정보관리연구, vol. 35, no. 4, 2004, pp. 81-105

유사어 사전을 이용한 웹기반 질의문의 자동 범주화에 관한 연구

A Study on Automatic Text Categorization of Web-Based Query Using Synonymy List

남 영 준* · 김 규 환**
Young-Joon Nam · Gyu-Hwan Kim

차 례

초 록

본 연구에서는 웹기반 질의문을 자동 범주화하는 방안에 대해 조사하였다. 질의문 범주화에 대한 실험은 SVM-light를 사용하여 범주자질로써 유사어 사전을 부여하기 전과 후를 비교하였다. 유사어는 학습을 통해 수작업으로 대상문서에서 713개를 추출하였다. 전체적으로 유사어 부여전과 부여후의 결과는 6개 범주에서 정도율은 -0.01%로 거의 변화가 없었으며 재현율은 8.53%가 향상되었다. F1-Measure 값도 4.58%가 향상되었다. 특히 범주내 정도율과 재현율의 표준편차가 18.39%나 개선되어 적정한 검색효율을 확보할 수 있었다.

키 워 드

자동범주화, 유사어 사전, 자동색인, 용어자질분석, 웹기반 질의문, 자동분류, 지지벡터기계

* 중앙대학교 문헌정보학과 교수
(Professor, Dept. of Library & Information Science, Chung-ang University, namyj@cau.ac.kr)
** 중앙대학교 문헌정보학과
(Dept. of Library & Information Science, Chung-ang University, quhwan@ms.cau.ac.kr)
• 논문접수일자 : 2004년 12월 3일
• 게재확정일자 : 2004년 12월 16일

ABSTRACT

In this study, the way of the automatic text categorization on web-based query was implemented. X2 methods based on the Supported Vector Machine were used to test the efficiency of text categorization on queries. This test is carried out by the model using the Synonymy List. 713 synonyms were extracted manually from the tested documents. As the result of this test, the precision ratio and the recall ratio were decreased by -0.01% and by 8.53%, respectively whether the synonyms were assigned or not. It also shows that the Value of F1 Measure was increased by 4.58%. The standard deviation between the recall and precision ratio was improve by 18.39%.

KEYWORDS

Automatic Categorization, Automatic Indexing, Feature Analysis, Query Text, Synonymy List, Automatic Indexing, Automatic Classification, SVM, Support Vector Machine

1. 서론

인간의 인지력은 외부 자극을 두뇌에 전달하기 위한 신호를 지각적 본능에 따라 구별하는 능력이라 할 수 있다. 인간은 태어나면서부터 사물을 좋은 것과 나쁜 것 혹은 위험한 것과 안전한 것을 선험적 지식으로 구별하는 습성을 갖고 있다. 이러한 선험적 구별 능력은 후천적 경험에 따라 보다 정확하게 사물을 판단할 수 있는 학습화된 인지능력으로 변화한다.

학습화된 인지능력은 지속적으로 이루어지는 학습과 경험에 의해 개선되며, 고등화된 인지능력으로 발전한다. 학습과 경험은 개인의 특성에 따라 매우 주관적으로 이루어지며, 인지능력은 개인에 따라 달라진다. 이러한 다양성 때문에 인간이 갖고 있는 분류능력을 표준화하거나 혹은 객관화하는 작업은 현실적으로 어려움이 수반된다.

전통적으로 도서관은 인간의 지적 산물을 새로운 지식 창조의 자원으로 활용하기 위해 분류작업과 편목작업은 물론 서지데이터를 상호교환하기 위해 각 지식을 분류할 수 있는 표준화 작업을 지속적으로 수행해 왔다. 또한, 소장된 모든 유무형의 자산을 주제로 구분하기 위해 분류표와 같은 지식관리 도구를 개발하였다. 분류작업은 정확하게 구분되는 자료를 특정 주제에 배정보다 대략적인 정보를 인간의 주관적 판단에 따라 객관적 분류표를 통해 객관화하는 고도의 지적 작업이다.

정보의 양적 증가와 이에 따른 분류작업의 양적 증가는 기존의 수작업으로 관리하기 어려

운 수준에 이르렀으며, 분류대상이 되는 정보의 유형도 기존의 책자형태의 정보와 함께 전자형태의 정보까지 확대되었다. 또한 정보서비스적인 측면에서 참고질의와 함께 도서관 이용질의와 같은 정보서비스가 웹을 통해 이루어지는 것이 일반화되었다. 이러한 질문에 대한 즉각적인 대응은 도서관의 새로운 업무가 되었으며, 물리적 공간에서의 서비스 못지않게 새롭게 많은 투자와 연구가 필요하게 되었다. 이러한 요구에 부응하여 미국 ERIC의 'Ask a Eric'과 한국과학기술정보연구원에서 제공하는 Question포인트와 같은 웹을 기반으로 하는 새로운 정보서비스 모델이 시행중에 있다. 이러한 서비스는 기존의 전화나 방문과 같은 음성기반의 요구보다 웹기반 질의메일이나 웹폼 등 웹기반의 질의 형태로 이루어지고 있다. 이에 따라 도서관은 질의자에 대한 최적의 서비스를 위해 해당 질의문에 대한 최적의 해답을 제공할 수 있는 사서를 연결하고 이용자의 질의 욕구를 해결하는 방안을 모색하는 것이 새로운 과제로 부상되고 있다.

본 연구에서는 이용자의 새로운 웹기반 정보서비스 모델 가운데 질의유형을 범주화하여, 도서관과 사서가 이를 효과적으로 처리할 수 있는 자동범주화 방안을 제안하고자 한다. 이를 위해 지지벡터기계(SVM)를 기반으로, 자체적으로 구축한 유사어 사전을 범주자질지식으로 활용하는 실험을 수행하고자 한다.

2. 자동분류 및 범주화

문헌정보학에서의 자동분류는 전산학 분야에서 자동 범주화와 매우 유사한 알고리듬과 가정 하에서 연구가 진행되었다. 단 대상 집단에서 문헌정보학 분야에서는 실험 집단이 책자형 문헌을 대상으로 하는 반면, 전산학 분야에서는 신문기사와 같이 실험대상의 정보량이 상대적으로 적은 것을 대상으로 하였다.

2.1 자동 분류의 연구대상

수작업으로 이루어지는 분류작업은 인간의 인지적 특성을 바탕으로 최소한의 분류기준인 분류체계에 의거하여 대상 자료를 유연하게 처리하는 방법이다. 자동 분류는 수작업으로 이루어지는 분류작업에 대응되는 의미로 사용되고 있다. 자동 분류는 분류과정에서 인간의 인지적 개입을 최소화하고 기계에 의존하여 작업을 수행하는 것이다. 자동 색인과 자동 분류에 사용되는 알고리듬은 단어에 기반하는 것(통계적 방법)과 문장에 의존하는 것(의미적 해석), 문헌구조에 의존하는 것(구조적 방법)으로 대별된다(남영준 1995). 이 가운데 문헌에 나타난 용어의 빈도를 바탕으로 하는 자동분류 시스템의 연구가 활발하게 이루어지고 있다.

한승희(2003)는 용어의 의미모호성을 해소하면서 용어간의 관계표현이 가능한 용어 자동분류기법의 대안으로 퍼지 클러스터링 기법을 사용하여 용어간의 계층관계를 표현할 수 있음을 보여주었다. 이 방법은 전통적인 정보접근 수단인 표제를 비롯하여 저자, 목차, 권말색인

정보 이외에, 하이퍼텍스트 기반의 전자도서에 대한 접근을 새롭게 할 수 있으며, 일련의 분류 계층인 웹사이트의 사이트맵 구축도 가능하다고 주장하였다. 이 기법은 분석 대상이 되는 정보원에 출현한 단어와 함께 질의확장을 위해 외부의 지적체계(시소러스 등)를 사용하는 것이다.

노정순(2004)은 온라인 목록시스템을 이용하여 이용자가 검색한 자료를 계층구조로 분류하여 열람할 수 있는 최적의 시스템을 제안하였다. 그는 이 연구에서 문헌정보학 분야의 단행본과 학위논문을 대상으로 서명을 이용한 자동색인기법을 비롯하여 용어가중치기법, 유사도계수, 클러스터링 기법을 변수로 실험을 실시하였다. 실험의 방법은 서명필드에 나타난 것과 저자(연구자)가 선별한 색인어 등 실험대상이 되는 문헌집단에 나타난 단어에 기반하여 연구를 수행하였다.

이재윤(2001) 등은 전자분야 신문기사를 중심으로 구축한 KT Set을 비롯하여 국제분야와 경제분야의 신문기사를 대상으로 자체 구축한 테스트 Set인 KFCM CL을 이용하여 자동분류(클러스터링) 실험을 수행하였다. 실험은 해당 신문기사에서 추출한 단어집단의 유사관계를 주요 알고리듬별로 분석하였다.

김현희와 이용례(1990)는 유기화학분야의 문헌을 대상으로 자동분류실험을 실시하였다. 분류의 기준은 Chemical Abstract에 수록된 카테고리 가운데 일부를 차용하였다. 분류의 기준으로 사용한 것은 해당 문헌을 실험집단과 검증 집단으로 구분하여 분석하였다. 실험은 판별분석기법을 사용하였으며, 용어 확장을 위해 어근을 하나의 키워드로 간주하여 용어 출현빈도를 고려하였다.

이경호와 김정현(2001)은 분류 자동화를 위한 주제어를 선정하기 위한 정보원으로써 문헌의 표제와 해당 자료에 대한 키워드를 이용하여 식물학 분야의 자료의 자동분류에 대한 실험을 수행하였다. 사용한 분석알고리듬은 CC의 카테고리를 이용하여 분류데이터베이스를 구축하고 탐색 용어에 대한 주제어의 출현빈도를 이용하여 자동분류를 시도하였다.

2. 2 자동 문서 범주화의 연구대상

문서 범주화와 문헌 분류를 구분하는 것은 분석 의도와 분석대상의 유형에 따라 차이가 발생한다. 문헌 분류 과정은 분석대상 자료의 내용적 혹은 형태적 특성에 따라 수집된 문헌들을 계층화하는 것이다. 이에 비해 문서 범주화 과정은 분류의 주제범위를 보통 주제범주가 사전에 할당하는 계층적으로는 단일계층에 여러 개의 카테고리 범주에 분석대상 자료를 배정하는 것이다. 즉, 전자가 계층위주의 분류라면, 후자는 수평구조의 범주화라고 할 수 있다. 또한 문헌자동분류는 책자형 자료를 중심으로 이루어진다면 문서자동범주화는 상대적으로 짧은 문장의 자료를 중심으로 이루어진다.

특히 자동범주화 연구가 활발하게 진행되는 것은 인터넷을 통해 생성되는 정보의 양이 수작

업으로 관리할 수 없는 수준에 이르렀기 때문에 기계를 통한 새로운 해결책으로 등장한 것이다. 따라서 자동범주화와 관련된 연구에서 그 분석 대상이 대부분 웹문서나 혹은 웹기반 질의메일, 전자문서 등을 대상으로 하고 있다.

유영순(2002)은 전자 메일 문서에 대해 자동 분류 성능을 향상시키기 위한 실험을 수행하였다. 그는 분류 성능에 영향을 끼치는 전자메일 문서의 특성들에 대한 전처리기 모듈을 도입하였다. 즉, 전자메일 문서의 특성들에 있어서 분류 성능 저하의 요인들을 제거하고, 분류성능 향상 요인들은 추가하여 결과적으로 전자메일 문서 분류성능을 향상시킬 수 있음을 보였다.

한광록과 선복근(2000)은 인터넷 문서 자동 분류시스템 개발에 관한 연구를 수행하였다. 이 실험에서는 범주별 인터넷 문서들을 수집하고 수집한 문서에 대하여 카이제곱 검정을 수행함으로써 범주화 자질을 추출하여 벡터의 유사도 계산을 통한 문서 분류방법보다 신경망의 역전파 네트워크 학습 시스템을 사용하는 것이 좀 더 좋은 분류 결과를 보임을 확인하였다.

이지행과 조성배(2000)는 다중 신경망을 이용한 한메일넷 질의를 자동 분류하는 시스템에 관한 연구를 수행하였고 홍진혁 등(2000)은 실세계의 FAQ 메일 자동분류를 위한 문서 특징 추출 방법에 따른 성능 비교 실험을 수행하였다. 이 두 연구는 웹 메일의 자동범주화를 위한 자질추출 방법과 문서 분류기에 대한 성능 개선 방안을 중심으로 연구하였다.

이지행과 조성배(2002)는 전자메일 문서의 자동분류를 위해 여러 분류기를 결합한 자동분류에 관하여 연구하였다. 이 방법에서 문서 분류를 위한 기계학습 알고리듬은 각각의 특성에 따라 다른 분류 양상을 가졌다. 따라서 분류기를 효율적으로 결합할 경우 전체적인 분류 성능의 향상을 확인하였다.

양근우와 허순영(2004)은 지식관리시스템에 등록된 전문가의 전문분야를 자동으로 수집하고 특정 주제분야에 대한 전문지식의 수준을 측정할 수 있는 전문가 분류 및 관리 자동화 방법론을 분류하기 위한 분류자동화 방법론을 지식관리시스템에 적용할 수 있는 알고리듬을 위한 방법을 제시하였다. 그들은 이를 위해서는 자동분류기 개발을 위한 학습이 필요하며, 그 과정에는 학습용 문서를 준비하고, 분야별 통합 문서 벡터값 계산이 선행되어야 함을 설명하고 있다. 이 때 학습용 문서는 전문가들이 생성한 경험데이터로써 1 2페이지로 작성된 분량의 자료이다.

May(1997)는 메시지 유형 기반의 자동문서 범주화 방법을 전자메일 문서의 문헌자동분류의 방법으로 제안하였다. 즉, 웹 메일을 메시지의 유형에 따라 범주화를 수행하는 것이다.

일차적으로 메시지의 주제별 프레임을 설정하고 분석대상, 메시지가 그 프레임에 어느 정도 일치하는지를 판단하여 주제범주화를 수행하는 것이다. 이를 위해서는 주제별 메시지 프레임을 결정짓는 요소(자질)를 선언하고, 분석대상 메시지의 범주자질을 추출하는 과정을 수행하였다.

2. 3 자동문서범주화 원리

2. 3. 1 자동문서범주화의 정의

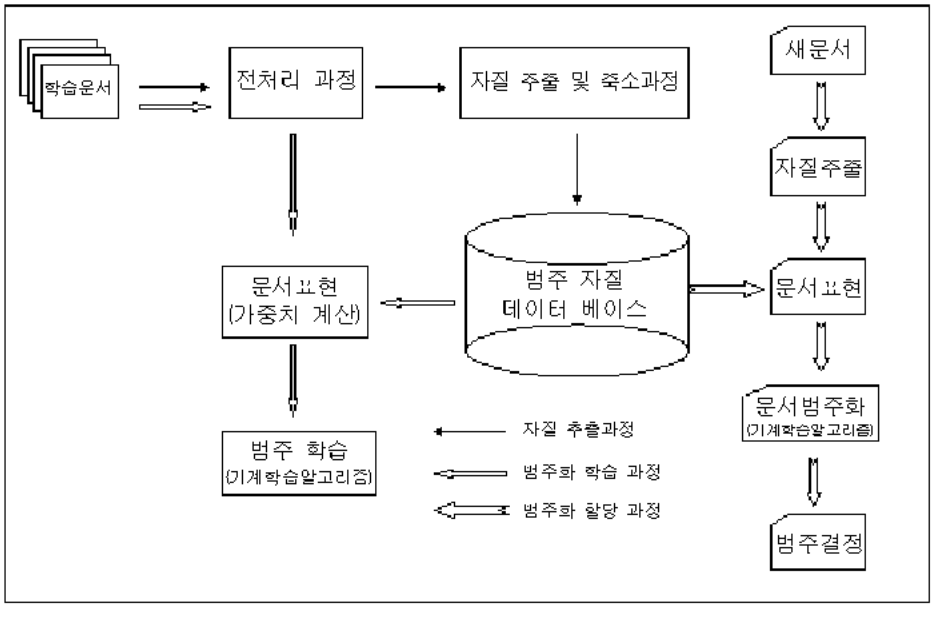
문헌분류과정과 문서범주화는 적용알고리듬과 함께 기법에서 유사성과 최소한의 상이성을 갖고 있다. 이를 구분하기 위해서는 다음과 같이 문서범주화에 대한 구체적인 범주화과정을 분석하였다.

문서범주화의 전체적 과정은 학습문서집합이 주어지고 자질 추출과정을 거치는 사전처리 과정을 갖는 것이다. 자질(feature)은 범주를 설정하기 위한 중요한 기준요소가 되기 때문에 초기 자질집합으로부터 유용한 자질만을 가려내는 자질선정이 선행되어야 한다. 최종적으로 선별된 자질과 가중치로 표현된 문서집단이 준비되면 특정 학습기법을 이용해 주제범주들을 학습시킨다. 이렇게 결정된 주제범주 프레임에 분류 대상문서가 입력되면 범주화 알고리듬에 따라 다양한 방법으로 범주화가 이루어진다.

일반적인 자동범주화 시스템은 〈그림 1〉과 같이 자질추출과정, 범주화 학습과정, 범주화 할당과정으로 이루어진다.

자동문서 범주화의 각 과정에서 처리되는 주요 내용을 요약하면 다음과 같다.

① 자질추출과정 : 사전에 분류가 이루어져 구조화된 주제범주의 특징을 나타내기 위해 두 단계로 구분할 수 있다. 하나는 주제범주에 할당된 학습문서를 통해 자질들을 추출하고, 나머지는 추출된 자질들 중에서 범주 구별능력이 높은 자질만을 선정하는 과정이다. 이 과정에서



〈그림 1〉 자동문서범주화 과정

추출된 자질들은 학습문서표현을 위하여 (범주, 단어)쌍에 대한 데이터베이스에 저장된다. 본 연구에서는 (범주, 단어)쌍에 대한 데이터베이스를 "범주자질 데이터베이스"라고 하였다. 범주자질 데이터베이스는 학습문서 표현을 위해서 사용할 뿐만 아니라 〈그림 1〉에서 신규문서로 표현된 검증문서를 표현할 때 사용한다.

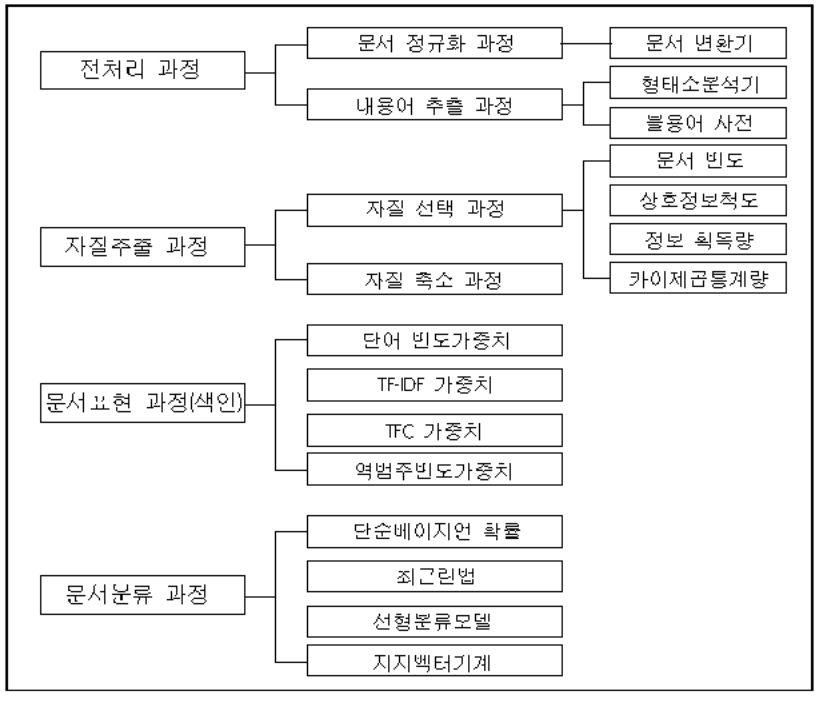
② 범주화 학습과정 : 범주별 학습문서들을 범주자질 데이터베이스에 의해 문서표현한 후 문서 벡터값을 문서분류기에 입력하여 범주를 학습하는 것이다. 이 때 문서표현과정에서는 자동범주화의 기본 모델인 벡터공간모델에 따라 학습문서들을 벡터로 표현한다. 문서벡터는 학습문서가 가지고 있는 단어와 문서 내에서 그 단어의 빈도수를 이용하여 각 단어들의 문서에서의 중요도를 계산한다.

③ 범주화 할당 과정 : 주제범주를 새롭게 할당해야 할 문서를 학습에서 사용한 방법과 동일하게 전처리 과정을 걸쳐 명사를 추출하고 단어와 단어의 가중치를 벡터로 표현하고 문서분류기에 입력하여 주제범주를 할당하는 과정이다.

2.3.2 자동문서범주화 기법

자동문서범주화의 단계별 처리 기법은 〈그림 2〉로 요약할 수 있다.

각각의 요소들에 분석대상에 따라 최적의 파



〈그림 2〉 자동범주화의 단계별 과정 및 처리 기법

라미터를 부여함으로써 그 범주화 과정과 결과가 크게 변화한다. 기존의 선행연구들은 각각의 파라미터를 조작하여 독특한 개별화된 결과를 도출하였지만, 아래 4단계별 과정과 기법들이 표준화된 것들이다.

1) 전처리과정

전처리과정은 분석대상이 되는 정보원에서 내용어를 선택하는 과정이다. 내용어는 문서의 내용이나 특징을 잘 반영하는 용어를 말한다. 내용어에 대한 품사는 여러 가지가 있으나, 범주화 연구에서는 그 가운데 명사(복합명사)와 명사구를 자연어 문장에서 추출하는 것이다. 이들 명사를 추출하기 위해서 형태소 해석기와 불필요한 명사나 무의미한 품사를 제외하기 위해서 불용어사전을 사용한다.

2) 자질 추출과정

자질추출과정에서 자질 선택의 파라미터로는 문서빈도와 상호정보척도, 정보획득량, 카이제곱통계량이 적용될 수 있다.

3) 문서표현과정

문서표현과정은 색인화 과정이며 이를 위해 가장 일반적으로 쓰이는 방법은 벡터공간모델이다. 이것은 문서 전체에 나타난 각 자질의 빈도를 이용하여 문서를 하나의 벡터로 표현하는 것으로 보통 자질의 빈도와 역문헌빈도 혹은 역범주빈도를 이용하여 가중치를 부여함으로써 문서를 표현한다. 자동범주화 분야에서는 〈자질:값〉 표현법을 사용한다(Salton, Fox and Wu, 1983). 각 단어는 범주화를 위한 자질이 되고, 문서 내에서의 빈도수 등을 이용한 단어 가중치가 값으로 처리된다. 대표적인 자질 가중치 부여기법은 이중가중치, 단어빈도 가중치, TF IDF 가중치, tfc 가중치, 역범주빈도 가중치 등이 있다.

4) 문서분류 과정

문서분류과정은 기계학습에 사용되는 알고리듬을 사용하여 입력된 문서를 정해진 범주로 분류하는 단계를 말한다. 우선은 기계학습기가 범주화의 규칙을 학습하여야 하기 때문에 실험문서 중 일부를 각 범주별 학습문서로 선정하고 범주별 학습문서를 통해 범주자질을 추출한다. 추출된 범주자질을 통해 학습문서를 벡터화한 후에 기계학습기에 삽입하여 학습시키게 되는데 이를 범주학습과정이라고 한다. 그리고 분류할 신규문서를 범주자질 집합을 통해 벡터화한 후 기계학습기에 삽입하여 학습된 규칙에 따라 신규문서를 범주에 할당하는데 이를 범주할당과정이라고 한다.

범주학습과정과 범주할당과정에는 기계학습 분야에서 사용되는 문서분류 알고리즘이 사용된다. 문서분류 알고리즘은 크게 규칙기반 모델(rule based model)과 연역적 학습 모델(inductive learning model), 그리고 유사도를 활용한 모델로 구분된다. 먼저 규칙 기반 모델은 학습 문서들에서 나타나는 범주간의 구별된 규칙을 전문가가 찾아주거나 학습을 통해 추출된 규칙을 이용하여 문서를 분류하는 모델이다.

연역적 학습모델로는 학습문서에서 자질을 추출하여 이를 확률적인 접근 방법으로 사용한 단순 베이지언 확률 모델과 학습 문서를 통해

생성된 양성 자질(positive feature)과 음성 자질(negative feature)을 벡터 공간으로 표현하고 이들의 차이를 극명하게 하는 지지벡터(support vector)를 찾는 지지벡터기계(support vector machine)가 있고 유사도를 활용한 방법으로는 최근린법(k nearest neighbor)과 선형 분류 모델(linear classification model)등이 있다(고영중 1999).

3. 유사어 사전을 이용한 문서 자동범주화의 실험

문서 자동범주화를 위해 다양한 알고리듬에 의해 최적의 결과를 얻기 위해 다양한 지식이 사용되고 있다. 본 장에서는 선행연구에서 최선의 자동범주화 기법을 선택하고 유사어 사전을 응용하여 새로운 문서 자동범주화 실험을 수행한다.

3.1 문서 자동범주화 모델 설계

자동문서범주화 실험에서는 일반적으로 정제되고 정형화된 문서를 대상으로 하여 분류를 수행한다. 그러나 현재 보편적으로 사용되고 있는 인터넷 상의 웹기반 질의메일들은 비정형적인 특성을 많이 포함하고 있다. 즉 속어 및 약어의 빈번한 사용과 자유로운 문체로 인해 문법적 오류 등 그 비정형성이 어느 다른 문서들보다도 크게 나타나고 있다. 이러한 문서들을 대상으로 하는 범주화 실험의 경우에는 분류의 정확도가 상대적으로 떨어지게 된다(유영순 2000).

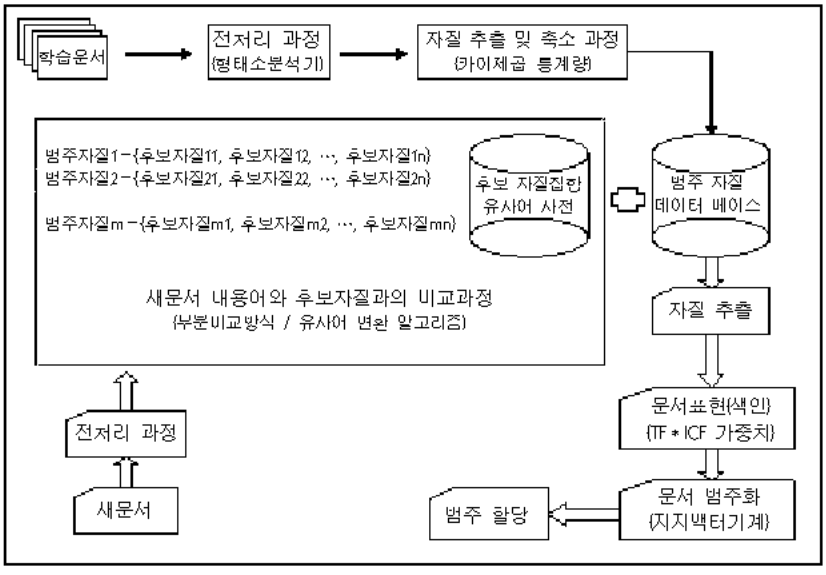
본 연구에서는 비정형화된 문서가 갖는 범주화를 위한 잡음요소를 해결하기 위해 비정형적 잡음 요소들을 추출하고 범주자질을 전거어로 활용하여 유사어 사전을 이용한 새로운 모델을 제안한다. 잡음 요소는 웹 문서 내에 포함된 품사가 명사인 단어 중에서 웹 상에서 사용되는 약어 · 속어, 문법적 오류, 동일 주제범주에서 동일한 의미를 나타내는 단어들이다. 다음 <그림 3>은 유사어 사전을 이용한 웹기반 질의메일 자동범주화 시스템의 전체적인 구성을 도식화한 것이다.

1) 먼저 학습문서인 웹기반 질의메일들을 전처리 과정(형태소 해석기와 불용어 사전)을 걸쳐 품사가 명사인 단어만을 추출하였다.

2) 전처리 과정을 통해 추출된 명사를 자질 추출기법인 카이제곱 통계량과 카이제곱 검증을 통해 범주 구분 능력이 높은 명사만을 자질로 추출하여 (자질, 범주)쌍으로 구성된 범주자질 데이터베이스를 구축하였다.

3) 범주자질에 대한 후보 자질집합 유사어 사전은 수작업 방식을 이용하여 웹 메일의 특징을 분석하여 잡음 요소들을 추출하여 범주자질을 전거어로 한 후보 자질집합 유사어 사전을 구축하였다.

4) 신규문서가 입력되면 전처리 과정을 통해 단어들을 추출하고 추출된 단어들은 유사어 사전의 후보자질과의 비교과정을 통해 잡음 요소들을 정제하여 범주자질로 전환하였다.



〈그림 3〉 웹기반 질의메일의 자동범주화 시스템 전체 구성도

5) 신규문서에 포함된 단어들이 유사어 사전을 통해 정제되어 범주자질로 변환되면 TF*ICF 가중치를 자질값으로 이용하여 신규문서를 〈자질:값〉으로 표현하고 학습된 지지벡터기계 알고리즘을 이용하여 범주를 할당하였다.

3. 2 자동범주화의 실험적 구현

3. 2. 1 실험데이터

본 연구에서 웹기반 질의메일의 자동범주화의 실험적 구현을 위해서 실험문서를 수집하였다. 실험문서는 기업 내에서 콜센터나 이메일을 통해 입수된 고객 상담정보와 신제품에 대한 아이디어, 미래제품 설계 제안, 품질/서비스 개선 제안, 그리고 인터넷에 등록된 고객 의견, 시장 · 유통으로부터의 반응과 같은 기업의 제품 및 서비스에 대한 고객의 다양한 목소리(불만, 제안, 칭찬 등)를 상품기획, 제품개발, 서비스, 품질개선 등의 업무에 연계 활용할 목적으로 수집된 VOC(Voice of Consumers : 고객의 목소리) 데이터로 특정한 포맷을 제공하지 않는 형태이다.

본 실험에서는 〈표 1〉과 같이 각 범주별로 문서를 수집하여 범주별 자질의 추출, 문서 범주학습, 문서 분류 성능 평가를 위한 문서집합으로 총 828개의 실험문서를 선정하였다. 그리고 828개의 실험문서를 문서범주학습에 사용할 학습문서와 실질적으로 자동범주화 시스템의 성능을 검증하기 위한 검증문서로 구분하였으며, 그 비율은 각각 70:30의 비율로 하였다. 이

〈표 1〉 실험문서의 구성

내용범주	범주별 실험문서	학습문서	검증문서
물류	100	70	30
요청	219	153	66
정보	123	86	37
제안	161	113	48
제품	165	116	49
칭찬	60	42	18
합계	828	580	248

는 자동범주화 실험에서 학습문서를 통한 범주 구별능력이 높은 자질을 추출하는 것이 자동범주화의 분류성능에 더 큰 영향을 미치기 때문이다.

수집된 828개의 실험문서들을 범주별로 사전 분류를 하기 위해서 범주 체계를 설정하였다. 본 연구에서는 실험문서의 범주 계층을 기업의 기존 부서의 요구사항과 VOC 자료의 유형(문의성, 제안성, 시장동향)을 조사하여 물류, 칭찬, 정보, 요청, 제안, 제품의 6가지 범주 체계로 구성하였다.

3.2.2 전처리과정

본 연구에서는 전처리 과정으로 형태소 해석기와 불용어 사전을 사용하였다. 형태소 해석기와 불용어 사전을 통해서 내용어를 수집하였으며, 추출 대상은 명사만을 추출하여 내용어로 사용하였다. 이는 명사는 개념을 도입하고 설명하는데 쓰이고 있으며 내용어로 가장 많이 등장하는 중요한 품사로서 특히, 한국어에서는 동작성 명사에 '하다', '되다' 등의 동사 파생 접미사가 붙어서 동사가 되는 경우가 많고 형용사의 경우에도 상태성 명사와 더불어 '하' 등의 형용사 파생 접미사가 붙어서 형용사가 되는 경우가 많으므로 명사의 비중이 그만큼 크기 때문이다.

한편, 본 연구에서는 형태소 분석기에 처리속도를 개선하기 위해 불용어 사전을 사용하여 내용어 추출에 있어 불필요한 용어가 선정되지 않도록 하였다. 불용어 사전은 본 실험에서 사용한 형태소 해석기에 부착된 내용을 기본으로 실험자가 판단하여 전처리 과정에서 내용어로 의미있는 용어를 제거하고, 포함되지 않은 불용어 가운데 잡음 요소로 판단되는 일반 용어들은 별도로 추가하였다. 이는 유사어 사전 구축과는 별도로 잡음 요소를 사전에 최대한 처리할 수 있도록 하는 작업이었다. 불용어 사전에 포함된 용어들은 형태소 해석결과의 태깅 과정에서 명사로 잘 못 분류될 수 있는 용어와 무의미 용어를 사용하였다. 특히 대명사가 포함된 명사구 등을 일차적으로 제거하였으며, 명사나 명사구를 수식하는 역할을 하는 용어들도 불용어사전

〈표 2〉 불용어 사전의 예

_가	_그런데서	_넷	_때	_별안간	* 256메가	_and
_가운데	_그런즐	_넷째	_때문	_볏	* 3-1절	_another
_갈	_그럴수록	_년	; _땡	_본	* 3.1절	_any

에 포함하였다. 불용어 사전은 형태소 해석기가 탑재되기 때문에 KS 완성형(KS C 5601 1987) 한글코드로 작성하였으며, 수록된 불용어의 수는 10,000개이다.

본 연구에서는 형태소 해석기와 불용어 사전을 이용하여 〈표 3〉과 같이 학습문서에서 범주화를 위한 자질값에 영향을 줄 수 있는 총 3,937개의 명사를 추출하였다.

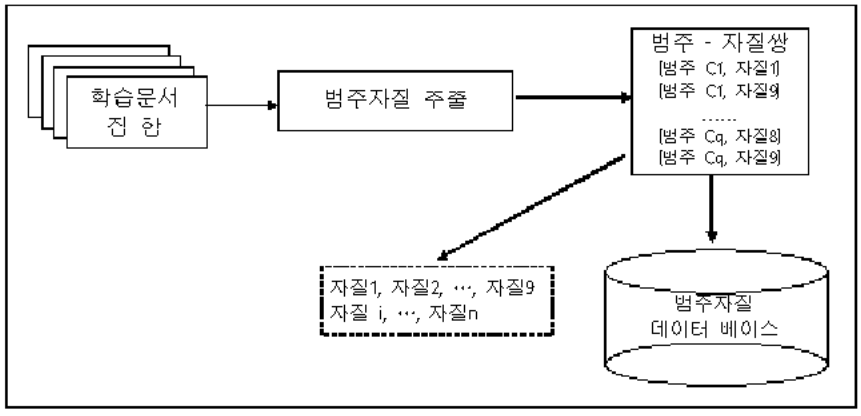
〈표 3〉 추출된 용어의 수 (대상 학습문서)

내용범주	전처리된 학습문서	
	학습문서의수	전체 단어수
물류	70	528
요청	153	213
정보	86	1216
제안	113	912
제품	116	580
칭찬	42	488
합계	580	3,937

3. 2. 3 자질 추출 및 축소

본 연구에서 자질추출은 전처리 과정을 거쳐 추출된 3,937개의 내용어를 이용하여 각 단어가 통계적으로 그 범주를 대표할 만한 특징을 가진 단어인가를 판단하기 위해 카이제곱 통계량과 카이제곱 검증을 수행하였다. 본 연구에서는 카이제곱 통계량을 계산을 하여 3,937개의 내용어를 카이제곱 통계량이 평균치(11.34366) 이상인 974개의 단어를 범주자질로 선정하였다. 이와 같이 카이제곱 검정을 통해 물류를 포함한 6개의 범주별로 통계적으로 대표성을 가지고 있는 범주자질 데이터베이스를 구축하였다. 〈그림 4〉는 범주자질 데이터베이스 구축과정을 도식화한 것이다.

이상의 과정을 통해 6개 주제범주에 범주자질로서 활용한 내용어는 〈표 4〉와 같다. 자질을 표현한 내용어군에 포함된 명사들은 일반 사전에 수록된 표제항에서 사용된 표준어나 외래어



〈그림 4〉 범주자질 데이터베이스의 구성도

〈표 4〉 주제범주에 할당된 범주자질의 예

주제범주	범 주 자 질
물류	설치, 배송, 주문, 시간, 오늘, 주문번호, 본사, 물류, 방문, …
요청	탈퇴, 승급신청, 회원탈퇴, 데이터케이블, 처리, 케이블명, 요청 …
정보	총액, 주가, 시가총액, 보유, 한국, LG전자, 대표주, 토지, 비즈니스위크 …
제안	아이디어, 제안서, 제휴, 발전, 제휴방안, 기반제휴, 개선, 성능, 추가 …
제품	출시, 업그레이드, 신버전, 모델명, 기계, 기기, 정품, 출시일, sk텔레콤 …
칭찬	서비스, 수리, 센터, 친절, 칭찬, 접수, 감사, 감동, 만족, 기분, 자랑 …

를 포함하는 한글 외에도 고유명사와 비속어 등이 포함되어 있다.

3. 2. 4 유사어 사전의 적용

자동 범주화 성능향상의 한 방안으로써 유사어 사전을 활용하기 위해 우선적으로 실험문서에 들어 있는 용어를 기준으로 잡음 요소를 분석하였다.

3. 2. 4. 1 웹기반 질의메일의 잡음 요소 추출

① 잡음 요소 추출

잡음 요소를 추출하기 위해 다음 두 가지 방법을 적용하였다. 첫째, 수집된 580개의 학습문서중에서 250개의 표본 집합을 전처리 과정을 통해 단어들을 추출하였다. 추출된 단어가운데 통신상 통용되는 약어와 속어, 특수기호와 문법적 오류가 쓰인 단어들을 선별하였다. 그리고 이들 단어들을 표준어로 변환되어야 할 단어군 (A)로 분류하였다. 둘째, 웹기반 질의메일의 잡음 요소를 선별하는 과정에서 동일 범주에서

동일한 의미로 사용되고 있는 단어임에도 다른 표현으로 쓰인 단어들의 사용 빈도가 높게 나타났다. 따라서 본 연구에서는 이들 단어들도 정제되어야 할 잡음 요소로 정의하고, 이들 단어들은 대표어로 변환되어야 할 단어군 (B)로 분류하였다.

② 잡음 요소 추출 결과

조사 분석 결과, 약어 · 속어 및 문법적 오류로 인하여 표준어로 변환되어야 할 단어군 (A)과 동일범주에서 동일한 의미를 가지는 단어군 (B)에 속하는 단어는 총 250개의 표본 집합에서 단어(1,640개) 중 4%정도 포함되어 있는 것으로 나타났다. 이들 단어군 (A)과 단어군 (B)에 대하여 본 연구에서는 범주자질에 대한 자질 변형으로 정의한다. 자질 변형은 자질로 가치가 있음에도 불구하고 그 형태가 변형되어 자질로 추출되지 못하는 용어들을 말한다.

약어 · 속어 및 문법적 오류로 인하여 표준어로 변환되어야 할 단어군 (A)을 추출한 예는 〈표 5〉와 같다. 즉, 웹기반 질의메일은 일반 사용자들이 특정한 형식에 제한 받지 않고 자유로운 문체로 작성하고 있음을 보여준다. 대부분 표준형이 완전하게 정착되지 않은 외국어와 모음값을 혼동하기 쉬운 외래어로 구성되어 있었다.

〈표 5〉 웹기반 질의메일의 잡음요소 단어군 일부 (A)

단어군 (A)	개수	단어군 (A)	개수
업그래이드	4	타비	4
업그래드	3	컴터	5
업글	4	배터리	3
데이타	6	서비스썬터	5
애라	5	써비스센터	4
애러	4	핸펀	4

〈표 6〉 웹기반 질의메일의 잡음요소 단어군 (B)

단어군 (B)	개수	단어군 (B)	개수
기종	4	운반	2
단말기	3	운송	8
무상서비스	3	배송	13
무료서비스	2	배달	9
무상수리	6	최신모델	5
반품	2	신모델	4

즉 이에 대한 표준어가 분명하지 않지만 수작업으로는 인지할 수 있는 용어들이다. 또한 키보드 입력에 있어 발생하기 쉬운 용어들이다. 예를 들면 서비스썬터는 서비스쎈터라고 입력할 것을 키워드 입력실패로 입력한 예이다.

〈표 6〉은 용어간에 동일한 의미를 갖고 있으되, 표현형태가 다양하게 표현된 단어군 (B)의 예이다. 웹 문서에서 일반 사용자들은 같은 내용에 대해서 용어 사용의 개인차로 인하여 다양한 표현을 쓰고 있다.

따라서 단어군 (A)와 (B)는 범주자질과 동일함에도 불구하고 서로 다른 종류의 자질로 취급됨으로써 출현빈도가 적어 자질로 추출되지 않는다. 또한 동일한 범주의 유사 자질임에도 불구하고 자질 가중치가 분산됨으로써 범주 모호성을 유발하고 있으며 그 밖에 서로 다른 종류의 자질로 인식되어 문서 자체를 다른 범주로 분류하게 되는 원인이 되고 있다.

〈그림 5〉는 본 연구의 학습문서 중에서 일반 사용자들이 작성한 웹기반 질의메일로서 문법상 오류 및 통신상의 약어 · 속어의 표현이 포함된 사례이다.

> 비번 찾기 위해서 메일로 바로 들어오는 걸로 알았는데 아니네요..텔레비젼 서비스 신청을 하려구요..근처에 있긴 한데 텔레비전이라 여자 혼자 들고 가기엔 벅차네요.. 직접 집으로 오셨으면 하는데.. 그러면 출장비를 제가 내야 하는건지.. 텔레비젼 기종은 couple 이구요.. 티비는 그럭저럭 나오는데.....멜로 연락주세요..

〈그림 5〉 단어군(A)을 포함한 문서의 예

다음 질의문 형태는 앞의 입력오류로 인해 발생하는 것과 달리 개인차로 인해 문장을 구성할 때 개인적으로 사용하는 용어의 차이에 따라 다르게 표현함으로써 자동범주화에 영향을 미칠 수 있다. 즉, 동일한 개념을 다른 용어로 표현하는 것도 범주화의 잡음요소로 간주한다.

즉, 개인차에 따른 다양한 표현방법은 자동문서범주화에 잡음 요소들로서 분류의 정확도

> 1) 주문번호는 잘 모르겠는데요. 제 이름은 000입니다. 확인하시고 빨리 배달해 주시기 바랍니다.
> 2) 안녕하세요! 배송이 순차적로 진행하는 것은 이해 할 수 있으나 일정확인이 안된다는 것은 이해가 되질 않습니다.
> 3) 이틀 전에 발송했다는 메일을 받았습니다. 발송 메일은 오는데 왜 물건은 배송되지 않 는 겁니까?

〈그림 6〉 단어군(B)을 포함한 동일범주 문서의 예

를 저하시키는 요인이 될 수 있다. 왜냐하면 웹 문서의 오분류(오범주화)의 가장 큰 이유로써 분석대상 문헌에서 정확한 자질이 추출되지 않아 벡터를 잘 못 계산하는 것이다. 즉 문서 내에 약어와 속어가 자질로 채택되었는지의 여부에 따라 범주 배정을 위한 변별력 있는 정보가 사용되지 않고, 이로 인해 주요 내용어를 의미적으로 해석으로 하지 않고 형태적으로 해석함에 따라 오범주화가 야기될 수 있다.

자질 변형은 자질과 동일하거나 유사한 의미를 가지면서도 문서 작성자들의 단어 선정능력의 개인차나 개별적 전문성에 따라 차이를 보이고 있었다. 따라서 본 연구에서는 변형된 자질을 표준어와 대표어로 전환시키기 위한 정제도구로써 범주자질에 대한 후보자질집합 유사어 사전을 구축하였다.

3. 2. 4. 2 웹기반 질의메일의 잡음 요소 제거

질의메일의 잡음요소를 판단하고 이를 제거하여 범주화효율을 높이기 위해 다음 세가지 선행작업을 수행하였다.

① 후보자질집합 유사어 사전 구축

유사어 사전의 구축 이유는 범주자질과 유사어 관계에 있는 단어들을 후보자질 데이터베이스에 추가함으로써 새로운 문서가 입력되었을 때 그 문서의 변형된 자질을 후보자질 유사어 사전을 통해 범주자질로 전환시키기 위함이다. 이는 자질변형으로 자질 가중치가 잘못 배정되거나 미배정됨으로 인해 발생할 수 있는 웹기반 질의메일의 자동범주화 시스템의 정도율과 재현율의 저하를 방지하고 그 분류성능을 높이기 위한 것이다.

본 연구에서는 전체 학습문서 580개에 대하여 잡음 요소를 추출하여 범주자질에 대하여 후보자질 유사어 사전을 다음과 같이 구축하였다.

첫 번째 단계, 학습문서 580개에 대해 잡음요소들을 추출하여 범주자질를 전거어로 선정하고 유사어 관계에 있는 단어(잡음 요소)를 선별하였다. 잡음 요소를 유사어로 선별하는 기준은 크게 두 가지로 구분하였다. 하나는 범주자질에 대해 동의어 관계로 동일한 개념을 서로 다르게 표현하고 있는 경우이고 또 다른 하나는 관련어 관계로 의미적으로 수평적 관계에 있는 표현들이다.

두 번째 단계, 선정된 범주자질을 대표어와 표준어를 나타내는 전거어로 삼고 하나의 범주자질과 함께 후보자질 유사어집합을 하나로 통합시켰다.

세 번째 단계, 본 연구에서 사용한 후보자질 유사어 사전은 응용 분야 맞춤형 사전으로서 실험문서로 수집한 웹기반 질의메일의 정제를 위한 것으로 한정하였다.

네 번째, 본 연구에서 웹기반 질의메일의 자동범주화 시스템의 성능향상은 정제된 후보자질 유사어 사전의 구성에 영향을 받는다고 가정하였다. 따라서 정제된 후보자질 유사어 사전이란 속어 및 약어에 대한 표준어(범주자질)와 다양한 표현의 통일된 대표어(범주자질)로 구성한다. 본 연구에서는 변환될 후보자질의 양적인 면으로 판단하여 사전의 내용을 최대한 확보하

〈표 7〉 유사어 사전

대표어 (범주자질)	후보자질집합
서비스기사	A/S기사, 애프터서비스기사, AS기사
흡집	기스, 긁힘
다운로드	다운, download, down
메모리	저장용량, memory, 매모리

여 정제의 정도를 높이고자 하였다.

이상의 과정을 통해 후보자질 유사어 사전을 구축한 결과, 후보자질 유사어 사전에 대표어와 표준어로 쓰인 범주자질을 제외하고 총 713개의 유사어가 선정되었다.

〈표 7〉는 본 연구에서 구축한 유사어 사전의 일부의 예이다.

② 유사어 사전을 이용한 잡음 요소 정제

본 연구에서는 학습문서를 통해 구축된 후보자질집합 유사어 사전을 이용하여 신규문서에 포함된 잡음 요소를 정제하기 위해 다음과 같은 과정을 수행하였다.

첫 번째 과정, 신규문서(d)의 전처리 과정을 통해 내용어를 추출하고 신규문서를 {단어$_1$, 단어$_2$, 단어$_3$,... 단어$_n$}으로 표현한다.

두 번째 과정, 신규문서(d)에서 추출된 단어 집합은 후보자질데이터베이스에 구축된 후보자질들과 비교하는 과정을 수행하였다. 비교는 부분비교 방식을 이용하여 변환율을 높였다. 이 과정을 통해 신규문서(d)에 포함된 단어들은 범주자질중 하나로 정제되어 변환된다. 이 때 후보 자질 유사어 사전에는 범주자질 자체도 포함시켜 신규문서에 포함된 범주자질이 변환과정에서 누락되지 않도록 하였다.

세 번째 과정, 이상의 과정을 통해 신규문서(d)는 {범주자질$_1$, 범주자질$_2$, 범주자질$_3$,..., 범주자$_p$}로 변환하였고 최종적으로 범주자질과 문서벡터값(TF * ICF)으로 표현된 문서를 문서분류기에 입력하여 해당 범주로 분류하였다.

③ 잡음 요소를 후보자질 유사어로의 변환

후보자질 유사어로의 변환 방식(부분비교 방식)

후보자질 유사어로 변환하는 방법으로 완전비교방식이 아니라, 부분비교방식을 사용함으로써 그 변환율을 높게 하였다. 변환율이 높은 것은 문서의 정제의 정도를 높게 하기 위함이다. 완전 비교를 하게 될 경우에는 변환이 상대적으로 어렵기 때문이다. 즉 문서의 정제 정도가 낮다는 것을 의미한다.

예를 들어, 본 연구에서 신규문서에 잡음 요소로 '애라와'를 선정하였다. 후보자질 유사어 사전에는 '애라'가 있고 범주자질 데이터베이스에 '에러'가 있을 경우, 신규문서에 '애라와'는 정제되어야 할 단어이지만, 완전 비교에 의

해 변환하게 되면, 완전 매칭에 이루어지지 않으므로 변환이 이루어지지 않았다. 그러나 부분 비교를 했을 경우 후보자질 유사어 사전의 문자열 기준으로 비교하면, 신규문서의 '애라와'는 후보자질 유사어 사전의 '애라'를 기준으로 부분 매칭되어 신규문서의 '애라와'는 후보자질 유사어 사전에 정의된 표준어(범주자질)인 '에러'로 변환되어 범주자질의 가중치를 부여받게 하였다.

비교부분방식의 처리

부분비교방식에 의한 후보자질 유사어로의 변환 알고리즘은 자동문서범주화의 정확도를 향상시키기 위한 방안으로 사용하다. 즉 웹기반 질의메일 자동범주화 시스템의 실험 환경에서 부차적으로 발생할 수 있는 오분류나 분류의 정확도를 저하시키는 요인들을 사전에 제거하기 위한 것이다.

변환 알고리즘의 처리 방식은 다음과 같은 순서에 의해 이루어졌다.

신규문서(d)에 포함된 단어집합 {단어$_1$, 단어$_2$, ..., 단어$_m$}은 대표어 및 표준어(범주자질)로의 변환을 위해 후보자질 유사어(S_{pm})들과 비교하여 최종적으로 범주자질(C_p)로 변환된다. 이 때 중복 비교횟수를 최소화하기 위해 신규문서(d)의 단어집합의 단어$_m$과 후보자질 유사어(S_{pm})를 정렬하였다. 두 개의 자질 벡터에 대한 비교는 신규문서(d)의 단어$_m$과 후보자질 유사어(S_{pm}) 각각을 비교하기 위해서 신규문서(d)와 후보자질 유사어 사전을 벡터화 하였다.

만약 신규문서(d) 벡터 i가 후보자질 유사어 사전 벡터 j보다 크면, 유사어 사전 벡터 j가 신규문서(d) 벡터에서 해당 정보를 발견할 때까지 후보자질 유사어 사전 벡터 j만 증가시킨다. 또한 반대의 상황에서는 신규문서(d) 벡터에서 유사어 사전 벡터의 단어를 찾을 때까지 문서벡터 i만 증가시켰다. 이 때, 유사어 사전 벡터는 내림차순으로 정렬되어 있으므로 동일한 문서 벡터를 반복 비교하지 않는다. 이런 변환 알고리즘 방식은 자동문서범주화 시스템의 처리속도를 저하시키지 않으면서 신규문서(d)의 단어들을 범주자질로 신속히 추가하는데 매우 효율적이었다.

3. 2. 5 범주자질 가중치 부여

본 연구에서는 후보자질 유사어 사전을 통해 웹기반 질의메일의 잡음 요소가 정제된 문서의 범주자질에 가중치를 부여하여 문서를 표현하였다. 범주자질에 대한 가중치 부여방법으로 용어 출현빈도와 역범주 빈도를 사용하였다. 전자는 해당 문서내에서 용어의 가중치를 부여하기 위함이고, 후자는 범주간의 분리도가 높은 용어에 높은 가중치를 주기 위함이다.

역범주빈도는 다음 공식(1)을 이용하여 계산하였다. t_i를 포함하는 범주의 개수는 CF_i이고 총 범주의 개수는 M이다.

$ICF_i = \log(M) \quad \log(CF_i)$ … 공식(1)

용어 출현빈도(TF_{ij})와 역범주빈도(ICF_i)를 이용해서 용어 t_i의 j번째 범주에서의 가중치 w_{ij}는 공식(2)와 같이 계산하였다.

$w_{ij} = TF_{ij} \times ICF_i = TF_{ij} \times (\log(M) - \log(CF_i))$ … 공식(2)

3.2.6 문서 분류 알고리즘

잡음 요소의 정제 도구로써 후보자질 유사어 사전을 이용하였다. 이 때 웹기반 질의메일 자동범주화 시스템의 성능 변화를 측정을 위해서 문서분류기로 지지벡터기계를 선정하였다. 지지벡터기계를 구현한 프로그램에는 Joachims에 의해 구현된 SVM light[1]를 본 연구에서는 사용하였다.

SVM light에서는 지지벡터기계의 성능은 몇 개의 실험 상수에 의해 최적화 될 수 있다. 즉 Polynomial, RDF(radial basis function)와 같이 비선형 학습이 가능하며, 결정 임계치 등을 조정할 수 있다. 그러나 본 연구의 목적은 서로 다른 분류 모델간의 성능 비교보다는 하나의 분류 모델에서 후보자질집합 유사어 사전을 부여하기 전과 후의 시스템의 성능변화를 비교하는 것이므로, 모든 실험 상수를 SVM light에서 제공하는 기본값으로 사용하였다. 기본값을 통한 실험은 선형 결정면을 학습하며, 결정 임계치는 학습문서의 선험적 확률값을 통해 결정하였다.

4. 유사어 사전을 이용한 웹기반 질의메일 자동범주화 방법의 성능평가

4.1 성능평가 방법

본 연구에서 지지벡터기계 문서분류기를 이용하여 하나의 문서는 하나의 범주로만 할당하였으며, 실험문서에 대한 범주 결정은 〈표 8〉과 같이 분할표[2]를 이용하여 정확하게 분류된 빈도수와 잘못 분류된 빈도수로 구분하여 처리하였다.

〈표 8〉 분할표

	positive example	negative example
긍정 예측치	a	b
부정 예측치	c	d

〈표 8〉에서 a는 해당 범주에 할당되어진 문서 중에 맞게 할당되어진 문서들(positive examples)의 도수이다. b는 해당 범주에 할당되어진 문서 중에 틀리게 할당되어진 문서들(negative examples)의 도수이다. c는 해당 범주에 할당되지 않은 문서 중에 해당 범주에

1) SVM-light : http://svmlight.joachims.org/
2) 분할표(contingency table) : 실험 결과를 두가지 이상의 특성에 따라 분류해서 각 범주에 속하는 빈도수를 기록한 정리표이다.

속하는 문서의 수이고, d는 해당 범주에 할당되지 않는 문서 중에 해당 범주에 속하지 않는 문서의 수이다. 이 분할표를 바탕으로 정도율과 재현율로 성능평가의 실험 효율을 측정하였다. 또한 F Measure 값을 이용하여 실험간[3] 및 범주간 효율을 다면적으로 측정한다.

4. 2 성능평가 결과

4. 2. 1 유사어 사전을 활용 후의 자동범주화 성능 평가

본 실험을 수행하기 위해서 248개의 검증문서(실험데이터)를 선정하였다. 선정된 검증문서는 해당 분야 전문가에 의해 사전에 배정 범주정보를 갖고 있도록 하였다. 이와 같은 실험데이터를 이용하여 본 연구에서 구축한 유사어 사전을 부여하기 전에 범주효율과 부여 후의 범주효율을 측정하였다. 특히 SVM의 기계처리를 위해 추출된 자질을 색인화된 단어의 출현값을 벡터로 표시하여 문서표현작업을 수행하였다. 수행된 결과값을 SVM을 이용하여 최종적으로 문서 범주화의 실험을 수행하였다.

<그림 7>은 웹기반 질의메일 자동범주화 처리과정을 도식화한 것이다.

일차적으로 실험데이터에 대해 유사어 사전을 부여하기 전의 전처리 과정을 아래와 같이 수행하였다.

1) 검증문서 248개를 전처리 과정(형태소 해석기 + 불용어 사전)을 통해 검증문서별 단어집합{단어$_1$, 단어$_2$, 단어$_3$,...,단어$_n$}을 구성하였다.

2) 검증문서는 범주자질 데이터베이스와의 매칭을 통해 단어집합{단어$_1$, 단어$_2$, 단어$_3$,..., 단어$_n$}을 {범주자질$_1$, 범주자질$_2$, 범주자질$_3$,..., 범주자질$_m$}로 변환하였다.

3) 검증문서의 범주자질에 TF * ICF 가중치를 부여하여 문서를 벡터화하였다.

4) 학습된 문서분류 알고리즘인 지지벡터기계를 이용하여 검증문서들을 분류하여 정도율을 비롯하여 재현율, F measure 값, 범주간 편차를 계산하였다.

최종적으로 유사어 사전을 부여하기 전과 부여한 후의 자동범주화 실험을 수행한 결과를 비교 분석하였다. 다음 <표 9>는 이상의 결과를 하나의 표로 비교한 결과이다.

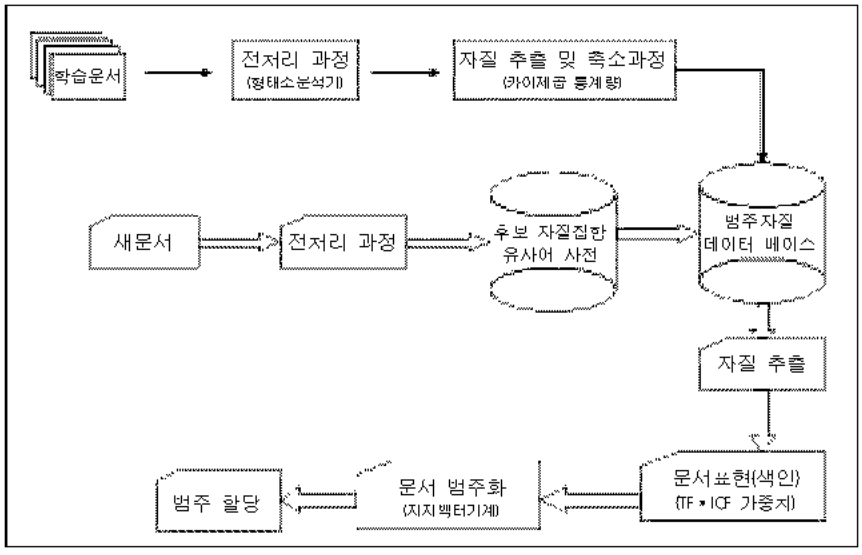
이 가운데 범주간 편차값은 범주별 정도율과 재현율의 차를 절대값으로 처리하여 양과 음의 값을 무시하였기 때문에 일반평균값과의 차이가 발생한다.

4. 3 결과 분석

유사어 사전을 이용한 SVM을 이용한 자동범주화 실험 결과는 다음과 같았다.

SVM만을 이용하여 실험한 결과에서 정도율은 87.05%로 조사되었다. 또한 재현율은

3) 실험간: 유사어 사전 부여 전과 유사어 사전 부여 후 검색효율 변화 측정



〈그림 7〉 유사어 사전 부여후 웹기반 질의메일 자동범주화 실험

〈표 9〉 유사어 사전 부여 후의 성능 평가 결과

주제 범주	유사어 사전 부여전				유사어 사전 부여후				유사어 사전 부여후의 성능 변화			
	Pr(%)	Re(%)	F_1(%)	편차	Pr(%)	Re(%)	F_1(%)	편차	Pr(%)	Re(%)	F_1(%)	편차
물류	93.75	41.66	57.68	52.09	85.10	81.81	83.42	3.29	-8.65	+40.15	+25.73	48.8
요청	61.76	61.76	61.76	0	84.12	81.64	82.86	2.48	+22.36	+19.88	+21.10	2.48
정보	100	92.11	95.89	7.89	91.49	88.22	89.82	3.27	-8.51	-3.89	-6.06	4.62
제안	99.05	86.84	92.54	12.21	87.23	84.56	85.87	2.67	-11.82	-2.28	-6.66	9.54
제품	67.74	100	80.76	32.26	85.81	89.01	87.38	3.20	+18.07	-10.99	+6.61	29.06
칭찬	100	77.77	87.49	22.23	88.52	86.04	87.26	2.48	-11.48	+8.27	-0.23	19.75
평균	87.05	76.68	81.53	21.28	87.04	85.21	86.10	2.89	-0.01	-8.53	+4.58	18.39

76.68%로 조사되었으나, 재현율과 정도율에 대한 편차는 0%(요청 범주)에서부터 52.09%(물류 범주)까지 불안정한 효율을 확인할 수 있었다. 예를 들면, 물류 범주의 실험 결과는 재현율이 41.66%에 불과하기 때문에 자동 범주화에 대한 신뢰도를 극히 저하시키는 수치이다. 또한 제품 범주도 정도율이 67.74%로 재현율(100%)에 비해 상대적으로 많은 차이를 보이고 있기 때문에 자동 범주화 알고리듬의 신뢰도를 저하시키고 있다. 즉, 재현율과 정도율의 기준값(F_1)을 기준으로, 유사어 사전을 부여하기 전에는 81.53%에 조사되었지만 내부 범주간 및 정도율과 재현율과의 표준편차는 21.28%로 매우 불안정한 수치로 판단한다.

한편 유사어 사전을 부여한 후에 정도율은 0.01%로 저하되었으나 이는 검색효율에 절대적인 영향을 미치지 않는 것이라 판단한다. 왜냐하면 재현율은 크게 상승하고, 범주간 정도율과 재현율간의 표준편차는 매우 안정적인 형태로 변화되었기 때문이다. 특히 요청 범주는 정도율과 재현율이 동시에 상승하였으며, 가장 큰 정도율과 재현율의 편차(52.09%)를 보였던 물류 범주가 유사어 사전 부여 후에 3.29%라는 안정된 편차로 변화되었다. 전체적으로 평균재현율은 유사어 사전 부여전에 76.68%였으나, 사전 부여 후에 85.21%로 상승하였다. 또한 정도율과 재현율과의 표준편차는 21.28%에서 2.89%로 크게 줄어들어 모든 범주에서 고른 정도율과 재현율을 확보할 수 있었다.

종합하면, 유사어 사전을 이용하여 재현율은 평균적으로 8.53% 향상되었으며, 정도율은 거의 동일한 수준(0.01%)을 유지하였다. 이와 같은 개선된 현상은 웹기반 질의메일이 갖고 있는 비정형적 잡음 요소로써 통신상의 약어와 속어, 문법적 오류등이 범주자질로 정제한 것에 기인한다. 즉, 웹기반 질의메일에 포함된 잡음 요소로 구분되는 단어들의 범주자질로의 추출 가능성을 보완함으로써 웹기반 질의메일 자동범주화 시스템의 분류성능을 높여주었기 때문이다.

단, 유사어 사전 부여 전에 상대적으로 정도율과 재현율이 높았던 정보, 제안 범주들은 정도율과 재현율이 모두 떨어지는 분류성능을 보이고 있다. 이는 유사어 사전이 기존 자동범주화 시스템에 부여되면서 자질 확장으로 인해 정보, 제안의 두 범주에 대해서 분류성능이 떨어진 것으로 분석된다. 그러나 범주간 정도율과 재현율간의 표준편차는 정보범주는 3.27%, 제안범주는 2.67%로 안정적인 형태로 변화되었다.

자동범주에 있어 적정한 재현율과 정도율은 자동범주화 알고리듬의 성능을 평가하는 매우 중요한 기준이다.

5. 결론

본 연구에서는 웹상에서 이루어지는 이용자 질의 메일을 6개의 범주로 자동범주화하는 실험을 수행하였다. 이 때 사용한 자동범주화 모델은 많은 선행연구에서 대체적으로 높은 효율을 보이고 있는 지지벡터기계(SVM) 알고리듬을 사용하였다. 또한 실험결과를 정규화하기 위해 자체적으로 구축한 유사어 사전을 사용하였다. 이 때 유사어 사전은 자동범주화를 위한 잡음 요소 정제를 위한 지식 베이스로 활용하였다. 최종적으로 SVM을 이용한 범주화 배정결과와 유사어 사전을 부여한 후의 범주화 배정 결과를 비교하였다.

이 실험의 환경과 일련의 과정 통해 얻어진 결과를 요약하면 다음과 같았다.

첫째, 실험은 국내 기업의 제품 사용 고객에게서 직접 입수한 질의문(웹메일 형태)의 828개 실험문서를 대상으로 하였다. 실험문서 가운데 580개를 학습문서로 하였으며, 248개

문서를 검증문서로 활용하였다. 실험분류기는 SVM Light모델을 사용하였으며, 유사어 사전은 자체적으로 구축한 713개의 용어를 사용하였다.

둘째, 실험은 각 범주에 속해있는 자질값을 벡터로 표현하고 이를 근거로 유사어 사전 부여 전의 범주화 효율을 측정하였다. 측정을 통해 정도율과 재현율, F_1값을 입수하였으며, 각 범주별 표준 편차를 측정하였다. 비교결과는 다음과 같이 정도율은 거의 차이가 없었으며, 재현율은 8.53%가 상승, F_1값은 4.57%가 상승하였다.

정도율의 차이 : 0.01%

재현율의 차이 : 8.53%

F_1의 차이 : 4.57%

이와 같은 결과로 SVM을 이용한 범주화에 있어 유사어 사전은 범주화에 매우 긍정적인 역할을 수행하는 것을 확인할 수 있었다.

한편, 본 연구에서는 각 범주별 표준 편차를 별도로 조사하였다. 이는 범주화 모델은 적정한 정도율과 재현율을 확보하여야 하기 때문에 표준편차가 너무 크게 나타나는 것은 신뢰할 수 없는 범주화 모델이라 할 수 있다.

유사어 사전을 부여하기 전에 표준 편차는 21.28%이고, 사전 부여 후 표준편차는 2.9%로 낮아지는 매우 긍정적인 수치를 획득할 수 있었다. 즉 SVM을 이용한 자동범주화 실험에서 정련된 범주자질로써 유사어 사전을 부여하여 정도율을 고정시키고, 재현율을 개선할 수 있음을 입증하였다. 특히 유사어 사전을 이용하여 범주별 정도율과 재현율의 표준편차를 크게 낮추어 자동범주화 모델에 대한 신뢰도를 확보하는 효과도 얻었다.

향후 과제로는 일차원적인 범주모델을 확대하여 계층적인 자동범주 및 자동분류 모델을 개발할 수 있는 대규모 실험 과 유사어 사전을 이용한 연구가 필요할 것이다. 이 때 문헌정보학 관련 질의를 중심으로 하는 실험 과 유사어 사전을 이용한 연구가 병행되어야 한다. 왜냐하면 현재 웹기반 환경으로 빠르게 변화되는 전자도서관시대에 이용자 질의를 효과적으로 처리할 수 있는 실제적인 자동 범주 및 자동분류 시스템 개발이 절실히 필요하기 때문이다. 궁극적으로 전자도서관에 접수되는 이용자 질의를 효과적으로 처리할 수 있는 질의 자동범주화 시스템을 구축하여 담당사서로 하여금 업무의 효율성과 이용자 요구사항에 대한 자료수집 및 분류 시간을 단축하는데 도움이 될 것이다.

참고문헌

고영중. 1999. 『비지도학습을 기반으로 한 자동 문서 범주화』. 석사학위논문, 서강대학교 대학원.

권호경. 1996. 『자동색인을 위한 가중치 불용어 사전의 구축』. 석사학위논문, 한양대학교 대학원.

김상범. 1999. 『범주간의 관계를 통한 자동 문서 범주화의 개선』. 석사학위논문, 고려

대학교 대학원.

김재훈, 김준홍. 2000. 도합유사도를 이용한 한국어 추출문서 요약. 『한글 및 한국어 정보처리 학술대회 발표논문집』, 12: 238 244.

김현희, 이용례. 1990. 문헌의 자동분류를 위한 판별 분류시스템 설계. 『도서관학』, 18집: 129 155

남영준. 1995. 『색인어 형태분석에 의한 한국어 자동색인기법연구』. 박사학위논문, 중앙대학교 대학원, 문헌정보학과.

노정순. 2004. OPAC에서 자동 열람을 위한 계층 클러스터링 연구. 『정보관리학회지』, 24(1): 93 117.

노현아. 2003. 『단어 빈도 가중치를 이용한 자동 문서 분류』. 석사학위논문, 전남대학교 대학원.

박수현. 1999. 한국어 정보검색시스템에서 시소러스를 이용한 검색 효율 향상. 『동서논문집』, 동서대학교: 335 344.

양근우, 허순영. 2004. 문서범주화를 이용한 지식관리시스템에서의 전문가 분류 자동화. 『경영정보학연구』, 14(2): 115 130.

유영순. 2002. 『전자메일분류를 위한 전처리 및 자질추출방법』. 석사학위논문, 충남대학교 대학원.

윤용욱. 2003. 『지지벡터 기계를 이용한 계층적 문서 분류』. 석사학위논문, 포항공과대학 대학원.

이경찬. 2003. 『용어가중치와 역범주 빈도에 의한 자동 문서 범주화』. 석사학위논문, 국민대학교 대학원.

이지행, 조성배. 2000. 다중 신경망을 이용한 한메일넷 질의 자동 분류 시스템. 『한국정보과학회 봄 학술발표논문집(B)』.

전미선, 박세영. 1991. 상호정보를 이용한 어의 모호성 해소에 관한 연구. 『제6회 한국 및 한국어 정보처리학술발표논문집』.

조광제, 김준태. 1997. 역카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류. 『한국정보과학회 봄 학술발표논문집(B)』.

한광록, 선복근. 2000. 인터넷 문서 자동 분류 시스템 개발에 관한 연구. 『정보처리학회논문지』, 7(9): 267 275.

한승희. 2003. 용어 자동분류를 위한 퍼지 클러스터링 기법 분석. 『제10회 한국정보관리학회 학술대회』, 2003년 8월 22~23일. [서울: 이화여자대학교].

홍진혁, 류종원, 조성배. 2001. 실세계의 FAQ 메일 자동분류를 위한 문서 특징 추출 방법의 성능 비교. 『한국정보과학회 봄 학술발표대회논문집(B)』.

Chapelle, O., P. Haffner, and V. Vapnik. 1990. "SVM for Histogram Based Image Classification." *IEEE Trans. On Neural Networks*. 10(5): 1055 1065.

Dumais, S.T., J. Platt, D. Heckerman and

M. Sahami. 1998. "Inductive Learning Algorithems and Representations for Text Categorization." *Proceedings of ACM CIKM98*: 148 155.

Joachims, T. 1998. "Text Categorization with support vector Machines : Learning with Many Relevant Feature." *Proc. In European Conference on Machine Learning*. : 137 142.

Luhn, H. P. 1958. "The Automatic Creation of Literature Abstracts." *IBMJRD*, 2(2): 159 165.

Salton, Gerard. 1989. Automatic Text Proceeding. Addison Wesley: 275 280.

Salton, Gerard, Edward A. Fox, and Harry Wu. 1983. "Extended Boolean Information Retrieval." *Communications of the ACM*, 26(11): 1022 1036.

Yang, Y. and J.O. Pedersen. 1997. "A Comparative Study on Feature Selection in Text Categorization." *Proc. Of the 14th International conference on Machine ICML 97*: 412 429.