인텔리전트마이너를 이용한 텍스트마이닝 시스템의 설계 및 구현

최윤정[U] 박승수
이 화 여 자 대 학 교 컴 퓨 터 학 과
{cris, sspark}@ai.ewha.ac.kr

Design and Implementation of a Text Mining System using Intelligent Miner

Yun-Jeong Choi[U] Seoung-Soo Park
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

데이터마이닝 기능은 문서의 구조화되지 않은 텍스트보다는 테이블과 일반적인 DB에 있는 구조화된 자료에 초점이 맞춰져 있다. 정보화의 과정속에서 많은 기업이나 조직들은 과거의 시스템을 DB로 구축하여 어느 정도 형태를 갖추게 되었지만, E-business, E-commerce가 활발해지면서 보유하고 있는 DB기반이 아닌 무작위의 새로운 데이터가 사용자들에 의해 생성되기도 한다.

본 논문에서는 이러한 텍스트 문서에 숨어있는 정보들을 발견하기 위한 텍스트마이닝 과정을 시나리오로 설정하고, 문서와 문서집합에 대해 분석도구를 적용하는 어플리케이션을 구현해 보았다. 대규모의 문서집합에 분석도구를 이용함으로써 빠른 문서처리가 가능하고 이는 사용자가 많은 양의 문서들을 다룰 때의 시간비용을 최소화시킬 수 있는 방법이 될 수 있다. 또한 마이닝과정을 통해 발견한 지식과 특징들을 기반으로 반구조화된 파일로 변환하여, 규칙발견, 데이터마이닝기법을 적용하여 의미있는 새로운 결론을 얻을 수 있을 것이다.

1. 서론

대부분의 기업과 웹사이트에 있어 기존의 데이터베이스 기반이 아닌 무작위로 드나드는 사용자들의 동선들로부터 생성되는 데이터들처럼 데이터베이스 구조를 가지지 않았지만 상당한 잠재적 가치를 지니고 있는 텍스트데이타들이 있다. e-mail이나 웹에서의 검색 결과, 관련된 문서들을 고려해볼 때, 이러한 대규모의 텍스트 데이타들을 사용자의 필요에 의해 개인적으로 DB로 구성하여 사용하기가 쉽지 않고, 정보들이 문서에 함축되어 있기 때문에 간단한 단어라도 찾기가 어렵다. 따라서 이러한 과정에 마이닝 기법이 적용될 수 있다.

텍스트마이닝은 문서정보마이닝,텍스트로부터 숨겨져 있거나 흥미있는 지식을 발견하고, 방대한 양의 문서집합에서 지식탐사를 위해 패턴을 추출하는 과정이라고 할 수 있다. 데이터마이닝 기능은 문서의 구조화되지 않은 텍스트보다는 테이블과 일반적인 데이터베이스에 있는 구조화된 자료에 초점이 맞춰져 있다. 텍스트마이닝 툴인 IBM의 인텔리전트 마이너에는 구조가 정해지지 않은 텍스트로부터 지식관리 응용프로그램을 구축할 수 있는 도구을 제공한다. 각 분석도구를 적절히 이용하여 문서에서 정보를 추출하고, 주제별로 문서를 구성한다. 또한 문서집합에서 주요주제를 찾고, 검색도구를 통해 관련문서를 검색하는 시스템을 구현하는데 사용할 수 있다.[2][4]

본 논문은 다음과 같이 구성되어 있다. 2장에서는 일반적인 텍스트 마이닝의 분석도구의 종류를 간단히 살펴보고 3장에서 인텔리전트마이너를 이용하여 웹에서 검색한 신문기사와 전자상거래 전자메일처리에 텍스트마이닝을 적용한 시스템 구현 예를 보인 후 4장에서는 결론을 내린다.

* 본 연구는 교육부 BK 사업의 지원으로 수행되었음

2. 관련연구

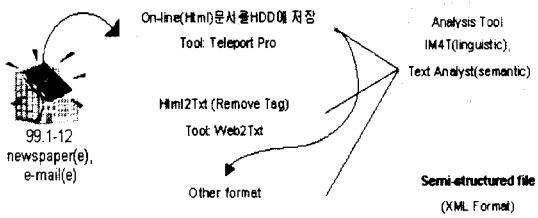
텍스트마이닝 기술체계는 자연어처리기법,시각화, 데이타베이스기술,기계 학습,그리고 데이타마이닝 분야를 포함하고 있다. 텍스트마이닝에서 개발된 많은 기술과 도구는 문서의 텍스트에서 정보나 지식을 발견하고 추출한다. 텍스트 분석도구는 텍스트 및 정보처리의 여러 분야에 도움이 되는 이미 개발된 기술(NLP)의 Tool Set으로 텍스트마이닝 툴에서 제공하고 있는 기능들을 참고하여 일반적으로 다음과 같이 정리된다.[1][2]

첫째, 작성된 문서의 언어를 자동으로 찾아내는 언어식별도구, 둘째, 문서를 미리 정의된 카테고리나 주제등을 자동으로 지정해주는 주제분류도구, 셋째, 국가별 사전 데이타를 이용하여 문서에서 미리 정의된 용어 없이 중요한 용어나 이름, 약어, 장소 등의 항목을 자동

으로 인식하는 특성추출도구, 넷째, 유사한 문서집합을 자동으로 그룹이나 '클러스터' 로 나눌수 있는 클러스터링도구, 다섯째, 문장을 분석하여 문서의 요약정보를 추출하는 요약도구등이 있다.

인텔리전트 마이너의 텍스트분석도구를 간단히 살펴보면, 우선 요약도구의 알고리즘은 요약을 작성하기 위해 문서와의 관계 및 문서에서의 위치에 따라 문장의 등급을 매긴다. 가령 긴 문단에서 최종문장과 긴 섹션에서 최종문단은 문서와 매우 관련이 높은 것으로 등급을 매기고, 가장 관련이 높은 문장을 추출하여 문서요약을 작성한다. 특성추출도구 기능을 사용하여 문서를 분석하고 문서내용에 대한 중요도에 따라 단어의 등급을 매긴 후, 문장에서 각 단어의 등급들을 모두 더해 문장의 총등급을 매긴다. 주제분류도구에서의 분류의 결과는 카테고리 이름의 목록과 각 문서에 대한 신뢰도 레벨로 나타난다. 문서는 하나이상의 카테고리에 지정될 수 있고, 신뢰도가 낮으면 보통 분류자가 최종결정을 내릴 수 있도록 문서를 따로 구분해 놓는다. 새로운 카테고리를 정의해야 할 경우 카테고리에 대한 예를 모아서 학습프로세스를 재실행한다. 클러스터링에는 최대차이지점에서 모음을 클러스터로 나누는 하향식 접근 방법과 비슷한 문서들을 그룹으로 계속 추가하는 상향식 접근방법으로 두가지 접근방법이 있는데, 이 방법들을 모두 사용하여 문서집합을 다른 시각으로 볼 수 있으며 다른 견해를 얻을 수 있다. 이러한 분석도구들을 여러 가지 방식으로 조합하여 반복적으로 수행함으로써 사용자가 마이닝 솔루션을 작성할 수 있다.

3. 시스템 구성



[그림1] 텍스트마이닝 적용과정

텍스트마이닝 진행순서는 어떠한 구조도 갖지 않은 텍스트 문서를 중간단계의 형태로 변환하는 텍스트정제과정과 그 중간형태로부터 패턴이나 지식을 추론해내는 지식발견 과정으로 진행한다.

본 연구에서는 99년 1-12까지의 영문기사와 전자상거래 웹사이트의 운영자수신메일을 대상으로 하여 해당 문서와 문서집합에 적절히 마이닝 기법을 적용한 후, 마이닝과정에서 추출해낸 정보들을 참고하여 반구조형태를 가진 파일로 변환하여 보았다.

3.1 데이터준비

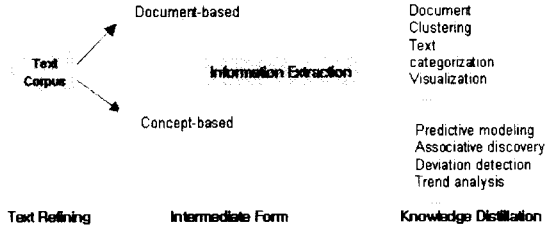
단계1: 뉴스사이트의 온라인 문서를 이미지와 기타 불필요한 코드를 제외한 순수 텍스트화일로 변환하여 하드디스크에 저장한다.
단계2: 메일박스의 메시지들을 첨부화일을 제거하여 각각의 개별화일로 나눈다.

3.2 텍스트분석도구 적용

단계 1: 온라인상에서 뉴스를 수집해 문서집합을 만든 경우, 우선 언어식별도구를 적용하여 기사에 쓰여진 주요언어를 구분할 수 있다. 요약도구를 적용하여 문서집합내의 문서마다 요약정보를 얻을 수 있고, 클러스터링도구를 적용하여 대규모 문서집합의 내용에 대한 개요를 얻을 수 있다. 특성추출도구를 통해 이름,용어,관계 등을 추출하고 이를 기반으로 주제분류도구에서 정의해야할 분류항목으로 사용될 수 있는 정보를 얻는다. 주제사용자의 관심사에 따라 항목을 추가로 지정한 후 주제분류도구를 적용한다. 분류의 결과는 분류항목과 각 문서에 대한 신뢰도 레벨로 나타난다. 문서는 하나이상의 분류항목에 지정될 수 있고, 신뢰도가 낮으면 보통 분류자가 최종 결정을 내릴 수 있도록 문서를 따로 구분해 놓는다. 미리 정의된 항목에 속하지 않는 주제를 포함하는 문서들이 자주 나타나면 새로운 분류항목을 정의해야하며, 새로운 예를 모아서 학습프로세스를 재실행한다.

단계 2: 광범위한 전자우편처리에 있어 각 수신인에 대한 샘플메시지세트를 수집하여 주제분류를 위한 학습데이터로 사용한다. 주제분류도구는 새로운 문서마다 신용도와 함께 수신인 목록을 제시하게된다. 만약, 주제분류과정 중 문서가 임계값을 넘는 신용도가 없다는 것은 결정을 내리기 위한 충분한 증거를 찾지 못한 경우이며, 직접 메시지를 읽어 분류해야함을 의미한다. 남아 있는 문서에 대해서 클러스터링도구를 적용하여 유사성별로 그룹화할 수 있으며, 문서와 클러스터간 관계와 각 클러스터를 특징짓는 키워드를 얻을 수 있다.

단계 3: 마이닝도구를 적용하여 이끌어낸 결과를 해석하고 시각화기능을 적용한다. 준비한 데이터를 재선택, 임의추출, 통합, 여과하거나 변환하여 반복수행을 통하여 놓쳐버릴 수 있는 특징들을 찾아낸다.



[그림2] 텍스트마이닝 수행에 있어 접근방식

단계3: 구조가 없는 텍스트문서들에 마이닝을 적용하여 얻은 지식과 특성 및 분류를 기반으로 반구조적인 형태를 갖는 파일로 변환한다. 이 과정에서 간단한 형식의 XML 형식으로 변환해보았다. 반구조적인 형태를 갖는 이 파일들에 데이터마이닝을 적용할 수 있다.

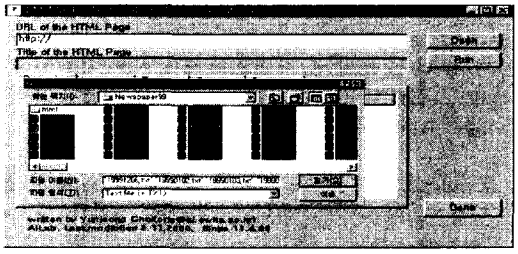
4. 시스템 구현

본 응용프로그램은 Windows NT와 Solaris 2.6상에서 인텔리전트마이너(IM4T)와 비주얼C++과 델파이를 사용하여 구현하였다. IM4T의 텍스트분석도구는 [그림 3]에서처럼 UNIX나 DOS의 커맨드라인 형식으로 프로그램에서 구현하기에 용이한 function을 제공한다.

```
command: "imzxrun -b 2 -f C -x n -o outfile text.txt"
<IMZ ID>text.txt</IMZ ID>
<IMZ TITLE>Local Education Outreach</IMZ TITLE>
<IMZ CONTENT>
NC 3 IBM ORG
NC 1 IBM Thomas J. Watson Research Center in Westchester County ORG
NC 2 James J. Smith PERSON
NC 1 Learning ORG
NC 1 Local Education Outreach Program ORG
NC 2 National Science and Technology Week ORG
NC 1 New York City PLACE
NC 1 President Clinton PERSON
NC 1 Somers , New York PLACE?
</IMZ CONTENT>
```

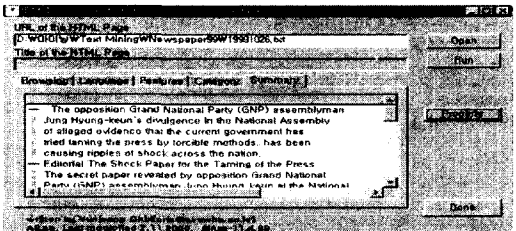
[그림 3] 인텔리전트마이너의 command 예

프로그램 실행과정은 [그림4]와 같이 브라우저 Tab에 마이닝할 대상을 loading 한 후, 3장에서 설정한 적용도구를 수행한다.



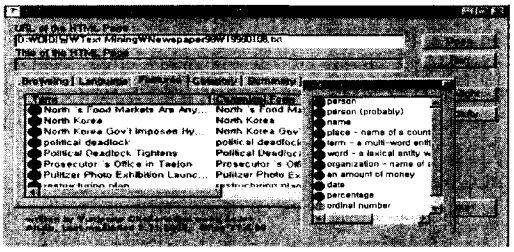
[그림 4] visual C++로 구현한 인텔리전트마이너의 GUI

[그림5]와 [그림6]은 해당문서에 요약도구와 추출도구를 사용한 예이고, [그림7]은 문서집합에 분류도구를

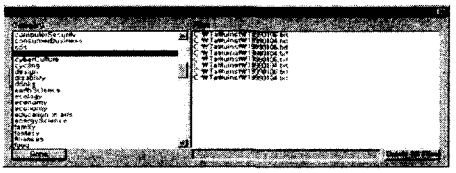


[그림 5] 특정 문서에 요약도구를 적용한 모습

사용하여 미리 정의한 항목으로 분류된 항목과 목록을 나타난 예이다.



[그림 6] 특정문서에 특성추출도구를 적용한 모습



[그림 7] 문서집합에 분류도구를 적용한 모습

5. 결론 및 향후 연구

본 연구에서는 구조의 형태를 갖지 않은 텍스트데이타를 대상으로 문서내에서와 문서집합에 대해 분석도구를 이용하여 마이닝을 수행하여 지식을 발견하는 과정을 설계 및 구현해보았다.

문서내에 분석도구를 적용하는 경우, 이는 full-text검색효과가 있었으며, 문서집합에 대해 클러스터링이나 요약도구를 수행함으로써 사용자가 광범위한 양의 문서들을 접할 때의 시간비용을 최소화시킬 수 있는 방법이 될 수 있었다. 또한 마이닝 수행과정을 통해 발견된 지식과 특성들을 기반으로 반구조화(semi-structured)된 파일로 변환하여 데이터마이닝 기법을 적용하여 규칙을 발견하거나 예측모델링등의 의미 있는 결론을 얻을 수 있을 것이다.

향후 연구는 인터넷의 문서집합을 대상으로, 제품과 모델을 추출하고 자주 언급되는 제조업체와 모델을 식별하는 과정에서 이를 인덱스구조의 파일로 변환하여 데이터마이닝을 적용하는 연구로 진행할 것이다.

5. 참고 문헌

[1] Jochen Dorre,Peter Gerstl and Roland Seiffert,"Text Mining:Finding Nuggets in Mountains of Textual Data" Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,1999

[2] Lee Hing Yan,"Text Mining - Knowledge Discovery from Text", Trend in Knowledge Discovery from Databases,29th June 1999 BIC_KRDL

[3] Kevin Knight, "Mining Online Text", Commun. ACM 42, 11 (Nov. 1999)

[4] IBM white paper for Intelligent Miner for Text