http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.22, No.2, May 2017, pp.61-77
https://doi.org/10.7838/jsebs.2017.22.2.061

학회 웹사이트의 토픽 정보추출을 이용한 주제에 따른 학회 자동분류 기법

Academic Conference Categorization According to Subjects Using Topical Information Extraction from Conference Websites

이수경(Sue Kyoung Lee)*, 김관호(Kwanho Kim)**

초 록

최근 온라인상에 게시된 학회정보가 급증함으로써 주제에 따른 학회정보의 자동분류는 연구자들에게 효율적인 관련 학회 탐색을 가능하게 한다. 그러나 대부분의 학회 목록 제공 서비스에서는 학회명칭, 날짜, 위치, URL 등의 정보만 제공하기 때문에 학회 주제를 파악할 수 있는 정보는 학회명칭에 국한된다. 따라서 본 연구에서는 URL을 통한 학회 웹사이트의 토픽정보를 추출함으로써 학회정보량의 부족문제를 해결하고, 동시에 양질의 정보로 학습의 성능을 향상시키는 기법을 제안한다. 구체적으로는 웹사이트 URL을 통해 수집한 HTML 문서로부터 주요 콘텐츠를 추출하고, 학회명칭과 유사한 토픽 키워드 정보를 선정하여 추가 가중치를 부여한다. 실 데이터를 활용한 실험 결과, 제안된 방법인 추가적인 웹 콘텐츠 정보의 사용은 주제에 따른 학회 분류의 성능을 성공적으로 향상시킬 수 있음을 확인하였다. 추후 연구에서는 웹 사이트의 구조를 고려한 토픽 정보추출을 통해 분류의 정확성을 더욱 향상시킬 계획이다.

ABSTRACT

Recently, the number of academic conference information on the Internet has rapidly increased, the automatic classification of academic conference information according to research subjects enables researchers to find the related academic conference efficiently. Information provided by most conference listing services is limited to title, date, location, and website URL. However, among these features, the only feature containing topical words is title, which causes information insufficiency problem. Therefore, we propose methods that aim to resolve information insufficiency problem by utilizing web contents. Specifically, the proposed methods the extract main contents from a HTML document collected by using a website URL. Based on the similarity between the title of a conference and its main contents, the topical keywords are selected to enforce the important keywords among the main contents. The experiment results conducted by using a real-world dataset showed

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2014R1A1A1006458).

* First Author, Department of Industrial and Management Engineering, Incheon National University (dltnrud1212@naver.com)

** Corresponding Author, Department of Industrial and Management Engineering, Incheon National University (khokim@inu.ac.kr)

Received: 2017-03-17, Review completed: 2017-05-12, Accepted: 2017-05-19

that the use of additional information extracted from the conference websites is successful in improving the conference classification performances. We plan to further improve the accuracy of conference classification by considering the structure of websites.

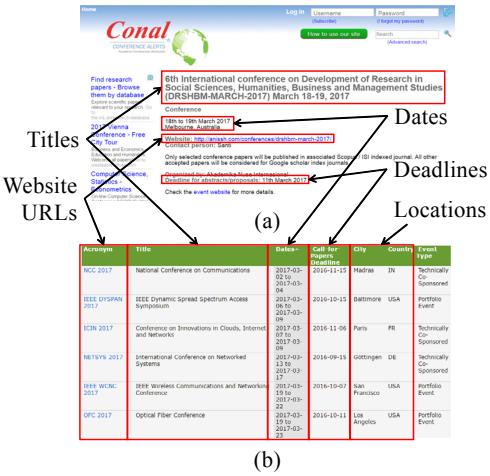
키워드 : 학회정보 분류, 토픽 정보추출, 텍스트 마이닝, 텍스트 분류, 웹 콘텐츠 분석
Academic Conference Classification, Topical Information Extraction, Text Mining, Text Categorization, Web Contents Analysis

1. 서 론

근래에 들어 온라인상에 게시되는 학회정보가 급증함으로써 연구 주제에 따른 학회정보 분류가 매우 중요해지고 있다. 연구자들은 대량의 학회정보로부터 연구 결과의 주제와 관련이 높은 학회 목록을 탐색하는 과정이 필수적이며, 주제별로 학회정보를 제공하는 서비스가 매우 중요하게 사용된다. 하지만 대부분의 학회 목록 제공 서비스의 경우 학회의 분야가 수작업으로 등록되기 때문에 관련 전문지식이 부족하면 올바른 학회분류가 어렵다. 또한 학문분야의 분류 체계가 변화함에 따라 다시 수작업으로 분류해야하는 문제점이 발생하기 때문에 많은 시간과 비용이 소모되며, 정확도가 떨어질 수밖에 없다.

따라서 이와 같은 비효율적인 과정을 해결하기 위해 자동화된 주제별 학회 분류기법에 대한 필요성이 대두되고 있다. 주제별 학회 자동분류는 시간과 비용을 절감하고 방대한 분류 체계 구축을 가능하게 하여 보다 빠르고 정확한 주제별 학회정보를 제공할 수 있다. 또한, 학회정보 기반의 응용 사이트에서 분류 서비스의 질을 향상시킴으로써 사용자의 편의성을 도모한다.

하지만 이러한 주제별 학회 자동분류 모델 개발 시 게시된 학회정보의 특성상 정보부족 문제가 발생한다. 대부분의 학회 목록 제공 서비스에서는 <Figure 1>과 같이 학회명칭 및 기간, 공식 웹사이트 URL 등의 정보만 제공하여, 학회의 주제를 파악할 수 있는 정보는 학회명칭에 국한되기 때문이다.



((a) Conference alerts and (b) ComSoc)

〈Figure 1〉 Examples of Academic Conference Categorization Services

따라서 학습 가능한 정보는 학회명칭만 존재하기 때문에 학습 정보량의 부족과 같은 전형적인 분류문제가 발생한다. 예를 들면, ‘The predict conference Dublin’ 학회의 경우, 주제 파악에 도움이 되지 않는 지역 정보 ‘Dublin’과 학회를 나타내는 ‘Conference’를 제외하면 ‘Predict’

만이 주제파악이 가능한 유의미한 단서로 남게 되는데 이를 통해 정확한 학회 분야를 파악하는 것은 매우 어렵다. 이외에도 '2017 International academic conference on business', '4th International HR conference 2017'의 학회에서도 각각 'Business', 'HR'과 같이 한 단어만이 학회의 주제 식별의 단서가 된다.

기존의 연구에서는 학회 분류기법을 개발하기 위해 새로운 정보를 추가한 학습 데이터를 구성하였다. Xia et al.[23]의 연구에서는 학회의 CFP(Call for paper) 정보와 특정 태그 정보를 활용하여, 학회의 분류기법을 제안한다. 하지만 해당 연구에서 사용한 태그 정보는 기 추출된 정보이기 때문에 태그 정보가 없는 새로운 학회에 대한 분류가 불가능하다. 그러므로 실제 환경에 적용하기에는 한계가 존재한다.

학회 웹사이트의 웹 콘텐츠는 정보량 부족 문제에 좋은 해결책으로 활용된다. 앞서 말한 'The predict conference Dublin' 학회를 예로 들면, 학회 웹사이트에서 'Data'와 'Artificial intelligence' 두 단어만 보더라도 컴퓨터 공학과 관련된 학회임을 쉽게 파악 할 수 있다. 즉, 학회의 웹 콘텐츠 활용은 정보의 확장뿐만 아니라 학회의 주제를 보다 명확히 파악할 수 있는 양질의 정보를 제공한다. 하지만 웹사이트는 비구조적인 형태로 출력되는 HTML 문서이기 때문에 분석하기 쉽지 않다. 또한, 대부분 자연어로 쓰여진 문장이기 때문에 학회주제와 의미적으로 연관된 정보를 추출하기가 어렵다.

따라서 본 논문에서는 정보부족 문제를 해결하기 위해 학회 웹사이트의 HTML 문서로부터 학회주제를 파악할 수 있는 토픽 정보를 추출하고, 추출된 정보를 기반으로 주제에 따른 학회 자동 분류기법을 제안한다. 제안된 기법은 다음과 같이 크게 3단계로 나뉜다.

첫째, 주요콘텐츠 추출을 통해 학회 웹사이트의 웹 콘텐츠 중 불필요한 부가정보를 제거한다. 둘째, 토픽 키워드를 선정하여 주요 콘텐츠 내의 정보 중 해당 학회주제와 연관성이 높은 정보를 선정한다. 또한, 선정된 키워드의 가중치를 설정함으로써 정보의 주제를 강화시킨다. 셋째, 기계학습(Machine learning)기법을 적용한 분류모델을 개발하여 새로운 학회문서를 선정된 클래스에 분류한다.

제안한 기법의 성능평가를 위해 3개의 데이터 집합을 기반으로 모델의 성능을 비교한다. 데이터집합은 학회명칭, 주요 콘텐츠, 토픽 키워드의 가중치 설정의 조합을 통해 구성되며, 이를 통해 추출된 정보가 모델의 성능에 어떤 영향을 미치는지 평가한다. 실험 결과, 학회명칭 및 주요 콘텐츠, 토픽 키워드를 모두 적용시킨 데이터 집합의 성능이 가장 우수하였으며, 이는 본 연구에서 목표로 한 새로운 정보의 추출이 성능향상에 긍정적인 영향을 끼쳤음을 보여준다.

본 논문의 구성은 다음과 같다. 제2장에서 기존 연구에 대한 고찰을 언급하고, 제3장에서 제안 기법에 대한 단계별 내용을 설명한다. 제4장에서는 제안한 방법에 따른 각 모델의 비교 결과를 분석하여 향상된 성능의 타당성을 검증하고, 마지막으로 제5장에서 결론을 기술한다.

2. 관련 연구

학회의 주요 정보추출과 관련된 기존연구는 연구목적, 추출 기법, 추출된 정보에 따라 <Table 1>과 같이 요약할 수 있다.

〈Table 1〉 Summary of Related Studies

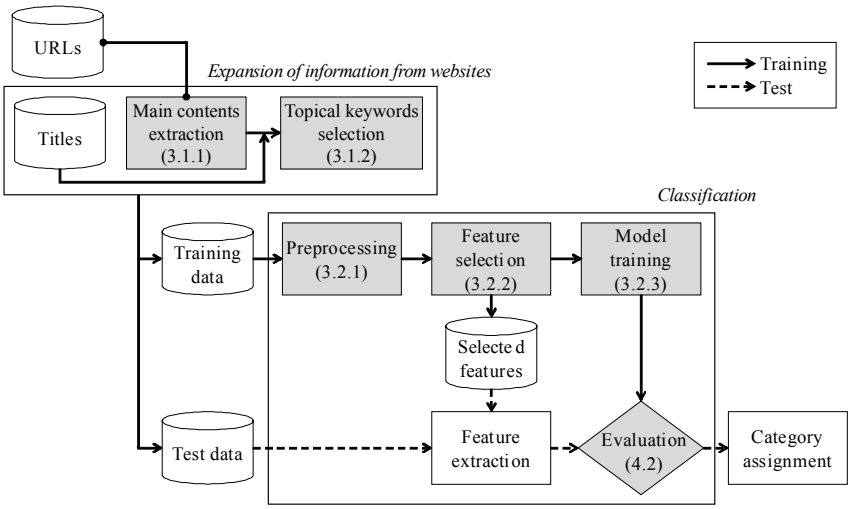
Objectives	Extraction approaches	Extracted information	References
Classification	-	Social tag	Xia et al.[23]
Information extraction	Machine learning	Date, CFP, Topic, Sponsor and Program committee	Xin et al.[24]
		Title, Date, Deadline, Location, Name, Title and URL	Schneider[19]
		Title, Date, URL and Acronym	Cox et al.[5]
		Title, Date, Deadline, Location, Homepage and Acronym	Kim et al.[8]
		Title, Date, Deadline, Location, URL and Contact number	Eom[6]
		Title, Date, Deadline, Location, Homepage and Acronym	Li et al.[14]
		Speaker, Location and Time	Ciravegna[2]
Information retrieval	Similarity between document and query	Date and Country	Lazarinis[10]

기존의 연구에서는 CFP로부터 학회의 명칭, 날짜, 위치, 기한 등의 필요정보를 추출하기 위해 다양한 기법들이 광범위하게 연구되었다. 예를 들면, CRF(Conditional random field) 모델[5, 19], 은닉 마르코프(Hidden Markov) 모델[6, 8], NLP(Natural language processing)[2], SVM (Support vector machine)과 퍼셉트론(Perceptron) [14] 등의 기법을 활용하여 학회의 필요정보들을 추출한다. Lazarinis[10]의 연구에서는 학회의 웹 콘텐츠로부터 개최지 및 날짜를 자동으로 식별하는 규칙을 제시한다. 이 규칙은 문서와 쿼리(Query) 간의 유사도 계산을 기반으로 생성된다.

Xin et al.[24]의 연구에서는 CFP로부터 정보를 추출한 이전연구들과는 달리 학회 웹사이트로부터 학회의 주요정보를 추출하는 기법을 제안한다. 이 연구에 따르면, 모든 학회에서 CFP 정보를 찾는 것은 불가능하며, CFP는 평문으로 정보추출의 성능을 향상시킬 수 있는 형식 및 구조를 잃게 된다. 반면에 96% 이상의 학회(2004~2008년)에서 학회의 필요한 정보를 제공하는 웹사이트가 존재한다. 따라서 본 연구에서도 실적용을 위한 주제별 학회 분류기법을 개발하기 위해 학회 웹사이트의 웹 콘텐츠를 활용하여 주요 콘텐츠를 추출한다.

하지만 위의 기존 연구들에서는 추출한 정보를 활용하여 학회분류를 수행하는 연구가 거의 존재하지 않았으며, 대부분의 연구가 학회의 연구주제를 파악하기 위한 의미정보가 아닌 날짜 또는 위치와 같은 단순 필요정보추출에 초점을 두고 있다.

예외적으로 Xia et al.[23]의 연구에서는 학회의 CFP 정보와 Social tag 정보를 결합하여 학회의 분류기법을 제안한다. 이 연구는 추출된 새로운 태그 정보를 활용하여 학회분류의 성능을 향상시켰다는 점에서 본 연구와 매우 유사하다. 하지만 해당 연구에서 사용된 태그 정보는 기 추출된 정보이기 때문에 태그 정보



〈Figure 2〉 Research Framework(Relevant Sections are in Parentheses)

가 없는 새로운 학회에 대한 분류가 불가능하므로 실제 환경에 적용하기에는 한계가 존재한다.

따라서 본 연구에서는 주제별 학회 분류의 성능을 높은 정확도로 최적화하기 위한 보완적인 데이터로써 주요 콘텐츠 및 토픽 키워드 정보를 추출하고, 이를 활용하여 실 환경에 적용가능한 일반적인 프레임워크를 제안한다.

3. 제안 기법

제3장에서는 제안된 시스템을 <Figure 2>와 같이 두 단계로 구분하여 설명한다. 이는 웹사이트로부터 학회정보를 확장시키는 과정, 그리고 적절한 텍스트 분석기법을 통한 분류 모델 구축 과정으로 진행된다.

본 연구에서 제안하는 주제별 학회 분류기법은 M개의 고유 단어로 구성된 N개의 문서를 L개의 클래스 중 하나의 클래스로 분류하는 문제이다. 구체적으로, i번째 단어는 w_i, $i=1, \cdots, M$, j번째 문서는 d_j, $j=1, \cdots, N$, l번째 클래스는 c_l, $l=1, \cdots, L$로 표현된다. 문서는 각 단어의 빈도수의 벡터형태로 구성되므로 문서 d_j는 $d_j = \langle f_{1j}, f_{2j}, \cdots, f_{Mj} \rangle$, $j=1, \cdots, N$로 표현된다. 여기서 f_{ij}는 j번째 문서의 i번째 단어의 빈도수를 의미한다.

3.1 콘텐츠와 키워드 추출을 통한 학회 정보 확장

웹사이트의 콘텐츠를 활용한 학회정보의 확장은 두 단계로 구분된다. 첫 번째 단계는 학회 웹사이트에 존재하는 웹 콘텐츠 중 주요 콘텐츠를 추출하는 과정으로 부족한 학회의 정보량을 보충한다. 이 과정은 웹 콘텐츠 중 불필요한 부가 정보를 제거하는 과정으로 특정 학회의 주제를 명확히 파악할 수 있는 정보만 추출하기는 어렵다는 한계점이 있다. 따라서 두 번째 단계에서 주요 콘텐츠 내의 정보 중 학회주

제와 직접적으로 연관된 토픽 키워드 정보를 선정하여 분류의 성능을 향상시키고자 한다.

3.1.1 주요 콘텐츠 추출

학회 웹사이트는 학회정보와 관련이 없는 부가적인 정보들을 포함하고 있기 때문에 학회 정보와 관련된 콘텐츠만 추출하는 과정이 반드시 필요하다. 예를 들면, 주로 웹을 이용하기 위한 메뉴, 관련학회 사이트 링크, 다양한 광고와 이미지 등을 포함한다. 대부분 이들은 학회의 분야를 파악할 수 있는 정보와 관련이 없어 웹 콘텐츠를 활용할 때 방해가 된다. 즉, 학회 웹사이트의 웹 콘텐츠 추출 시, 주제 관련 내용과 함께 이러한 잡음도 같이 수집되어 시스템의 성능뿐 아니라 분류의 효율성도 저하시킨다[12].

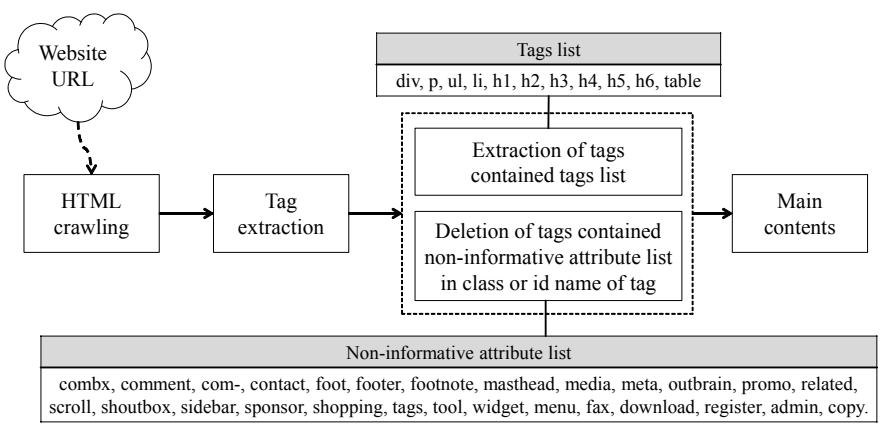
따라서 이를 해결하기 위해 HTML의 태그들 중 정보성 내용을 포함하는 태그를 추출하여 활용한다. 추출 기법은 기 개발된 HTML 콘텐츠 추출 알고리즘을 활용하며, 이는 <Figure 3>과 같이 진행된다[16].

먼저, 학회 웹사이트 URL을 통해 수집된 HTML 문서에서 태그 목록(Tags list)에 포함된 태그만 추출한다. 태그 목록은 문단, 문장을 구성하는 텍스트 태그인 , <p>, , 테이블 태그인 <table>, 헤더 태그인 <h>, 목록 태그인 과 을 포함한다. 학회소개글 및 토픽 목록과 같은 주요 정보는 모두 텍스트로 이루어져있기 때문에 텍스트와 관련된 태그 속성은 모두 포함한다.

추출된 태그 내에서도 태그의 클래스나 ID 속성에 'Footer' 또는 'Copyright'와 같은 단어가 포함될 경우, 학회의 연구 주제와는 무관한 불필요한 콘텐츠일 가능성이 높다. 따라서 'Non-informative attributes list'를 선정하여, 불필요한 속성 정보를 포함한 태그를 삭제한다. 최종적으로 추출된 태그를 주요 콘텐츠로 정의한다.

3.1.2 토픽 키워드 선정

학회명칭은 짧은 정보이지만 학회의 주제를 가장 직접적으로 대표하는 유의미한 정보이기 때문에, 앞서 추출된 주요 콘텐츠로부터 토픽 키워드를 식별하기 위한 기준 정보로 활용한다.



〈Figure 3〉 Main Contents Extraction

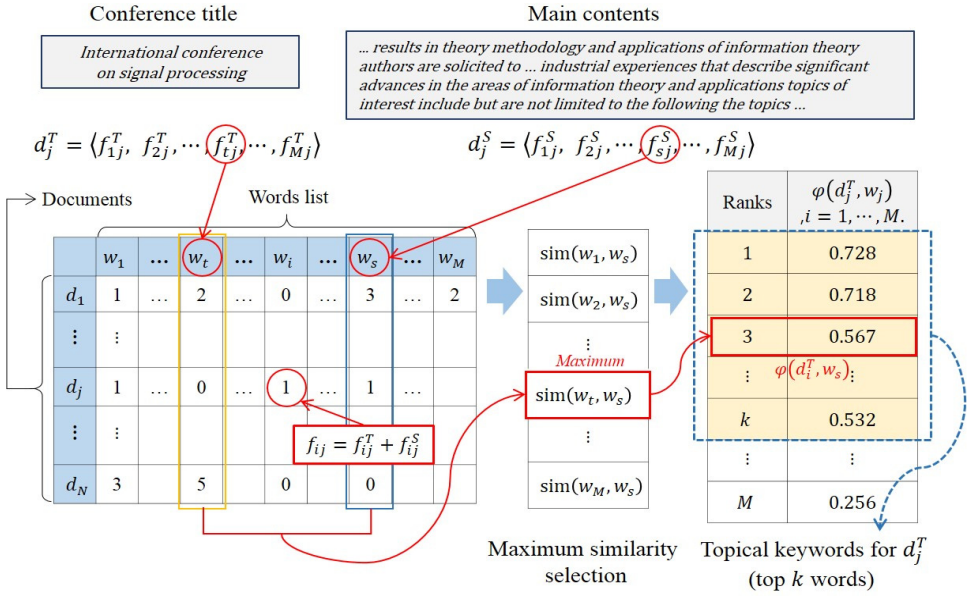
따라서 본 절에서는 주어진 학회명칭을 기준으로 주요 콘텐츠의 단어들 중 유사도가 높은 k개의 단어를 선정하여 토픽 키워드로 정의한다.

구체적으로 이 유사도는 학회 문서집합 내에서 학회명칭의 단어와 주요 콘텐츠의 단어 간의 정보량을 고려해야 하므로 다이스 유사도 기법을 활용한다. 다이스 유사도 기법은 단어의 동시 출현 빈도를 기반으로 두 단어의 의미적인 유사도를 계산하기 때문에 문서집합에서 두 단어의 정보량을 고려한 유사도 측정이 가능하다[1].

두 단어의 유사도는 <Figure 4>와 같이 수집한 N개의 학회정보 문서집합을 기준으로 계산되며, 문서 d_j, $j=1,\cdots,N$에서 학회명칭으로만 이루어진 문서는 d_j^T, $j=1,\cdots,N$ 주요 콘텐츠의 문서는 d_j^S, $j=1,\cdots,N$로 표현한다. 문서 d_j^T와 d_j^S는 각 문서에서 출현하는 단어의 빈도수로 구성되며, $d_j^T=\langle f_{1j}^T, f_{2j}^T, \cdots, f_{Mj}^T\rangle$, $d_j^S=\langle f_{1j}^S, f_{2j}^S, \cdots, f_{Mj}^S\rangle$와 같이 표현된다. 즉, 문서 d_j의 단어 w_i의 총 빈도수 f_{ij}는 학회명칭의 i번째 단어 빈도수 f_{ij}^T와 주요 콘텐츠의 i번째 단어 빈도수 f_{ij}^S의 합과 같다. 이를 기반으로 t번째 단어 w_t, $t=1,\cdots,M$와 s번째 단어 w_s, $s=1,\cdots,M$의 유사도 점수는 식 (1)과 같이 계산된다.

$$\mathrm{sim}(w_t,w_s)=\frac{2\sum_{j=1}^{N}\delta(f_{tj})\cdot\ \delta(f_{sj})}{\sum_{j=1}^{N}\delta(f_{tj})+\sum_{j=1}^{N}\delta(f_{sj})} \quad (1)$$

여기서, $\delta(\cdot)$는 주어진 값이 0 이하이면 0, 그렇지 않으면 1을 반환하는 판별함수이다. 따라서 식 (1)의 $\sum_{j=1}^{N}\delta(f_{tj})$는 단어 w_t가 출현한 문서의



〈Figure 4〉 Topical Keywords Selection based on Dice Similarity

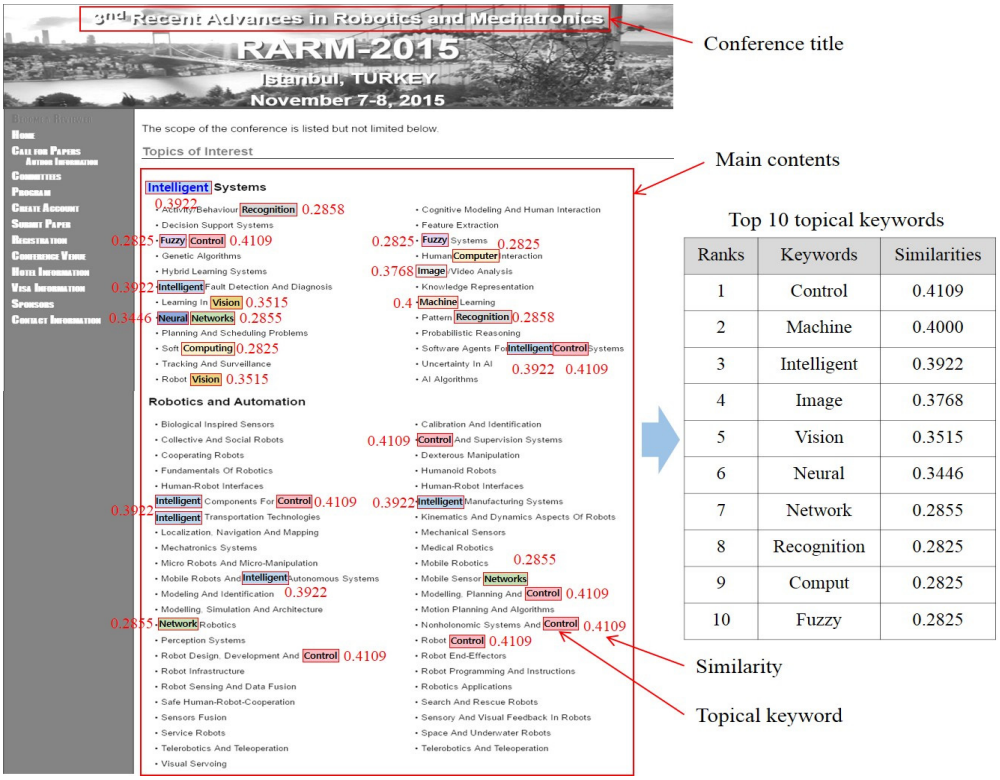
개수를 의미하고, $\sum_{j=1}^{N}\delta(f_{sj})$는 단어 w_s가 출현한 문서의 개수를 의미한다. 그리고 $\sum_{j=1}^{N}\delta(f_{tj})\cdot\delta(f_{sj})$는 두 단어가 함께 출현한 문서의 개수를 의미한다. 그러므로 $\mathrm{sim}(w_t, w_s)$는 단어 w_t 또는 w_s가 출현하는 문서들 중 동시에 출현하는 문서의 비율을 측정하는 척도로써, 두 단어가 수집된 학회 문서집합에서 얼마나 연관된 정보인지 파악할 수 있다.

학회명칭 문서 d_j^T, $j=1,\cdots,N$와 주요 콘텐츠의 단어 w_s,$s=1,\cdots,M$ 간의 최종 유사도 점수 $\varphi(d_j^T, w_s)$는 식 (1)에서 계산된 문서 d_j^T 내의 모든 단어와 주요콘텐츠 단어 w_s의 유사도 점수 중 가장 높은 점수로 정의하며, 식 (2)와 같이 표현된다. 여기서 $\delta(f_{tj}^T)\cdot\delta(f_{sj}^S)$는 단어 w_t와 w_s가 각각 j번째 학회명칭 문서 d_j^T와 주요 콘텐츠 문서 d_j^S에서의 출현 여부를 나타낸다.

$$\varphi(d_j^T, w_s) = \max_t\{\mathrm{sim}(w_t, w_s)\cdot\delta(f_{tj}^T)\cdot\delta(f_{sj}^S)\} \quad (2)$$

최종적으로 각 문서의 토픽 키워드는 유사도 점수 $\varphi(d_j^T, w_i)$, $j=1,\cdots,N$, $i=1,\cdots,M$에 따라 상위 k개의 단어가 선정된다.

<Figure 5>는 실제 특정 학회 웹사이트에서 추출된 주요 콘텐츠 및 토픽 키워드를 나타낸다.



Ranks	Keywords	Similarities
1	Control	0.4109
2	Machine	0.4000
3	Intelligent	0.3922
4	Image	0.3768
5	Vision	0.3515
6	Neural	0.3446
7	Network	0.2855
8	Recognition	0.2825
9	Comput	0.2825
10	Fuzzy	0.2825

〈Figure 5〉 Example of Extracted Main Contents and Topical Keywords

토픽 키워드는 상위 10개의 키워드를 추출했을 때의 결과를 나타낸다. 예를 들어, 학회명칭의 단어인 'Robotic', 'Mechatronics'와 유사도 점수가 높은 키워드는 'Control', 'Machine', 'Intelligent'라고 할 수 있다.

3.2 주제별 학회 자동 분류

3.2.1 토픽 키워드의 단어 가중치 설정

앞의 절에서 선정된 k개의 토픽 키워드는 가중치 부여를 통해 학습 데이터의 주제적 측면을 강화하고자 한다. 기본적으로 문서 내부에 존재하는 단어의 상대적인 가중치 z_{ij}, $i=1, \cdots, M$, $j=1, \cdots, N$를 설정하기 위해 TF-IDF 가중치 모델을 사용하여 식 (3)과 같이 표현된다[13].

$$z_{ij}=\lambda_{ij}\cdot \frac{f_{ij}}{\sum_{i=1}^{M} f_{ij}}\cdot \log\left(\frac{N}{1+\sum_{j=1}^{N}\delta(f_{ij})}\right) \quad (3)$$

식 (3)의 λ_{ij}, $i=1, \cdots, M$, $j=1, \cdots, N$는 강화계수를 나타내는데 이는 단어가 학회의 주제와 연관된 정도를 나타낸다. 주제와의 연관성은 선정된 토픽 키워드를 기준으로 결정되며, 문서 d_j, $j=1, \cdots, N$의 단어 w_i, $i=1, \cdots, M$가 토픽 키워드일 경우 λ_{ij}는 특정 값을 설정하고, 아니면 1의 값을 갖게 된다.

그리고 단어 빈도를 의미하는 TF는 문서 d_j에 모든 단어의 출현 횟수 중 단어 w_i가 출현한 횟수 f_{ij}를 나타낸다. 다음으로 역문서 빈도를 의미하는 IDF는 전체 문서의 수 N을 w_i를 포함한 문서의 수로 나눈 뒤 로그를 취함으로써 얻어진다[11]. 여기서 분모가 0이 되는 것을 방지하기 위해 분모에 1을 더해준다[22].

마지막으로 각 단어의 가중치로 벡터 공간 모델(Vector space model)에 단어 벡터를 생성한다. 벡터 공간 모델은 텍스트 문서를 벡터로 표현하는 모델로써 문서 d_j의 벡터 $Z_j \in R^M$는 M차원의 벡터로, 단어 가중치를 활용해 $Z_j=\langle z_{1j}, z_{2j}, \cdots, z_{Mj}\rangle$, $j=1, \cdots, N$로 정의된다.

3.2.2 특성추출(Feature Selection)

특성추출은 전체 문서집합 R^M에서 출현하는 모든 M개의 단어 중 분류에 결정적인 영향을 미치는 M'개의 단어를 찾는 기법이다. 본 연구에서는 상관관계 추론 기법인 피어슨 상관계수의 우선순위에 따라 선정된다.

상관관계 추론 기법은 단어 w_i, $i=1, \cdots, M$의 특정 클래스 c_l, $l=1, \cdots, L$에 대한 상관관계를 측정하여 각 단어의 각 클래스에 대한 연관정도를 판단한다. 상관관계 분석에 의해 얻어지는 상관계수는 -1에서 1 사이의 값을 갖으며, 상관계수의 절댓값이 1과 가까울수록 완전한 상관관계라고 할 수 있다[18]. 이는 식 (4)와 같이 계산되며, 여기서 c_{lj}, $l=1, \cdots, L$, $j=1, \cdots, N$는 d_j가 클래스 c_l에 속하는지 여부에 따라 참이면 1, 아니면 0의 값을 갖는다.

$$\rho(w_i, c_l)=\frac{N\left(\sum_{j=1}^{N} z_{ij}c_{lj}\right)-\left(\sum_{j=1}^{N} z_{ij}\right)\left(\sum_{j=1}^{N} c_{lj}\right)}{\sqrt{\left[N\sum_{j=1}^{N} z_{ij}^2-\left(\sum_{j=1}^{N} z_{ij}\right)^2\right]\left[N\sum_{j=1}^{N} c_{lj}^2-\left(\sum_{j=1}^{N} c_{lj}\right)^2\right]}} \quad (4)$$

그리고 단어 w_i, $i=1, \cdots, M$의 피어슨 상관관계 점수 $\rho(w_i)$는 식 (5)와 같이 L개의 점수 중 가장 큰 값이 된다.

$$\rho(w_i)=\max_l\{\rho(w_i,\ c_l)\},\quad l=1,\ \cdots,\ L. \tag{5}$$

최종적으로 $\rho(w_i)$ 값이 높은 단어순으로 M' 개의 단어를 선정하고, 선정되지 않은 단어의 가중치는 0의 값이 부여됨으로써 분류에서 제외된다.

3.2.3 분류(Classification)

본 연구에서는 선형 SVM(Support vector machine)을 사용하며, 이는 입력 데이터가 많은 차원을 가진 경우에 좋은 결과를 보이고 선형 커널이 텍스트 분류에 있어서 매우 적합한 것으로 알려져 있다[7].

하지만 본 연구에서는 $L(>2)$개의 클래스 중 하나로 분류하는 문제를 다루고 있으므로 이진 분류기인 선형 SVM을 여러 개를 조합하여 사용한다. 그러므로 각 쌍마다 분류기를 생성하는 방법을 활용하여, $L(L-1)/2$개의 모든 조합에 대하여 SVM 분류기를 학습시킨다[9]. 클래스 c_g와 c_l에 대한 판별함수는 두 클래스로 이루어진 데이터 집합 간의 최대 여백을 갖는 판별 경계를 찾아 학회 문서를 분류한다. 새로운 문서 d에 대한 입력벡터를 Z이라 했을 때, 판별함수 $F_{gl}(Z)$, $g=1,\ \cdots,\ L$, $l=1,\ \cdots,\ L$, $g\neq l$는 식 (6)과 같이 정의된다[4].

〈Table 2〉 Assignment Matrix

l \ g	1	2	3	⋯	L
1	-	-1	1	⋯	-1
2	1	-	-1	⋯	-1
3	-1	1	-	⋯	-1
⋯	⋯	⋯	⋯	-	1
L	1	1	1	-1	-

$$F_{gl}(Z)=\mathrm{w}_{gl}\cdot Z+b_{gl} \tag{6}$$

여기서 w_{gl}은 클래스 c_g와 c_l에 대한 판별 경계의 법선 벡터, b_{gl}는 기준치를 나타내며, 식 (7)과 같이 학습 데이터로부터 최적의 w_{gl}와 b_{gl}를 찾는 조건부 최적화 문제를 통해 생성된다. 식 (7)에서 y_j, $j=1,\cdots,N$의 값은 j번째 문서가 c_g에 속하면 +1, c_l에 속하면 -1 값이 부여된다.

$$\begin{aligned}&\arg\min_{(\mathrm{w}_{gl},b_{gl})}\|\mathrm{w}_{gl}\|\\&\text{s.t. } y_j(\mathrm{w}_{gl}\cdot Z_{glj}-b_{gl})\geq 1\end{aligned} \tag{7}$$

이렇게 얻어진 판별함수 $F_{gl}(Z)$에서 새로운 문서 d가 c_g로 분류되면 +1, 아니면 -1의 값이 할당된다. <Table 2>는 특정 한 문서에서 모든 클래스의 조합에 따른 분류 결과를 나타내며, 새로운 문서 d에 할당되는 최종 클래스 c_{g^*}는 식 (8)을 통해 도출된다. 즉, <Table 2>의 g, $g=1,\ \cdots,\ L$에 대한 각 열의 합계 중에서 가장 큰 값을 갖는 g^*가 최종 클래스 c_{g^*}로 할당된다.

$$\begin{aligned}&c_{g^*}=\arg\max_{g^*}\sum_{l=1}^{L}F_{gl}(Z),\\&g=1,\ \cdots,\ L,\ l=1,\ \cdots,\ L,\ g\neq l\end{aligned} \tag{8}$$

4. 실험 및 구현

4.1 실험 환경

본 연구에서 제안한 주제별 학회 분류기법을 평가하기 위해 실제 학회 검색 사이트에서 제공하는 학회 분야 카테고리에 따른 데이터

를 수집하였다[21]. 데이터는 크게 9개의 카테고리로 구성되며, 분류모델의 정확한 성능 측정을 위해 카테고리별 데이터의 개수를 동일한 비율로 구성하였다(<Table 3> 참고). 수집한 문서 집합은 총 7,875개이며, 문서집합의 80%인 6,300개를 무작위로 추출하여 학습을 위한 학습 데이터로, 나머지 20%는 분류기의 성능 평가를 위한 실험 데이터로 활용하였다.

〈Table 3〉 Description of Dataset According to Categories

Class No.	Categories	Numbers of instances
1	Mechanical Engineering	875
2	Computer Science	875
3	Electrical and Electronic Engineering	875
4	Industrial Engineering	875
5	Civil and Environmental Engineering	875
6	Medicine	875
7	Education	875
8	Nature Science	875
9	Social Science	875

또한, 텍스트 분류 알고리즘은 정형화된 데이터 구조를 대상으로 수행되기 때문에 수집된 비정형 데이터를 분석 가능한 데이터로 정제하는 과정이 필요하다[15]. 따라서 본 연구에서는 데이터 정제를 위해 다음과 같은 과정을 거쳤다. 우선, 문장 형태의 정보들을 Alphabetic tokenizer를 사용하여 숫자나 기호와 같은 단어를 제거하고 모든 문자를 소문자로 변환하여 단어형태로 분리하였다. 다음으로 Porter stemming algorithm을 사용하여 단어에서 어근을 추출하거나 접사를 제거하는 작업인 어간 추출(Stemming) 단계를 거쳤다[20]. 마지막으로, 의미를 가지는 어휘만을 추출하기 위해서 불용어(Stopword)를 제거하였다. 입력 데이터는 모두 영어로 구성되어 있으며, 관사, 조동사, 전치사, 접속사, 대명사 등은 불용어로 처리되었다.

4.2 평가 방법

본 논문의 성능 평가를 위해서는 텍스트 분류 문제에서 일반적으로 많이 사용되는 정확도(Accuracy)를 사용하였다[17]. 예측 결과는 <Table 4>와 같이 각 문서의 예측된 클래스가 실제 클래스와 일치하는지 여부에 따라 가능한 모든 경우를 4가지 범주로 분류할 수 있다. 정의된 혼동행렬(Confusion matrix)의 표기법에 기초하여, 분류의 성능은 식 (8)과 같이 분류의 정확도 관점에서 평가된다.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{8}$$

〈Table 4〉 Confusion Matrix

	Predicted positive	Predicted negative
Actual positive	True positive (*TP*)	False negative (*FN*)
Actual negative	False positive (*FP*)	True negative (*TN*)

4.3 분류기법 설정

본 절에서는 추출한 주요 콘텐츠와 토픽 키워드의 선정의 효과를 비교하기 위해 제안한 방법들을 다음과 같이 세 가지 기법으로 구성

하였다. 첫 번째 α-기법은 학회명칭을 활용한 분류기법이며, 다른 기법의 성능을 평가하기 위한 기준 분류기법이다. 두 번째 β-기법은 학회명칭과 <Figure 3>을 통해 제안된 주요 콘텐츠를 활용한 분류기법이다. 즉, 토픽 키워드를 고려하지 않으므로 $k=0$인 기법이다. 마지막으로, γ-기법은 학회명칭과 주요 콘텐츠를 포함하며, <Figure 4>를 통해 선정한 토픽 키워드까지 모두 고려한 분류기법으로 토픽 키워드의 선정이 모델 성능에 미치는 영향을 평가하고자 하였다.

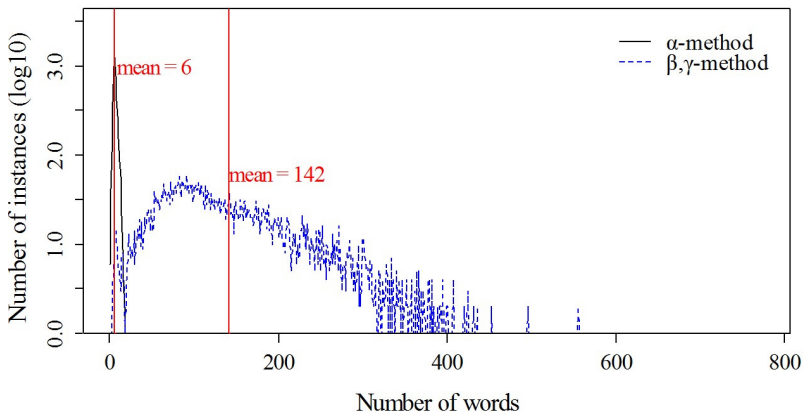
4.4 실험 결과 및 분석

<Figure 6>는 기존 학회정보에서 정보부족 문제가 존재함을 보여주며, 제안한 기법을 통해 이를 완화했음을 확인할 수 있다. 그래프의 가로축은 각 문서에 출현하는 고유 단어의 개수를 의미하고, 세로축은 문서 개수의 로그 값을 의미한다. 학회명칭을 기반으로 한 α-기법의 6,300개의 문서가 평균적으로 6개의 단어를 포함한다. 반면에 β와 γ-기법의 문서의 경우 평균적으로 142개의 단어를 포함함을 확인할 수 있다. 이는 웹 콘텐츠에서 추출한 주요 콘텐츠가 풍부한 데이터양의 확보에 충분한 역할을 수행했다고 판단할 수 있다.

다음으로 <Table 5>는 제3.2.2절에서 선택된 단어의 개수 M'와 토픽 키워드의 개수 k의 조합에 따라서 γ-기법의 성능변화를 보여준다. 여기서 토픽키워드에 대한 강화계수 λ_{ij}는 경험상 가장 좋은 성능을 보였던 2로 설정하였다.

<Table 5>에서 M'이 3,000개일 때 가장 높은 정확도를 보이고, 3,000개 이상일 때 다시 감소하는 것을 확인할 수 있다. 이는 무조건 많은 양의 데이터를 학습시키는 것 보다는 원본 데이터에서 각 클래스와 상관관계가 낮은 특성들을 제외시키고 높은 순의 특성들을 기반으로 학습하는 것이 보다 정확한 분류기의 생성을 가능하게 함을 보여준다.

또한, γ-기법은 $k=0$인 β-기법에 비해 대부분이 높은 정확도를 보였다. 따라서 토픽 키워드의 선정으로 학습 정보의 주제적 측면을 강화한 γ-기법이 주요 콘텐츠를 그대로 사용하는 β-기법보다 좋은 성능을 보였다고 판단



〈Figure 6〉 Instance Distribution According to Number of Words

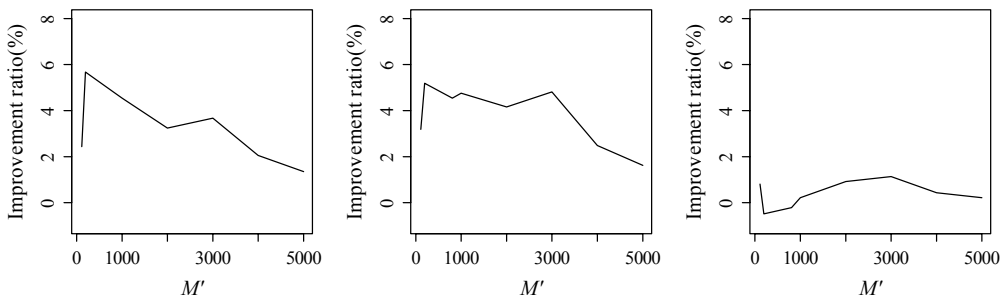
〈Table 5〉 Classification Performances of γ-Method by Varying M' and k

M'	k	Accuracy	M'	k	Accuracy	M'	k	Accuracy
100	0	0.662	800	0	0.753	3000	0	0.778
	5	0.667		5	0.752		5	0.786
	10	0.669		10	0.750		10	0.783
	15	0.660		15	0.752		15	0.784
	20	0.667		20	0.755		20	0.783
300	0	0.724	2000	0	0.766	5000	0	0.771
	5	0.729		5	0.772		5	0.772
	10	0.731		10	0.774		10	0.774
	15	0.721		15	0.771		15	0.774
	20	0.719		20	0.769		20	0.772

할 수 있다.

특히 M'= 3000일 때, k=5에서 가장 높은 정확도를 보였으며, k= 20일 때는 오히려 정확도가 감소하였다. 이는 k가 20보다 더 높아지더라도, 분류의 성능 향상에 미치는 영향이 미비할 것이며, 오히려 과도하게 많은 양의 단어를 선정할 경우 분류에 부정적인 영향을 끼침을 함축한다. 그러므로 주어진 환경에서는 파라미터 k의 값을 5로 설정하는 것이 적절하였으며, 분류의 성능을 향상시킬 뿐만 아니라 계산 비용도 절감할 수 있다.

<Figure 7>은 분류기법 간의 성능향상 정도를 분석하고자 하였다. 여기서, α-기법은 최대 단어의 개수가 2,162개 이므로 M'이 그 이상이 되더라도 단어의 개수는 2,162개이다. <Figure 7> (a)와 (b)의 그래프에서 보면, M'이 커질수록 α-기법 대비 β와 γ-기법의 개선 정도가 낮아지는 경향이 있었다. 이는 M'이 증가함에 따라 α-기법도 충분한 정보를 얻음으로써 양질의 데이터 학습이 가능해지기 때문에 α-기법의 분류 성능이 향상되는 것이다. 하지만 학습 데이터양이 적은 상황에서는 β와



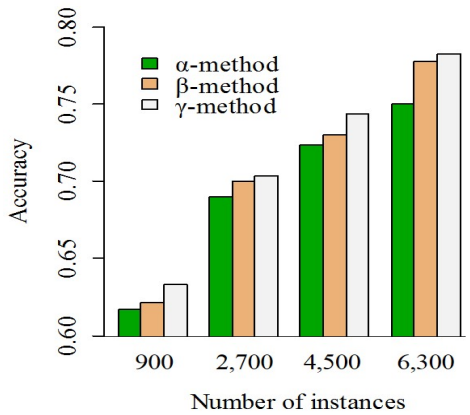
(a) β over α-method (b) γ over α-method (c) γ over β-method

〈Figure 7〉 Improvement Ratio According to the Numbers of Selected Features in Terms of Accuracy Under k= 5

γ-기법의 개선 정도가 매우 높아 분류의 성능을 크게 향상시킬 수 있다.

다음으로, <Figure 7> (c)는 β-기법 대비 γ-기법의 개선 정도가 미비하였다. γ-기법은 단어의 개수가 늘어나는 것이 아니라, 토픽 키워드에 대한 가중치에 차별성을 주는 것이기 때문에 단어의 개수 M'과 상관없이 성능 변화가 상대적으로 작게 나타남을 확인할 수 있다.

마지막으로 <Figure 8>는 데이터의 개수에 따른 성능을 분석하기 위해 학습 데이터를 6,300개, 4,500개, 2,700개, 900개로 다르게 구성하였다. 학습 데이터의 개수가 증가함에 따라 분류성능이 증가함을 확인할 수 있으며, 6,300개일 때 α, β, γ-기법은 각각 0.750, 0.778, 0.783로 각 기법에서 가장 높은 정확도를 보였다.



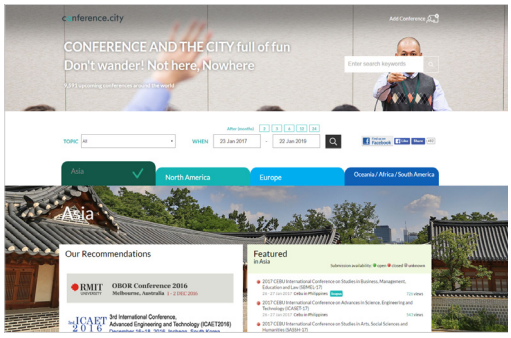
〈Figure 8〉 Accuracy of Classification used According to the Number of Training Instances under $k=20$

또한, 학습 데이터 개수에 상관없이 β와 γ-기법이 α-기법에 비해 항상 성능이 좋았으며, 이는 추가적인 정보의 활용이 분류에 긍정적인 영향을 끼쳤음을 보여준다. 따라서 제안된 기법은 정보 부족문제를 해결하였을 뿐만 아니라 질이 좋은 데이터로써 분류의 성능을 향상시켰다고 판단할 수 있다.

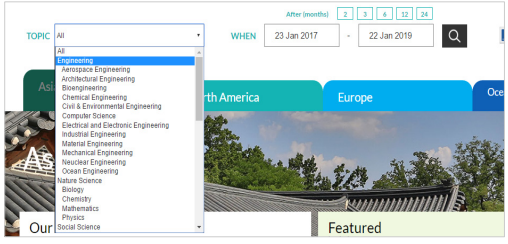
4.5 서비스 적용 사례

제안한 분류기법은 <Figure 9> (a)와 같이 전 세계의 학회정보 검색 서비스인 'Conference.City' (http://www.conferece.city)에 적용되었다[3].

적용된 분류기법은 새로 등록된 학회정보에 대해 자동으로 주제별 분류를 수행함으로써 분류의 서비스 질을 향상시켰으며, <Figure 9> (b)와 같이 사용자들에게 학문 분야에 따른 주제별 학회정보 검색 서비스를 제공하고 있다.



(a)



(b)

〈Figure 9〉 Screenshots of the Academic Conference Categorization Service Using the Proposed Methods

5. 결 론

학회의 공식 웹사이트의 웹 콘텐츠들은 주제별 학회 분류의 성능을 향상시키는데 효과적인 정보를 제공한다. 본 논문에서는 학회분류의 성능 향상을 위한 보완적인 데이터로 웹 콘텐츠로부터 학회의 주제와 연관된 정보를 추출하는 기법을 제안한다.

제안된 방법의 성능 평가를 위해 학회정보 검색 사이트에서 수집한 실 데이터를 토대로 실험한 결과, 주요 콘텐츠와 토픽 키워드를 모두 포함한 데이터 집합의 성능이 가장 우수했다. 이는 웹 콘텐츠로부터 학회주제와 관련된 추가적인 데이터 추출이 주제별 학회 자동 분류에 긍정적인 영향을 끼쳤다고 판단할 수 있다.

이를 바탕으로 본 연구는 두 측면에서 다음과 같은 가치를 지닌다. 첫째, 방법론적인 측면에서 주요 콘텐츠와 토픽 키워드 추출기법을 통해 학회분류의 정보 부족문제를 완화하고, 분류의 성능을 성공적으로 향상시켰다. 또한, 정보 부족 문제와 같은 비슷한 상황에 직면한 다양한 텍스트 분류 연구의 지표가 될 것이라 기대한다. 둘째, 응용적인 측면에서 학회정보 검색 서비스의 적용을 통해 실질적인 사용성 향상에 기여하였으며, 학회 분야에 특화된 정보를 사용한 분류기법을 개발하여 연구주제에 따른 관련학회 탐색 과정에서 효율성 제공하였다.

References

[1] Cho, J., "A New Word Semantic Similarity Measure Method based on WordNet," Journal of Korean Institute of Information Technology, Vol. 11, No. 7, pp. 121-129, 2013.

[2] Ciravegna, F., "$(LP)^2$, An Adaptive Algorithm for Information Extraction from Web-related Texts," Proceeding of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining, 2001.

[3] Conference.city, "International Conference Search Engine," [URL] http://www.conference.city/.

[4] Cortes, C. and Vapnik, V., "Support Vector Networks," Machine Learning, Vol. 20, No. 3, pp. 273-297, 1995.

[5] Cox, C., Nicolson, J., Finkel, J. R., Manning, C., and Langley, P., "Template Sampling for Leveraging Domain Knowledge in Information Extraction," Proceeding of PASCAL Challenges Workshop, 2005.

[6] Eom, J., "Information Extraction Using a Hidden Markov Model," Thesis of Graduate School of Seoul National University, 2001.

[7] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proceeding of the 10th European Conference on Machine Learning, Vol. 1398, pp. 137-142, 1998.

[8] Kim, J., Park, S. B., and Lee, S. J., "Information Extraction from Call-for-Papers Using a Hidden Markov Model," Proceeding of 2005 Conference on the HCI Society of Korea, Vol. 2005, No. 1, pp. 967-972, 2005.

[9] Kreßel, U., "Pairwise Classification and Support Vector Machines," Advances in

Kernel Methods Support Vector Learning, pp. 255-268, 1999.

[10] Lazarinis, F., "Combining Information Retrieval with Information Extraction for Efficient Retrieval of Calls for Papers," Proceeding of IRSG'1998, 1998.

[11] Lee, S. and Kim, H., "Keyword Extraction from News Corpus using Modified TF-IDF," The Journal of Society for e-Business Studies, Vol. 14, No. 4, pp. 59-73, 2009.

[12] Lee, Y., "A Study on Extracting News Contents from News Web Pages," Journal of the Korean Society for Information Management, Vol. 26, No. 1, pp. 305-320, 2009.

[13] Leopold, E. and Kindermann, J., "Text Categorization with Support Vector Machines: How to Represent Texts in Input Space?," Machine Learning, Vol. 46, pp. 423-444, 2002.

[14] Li, Y., Bontcheva, K., and Cunningham, H., "Using Uneven Margins SVM and Perceptron for Information Extraction," Proceeding of the 9th Conference on Computational Natural Language Learning, 2005.

[15] Munková, D., Munk, M., and Vozár, M., "Data Pre-Processing Evaluation for Text Mining: Transaction/Sequence Model," 2013 International Conference on Computational Science, Vol. 18, pp. 1198-1207, 2013.

[16] ReadabilityBUNDLE Library, [URL] https://github.com/srijiths/readabilityBUNDLE.

[17] Roh, J.-H., Kim, H.-j., and Chang, J.-Y., "Improving Hypertext Classification Systems Through WordNet-based Feature Abstraction," The Journal of Society for e-Business Studies, Vol. 18, No. 2, pp. 95-110, 2013.

[18] Ryu, J., "Real-world Pattern Classifications Using Optimal Feature/Classifier Ensemble," Master's Theses for Graduate School of Seoul National University, 2002.

[19] Schneider, K., "Information Extraction from Calls for Papers with Conditional Random Fields and Layout Features," Artificial Intelligence Review, Vol. 25, No. 1, pp. 67-77, 2006.

[20] Sebastiani, F., "Machine Learning in Automated Text Categorization," ACM Computing Surveys, Vol. 34, No. 1, pp. 1-47, 2002.

[21] WikiCFP, "A Semantic wiki for Calls For Papers in Science and Technology Fields," [URL] http://www.wikicfp.com/cfp/.

[22] Wikipedia, "TF-IDF," [URL] https://ko.wikipedia.org/wiki/TF-IDF.

[23] Xia, J., Wen, K., Li, R. and Gu, X., "Optimizing Academic Conference Classification using Social Tags," 2010 13th IEEE International Conference on Computational Science and Engineering, pp. 289-294, 2010.

[24] Xin, X., Li, J., Tang, J., and Luo, Q., "Academic Conference Homepage Understanding Using Constrained Hierarchical Conditional Random Fields," In Proceeding of International Conference on Information and Knowledge Management, pp. 1301-1310, 2008.

저 자 소 개

이수경 (E-mail: dltnrud1212@naver.com)
2012년 인천대학교 산업경영공학과 (학사)
2016년~현재 인천대학교 산업경영공학과 (석사과정)
관심분야 텍스트 마이닝, 통계적 기계학습

김관호 (E-mail: khokim@inu.ac.kr)
2006년 동국대학교 정보시스템전공 (학사)
2012년 서울대학교 산업공학과 (박사)
2013년 경희대학교 (연구박사)
2014년~현재 인천대학교 산업경영공학과 교수
관심분야 통계적 기계학습, 빅데이터