 받아 이를 처리해서 결과를 사용자에게 다시 제공하는 종류의 페이지를 위한 부류이다.

(I) 부류는 뉴스와 사설을 위한 것이다. 본 연구에서는 [11]의 'reports' 부류를 (J), (K), (L)의 3부류로 세분화하였다. (J)는 형식을 갖춘 연구 문서를 위한 부류이다. 부류 (N)은 사용자의 의견이나 질문, 답변 문서 자체로, (N) 부류 문서의 링크 모음이 (D) 부류의 문서가 되는 것이다. 부류 (O)는 제품 상세 정보를 나타내는 문서를 위한 것이다. 인터넷 쇼핑몰 사이트들이 꾸준히 증가함에 따라 검색결과에도 제품에 관한 설명을 담고 있는 문서가 많이 나온다. 제품 상세 정보의 부류가 구분이 되면 사용자는 제품 상세 정보에 관심이 있는지의 여부에 따라 이 스타일을 화면에 보이게 하거나 보이지 않게 선택할 수 있을 것이다. 마지막으로 부류 (P)는 개인적인 일기나 감상을 기록한 문서를 위한 것이다.

3.2 웹 문서 코퍼스 구축

학습 및 실험을 위해서 3.1절에서 정의한 웹 부류에 맞는 말뭉치를 구축한다. 인터넷으로부터 모두 1,224개의 웹 문서 말뭉치를 구축하였다. 일반 문서가 아닌 웹 문서를 대상으로 하기 때문에 문서 내용뿐만 아니라 이 문서를 가리키는 URL(Uniform Resource Locator)도 하나의 자질로 사용된다. 그렇기 때문에 이 URL정보 또한 문서의 내용과 함께 저장한다. 그리고 수집한 문서들이 속한 도메인(domain)이 한 곳에 집중되면 그 자질이 편향될 수 있으므로 동일 도메인 내에서 동일 부류를 5문서 이상 수집하지 않도록 제한하였다. 어떤 문서는 여러 개의 프레임(frame)으로 구성되어 있다. 프레임은 하나의 웹페이지를 여러 개의 웹 문서로 나누는 것으로 전체화면을 다시 읽어 들이지 않고 일부분만 갱신할 수 있는 등의 장점이 있다. 각 프레임은 독립적인 문

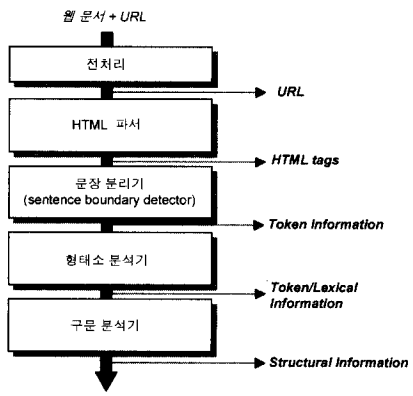
서이므로 각각의 URL을 가진다. 그래서 최종 수집된 코퍼스의 문서 개수는 1,224개의 페이지이지만 프레임을 고려하면 1,328개의 페이지이다. 전체 문서 중에서 45개의 문서는 PDF/PS 파일 형태이고 이들은 모두 'research report' 부류에 속했다. <표 2>는 각 부류별 문서와 문서의 출처 도메인의 수를 나타낸다. 이 테이블에서 '도메인 수'열의 합은 888이고 중복되지 않는 고유한 도메인의 수는 729개다.

〈표 2〉 각 부류별 문서 개수 및 도메인 개수

비문장위주(non-textual)			문장위주(textual)		
웹 스타일 부류	문서 개수	도메인 수	웹 스타일 부류	문서 개수	도메인 수
(A) personal homepage	37	23	(I) Journalistic materials	117	43
(B) public homepages	92	92	(J) Research report	97	41
(C) commercial homepages	73	71	(K) Official materials	150	107
(D) Bulletin collection	74	69	(L) Informative materials	123	97
(E) Link collection	61	54	(M) FAQs	54	52
(F) Image collection	60	48	(N) Discussions	53	19
(G) Simple table/list	32	28	(O) Product Specification	114	62
(H) Input page	48	46	(P) others(informal texts)	39	37
전체 문서의 개수: 1,224 고유한 도메인 개수: 729					

4. 웹 문서의 자동 분류를 위한 자질들

본 논문에서 웹 문서의 스타일을 분류하기 위해 사용하는 자질은 크게 다섯 개의 종류로 나누어 진다. 이 다섯 종류의 자질들은 URL, HTML태그, 토큰(token) 정보, 어휘(lexical) 정보, 그리고 구조적(structural)정보로부터 각각 추출된다. 이들 중 URL과 HTML 태그는 일반 문서에는 없고, 웹 문서만 가지고 있는 자질들이다. 다른 자질들은 일반 문서에서와 웹 문서 모두에서 공통으로 사용되는 것들이다. 자질 집합들을 추출하기 위해 (그림 1)과 같은 추출 과정을 수행한다.



(그림 1) 자질 추출 과정

URL 정보는 웹 문서와 함께 보관된다. 웹 문서에 대해 HTML 분석(parsing)을 하여 HTML 태그에 관련된 자질들을 추출한다. 그 후에 HTML 태그가 제거된 내용이 추출되면 이 내용에 문장분리기(sentence-boundary detector)를 적용한 후, 문장의 개수, 문장당 단어의 평균 개수 등의 통계 정보를 추출한다. 품사(part-of-speech tags) 정보는 형태소분석을 적용한 결과에서 추출한다. 어휘 정보의 자질은 한국어에서의 실질어(content word)와 기능어(function word)를 구분하기 위해 형태소분석 이후에 추출한다. 구문 분석의 결과로부터 구정보(phrase information)와 말덩어리(chunk) 정보를 추출한다. 다음 절에서 웹 문서의 스타일 분류에 사용된 자질에 대해 자세히 기술한다.

4.1 URL

HTTP URL은 인터넷에서 접근 가능한 자원의 주소를 일관되게 표현할 수 있는 형식을 말한다. 하나의 URL은 웹에서 하나의 문서의 위치를 지정한다. URL은 호스트 컴퓨터의 이름(도메인), 디렉터리 경로와 파일이름으로 구성된다. 일반적인 URL의 형식은 아래와 같다[12].

http://<host>:<port>/<path>?<searchpart>

<*port*>번호는 생략 가능하며 <*path*>, '?', <*searchpart*>도 선택 가능하다. <*searchpart*>는 데이터베이스를 접근하기 위한 질의부분이다. URL의 깊이(depth)는 <*path*>에 포함된 디렉터리의 개수로 정의한다. 이 URL의 깊이는 문서의 특성에 관한 유용한 정보를 내포하고 있다. 정보를 찾아가는 입구(portal)가 되는 엔트리 페이지(entry page)의 문서들은 대개 경로 정보없이 호스트 이름만으로 사용된다[17]. <표 3>은 URL 자질들과 그에 대한 간단한 설명이다.

〈표 3〉 URL로부터 추출한 자질 집합

자 질	설 명
U1	URL 깊이
U2	문서 형식[1] (문서이름의 확장자) ; *HTML, SCRIPT, DOC, OUTPUT, MIX*
U3	URL에 '~'가 사용되었나?
U4	파일이름에 다음의 것들이 포함되어 있는가? { *index, default, main, home, main_default* }
U5	도메인 ; *com, org, edu, net, gov, ac.kr, co.kr, go.kr, re.kr, ne.kr, or.kr, pe.kr, etc*
U6	URL의 수(문서의 프레임 개수와 동일)
UL1~UL35	URL에 다음의 35개 어휘가 사용되었는지의 여부 ; *faq, news, board, detail, list, qna, index, shop, data, go, view, front, main, company, item, paper, bbslist, product, read, papers, start, file, gallery, introduction, info, login, search, research, bbs, link, intro, people, profile* …

1) *HTM*은 확장자 'html', 'htm', 'xml'일 경우이고, *SCRIPT*는 'jsp', 'asp', 'php'일 경우이다. *DOC*의 경우에는 'pdf', 'doc', 'ppt'와 같은 응용 프로그램 파일의 확장자를 표시한다. *OUTPUT*은 scipt의 결과 페이지를 의미하며, *MIX*는 여러 개의 framed을 사용할 경우, 다른 종류의 확장자가 복합적으로 사용된 경우이다.

4.2 HTML 태그

HTML(Hypertext Markup Language)은 웹 브라우저 상에 정보를 표시하기 위한 마크업 언어이다. 제목, 문단, 리스트, 하이퍼링크 등은 모두 마크업의 태그를 사용하여 구조화된다. 이 태그는 웹 문서의 구성에 관한 정보를 가지고 있다. HTML 태그에 관련된 자질은 <표 4>와 같다.

〈표 4〉 HTML 태그로부터 추출한 자질 집합

자 질	설 명
H1	동일 도메인으로 나간 out-link 개수 / 문서에 사용된 전체 HTML 태그 개수
H2	다른 도메인으로 나간 out-link 개수 / 문서에 사용된 전체 HTML 태그 개수
H3	링크의 총개수(H1+H2) / 문서에 사용된 전체 HTML 태그 개수
H4~H75	HTML 태그의 개수 / 문서에 사용된 전체 HTML 태그 개수(72개의 HTML 태그에 대해서 ; *col, textarea, input, frame, iframe, select, img, area* …)

4.3 토큰 정보(Token Information)

하나의 문서에는 대체로 문자와 숫자 그리고 심볼 등이 혼합되어 사용된다. 이때 이들의 사용 비율이 문서의 스타일과 연관이 있을 수 있다. 기술문서, 뉴스, 논설 등의 형식을 갖춘 문서에서는 한글 이외에 한자와 알파벳 문자가 특히 많이 쓰인다. 그러므로 문서에서 한자와 알파벳 문자의 사용 비율이 스타일을 분류하기 위한 자질로 사용될 수 있다.

〈표 5〉 토큰 정보로부터 추출한 자질 집합

자 질	설 명
F1	문서의 글자수
F2	문서의 어절수
F3	문서의 문장 후보수
F4	확인된 문장 수 / 문장 후보수
F5	문장당 평균 어절수
F6	어절당 평균 글자수
F7	문장 후보수 / 글자수
F8~F13	각 문자형식의 빈도수 / 전체 단어의 개수 (문자형식 : *hangul, hanja, alpha, digit, punct, symbol*)
T1~T9	품사의 단어 빈도수 / 전체 단어 수 (9개의 POS에 대하여 ; 명사, 대명사, 형용사, 동사, 부사, 감탄사, 관형사, 조사, 어미)
T10	단어의 평균 형태소 모호성 개수
T11~T15	특정종류의 단어 빈도수 / 전체 단어의 개수 (특정종류의 단어 : 한자어, 외래어, 고유명사, 의성어/의태어, 직함)

<표 5>의 자질들은 형태소 분석과 사전 참조를 통해 추출한다. T11(한자어)자질은 외래어로 한글로 표기된다는 점에서 한자(漢字)인 F9(hanja word)자질과 다르다. T15(직함)자질은 '교수', '장관'과 같은 직함을 나타내는 단어에 관한 자질이다.

4.4 어휘 정보(Lexical Information)

어휘 정보 자질에는 가장 발생빈도가 높은 단어(most-frequently used words)의 어휘 빈도 비율 또는 얼마나 다양한 어휘를 구사하는가를 측정하는 어휘 다양성(vocabulary richness) 등이 속한다. 한국어의 경우, 어휘 정보는 기능어와 실질어를 나누어서 처리하였다.

〈표 6〉 어휘 정보로부터 추출한 자질 집합

자 질	설 명
MC1~MC50	실질어의 빈도수 / 전체 단어 수 (50개의 최빈도 실질어에 대해…)
MF1~MF50	기능어의 빈도수 / 전체 단어 수 (50개의 최빈도 기능어에 대해…)
MP1~MP32	구두점(punctuation mark)의 빈도수 / 전체 단어 수 (32개의 구두점에 대해)
S1	흔히 쓰이는(usual) 단어의 개수 / 전체 단어 수 (흔히 쓰이는 단어 : 빈도가 1000이상인 단어)
S2	흔히 쓰이지 않는(unusual)단어 개수 / 전체 단어 수 (흔히 쓰이지 않는 단어는 빈도가 1인 단어)
V1	고유한 단어 개수 / 전체 단어 수(어휘 다양성)

4.5 구조 정보(Structural Information)

구문 분석기의 결과로부터 구 정보를 추출한다. 구조 정보에서는 <표 7>에서 보는 바와 같이 50개의 자질들을 사용한다.

〈표 7〉 구조 정보로부터 추출한 자질 집합

자 질	설 명
P1	평서문(declarative sentence)의 수 / 문장 후보의 수
P2	명령문(imperative sentence)의 수 / 문장 후보의 수
P3	의문문(question sentence)의 수 / 문장 후보의 수
P4	분석 실패(parsing failure) 문장의 수/ 문장 후보의 수
P5	문장당 평균 구문 트리(syntatic tree)의 개수(구조적 중의성)
P6~P22	구(phrase)개수 / 문서의 전체 구 개수 (17개 구에 대해 : 명사구, 동사구, 형용사구, 보조용언구, 부사구, …)
P23~P39	구(phrase)당 평균 단어 수 (17개 구에 대해 : 명사구, 동사구, 형용사구, 보조용언구, 부사구, …)
C1~C11	11개의 종류에 대한 말덩어리(chunk)의 사용 빈도수 (날짜, 시간, 주소, 전화번호, 화폐, 이메일, 인명, 약자…)

5. 실험 결과

5.1 분류기와 실험 환경

실험에는 TiMBL 버전 4.0[13]을 분류기(classifier)로 사용한다. TiMBL은 패턴 분류에 효과적이라고 알려진 k-최근접 이웃 알고리즘(k-NN, k-Nearest Neighbor Algorithm)의 한 종류인 기억기반 학습기법(Memory-Based Learning)을 사용한다. 결과 평가는 'leave-one-out' 교차검증(cross validation) 방법을 사용한다.

모든 HTML 문서는 정해진 표준 HTML 태그를 가지고 있다. 각 문서는 머리(head)부분과 본문(body) 텍스트로 구성된다. 본문 부분에는 실제 문단, 리스트 등의 구성 요소들로 구성된 실제의 내용이 들어간다. 머리부분에는 문서의 제목(title)과 메타(meta) 데이터가 있다. 제목은 사용자 윈도우 속에 나타나지 않지만, 보통 윈도우 창의 제목에 표시되며, 북마크(bookmark)에도 이용되고, 검색 엔진(search engine)에서도 중요한 정보로 사용된다. 메타 태그는 웹페이지에 관한 정보를 제공하는 특별한 HTML 태그이다. 일반적인 HTML 태그들과는 달리, 메타 태그는 웹페이지의 표현에는 영향을 미치지 않는다. 그 대신에 누가 그 페이지를 만들었으며, 얼마나 자주 갱신되는지, 그 페이지는 무엇에 관한 것인지 등과 같은 정보를 제공하며, 그 페이지의 내용을 함축적인 키워드로 표시한다. 많은 검색엔진들이 인덱스를 만들 때, 이 정보를 이용하게 된다. 본문 부분에는 실제의 내용과 더불어 하이퍼텍스트 링크(hypertext link, hyperlink or link)가 부착될 수 있다. 앵커(anchor)라고 불리는 이 링크는 다른 문서나 그림 등을 웹 문서에 연결시켜서 웹 문서의 특징을 가장 잘 나타내 줄 수 있다. 앵커에는 연결되는 문서에 대한 설명문이 포함된다. 본 논문에서는 웹 문서를 제목과 메타데이터(TM), 앵커 문장(ANCH), 본문(BODY)의 3부분으로 나누고 이 3부분의 조합으로 만들어 지는 6가지(TM, ANCH, BODY, TM + ANCH, ANCH + BODY, TM + ANCH + BODY)의 대상에 대해 문서 분류를 수행하고자 한다. 이렇게 하는 이유는 웹 문서의 각 부분들이 자동 스타일 분류에 미치는 영향을 파악하고, 웹 문서의 어떤 부분이 가장 유용한지를 알아보기 위해서이다.

실험 말뭉치는 1,224개의 웹 문서로 구성된다. 실험에 앞서, 구축된 학습 말뭉치에 포함된 제목과 메타데이터에 대해 살펴보면, 230개의 문서만이 메타데이터를 가지고 있으며, 126개 문서에는 제목이 'Untitled Document'이거나 제목 자체가 아예 없었다. 제목이 있는 159개 문서에도 자기 자신의 제목을 가지는 대신 상위 레벨의 문서의 제목을 복사하여 가지고 있었다. 제목과 메타데이터는 문서를 가장 잘 설명하는 문장이나 단어를 사용하도록 설계되었으나, 현재 웹 문서에는 기대보다 그 사용이 부족하였다.

5.2 자질 집합에 따른 실험 결과

<표 8>에 요약된 결과는 문서의 각 부분에 대한 자질집합에 따른 정확률이다. URL과 HTML태그에 관련된 자질은 문서 내부의 자질이 아니므로 문서의 부분과는 관계없는 결과를 보이고 있다.

BODY를 포함했을 때와 포함하지 않았을 때를 비교해 보면 자질의 종류와 관계없이 BODY를 포함했을 때의 정확률이 더 높게 나타난다. 웹 문서의 주제(topic)별 분류에서는 본문에서 추출한 자질을 메타부분에서 추출한 자질에 추가했을 때, 오히려 정확률이 감소되었기 때문에, 본문의 자질을 사용하지 않는 경우가 많다[14]. 또 [15]의 연구에서도 본문의 자질이 주제별 자동 분류의 정확률을 감소시켰다. 그러나, 이와는 반대로 스타일 분류에서는 본문에서 추출한 자질이 자동 분류에 도움이 되었다. 특정 스타일의 특징과 관련된 문서 내의 통계정보는 전체 본문에서 얻을 수 있기 때문에 본문 내용의 자질이 스타일 분류에 도움을 주고 있다.

〈표 8〉 자질 집합에 대한 정확도 결과

자질 집합	사용 자질 개수	TM	ANCH	BODY	TM - ANCH	ANCH+ BODY	TM+ANCH +BODY
U(Url)	6	39.8					
UL(Url Lexical)	35	43.5					
H(Html)	74	**55.1**					
F(Frequency)	13	38.3	43.4	46.4	44.5	43.2	43.1
T(Token)	15	31.6	44.6	36.9	36.1	38.4	39.1
MC(MostFreqCont)	50	16.3	28.8	38.7	30	37.2	37.4
MF(MostFreqFuncw)	50	21.6	29.5	42.9	31.3	44.2	44.2
MP(MostFreqPunct)	32	25.0	35.9	45.6	38.1	**46.9**	45.8
S(Usual/Unusual)	2	15.0	18.1	17.4	18.6	16.4	16.4
V(VocabularyRich)	1	22.6	18.6	12.7	16.1	13.6	11.5
P(Phrase)	37	29.1	33.3	38.6	35	38.2	37.3
C(Chunk)	11	20.4	33.2	37	35	40.6	43.4

<표 8>에서, HTML 태그 자질을 제외한 모든 자질들은 모두 정확률이 50%를 넘지 못하고 있다. 단독 자질 중에서는 HTML태그가 가장 높은 정확률을 보이며 최빈도 구두점(MP)이 두 번째로 높은 정확률을 보인다. 또한, 자질 MF(최빈도 기능어)가 MC(최빈도 실질어)보다 문서의 스타일 분류에 있어서는 좋은 결과를 보였다.

5.3 모든 자질 집합에 대한 실험 결과

<표 9>는 'ANCH+BODY'와 'TM+ANCH+BODY' 조합에 대한 실험 결과이다. 우선, 웹 고유의 자질만을 사용한 결과 (1)가 일반문서의 자질만을 사용한 경우 (2)보다 좋은 성능을 발휘함을 볼 수 있다. 또한, 모든 자질들을 함께 사용하면 74%의 정확률을 얻을 수 있었다. 사용 가능한 모든 자질을 한꺼번에 사용하는 것이 좋은 방법만은 아니다. 중복이 되거나 불필요한 자질을 제거하고 꼭 필요한 자질만을 사용할 때, 전체 성능을 향상시킬 수 있다.

〈표 9〉 선택적 실험에 대한 정확률

사용된 자질들	자질 개수	ANCH+ BODY	TM+ANCH +BODY
(1) U(Url) + UL(Url Lexical) + H(Html tags)	116	64.9%	64.9%
(2) (1)의 자질을 제외한 모든 자질들	213	61.4%	60.0%
(3) 모든 자질 (1)+(2)	329	73.9%	74.3%
(4) 가장 좋은 성능을 발휘하는 자질 집합 U + UL + H + T(Token) + MF(MostFreqFuncw) + MP(MostFreqPunct) + S(Usual/Unusual) + C(Chunk)	226	**75.7%**	75.6%

이와 같이 자질선택(feature selection) 방법[16]을 이용해

서 문서 분류에 유용한 자질 집합을 선정한다. 본 연구에서는 포워드 순차 선택(forward sequential selection method) 자질 선택 방법을 사용하여 가장 성능이 좋은 자질의 조합을 구한다. 자질을 선택하여 75.7%까지 성능이 향상될 수 있었다. 가장 좋은 결과를 보인 조합은 { U, UL, H, T, MF, MP, S, C }이다. 선택에서 제외된 자질 집합은 F(Frequency), P(Phrase), MC(Most frequently used Content word), 그리고 V(어휘 다양성)이다.

〈표 10〉 가장 좋은 결과를 보인 자질 집합에 대한 혼동 행렬(P : 정확률, R : 재현율)

	비문장위주(non-textual)								문장위주(textual)							
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	**24**	2	0	0	2	1	1	0	0	0	2	1	1	0	3	0
B	0	**61**	**22**	0	1	0	0	1	4	0	3	0	0	0	0	0
C	0	**14**	**52**	1	3	2	0	0	0	0	1	0	0	0	0	0
D	0	2	1	**59**	2	0	3	1	1	0	0	0	2	1	2	0
E	1	3	6	1	**37**	1	2	0	1	1	4	3	0	0	1	0
F	1	1	2	2	1	**41**	7	0	0	0	3	1	0	0	1	0
G	0	0	0	2	1	0	**21**	1	0	0	4	0	0	0	3	0
H	0	3	0	2	2	4	8	**17**	2	0	5	2	0	0	3	0
I	0	1	0	0	0	0	0	0	**111**	0	0	3	0	0	2	0
J	0	0	0	0	0	0	0	0	0	**93**	0	4	0	0	0	0
K	2	3	0	2	1	1	2	3	5	0	**104**	**16**	4	0	7	0
L	1	2	0	0	0	0	0	0	2	7	**12**	**90**	3	1	3	2
M	0	1	0	2	1	0	0	0	0	0	4	4	**41**	0	1	0
N	0	0	0	2	0	0	0	0	0	0	0	0	0	**51**	0	0
O	0	0	0	4	1	2	3	4	1	0	3	2	0	1	**93**	0
P	0	0	1	1	0	0	2	1	1	1	2	10	5	1	1	**13**
P/R	83/65	65/66	62/71	76/80	71/61	79/68	43/66	60/40	87/95	91/96	71/69	66/73	73/76	93/96	78/82	87/33

A : personal homepage **B** : public homepage **C** : commercial homepage
D : bulletin collection **E** : link collection **F** : image collection
G : simple table **H** : input page
I : journalistic materials **J** : research report **K** : official materials
L : informative materials **M** : faq **N** : discussion
O : product spec **P** : others

<표 10>은 각 스타일에 따른 정확률과 재현율을 보인다. 전체적으로 문장위주의 스타일에 속한 문서의 정확률/재현율이 비문장위주의 스타일에 속한 문서보다 좋은 결과를 보이고 있다. 문장위주의 스타일 중에서는 'research', 'journalistic', 'discussion' 스타일의 정확률/재현율이 제일 높다. 'research report'의 경우, 모든 자질 집합에서 뚜렷이 변별되는 값들을 갖기 때문에 이 스타일의 문서는 다른 스타일보다 쉽게 분류가 된다.[2] 'Journalistic materials'에 포함된 뉴스 기사들은 정형화된 형식에 따르는 경향을 보이기 때문에 스타일 분류가 상대적으로 쉽다. 'Discussion' 스타일에 속한 문서는 문서가 속한 도메인은 다르지만 우연히도 비슷한 형식을 지니고 있어서 좋은 결과를 보인다. 'input page', 'simple table/list'와 'others' 스타일은 매우 실망스러운 성능을 보인다. 이것은 이들 스타일의 문서들이 뚜렷한 특성이 없다는 의미이다. 또한 많은 웹 문서들에는 테이블이나 입력 윈도우들을 기본적으로 가지고 있기 때문에 테이블이나 입력 윈도우만을 가진 문서를 찾기 어려웠다. 예상했던 대로 'public'과 'commercial' 스타일은 자질이 유사하여 자동으로 구별하기 어려웠다. 또한 'official'과 'informative' 스타일도 마찬가지로 구별하기 어려웠다. 이들 스타일을 분별하기 위해서는 다른 종류의 자질에 대한 연구가 더 필요할 것이다.

6. 결 론

본 연구는 웹 환경의 문서를 위한 웹 스타일을 정의하고 웹 문서의 자질을 추출하여 자동으로 스타일을 분류하였다.

웹 문서 스타일은 일반 문서와 달리 다양한 형식을 취하고 있다. 본 연구에서는 스타일 구분의 기준으로 사용자의 요구 사항에 초점을 두었다. 웹 문서는 일반 문서와 달리 자신의 URL을 가지고 있으며 문서의 구조를 나타내는 HTML 태그를 가지고 있다. 기존의 일반 문서에서 사용했던 자질과 웹 문서에서 사용되는 자질들에 대해 각 자질의 영향을 살펴 보았다. 어휘 다양성이나 어휘 자질보다는 HTML태그 자질과 최빈도 구두점 자질이 스타일 분류의 정확률을 향상시켰다. 웹 문서는 본문 이외에도 제목과 메타데이터, 앵커 텍스트를 가지고 있다. 문서의 본문은 주제 분류에서는 종종 사용되지 않았지만 스타일 분류에서는 스타일에 관한 유용한 자질을 제공하기 때문에 성능 향상에 큰 역할을 했다.

현재는 하나의 웹 문서는 하나의 스타일로만 분류된다. 하

2) 'Research' 스타일의 문서는 97개이다. 이 중에서 45개는 PDF파일이고 45개는 HTML문서이다. 나머지 7개는 script 실행결과 문서이다. 또한 'Research' 스타일의 평균 단어 개수는 12,979이지만, 'Research'를 제외한 다른 스타일에서의 평균 단어개수는 653개이다.

지만 하나의 문서가 여러 스타일의 특징을 모두 포함할 수 있으므로 여러 스타일에 중첩되어 포함될 수 있다. 실용적인 자동 분류 시스템의 개발에서는 이에 대한 고려가 필요할 것이다.

참 고 문 헌

[1] Lee, Yong-Bae and Myaeng, Sung Hyon, "Text genre classification with genre-revealing and subject-revealing features," In *Proceedings of the 25th Annual International ACL SIGIR Conference on Research and Development in Information Retrieval*, pp.145-150, 2002.

[2] Biber, Douglas, "Spoken and written textual dimensions in English : Resolving the contradictory findings," *Language*, Vol.62, No.2, pp.384-413, 1986.

[3] Biber, Douglas, "The multidimensional approach to linguistic analyses of genre variation: An overview of methodology and finding," *Computers in the Humanities*, Vol. 26, No.5-6, pp.331-347, 1992.

[4] Karlgren, Jussi, Cutting and Douglass, "Recognizing text genres with simple metrics using discriminant analysis," In *Proceedings of the 15th International Conference on Computational Linguistics*, pp.1071-1075, 1994.

[5] Biber, Douglas, *Dimensions of register variation : A cross-linguistic comparison*. Cambridge University Press. Cambridge, England, 1995.

[6] Michos, Stefanos, Stamatatos, Efstathios, Fakotakis, Nikos, Kokkonakis and George, "An empirical text categorizing computational model based on stylistic aspects," In *Proceedings of the Eighth International Conference on Tools with Artificial Intelligence*, pp.71-77, 1996.

[7] Kessler, Brett, Numberg, Geoffrey, Schütze and Hinrich, "Automatic detection of text genre," In *Proceedings of 35th Annual Meeting ACL*, pp.32-38, 1997.

[8] Stamatatos, Efstathios, Fakotakis, Nikos Kokkinakis and George, "Automatic text categorization in terms of genre and author," *Computational Linguistics*, Vol.26, No.4, pp. 471-495, 2000.

[9] Stamatatos, Efstathios, Fakotakis, Nikos Kokkonakis and George, "Text genre detection using common word frequencies," In Proceedings *of the International Conference on Computational Linguistics*, pp.808-814, 2000.

[10] Karlgren, Jussi, Bretan, Ivan, Dewe, Johan, Hallberg, Anders Wolkert and Niklas, "Iterative information retrieval using fast clustering and usage-specific genres," In *Proceedings of the Eighth DELOS Workshop on User Interfaces in Digital Libraries*, pp.85-92, 1998.

[11] Dewe, Johan, Bretan, Ivan Karlgren and Jussi, "Assembling a balanced corpus from the internet," In *Proceedings of 11th Nordic Computational Linguistics Conference*, 1998.

[12] Berners-Lee, Tim, Masinter, Larry McCahill and Mark, Uniform resource locators, Internet RFC 1738, 1994.

[13] Daelemans, Walter, Zavrel, Jakub, Ko van der Sloot and Antal van den Bosch, "TiMBL: Tilburg Memory-Based Learner version 4.0 reference guide," 2001.

[14] Wang, Yitong Kitsuregawa and Masaru, "Evaluating contents-link coupled web page clustering for web search results," In *Proceeding of 11th International Conference on Information and Knowledge Management*, pp.499-506, 2002.

[15] Pierre, John M., "Practical issues for automated categorization of web sites," *ECDL 2000 Workshop on the Semantic Web*, 2000.

[16] Caruana, Rich Freitag and Dayne, "Greedy attribute selection," In *International Conference on Machine Learning*, pp.28-36, 1994.

[17] Kraaij, Wessel, Westerveld, Thijs Hiemstra and Djoerd, "The importance of prior probabilities for entry page search," In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.27-34, 2002.

이 공 주

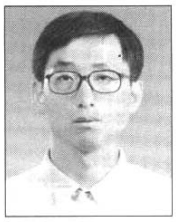
e-mail : kjoolee@kic.ac.kr

1992년 서강대학교 전자계산학과(학사)
1994년 한국과학기술원 전산학과(공학석사)
1998년 한국과학기술원 전산학과(공학박사)
1998년~2003년 (주)한국마이크로소프트 연구원
2003년~이화여자대학교 컴퓨터학과 대우전임강사
2004년~현재 경인여자대학 전산정보과 전임강사
관심분야 : 자연언어처리, 자연어인터페이스, 기계번역, 정보검색

임 철 수

e-mail : cslim@csone.kaist.ac.kr

1992년 경북대학교 컴퓨터 공학과(학사)
1994년 한국과학기술원 전산학과(석사)
1994년~1996년 한국무역정보통신(주) 연구원
1996년~현재 한국과학기술원 전산학과 박사과정,(주)시멘틱퀘스트 연구원
관심분야 : 자연언어처리, 기계번역, 정보검색



김 재 훈

e-mail : jhoon@mail.hhu.ac.kr

1986년 계명대학교 전자계산학과(학사)
1988년 한국과학기술원 전산학과(공학석사)
1996년 한국과학기술원 전산학과(공학박사)
1988년~1997년 한국전자통신연구원, 선임연구원
1997년~1999년 한국해양대학교, 컴퓨터공학과, 전임강사
2000년~2002년 한국과학기술원 첨단정보기술연구센터, 연구원
2001년~2002년 USC, Information Sciences Institute, 방문연구원
1999년~현재 한국해양대학교, 컴퓨터공학과, 부교수
관심분야 : 자연언어처리, 한국어 정보처리, 정보검색, 정보추출