ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

역사객체 기반의 기계학습 기법을 활용한 웹 문서의 시간정보 추출 방안 제안☆

A Proposal of Methods for Extracting Temporal Information of History-related Web Document based on Historical Objects Using Machine Learning Techniques

이 준[1*] 권 용 진[1]
Jun Lee KWON, YongJin

요 약

최근 검색엔진을 통한 정보검색 과정에서 특정 시구간 상황에 대응하는 문서를 검색하고자 하는 경우가 있다. 예를 들면, 임진왜란 이전의 시대적 상황과 관련된 문서를 검색하기 위해, 키워드 '임진왜란'으로 검색하면 시간에 관계없이 임진왜란 당시나 전후의 모든 문서가 검색되어 추가적인 작업이 요구된다. 또한, 역사관련 문서의 경우는 문서내용에 대응하는 시간 정보가 문서 생성시간과 일치하지 않는 경우가 대부분이다. 만약 웹 문서의 내용에 대응하는 시간 정보를 추출 할 수 있다면 효과적인 정보검색은 물론 다양한 응용에 적용 가능할 것이다. 따라서 본 논문은 문서 내용에 대응하는 시간정보 추출을 목적으로, 조선시대를 대상으로 한 역사문헌을 활용하여 조선시대 역사관련 문서의 시간추출에 대한 연구를 진행한다. 역사 문헌과 웹으로부터 수집된 역사관련 문서를 바탕으로 역사객체를 정의하고, 이를 기반으로 다양한 기계학습 기법을 활용하여 웹 문서의 시간정보 추출에 대한 가능성을 확인한다. 또한 기계학습 과정에 있어서 객체의 유사도에 기반 한 여과과정을 제안하고 이를 적용한 효율적인 시간정보 추출 및 정확도 향상에 대한 결과를 비교 분석한다.

☞ 주제어 : 시간정보, 시간추출, 기계학습, 유사도 여과법, 역사정보, 역사객체

ABSTRACT

In information retrieval process through search engine, some users want to retrieve several documents that are corresponding with specific time period situation. For example, if user wants to search a document that contains the situation before `Japanese invasions of Korea era', he may use the keyword `Japanese invasions of Korea' by using searching query. Then, search engine gives all of documents about `Japanese invasions of Korea' disregarding time period in order. It makes user to do an additional work. In addition, a large percentage of cases which is related to historical documents have different time period between generation date of a document and record time of contents. If time period in document contents can be extracted, it may facilitate effective information for retrieval and various applications. Consequently, we pursue a research extracting time period of Joseon era's historical documents by using historic literature for Joseon era in order to deduct the time period corresponding with document content in this paper. We define historical objects based on historic literature that was collected from web and confirm a possibility of extracting time period of web document by machine learning techniques. In addition to the machine learning techniques, we propose and apply the similarity filtering based on the comparison between the historical objects. Finally, we'll evaluate the result of temporal indexing accuracy and improvement.

☞ keyword : Temporal information, Termporal Extraction, Machine learning. Similarity filtering, Historical information, Historical Object

1. 서 론

최근 인터넷과 웹의 급속한 발달로 인하여 웹 상에는 방대한 양의 문서가 생성되어 축적되고 있고, 이러한 웹을 기반으로 탄생한 인터넷 검색엔진을 이용하여 사용자

[1] Dept. of Telecomm. and Info. Engineering, Korea Aerospace University. Goyang-si, Gyeonggi-do, Korea
* Corresponding author (jun@grrc.kau.ac.kr)
[Received 26 April 2015, Reviewed 28 April 2015, Accepted 19 June 2015]

☆ This work was supported by the GRRC program of Gyeonggi province. [GRRC2015-B01 Ambient mobile broadcasting service system development]

는 원하는 정보에 대한 검색을 수행한다. 현재 구글로 대표되는 인터넷 검색엔진은 사용자가 찾고자 하는 정보에 대한 대표 키워드를 쿼리로 입력 받아 그 키워드가 포함된 웹 문서들을 일정한 규칙에 따라 나열하는 방식으로 검색결과를 제공한다. 하지만 검색결과로 제공된 문서들은, 물론 해당 키워드가 포함되거나 관련된 문서일지라도 문서가 배경으로 하는 시점에 따라 문서의 내용은 서로 다른 경우가 많다.

예를 들어 최근에 유행하는 가요에 대한 정보를 얻기 위하여 최신가요라는 키워드를 입력하여 검색을 수행하면, '최신가요' 라는 키워드를 포함하는 모든 웹 문서가 검색결과로 제공된다. 물론 사용자가 의도했던 현재 시점의 최신가요에 대한 웹 문서가 검색 될 수도 있지만, 일반적으로 현재 검색엔진들은 웹 문서내용상의 시간적인 배경을 고려하지 않기 때문에 사용자는 검색결과로 제공된 웹 문서들에서 자신이 원하는 현재 시점의 최신가요에 대한 웹 문서를 다시 일일이 살펴봐야 한다.

한편 웹 문서에서 최신가요가 나타난 내용을 살펴보면 '2010년 5월 최신가요', '1990년대 최신가요' 등과 같이 그 당시의 시간적 배경을 가진 상황에서의 최신가요에 대한 내용이 작성되어 있다. 그러나 과거 시점에 대한 최신가요에 대한 내용이 작성되어 있을지라도 그 웹 문서를 작성한 시점은 과거의 그 시점이라고 특정할 수 없다. 왜냐하면 현재 관점에서 과거의 최신가요에 대한 내용을 기술한 웹 문서일 수도 있기 때문이다.

이처럼 웹 문서는 문서가 생성된 시점을 시간적 배경으로 갖는 것이 아니라, 그 내용에 따라 시간적인 배경이 다른 내용상의 특징을 갖는다. 웹 문서의 이런 내용상의 특징은 웹 문서의 내용이 특정시점이나 시구간과 같은 시간특성이 배경이 되어 기술되는 것으로부터 기인한다. 이런 시간특성 때문에 동일한 주제를 기술함에 있어서 패션, 음악과 같은 대중적인 유행이나 3G, 4G와 같은 통신기술과 같이 일정한 주제에 대한 세부적인 내용이 시대적 배경이나 시간적 흐름에 따라서 변화하는 것을 많은 문서에서 발견할 수 있다.

이와 같이 특정 주제에 관련된 웹 문서는 특정시점이나 시구간에 따라 그 내용이 달라질 수 있으므로, 그 내용의 배경이 되는 시간특성을 분명히 고려해야만 한다. 이렇게 문서의 내용이 내포하거나 배경으로 하는 특정시점 또는 시구간을 문서의 시간특성(Temporal Characteristic)이라고 하며, 이러한 문서의 시간특성은 그 문서의 시간정보(Temporal Information)로써 정보검색 분야에 중요한 요소로 활용 될 수 있다. 이러한 시간정보는 단순하게 문서의 메타데이터나 시간태그을 통해 파악할 수 있는 문서의 생성 및 수정 연, 월, 일과 같은 시간속성과는 다르게 문서의 내용이 실제로 내포하는 시간정보를 나타낸다는 점에서 의미가 있다.

그러나 현재 키워드 색인을 기반으로 관련문서들을 일괄적으로 나열하는 검색엔진들의 정보제공 방식으로는 이러한 문서의 시간정보를 충분히 반영하기 어렵다. 키워드 색인 기반 정보검색을 통해 제공된 검색결과에 대한 관점을 키워드에서 시간으로 달리하여 웹 문서의 내용이 포함하는 시간정보의 관점에서 바라본다면, 이는 시간으로 구분되지 않은 과거로부터 현재까지 전체기간 동안의 검색어와 관련 있는 모든 웹 문서들을 모아놓은 것에 불과하다. 이로 인해 검색시스템에서 시간에 따라 변하는 패션이나 특정 제품에 대한 변천사와 같이 특정 기간이나 특정 시점을 배경으로 한 정보를 검색하거나 시간에 따라 변화된 내용을 검색하고자 하는 사용자는 단지 키워드만을 이용하여 반복적인 정보검색 및 복잡한 여과과정을 수행하여야 하고, 이는 사용자에게 비효율적인 시간과 노력의 소모를 요구한다.

이러한 기존 검색엔진의 키워드 색인 방식과 웹 문서의 시간정보를 함께 고려하여 시간축을 기준으로 한 정보검색을 수행한다면 일정한 주제에 대한 시간에 따른 내용의 변화를 검색할 수 있을 것이다. 예를 들어 임진왜란 전후의 시대상황, S사와 A사의 스마트폰 개발 경쟁 비교 같은 새로운 형태의 정보검색이 가능할 것이다. 더불어 이러한 검색은 웹 문서가 가진 시간정보를 반영하기 때문에 사용자가 실제로 찾고자 하는 정보에 더욱 정확하게 접근 할 수 있도록 도와주며, 나아가 기존의 검색엔진이 제공하지 못하는 시간의 흐름에 따른 정보의 내용적인 변화를 검색하는데 있어서 도움을 줄 수 있다. 특히, 구체적으로 1년 혹은 10년과 같은 시간단위로 시구간을 선택하여 검색을 수행할 수 있다면, 예를 들어 한미 FTA시행 전 1년과 후 1년과 같은 시간적인 검색구간을 설정하여 시간에 따른 정보의 내용의 변화를 검색 가능할 것이다. 더불어 검색어에 대한 정보의 내용이 변하는 변곡점에 해당하는 시점과 같은 관련 핵심정보를 검색하기 용이 할 것이다.

이와 같은 웹 문서의 시간정보를 반영한 정보검색을 위해서는 기존 검색엔진에서 키워드를 기반으로 웹 문서들이 색인되는 것처럼, 먼저 웹 문서가 내포하는 시간정보를 추출하고 추출된 시간정보로 문서를 색인하는 과정이 필수적으로 선결되어야 한다. 이렇게 웹 문서의 시간정보를 추출하고 색인하는 과정을 시간색인(temporal

indexing)이라고 부르며, 최근 들어 이 분야에 대한 많은 연구들이 이루어지고 있다.

하지만 현재 시간색인에 대한 연구는 주로 정확한 시간이 명시된 뉴스기사나 웹 문서의 메타데이터 상의 시간 속성, 또는 'YYYY-MM-DD'와 같은 형태로 정규화 된 시간적인 표현을 추출하여 시간정보로 이용한 연구들이 대부분으로써, 문서가 내포하고 있는 실제 시간정보를 활용하는 것에는 한계가 있다. 만약 웹 문서의 내용을 분석하여 그 내용에 해당하는 시간정보를 추출하고 정보검색의 요소로써 사용할 수 있다면 실제 문서가 내포하는 시간정보를 반영하여 더욱 정확한 시간색인이 가능함과 더불어 차후 정보검색 분야에 상당한 도움이 될 것이다.

이와 같은 이유로 본 논문에서는 웹 문서 내용의 시간 추출에 대한 초기연구로써, 웹 상의 조선시대 역사 관련 문서를 대상으로 하여 해당 문서 내용의 정확한 시간정보 추출을 목표로 한다. 이를 위하여 조선시대 역사 문헌에서 추출한 역사객체를 기반으로 기계학습을 활용한 역사관련 웹 문서의 시간정보 추출 방법을 제안한다.

웹 상에서 역사에 대해 기록된 문서의 유형은 고문서나 고지도와 같이 과거에 작성된 자료를 디지털화 한 역사문헌들과 일반적인 사용자들에 의해 작성된 역사관련 웹 문서로 구분할 수 있다. 역사문헌은 비록 디지털화 되어 작성시점이 달라졌다고 생각할 수도 있지만 과거의 문헌을 그대로 디지털화 한 것뿐이므로 일반적으로 그 문헌이 실제 작성된 시기와 내용이 거의 일치한다고 볼 수 있다. 하지만 역사관련 웹 문서와 같은 경우에는 현재 시점에서 과거의 역사를 기술했으므로 작성시점과 내용상의 시간배경이 분명히 다르다. 이와 같이 역사정보는 웹 문서상의 작성시점과 내용상의 시간적 배경이 뚜렷하게 구분되는 특징으로 인하여 본 연구의 적합한 자료로써 고려되어 기반데이터로써 사용하였다. 특히 역사문헌은 이미 시간이 정확하게 알려진 과거의 사건이나 사실이 기록되어 있기 때문에 시간정보에 대한 파악이 용이하며, 본 논문에서 사용한 조선왕조실록과 같은 역사문헌은 일기와 같이 일별로 내용이 기술되어 있어 정확한 시간정보를 추출하고 판단하기 위한 훈련데이터로 매우 적합하다.

본 논문에서는 인간이 역사관련 웹 문서의 내용을 보고 그 내용이 내포하는 시간을 유추하는 방법에 기초하여 역사문헌 상에서 시간정보 판단에 주요요소가 되는 인물, 사건, 문화재, 지명을 역사객체로 정의하고, 정의된 역사객체를 기반으로 데이터 전처리 및 기계학습을 수행한다. 정확한 시간정보의 추출에 대한 가능성을 확인하기 위하여 기계학습의 이론적 유형별로 확률기반의 지도학습의 일종인 나이브 베이즈(Naïve bayes) 분류기, 반지도(semi-supervised)학습 알고리즘인 Co-EM, 객체기반의 K-NN을 각각 적용하여 학습을 수행하고 각각 추출된 시간정보의 정확성을 비교한다. 또한 기계학습 과정에 있어서 객체의 유사도에 기반 한 여과과정인 Similarity Filtering을 제안하고 이를 적용한 효율적인 시간정보 추출 및 정확도 향상에 대한 결과를 비교 분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로 시간정보를 활용한 정보검색과 시간정보 추출 및 색인에 대한 기존연구를 살펴본다. 3장에서는 기계학습에 사용된 역사문헌 데이터 집합의 구성 및 역사관련 웹 문서의 시간정보 추출에 필수적인 요소로 제안된 역사객체에 대하여 설명한다. 4장에서는 시간정보 추출을 위한 기계학습 알고리즘의 적용에 대하여 기술 하고 추출결과로써 정확도를 비교 분석한다. 5장에서는 본 논문에서 제안한 객체의 유사도에 기반 한 여과과정에 대해 소개한다. 6장에서는 본 논문의 결론 및 향후 연구를 논의 한다.

2. 관련 연구

본 논문은 시간정보(Temporal Information)의 활용을 위한 시간색인(Temporal indexing)에 대한 선행 연구로써 웹 문서의 시간정보 추출에 중점을 두고 있다. 특히 정보검색 서비스에 있어서 시간속성 활용의 이점에 대해 주목하고 있으며, 시간정보의 활용에 있어서 가장 선결되어야 할 분야인 시간정보 추출에 대한 연구에 중점을 둔다. 그래서 본 논문에서는, 역사를 대상으로 기술된 웹 문서를 이용하여 웹 문서의 내용이 내포하고 있는 시간정보를 추출하고 그 정확도를 비교 분석하여 시간정보 추출에 대한 가능성을 확인한다.

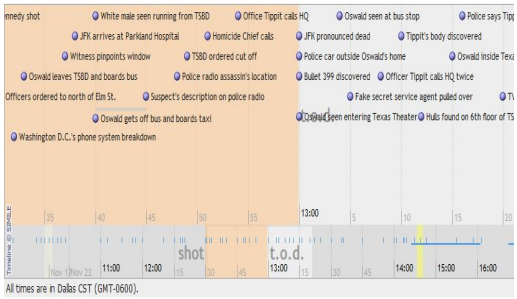
이 장에서는 정보검색 분야에서 시간정보를 활용한 최근 연구동향과 시간정보가 활용되는 응용분야에 대하여 살펴본다. 또한 본 논문이 목표로 하는 시간정보 추출 및 색인에 대한 연구 동향을 살펴본다.

2.1 시간정보를 활용한 정보검색

시간은 정보공간에 있어서 중요한 하나의 차원으로 고려될 수 있으며, 정보추출(Information extraction), 주제탐색(Topic-detection), 질의응답 시스템(Question-answering) 등 정보검색의 여러 분야에서 매우 유용하게 사용될 수 있다. 특히 웹 문서의 시간정보를 활용하여 정보검색의

결과를 재구성하거나, 검색결과 순위의 계산에 웹 문서의 생성시간 및 수정시간과 같은 속성을 반영하여 사용자가 검색한 정보를 더욱 정확한 검색결과로써 제공할 수 있다[1]. 그러나 기존의 검색엔진을 비롯한 정보검색 시스템에서는 웹 문서의 시간정보의 장점을 최대한 활용하지 못한 정보검색이 이루어지고 있다. 하지만 최근 몇 년 동안 정보검색에 있어서 시간정보의 중요성을 인식하고 이를 분석하고 개척하는데 있어서 흥미 있는 연구들이 진행되고 있다[2]

M. Gertz[6] 등은 정보검색 결과로 제공된 웹 문서들의 메타데이터 상에 명시된 시간속성을 추출하고 이를 기준으로 군집화를 수행하여 유사한 시간정보를 갖는 웹 문서들끼리 군집화한 후 검색결과를 재구성하는 연구를 수행하였다. 또한 유사한 연구로써 J.Makkonen[7] 등은 웹 문서의 시간정보가 정형화된 형태로써 기록된 뉴스기사들을 활용하여 각 기사가 작성되거나 수정된 시간을 추출하고, 추출된 시간정보들의 비교를 통해서 기사들 간 유사성을 측정하는 연구를 수행하였다. 검색결과 제공의 또 다른 측면에서 검색결과로 제공되는 문서의 발췌구문(snippet)의 생성에 있어서도 시간정보가 활용될 수 있다. O.Alonso [8] 등은 문서의 메타데이터상의 시간속성과 본문에 명시된 시간정보를 추출하고 이를 문서의 주제와 함께 고려하여 가장 적합한 본문의 일부를 발췌구문으로 생성하는 연구를 진행하였다.



(그림 1) Timeline을 활용한 JFK저격사건의 정보 표현
(Figure 1) Representation of John F Kennedy Assassination using Timeline

정보검색 결과제공에서의 활용뿐만 아니라 데이터 마이닝을 통해 추가적인 유용한 정보를 추출하기 위해 블로그의 본문과 더불어 블로그의 생성 및 수정 시간과 같은 시간정보를 추가적으로 활용한 연구[9], [그림 1]와 같이 문서의 주제를 검출하고 검출된 주제를 해당 기간(Timeline)상에 표현해주는 연구[10], Future Retrieval 분야와 결합하여 미래의 어떤 사실이나 객체간의 관계를 예측하는데 있어서 시간정보를 활용하는 연구[11] 등과 같이 다양한 정보검색 분야에서 시간정보가 활용 중에 있다.

더불어 최근 진행되고 있는 시간정보를 활용해서 정보의 일정한 출현 및 소멸시기와 같은 패턴을 분석하거나 반복되는 특징을 인식하는 Temporal pattern[12] 분야, 웹 문서 내용의 시간정보를 파악하여 정보의 최신성을 파악하고 이를 정보검색에 활용하는 Fresh Information Retrieval[13] 분야, 시간정보와 지리적인 공간정보를 동시에 고려하여 지도상에 시간의 순서에 따라 정보를 매핑 하거나 이동체에 대한 정보검색에 활용되는 Spatio-Temporal exploration [14] 등의 정보검색 관련 연구 분야에서 시간정보는 중요한 하나의 요소로써 활용되고 있다.

2.2 시간정보의 추출 및 색인

정보검색 분야에 시간정보를 활용하기 위해서는 웹 문서의 내용이 내포하는 시간정보를 추출해야 한다. 일반적으로 웹 문서의 시간속성은 문서의 생성일자나 최종 수정일과 같은 데이터를 통해 간단하게 추출할 수 있다.

하지만 정보검색의 목적을 위해서 정확한 시간정보를 이용하려면 단순한 문서의 메타정보뿐만 아니라, 문서의 내용상에 존재하는 시간정보를 추출 할 수 있어야 한다.

문서의 내용상에 존재하는 시간정보의 추출을 위해서 J.Pustejovsky[3] 등은 XML기반의 TimeML의 제안을 통해서 문맥상의 시간을 나타내는 표현(Temporal Expression)을 일자(Data), 시간(Time), 기간(Duration), 집단(Set)등의 4가지 유형으로 분류하였다. 또한 문장 속에서 시간표현에 해당하는 부분을 실제 시간을 나타내는 표준적인 형태, 즉 'YYYY-MM-DD'와 같은 형태로 정규화하여 시간정보를 사용하기 용이하도록 하는 시간정보의 정규화 표현을 제안하였다.

하지만 실제 문서에서 시간 표현은 위와 같은 4가지 유형뿐만 아니라 매우 다양하고 비정형적인 형태로 발생하게 때문에 정규화는 간단하게 수행되기 어렵다. 비 정형적인 웹 문서의 시간정보를 정확하게 추출하기 위해서 O.Alonso[4] 등은 웹 문서에 나타나는 시간표현 형태를 3가지로 구분하였다. 구분된 표현의 유형은 'December 2004'와 같이 직접적으로 시간과 대응되는 Explicit 표현, 'Labor Day 2008'과 같이 기념일이나 사건과 같은 특정 이름으로 명명되는 Implicit 표현, 'Last week'와 같이 문서의 메타데이터와의 비교를 통해 시간정보를 알 수 있

는 Relative 표현으로 정의하였다. 이렇게 시간표현을 세 가지 유형으로 정의함으로써 문서의 메타데이터상의 시간속성 뿐 만 아니라 문맥상에 존재하는 시간정보에 대하여 더욱 정확한 추출이 가능하도록 하였다.

이와 같은 시간정보의 추출과 정규화의 전반적인 과정을 시간 색인(Temporal indexing)이라고 부르며 시간 색인을 수행하는 기계를 시간 태거(Temporal tagger)라고 한다. 시간 태거들은 형태소 분석과 같은 기본적인 텍스트 처리 과정과 일련의 자연언어처리 과정을 거쳐 문장을 전 처리하고, 전처리 된 문장에서 기계학습 기술들을 활용하여 시간의 범위(boundary)를 확인한 후 시간정보를 추출하고 이를 정규화한다[5].

하지만 이러한 기존의 연구는 웹 문서에 존재하는 Time Stamp 또는 신문기사 상의 날짜정보 등과 같이 문서가 작성됨과 동시에 현재 날짜가 저장되는 속성을 가지고 수행하거나 내용상에 시간을 표현하는 정형화된 몇 가지 형식(YYYY-MM-DD)만을 사용해서 시간정보를 추출하는 방법만이 일반적으로 제안돼왔다. 또한 웹 문서상에 시간을 표현하는 표현이 존재하지 않으면 시간정보를 추출할 수 없는 단점이 있다. 따라서 웹 문서의 내용이 내포하거나 나타내고 있는 시간정보를 추출하는 방법에 대한 연구는 아직 연구가 많이 필요한 분야이며, 본 논문에서는 이러한 메타데이터상의 시간속성이나 내용상의 시간적인 표현 없이 문서의 시간정보를 추출하는 방법에 대한 연구를 진행하였다.

3. 기본 데이터 집합

본 장에서는 기계학습에 사용된 역사문헌 집합을 소개하고, 데이터 전 처리에 기준이 되는 역사객체에 대하여 정의하고 설명한다. 역사문헌은 과거의 그 당시 시점에서 사건이나 사실을 기록하기 때문에 비교적 시간정보에 대한 표현이 명확하며, 본 논문에서 사용한 조선왕조실록은 일기와 같이 일별로 내용이 기술되어 있어 시간정보를 추출하고 판단하기 위한 훈련데이터로 매우 적합하다.

본 논문에서는 기계학습을 위한 데이터의 정리 및 가공을 위해 역사객체를 정의하여 이를 기준으로 데이터를 전 처리한다. 적용할 훈련 및 실험 데이터는 역사객체추출을 위한 형태소 분석, 불용단어 제거 등의 전 처리가 이루어지며 역사관련 웹 문서의 수집 및 기계학습을 통한 시간정보 추출 또한 역사객체를 기반으로 수행한다.

3.1 기계학습을 위한 데이터 집합

기계학습에 사용할 기본 데이터 집합은 학습에 사용할 훈련 데이터와 추출된 시간정보의 정확도를 테스트하고 실제 결과를 살펴볼 실험데이터 두 부분으로 구성된다. 먼저 학습의 바탕이 되는 훈련데이터는 조선시대(1394 ~ 1911)를 대상으로 한 대표적인 역사서인 '조선왕조실록'을 사용하였다. 조선왕조실록은 조선시대 역대 임금들의 실록으로써, 태조부터 철종까지의 472년간에 걸친 25대 임금들의 1,893권의 실록을 말한다. 이 실록은 조정에서 일어나거나 보고되는 일들을 연, 월, 일 순서인 편년체 형태로 기록함에 따라 일정한 시간적인 흐름을 가지고 기술된다. 특이 매일 발생한 일에 대하여 그날의 날짜에 일기처럼 기록함으로써 해당내용에 대한 정확한 시간정보를 파악하는데 용이하다.

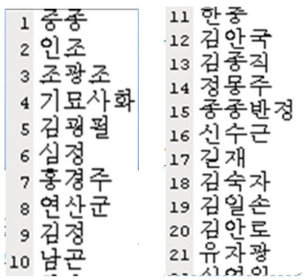
현재 실록 전권이 국문으로 디지털화 되어 웹 상에서 각 페이지를 열람 가능하며, 일정한 편집방법에 따라 디지털화 되어 있기 때문에 본 시스템에서는 조선왕조실록에 적합한 Wrapper를 제작하여 디지털화 된 조선왕조실록의 국역본을 수집하였다. 조선왕조실록은 제목, 본문, 서명, 분류, 용어풀이로 구분되어 있어 이에 적합한 Wrapper를 제작하여 수집하고 데이터베이스화 하였다. 이를 통해서 매우 정확한 시간정보를 가진 조선왕조실록을 기계학습의 훈련데이터로 확보 할 수 있다.

두 번째로 시간정보를 추출하고 추출결과로써 정확도를 비교할 테스트 데이터로 역사정보를 기술하고 있는 웹 문서를 수집하였다. 역사객체를 검색 쿼리로 하여 검색결과로 제공되는 웹 문서를 쿼리 당 100개씩 수집하고, 수집된 웹 문서는 각각 문단단위로 나누어 형태소 분석 등의 전 처리를 수행하였다. 특히 수집된 웹 문서들 중에서 전 처리를 수행한 후에 해당 문서의 역사객체가 10개 이하로 출현한 문서는 동음이의어와 같은 이유로 검색되어 수집된 문서로써, 역사와 관련된 내용을 기술하고 있다고 보기 어렵다고 판단하여 역사관련 웹 문서집합에서 제외하였다. 본 논문에서는 이와 같은 과정을 통해 전처리 되어 나누어진 역사관련 웹 문서의 문단 중에서 임의로 50개를 선정하여 테스트 데이터로 설정하고, 이 문단이 내포하는 시간정보를 기계학습을 통해 추출하고 문단의 실제 시간정보와 비교하여 정확도를 판단하였다.

3.2 역사객체 기반의 데이터 전처리

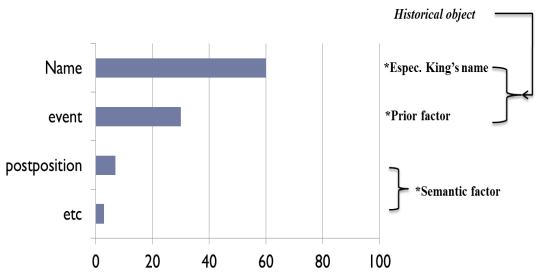
데이터 집합의 전 처리에 있어서 모든 데이터는 역사문헌에 기술되고 있는 객체들을 기준으로 처리된다. 여

기서 객체란 역사정보의 검색에 있어서 검색엔진을 통한 키워드 검색에서 쿼리로 사용되며, 주로 정보를 얻고자 하는 주요 대상이다. 이렇게 역사정보로써 검색이 요구되는 객체는 유형이 특정범위로 한정된다는 특징이 있다. 예를 들어, 대중적으로 많이 사용되는 '한국 역사정보 통합시스템' 은 디렉터리 분류를 통하여 역사정보를 제공한다. 일반적으로 디렉터리 분류는 검색이 요구되는 유형별로 수행하는 것이 대부분이다. '한국 역사정보 통합시스템' 의 디렉터리 분류의 경우 '인물, 지도, 유물, 도서, 연구자료' 등으로 분류되어 있고, 이를 객체의 점유비율로 정리하면 '인물(60%), 유물, 도서(25%), 지도, 지명(10%), 기타(5%)'로 구성됨을 알 수 있다. 또한 대중적으로 사용되는 역사백과사전인 '한국 민족문화 대백과사전' 의 콘텐츠 색인 분류는 '인명, 지명, 서명, 잠재(사건, 관직 등)'으로 분류되어 있다. 이러한 대중적인 역사정보 제공시스템들의 디렉터리 분류 또는 콘텐츠 색인으로 볼 때, 역사정보는 주로 '인명, 문화재, 지명, 사건'을 중심으로 기술되고 검색된다는 사실을 발견하였고, 이와 같은 사실을 통하여 본 논문에서는 넓은 의미의 역사정보를 '역사객체'로 정리하여 사용한다. 데이터의 가공을 위하여 한국어 형태소 분석기(KLT)[15]을 사용하였고, 이를 통해 다른 어형을 제외한 일반명사만을 선별하였다. 또한 선별된 일반명사 중에서 정의된 역사객체를 제외한 일반명사를 제거하였다. 이렇게 얻어진 역사객체는 약 28,000개 이며 [그림 2]는 역사객체의 일부 예이다.



(그림 2) 역사객체의 예
(Figure 2) Example of Historical Object

본 논문에서는 역사객체를 시간정보의 추출에 있어서 기준 요소로 사용한다. 이에 대한 근거는 설문을 통하여 인간이 역사문헌을 보고 해당 시간정보를 판단하는 요소를 조사함으로써 확보하였고, [그림 2]와 같은 결과를 나타내었다.



(그림 3) 인간의 역사문헌에 대한 시간정보 판단 기준 조사결과
(Figure 3) The survey result for criteria for discrimination of temporal information in human

설문조사는 역사에 대한 전문적인 지식을 가지지 않은 일반인 10명을 대상으로 수행하였으며, 개인당 역사를 기술한 임의의 문단을 10개씩 제공하고, 해당 문단이 내포하는 시간정보를 판단하는 기준으로써 사용한 주요 요소를 조사하였다. 조사결과로써 [그림 2]에서와 같이 문단 안에 포함된 인명, 특히 왕의 이름으로 시간을 유추하는 경우가 대부분이었고, 두 번째로 역사적인 사건이 시간 유추에 사용되는 요소로써 고려되었다. 그 외에 동사 또는 전, 후치사 등으로 시간을 판단하는 경향을 보였다. 이와 같은 결과로 볼 때, 인간은 주로 역사문헌에 기술된 역사객체를 기준으로 시간을 유추한다고 판단할 수 있고, 이에 따라 본 논문에서는 역사객체를 기준으로 형태소분석 등의 전 처리를 수행하고 기계학습을 통해 테스트 데이터의 시간정보를 추출한다.

4. 기계학습기법을 활용한 역사관련 웹 문서의 시간정보 추출

본 장에서는 기계학습 알고리즘 적용 및 시간정보 추출방법에 대하여 기술 한다. 정확한 시간정보의 추출에 대한 가능성을 확인하기 위하여 기계학습의 이론적 유형별로 확률기반의 지도학습의 일종인 나이브 베이즈 분류기, 반지도(semi-supervised) 학습 알고리즘인 Co-EM, 객체기반의 K-NN을 각각 적용하여 학습을 수행한다. 또한 본 논문에서 제안한 객체의 유사도에 기반 한 여과과정에 대해 소개하고 적용한다. 학습을 통해 추출된 역사관련 웹 문서의 시간정보는 구체적으로 조선시대 (1394~1911)년을 10년 단위로 구분한 연도로 결정되며, 50개의 실험데이터에서 추출된 시간정보를 실험데이터의 실제 시간정보와 비교하여 정확도를 판단한다.

4.1 나이브 베이즈 분류기를 통한 시간정보 추출

나이브 베이즈 분류기는 사건들 간에 강한 독립(independence)을 가정한 베이즈 정리(bayes' theorem)를 기반으로 하는 기본적이지만 대체적으로 좋은 성능을 보장하는 확률 분류기이다. 각 사건들이 서로 영향을 주지 않는 독립적인 관계라고 가정한다면 베이즈 정리에 의해 다음 (1)식과 같이 표현될 수 있다.

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)} \quad (1)$$

본 논문에서는 조선시대를 각 10년 단위로 나누고, 10년 단위의 연대를 하나의 분류로 하여 63개의 분류(1390-1910)를 갖는 다중분류 경우(Multiclass-case)로 적용한다. 조선왕조실록 상에서 각 10년 단위 연도마다 출현한 역사객체 출현횟수를 합산하여 각각의 역사객체에 대한 해당 연대의 베이즈 확률을 계산하고, 테스트 데이터에서 출현한 역사객체에 대하여 각 연대별 베이즈 확률을 종합한다. 이렇게 종합된 역사객체의 베이즈 확률이 최대의 확률 값을 갖는 연도를 해당 테스트 데이터의 시간정보로써 결정한다.

$$\text{classify}(f_1,\ldots,f_n) = \underset{c}{\text{argmax}}\ p(C_k{=}\ c_j)\prod_{i=1}^{n} p(F_i = f_i|C_k{=}\ c_j).$$

$$C_k: year, F_i: \text{objects} \quad (2)$$

즉 테스트 데이터에서 발생한 역사객체들에 대하여 조선왕조실록을 통해 미리 계산된 확률 값을 각 연대별로 곱하여 가장 높은 확률을 갖는 연대가 테스트 데이터가 분류될 연대가 된다. 그러므로 테스트 데이터에서 발생한 역사객체의 총 출현횟수가 많고, 발생된 역사객체의 종류가 가장 유사하게 출현한 연대가 결론적으로 높은 확률 값을 가지게 된다. 이에 따라 다음 (2)식을 만족하는 연대를 테스트 데이터의 시간정보로써 추출하며, 각 연대에서의 확률계산은 구체적으로 식(3)과 같이 이루어진다.

$$P(C|F_n) = argmax\left(\prod_{i=1}^{n} P(F_i|C_k)\right)_{k=1390}^{1910} \quad (3)$$

나이브 베이즈 분류기를 시간추출에 적용함에 있어서 고려사항으로써 모든 역사객체의 각 연도에서의 출현은 독립적이라는 가정을 가진다. 또한 테스트 데이터에서 출현하였지만 해당 연대에 한번도 출현하지 않음으로써 확률 값이 0으로 계산되는 0-확률문제(o probability value problem)를 방지하기 위해, 특정 연대에 출현하지 않음에도 역사객체의 출현 횟수를 확률계산의 결과에 영향을 최소화 하는 1로 고정한다.

이와 같은 방법으로 시간정보를 추출한 결과 전체 50개의 테스트 데이터 중에 23개의 시간정보를 정확하게 추출해내어 46%의 정확도를 보였다. 웹 문서의 시간정보를 추출함에 있어서 다소 낮은 정확도로 여겨질 수 있지만 문서가 내포하고 있는 시간을 약 50% 확률로 추출할 수 있는 가능성을 보였다는 점에서 의의가 있다.

나이브 베이즈 분류기를 적용한 결과를 살펴보면, 23개의 정확하게 추출된 테스트 데이터 중에서 문단에 출현한 역사객체의 개수가 10개 이상인 경우가 22개로써 많은 수의 역사객체가 출현한 경우에 높은 정확도를 보임을 확인 할 수 있다. 즉, 다수의 역사객체가 출현하는 데이터 일수록 높은 정확도로 해당 역사문헌의 시간정보를 추출할 수 있고 이는 2.2절에서 살펴본 바와 같이 인간이 문서의 내용에서 시간정보를 추출하는 방법과 유사함을 확인 할 수 있다.

4.2 Co-EM 알고리즘을 통한 시간정보 추출

본 논문에서는 50개의 테스트 데이터로 실험을 실시하였지만, 향후에 많은 양의 웹 문서로 구성된 비지도 데이터(Unsupervised data)의 활용을 위하여 반지도 학습(semi-supervised learning) 알고리즘 중 하나인 Co-EM을 적용하여 학습을 수행하였다. Co-EM 알고리즘은 확률모델에서 많이 사용되는 EM(expectation maximization) 알고리즘을 반지도 학습에 적용한 알고리즘으로써, 데이터를 분류하기 위하여 두 개의 집합을 서로 상호 교차적으로 학습시키면서 데이터를 분류한다. 세부적인 적용 알고리즘은 [표 1]과 같이 동작한다.[16]

먼저 훈련데이터로부터 하나의 분류기(h1)를 학습하고, 학습된 분류기로부터 비지도 데이터를 학습한다. 학습된 비지도 데이터로부터 다른 하나의 분류기(h2)를 학습하고 이 분류기로부터 다시 데이터를 학습하는 방식으로 교차 수행된다.

(표 1) Co-EM 알고리즘
(Table 1) Co-EM Algorithm

Co-Em Algorithm

1 Labeled data set L, Unlabeled data set U, Let U1 be empty, Let U2=U
2 Iterate the following
  1 Train a classifier h1 from the feature set (instance) X1 of L and U1
  2 Probabilistically label all the unlabeled data in U2 using h1
  3 Train a classifier h2 from U2
  4 Let U1=U, probabilistically label all the unlabeled data in U1 using h2
3 Combine h1 and h2

Co-EM 알고리즘은 본래 2개의 집합으로 분류하는 알고리즘으로 설계되었지만, 본 논문에서는 조선시대의 연도(1394~1911)를 10년 단위로 나눈 연대로 시간정보를 결정해야 하기 때문에 2개의 집합이 아닌 다수의 집합으로 분류하는 방법으로 수정하여 적용한다.

Update Rule

$$P(Y|X_1 = k) = \sum_j P(Y|X_2 = j)P(X_2 = j|X_1 = k)$$
$$P(Y|X_2 = j) = \sum_j P(Y|X_1 = k)P(X_1 = k|X_2 = j)$$

X_1 = unlabeled data (historical objects in web document's paragraph)
X_2^r = labeled data (each year's historical objects), r=1390~1910(per 10 year) (4)

$$classify(X) = argmax P(Y|X_1) \times P(Y|X_2) \quad (5)$$

먼저 분류해야 할 테스트 데이터를 하나의 집합(h2)으로 정의하고 나머지 조선시대 연대(1390~1910)를 또 다른 각각의 집합(h1)으로 구성한다. 테스트 데이터에서 출현한 역사객체를 기반으로 하여 객체의 발생확률을 식(4)의 갱신규칙에 적용하여 각 연대집합에 대해 계산함으로써 각 연대 집합의 확률 값을 최대화하고, 식(5)에 따라 최대화된 연대 집합의 확률 값과 테스트 데이터의 확률 값을 곱하여 연대를 결정한다.

Co-EM 알고리즘으로 시간정보를 추출한 결과 전체 50개의 테스트 데이터 중에 32개의 시간정보를 정확하게 추출해내어 64%의 정확도를 보였다. 시간정보를 추출함에 있어서 나이브 베이즈 보다 좀 더 좋은 성능을 보였으나, 테스트 집합에 출현한 역사객체를 대상으로 모든 연대에 대하여 Co-EM 알고리즘을 각각 수행하여 확률 값을 계산해야 하는 높은 계산상의 복잡도를 가진다.

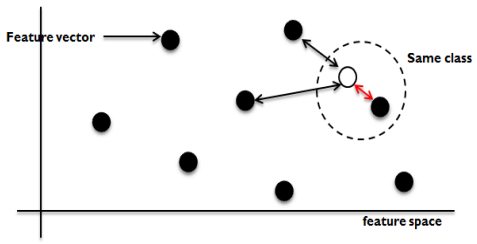
하지만 64%의 비교적 높은 정확도로 시간정보를 추출해내는 성능을 보임으로써 실제 웹 문서의 시간정보 추출에 사용할 수 있을 것으로 보인다. 본 절에서 적용한 Co-EM과 같은 반지도 학습은 비지도 데이터의 수량에 많은 영향을 받는 것으로 알려져 있어, 향후 기존에 수집된 많은 양의 비지도 데이터를 사용하여 학습을 수행하면 더 높은 정확도를 기대할 수 있을 것으로 보인다. 이는 향후 연구로 수행할 예정이다.

4.3 객체기반 학습을 통한 시간정보 추출

앞 절에서 적용한 확률기반의 알고리즘과는 다르게 본 절에서는 객체기반(Instance based)의 기계학습 알고리즘을 적용하여 웹 문서의 시간정보를 추출한다.

객체기반의 기계학습 알고리즘에는 대표적으로 Lazy 알고리즘을 비롯하여 다양한 알고리즘이 존재하지만, 본 논문에서는 여러 개의 목표 집합을 가지고 다차원에서 분류를 수행해야 하기 때문에 계산량에 이점이 있으면서 비교적 좋은 성능을 보장하는 K-NN(nearest neighbor) 알고리즘을 적용한다.

K-NN 알고리즘은 [그림 3]에서 보는 바와 같이 각 목표 집합을 특성벡터로 구성하고 구성된 특성벡터 값을 통해 특성공간에 배치한 후, 특성벡터로 구성된 테스트 데이터를 특성공간에 배치하여 가장 가까운 거리에 있는 목표 집합으로 분류하는 방법이다.



(그림 4) K-NN 알고리즘
(Figure 4) K-NN Algorithm

여기서 K는 테스트 데이터가 분류 될 수 있는 목표 집합의 개수를 의미하며, 본 논문에서는 1개의 연도로 시간정보를 추출하므로 K=1인 1-NN 알고리즘을 적용한다.

K-NN 알고리즘의 적용에 있어서 가장 중요한 점은 목표 집합을 특성 벡터로 변환하고 특성 공간에 배치하는 방법이다.

본 논문에서는 정보검색에 널리 사용되는 TF/IDF (Term frequency/Inverse document frequency) 알고리즘에 착안하여 목표 집합의 특성벡터를 구성한다. TF/IDF에서 TF는 특정 항목이 문서 내에 얼마나 자주 발생하는지를 나타내는 값으로써 항목이 특정문서 내에서 얼마나 중요한 것인지 나타내는 수치이다. 하지만 특정 항목 자체가 문서집합 내에서 자주 사용되는 경우, 이것은 그 항목이 흔하게 등장하여 별로 중요하지 않다는 것을 의미한다. 이것을 DF(document frequency)라고 하며, 이 값의 역수를 IDF라고 한다.

이러한 TF/IDF의 성질에 착안하여 전체 조선시대 기간 동안 고르게 많이 분포한 단어에 IDF특성을 적용하여 역 가중치를 부여하고, 특정 연대에 많이 출현한 단어에 TF개념을 적용하여 역사객체의 출현 빈도를 개체의 가중치로 설정한다. 이렇게 구성된 집합의 특성 벡터는 식(6)과 같이 표현된다.

$$F_y = w_1x_1, w_2x_2, \ldots, w_nx_n$$

$$\text{Weight } w_i = 1 - |norm(x_i\ total\ sum\ frequency)| \quad (0 \text{ to } 1)$$

$$\text{Instance value } x_i = |norm(x_i\ current\ year\ frequency)| \quad (0 \text{ to } 1) \qquad (6)$$

여기서 Weight w_i는 IDF의 의미와 동일하며 전체 기간 동안에 x라는 객체가 많이 출현했다면 그 출현빈도에 음의 가중치를 준다는 의미로 사용됐으며, 또한 Instance value x_i는 TF의 의미로써 현재 년도에 많이 출현 할수록 높아지는 가중치를 의미한다.

식(6)으로 구성된 각 연대의 특성 벡터를 특성공간에 배치하고 테스트 데이터에 출현한 역사객체를 이용하여 같은 방법으로 특성 벡터를 구성하여 두 벡터 사이의 거리가 가장 가까운 특성 집합을 해당 테스트 데이터의 집합으로 분류한다.

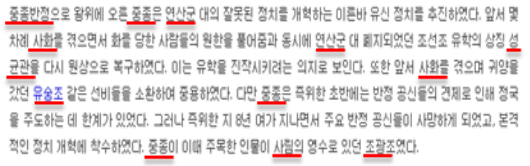
이렇게 1-NN 알고리즘을 통한 시간정보의 추출의 정확도는 24%(12/50)로 나타났다. 앞서 수행한 확률 기반의 학습에 비하여 상당히 낮은 정확도를 보였는데, 이러한 결과는 객체기반 학습의 특성에 기인한 것으로써 테스트 데이터에 충분한 양의 역사객체의 수가 확보되지 않아 나타난 결과로 추론된다. 이에 따라 실제 시간추출에 사용하는데 부적합한 것으로 보이나, 1-NN으로 구성된 알고리즘은 K=3인 K-NN으로 구성하여 학습을 수행한 후 그 결과에서 다시 시간정보를 추출하는 식의 알고리즘 개량을 통한 개선의 여지가 있다. 또한 주로 학습에 중심이 되는 역사객체로써 인명이 사용되는데, 조선시대 사람의 수명은 보통 50~60세 전후로 실제 역사에 나타나는 기간은 약 30년 정도에 해당한다. 이에 따라 연대의 약 3구간에 걸쳐 높은 빈도로 발생하는 결과로 인해 본 논문에서 적용한 1-NN의 적용에 약점을 보인다. 실제로 1-NN을 적용하고 후순위인 2, 3번째 특성벡터의 연도를 조사한 결과 65%의 확률로 정확한 시간정보를 가지는 후보군이 소속됨을 확인 할 수 있었다. 이러한 여러 가지 개선점은 향후 연구로 수행할 예정이다.

5. Similarity filtering을 적용한 시간정보의 추출

본 장에서는 테스트 데이터에서 발생한 역사객체들의 관계에 주목하여 이를 활용한 Similarity filtering방법을 제안하고 4.1절과 4.2절에서 소개한 나이브 베이즈 알고리즘과 Co-EM 알고리즘을 통한 기계학습 이전에 제안한 Similarity filtering 방법을 적용하여 정확도 향상 및 계산상의 이점에 대하여 확인한다.

5.1 Similarity filtering

역사정보 기술상의 특징으로써 역사정보는 관련 객체들과 함께 기술된다는 점이 있다. 예를 들어 [그림 4]에서 보는 바와 같이, 중종, 연산군, 조광조 등 관련 인물들과 사건들이 함께 관련 지어 기술되는 것을 볼 수 있고, 이러한 기술 방식을 통해 역사정보는 관련 있는 인물이나 사건 등의 객체들이 함께 기술됨을 확인 할 수 있다.


중종반정으로 왕위에 오른 중종은 연산군 대의 잘못된 정치를 개혁하는 이른바 유신 정치를 추진하였다. 앞서 몇 차례 사화를 겪으면서 화를 당한 사람들의 원한을 풀어줌과 동시에 연산군 대 폐지되었던 조선조 유학의 상징 성균관을 다시 원상으로 복구하였다. 이는 유학을 진작시키려는 의지로 보인다. 또한 앞서 사화를 겪으며 귀양을 갔던 유숭조 같은 선비들을 소환하여 중용하였다. 다만 중종은 즉위한 초반에는 반정 공신들의 견제로 인해 정국을 주도하는 데 한계가 있었다. 그러나 즉위한 지 6년 여가 지나면서 주요 반정 공신들이 사망하게 되었고, 본격적인 정치 개혁에 착수하였다. 중종이 이때 주목한 인물이 사림의 영수로 있던 조광조였다.

(그림 5) 연관 있는 객체들과 함께 기술된 역사정보의 예
(Figure 5) Example of historical information with the relation historical objects

본 논문에서는 이러한 역사정보의 기술상의 특징에 착안하여 함께 기술된 역사객체간의 관계로써 지도 데이

터를 여과하여 학습결과의 정확도를 높이고 계산상의 복잡도를 줄이는 것을 목표로 한다. 이에 대한 적용 알고리즘은 다음의 [표 2]와 같다.

(표 2) Similarity filter 적용 알고리즘
(Table 2) Applying similarity filter algorithm

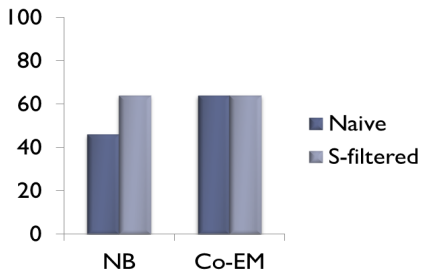
Applying Similarity filter
1. Selecting time-period in which all of test object exist (eliminating another time-period)
2. Using similarity between objects
   2.1. Calculate Euclidian distance between objects (selected time-period)
   2.2. Sum all distances (each time-period)
   2.3. Select top K high value time-period (if selected time-period >3, then k=3)
3. Applying ML for top k time-period

Similarity filtering 알고리즘은 먼저 테스트 데이터에서 발생한 역사객체간의 관계를 거리측도를 통하여 계산한 후, 역시 테스트 데이터에서 발생한 역사객체만을 대상으로 각 연대별로 연관도를 계산하여 가장 유사한 발생 경향을 가지는 연대를 선택한다. 이때 각 객체간의 연관도는 거리측도 중 하나인 유클리디안 거리(Euclidian distance) 를 사용한다(식7).

이러한 Similarity filtering 알고리즘의 목표는 해당 테스트 데이터와 가장 유사한 역사객체의 발생 경향을 가지는 연대를 K개 선택하여, 해당 K개의 연대만을 목표 집합으로 하여 분류를 수행 할 수 있도록 하는 것이다. 이에 따라 시간정보의 추출에 있어서 테스트 데이터의 목적 연대와 크게 벗어나는 연대를 기계학습 전에 제외하여 정확도를 향상시킴과 동시에 계산상의 복잡도에서 큰 이점을 가질 수 있다.

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \quad (7)$$

본 연구에서는 K=3으로 설정하여 테스트 데이터와 가장 유사한 역사객체의 출현빈도를 가지는 3개의 연대를 계산을 통해 선택하고, 해당 연대만을 대상으로 4.1절과 4.2절에서 기술한 기계학습 알고리즘을 적용한다. 적용후의 정확도에 대한 결과는 다음 [그림 5]와 같다.



(그림 6) Similarity filtering의 적용 전후 정확도 비교
(Figure 6) Comparing accuracy of before and after applying similarity filtering

[그림 5]의 결과에서 나이브 베이즈의 경우에서는 기존의 46%의 정확도에서 Similarity filter를 적용한 결과 64%의 정확도로 약 18% 향상됨을 보였다. 이는 기존의 나이브 베이즈의 계산에 있어서 과도하게 자주 출현한 역사객체가 정확한 시간정보의 추출에 크게 어긋나도록 미치는 영향을 3개의 연대로 제한함으로써 미연에 방지하는 효과로 얻어지는 결과로 보인다. Co-EM의 경우에는 정확도의 향상에는 영향이 없었으나, Co-EM 알고리즘 적용의 특성상 각 연대에 대하여 출현한 역사객체의 수만큼 반복해서 계산을 수행해야 하므로 상당한 복잡도의 계산량이 요구되었으나, Similarity filter를 적용하면 기존의 63개의 연대(1390~1910, 10년 단위)를 3개 연대로 축소하므로 상당한 계산적 이점을 얻을 수 있다.

6. 결론 및 향후 계획

본 논문에서는 웹 상에 존재하는 문서가 내포하고 있는 시간정보를 추출하기 위한 방법으로써, 역사문헌인 조선왕조실록을 기준으로 하여 역사관련 웹 문서를 대상으로 확률 및 객체 기반의 기계학습 알고리즘인 나이브 베이즈, Co-EM, K-NN을 각각 적용하고 제안한 Similarity filtering을 통하여 시간정보 추출에 대한 가능성을 확인하였다. 이는 아직 초기 연구 분야인 시간정보(Temporal information)에 대한 연구 분야 중 시간색인에 대한 연구의 가능성을 확인하기 위하여 다양한 이론에 기반 한 기계학습 기법을 적용하여 그 가능성을 찾고자 하였다.

비록 시간정보 추출에 대해 높은 정확도를 보이지는 못하였지만, 정형데이터인 조선왕조실록을 기반으로 하여 비정형데이터인 역사관련 웹 문서가 내포하고 있는 시간정보를 추출하는 방법론을 제안하고 결과를 비교 분

석하였다. 또한 제안한 Similarity filtering의 적용을 통하여 정확도의 향상과 더불어 계산상의 복잡도에 이점을 가지는 것 또한 실험을 통하여 확인하였다. 본 논문에서는 인간이 정보를 보고 해당 정보가 내포하는 시간을 유추하는 방법에 기초하여 역사정보의 인물, 사건 등의 역사객체를 정의하였고, 이를 기반으로 기계학습 알고리즘을 수행하였다.

향후 연구로써 소수의 정형데이터를 가지고 많은 양의 비정형데이터를 함께 분류 할 수 있는 반지도학습에 시간정보 추출 및 시간정보 색인에 적용할 수 있는 연구를 진행하여 많은 양의 역사에 대한 웹 문서를 학습하고 이를 통해 시간정보 추출의 정확도를 높일 수 있는 방법에 대한 가능성을 확인 할 것이다. 또한 본 논문에서 역사정보 내에서의 중요한 요소로써 정의한 역사객체를 활용한 객체기반 학습(instance-based learning) 알고리즘에 대한 개량을 통하여 객체를 기반으로 시간정보 추출을 보다 정확하게 수행할 수 있는 방안을 연구 할 것이다.

나아가 궁극적으로 본 논문에서 사용한 역사문헌 뿐만 아니라 뉴스, 블로그 등과 같은 일반적인 비정형 웹 문서의 내용에서 정확한 시간정보 추출할 수 있는 방안에 대하여 연구를 진행 할 것이다. 이러한 비정형 웹 문서의 시간정보 추출을 통하여 현재의 키워드 기반 정보검색에서 벗어난 시간축을 기준으로 한 정보검색 서비스의 개발이 가능할 것이며, 정보의 내용이 변하는 변곡점에 해당하는 시점 추출과 같은 시간관련 연구에 많은 활용이 가능할 것으로 예상된다.

참 고 문 헌 (Reference)

[1] J. F. Allen. Maintaining Knowledge about Temporal Intervals. In Communications of the ACM, 26(11):832 – 843, 1983. http://dx.doi.org/10.1145/182.358434

[2] O. Alonso, M. Gertz, and R. Baeza-Yates. On the value of temporal information in information retrieval. SIGIR Forum, 41(2):35{41, 2007. http://dx.doi.org/10.1145/1328964.1328968

[3] J. Pustejovsky, J. M. Casta˜no, et al. TimeML: Robust Specification of Event and Temporal Expressions in Text. In Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, pages 28 – 34, 2008. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.161.8972

[4] O.Alonso, J. Strotgen, R. Baeza-Yates, and M. Gertz. Temporal information retrieval: Challenges and opportunities. In International Temporal Web Analytics Workshop, pages 1-8, 2011 http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.363.4483

[5] O. Kolomiyets and M.-F. Moens. Meeting TempEval-2: Shallow Approach for Temporal Tagger. In Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW '09), pages 52 – 57, 2009. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.164.9479

[6] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and Exploring Search Results Using Timeline Constructions. In Proceedings of the 18th ACM International Conference on Information and Knowledge Management (CIKM '09), pages 97 – 106,2009. http://dx.doi.org/10.1145/1645953.1645968

[7] J. Makkonen and H. Ahonen-Myka. Utilizing Temporal Information in Topic Detection and Tracking. In Proceedings of 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '03), pages 393 – 404, 2003. http://dx.doi.org/10.1007/978-3-540-45175-4_36

[8] O. Alonso, R. Baeza-Yates, and M. Gertz. Effectiveness of Temporal Snippets. In Proceedings of the Workshop on Web Search Result Summarization and Presentation (WSSP 09), pages 1 – 4, 2009. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.160.5485

[9] A. Qamra, B. Tseng, and E. Chang. Mining Blog Stories Using Community-based and Temporal Clustering. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06), pages 58 – 67, 2006. http://dx.doi.org/10.1145/1183614.1183627

[10] R. Swan and J. Allan. TimeMine: Visualizing Automatically Constructed Timelines. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00), page 393, 2000. http://dx.doi.org/10.1145/345508.345674

[11] A. Jatowt, K. Kanazawa, S. Oyama, and K. Tanaka. Supporting Analysis of Future-related Information in News Archives and the Web. In Proceedings of the 9th Joint Conference on Digital Libraries (JCDL '09), 2009. http://dx.doi.org/10.1145/1555400.1555420

[12] B. Shaparenko, R. Caruana, J. Gehrke, and T. Joachims. Identifying Temporal Patterns and Key Players in Document Collections. In Proceedings of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDM '05), pages 165 – 174, 2005. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.8382

[13] Toyoda, M., & Kitsuregawa, M. What's Really New on the Web? Identifying New Pages from a Series of Unstable Web Snapshots. In WWW2006: Proceedings of the 15th International World Wide Web Conference (pp. 233 – 241). Edinburgh, Scotland. May 23 – 26: ACM Press. http://dx.doi.org/10.1145/1135777.1135815

[14] J. Strotgen, M. Gertz, and P. Popov. Extraction and Exploration of Spatio-temporal Information in Documents. In Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR '10), pages 1 – 8, 2010. http://dx.doi.org/10.1145/1722080.1722101

[15] Seung-Shik Kang, Byoung-Tak Zhang, A General Morphological Analyzer and Spelling Checker for the Korean Language Using Syllable Characteristics , Journal of KIISE (B), Vol.23. No.5, 1996 http://www.dbpia.co.kr/Journal/ArticleDetail/444487

[16] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of cotraining.In Proceedings of the Workshop on Information and Knowledge Management, 2000. http://dx.doi.org/10.1145/354756.354805

◐ 저 자 소 개 ◑

이 준 (Jun Lee)
2009년 한국항공대학교 정보통신 공학과 학사
2011년 한국항공대학교 정보통신 공학과 석사
2011년~현재 한국항공대학교 정보통신 공학과 박사과정
2008년~현재 한국항공대학교 차세대방송미디어기술연구센터 연구원
관심분야: 정보처리, 정보검색, 데이터 마이닝, 웹 애널리틱스
email: jun@grrc.kau.ac.kr

권 용 진 (KWON, YongJin)
1986년 한국항공대학교 항공전자 공학과 학사
1990년 일본 교토대학 대학원 정보 공학과 석사
1994년 일본 교토대학 대학원 정보 공학과 박사
1994년~현재 한국항공대학교 항공전자 및 정보통신공학부 정교수
2007년~현재 경기도지역협력연구센터(GRRC) 센터장
관심분야: 논리회로 설계 및 합성, 알고리즘 개발, 정보보호, 정보검색, LBS
email: yjkwon@tikwon.hangkong.ac.kr