텍스트마이닝 기법의 기술정보분석 적용 가능성 연구

배상진(KISTI) · 박철균(아주대)
sjbae@kisti.re.kr

< 목 차 >

I. 서론

인터넷을 통하여 자기의 생각(지식)을 문서로 정리하여 전 세계인에게 전달하는 것이 손쉬워졌다. 그만큼 우리가 접할 수 있는 정보의 양이 폭발적으로 증가하고 있다는 것을 의미한다. 특정 키워드를 가지고 검색을 하였을 때 너무 많은 정보가 쏟아져 나와 정보의 질적 평가가 새로운 과제가 되었다. 한편, 학술정보의 주 전달 매체도 이전의 학술잡지에서 전자저널로 바뀌었다.

정보를 저장하는 가장 자연스러운 기본적인 형태는 결국 "텍스트"이고, 최근 조사에 따르면 기업체 정보의 80%가 텍스트 문서 형태로 보관되고 있다.[1] 이러한 환경 속에서 텍스트에서 의사결정에 수렴할 만한 가치 있고 유용한 정보를 찾아내어 분석하는 작업의 중요성이 높아지고 있다.

최근, 기업에서 유용하고 잠재적인 정보를 발견해내기 위해 많이 사용하는 데이터마이닝 기술은 정형화된 형태의 데이터를 주 대상으로 하고 있다. 텍스트 데이터는 수치 데이터와 달리 자연어로 구성된 비구조적 데이터이다. 컴퓨터가 텍스트에서 함축된 정보를 추론해 내기 위해서는 텍스트를 구조적 데이터로 표현해야 할 필요가 있다. 이러한 기능을 수행하기 위해서 텍스트마이닝 기술이 필요하게 되었다.

시시각각 발생하는 수많은 정보 중에서 특정 이용자가 필요로 하는 정보를 주

기적으로 전달할 수 있는 대표적인 방법으로서는 검색엔진에 미리 질문식(키워드 연산식)을 등록해 놓고 자동으로 검색 제공하는 SDI(Selective Dissemination Information) 시스템을 들 수 있다. 그러나 SDI 시스템은 키워드 매치 방식에 의해 정보를 선별하기 때문에 검색된 정보의 질적 평가에 대해서는 검증이 어렵다는 문제와 제공하는 정보가 많을 때 이를 자동 분류하여 제공하는데 문제를 안고 있다. 텍스트마이닝은 텍스트 분류, 텍스트 군집화, 텍스트 요약, 텍스트 분할 등을 통하여 이러한 문제에 대한 해결방안을 제시해준다.

최근, 텍스트마이닝 기법은 지식관리시스템, 전자도서관, 정보 필터링, 정보검색 엔진의 기능강화, 전자상거래 등 다양한 분야에 적용하여 유용하게 활용되고 있다.

그러나, 국내에서 기술정보분석에 텍스트마이닝 기법을 적용하는 것을 연구하기 시작한 것은 최근의 일이다.

구체적으로 보면, 한국과학기술정보연구원의 문영호 외(2000)에서 문헌DB로부터 수집된 기술정보를 계량화하는 알고리즘을 개발해 KITAS라는 프로그램을 만들어 기술정보분석에 활용하고 있으나, 초기 단계의 프로그램으로 개선의 여지가 많고, 텍스트 마이닝 내지 데이터마이닝 기법을 구체적으로 적용한 것으로 보기는 어렵다.

한남대학교의 설성수(2002)에서는 최근 기술분석에 있어서 새로운 추세가 등장하고 있는데, 기술특허분석이 보다 체계적으로 시도되고 있고, 기술정보분석과 특허분석이 결합하는 추이를 보이고 있고 기술분석과 관련되어 무언가 새로운 흐름이 관측되고 있는 것으로 파악하고 있다. 또한, 기술정보분석이나 기술시장분석은 정보학의 발전에 의해 새로운 전기를 맞고 있으며, 그 중에서 텍스트마이닝은 아직도 발전 중에 있고 많은 과제를 가지고 있는 것으로 보고 있으며 현재까지는 전문가들을 대상으로 한 텍스트마이닝 툴 정도가 개발되어 있다고 하고 있다.

그러나, Yoon, Yoon & Park(2002)에서는 텍스트마이닝기법을 이용한 특허분석을 수행하고 있고, Morris(2002) 역시 텍스트마이닝기법을 이용하여 기술문헌분석과 특허분석을 실시하고 있다.

앞으로 텍스트마이닝 기법은 기술정보분석에 있어서 중요한 부분을 차지하면서 다양하게 이용될 것이 분명하다.

따라서, 본고에서는 텍스트마이닝의 정의와 응용동향을 살펴 보고, 텍스트마이닝을 구체적으로 정보분석에 활용가능한 모델을 검토해 보고자 한다. 본 논문의 구성은 다음과 같다. 2장에서는 텍스트마이닝의 정의를 살펴보고, 3장에서는 최근 응용동향을 기술한다. 4장에서는 2장과 3장의 내용을 토대로 텍스트마이닝을 이용한 전문정보 수집 및 분석 모델을 제시하고, 마지막으로 5장에서는 결론을 기술한다.

II. 텍스트마이닝의 정의

대량의 텍스트 문서를 효율적으로 다루기 위한 기술로서는 문서검색(Search Document), Organize Document, Discover Knowledge를 들수 있다. <표 1> 각 기술의 특징을 나타내었다.[2] 텍스트마이닝은 Organize Document와 Discover Knowledge의 혼합 기술로서 궁극적으로는 대량의 텍스트를 분류하고, 유용한 텍스트를 선별하고, 각 텍스트에 대한 요약 정보를 발췌하고, 검색의 정확도를 높이기 위한 키워드를 발췌하는데 유용하다. 또한 문서들 간에 공존하는 지식을 유추함으로써 새로운 지식을 창출하거나 미래를 예측하는데도 적용될 수 있다.

<표 1> 대량의 텍스트 문서 처리 기술

기능	목적	기술	데이터 표현 (Data Repres-entation)	자연어처리	결과물
문서검색 (Search document)	특정 주제와 관련된 데이터에 촛점	정보검색	문자열, 키워드	키워드 추출 (어간형태로 변환)	문서 세트
문서조직화 (Organize Document)	주제에 대한 전반적 동향파악	클러스터링,분류 (Classification)	키워드 집합 (Vector Space Model)	키워드 분산 분석	문서들의 집합
지식발굴 (Discover Knowledge)	컨텐트에서 관심 정보 추출	NLP, 데이터마이닝, Visualization	의미 개념	어의분석, 의도 분석	요약된 정보 (Trend Patterns, Association Rules 등)

1. 텍스트마이닝의 프레임웍

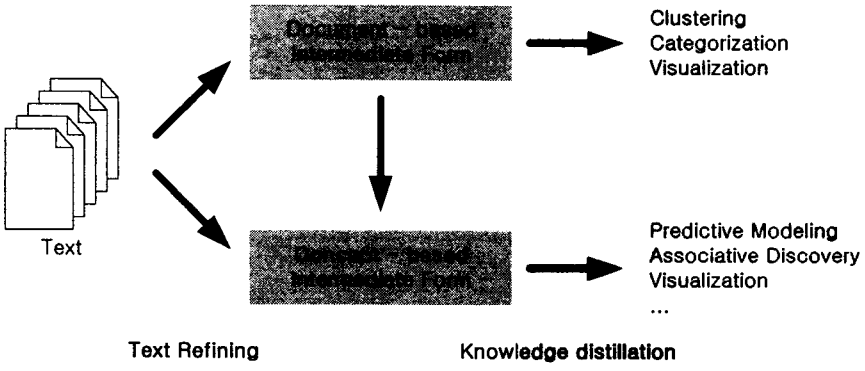
데이터마이닝이 구조적인 데이터를 대상으로 유용하고 잠재적인 패턴을 끌어내는 것이라고 한다면, 텍스트마이닝은 자연어로 구성된 비구조적인 텍스트 안에서 패턴 또는 관계를 추출하여 지식을 발견하는 것으로, 주로 텍스트의 자동 분류작업이나 새로운 지식을 생성하는 작업에 활용되고 있다. 오늘날 우리가 사용하는 대다수의 정보는 확실히 구조가 잡히지 않은 텍스트의 형태로 존재하기 때문에 자연어로 된 텍스트문서의 자동화되고 지능적인 분석은 매우 중요하다. [3]

텍스트마이닝은 데이터 준비 단계(Text Refining)와 지식 추출(Knowledge Distillation) 과정으로 나타낼 수 있다. 데이터 준비단계는 다양한 정보원(Information source : 인터넷, 인트라넷, 이메일 등)에서 수집한 자유로운 형태의

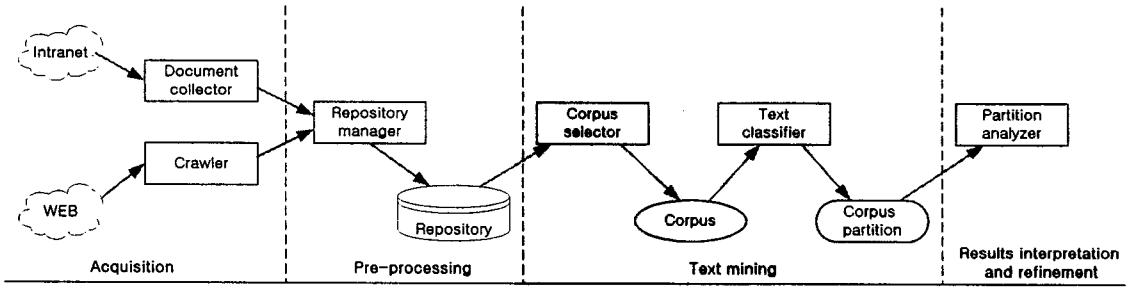
텍스트 문서를 Intermediate Form으로 바꾸는 단계로서 자연어처리, 웹문서의 경우 태그 제거, URL과 타이틀 추출 등의 기술체계를 포함하고 있다.

지식추출과정은 Intermediate Form의 문서에서 의미있는 패턴과 지식을 유추해 내는 과정으로서 클러스터링, 분류(Classification), 시각화(Visualization), 문서요약, 기계학습 등의 기술체계를 포함하고 있다.

<그림 1> 텍스트마이닝의 프레임웍



<그림 2> 텍스트마이닝 진행 과정



2. 단계별 상세 구조

1) 문서수집(Document acquisition)

텍스트마이닝은 기업(기관) 내외부에서 발생하는 모든 텍스트 문서를 수집 대상으로 한다. 웹게시물, 이메일, 특허정보, 문헌정보, 매뉴얼, 내부 보고서, 노하우 문서, 또는 웹 문서 등 모든 텍스트 파일형태의 문서를 대상으로 할 수 있다. 텍스트마이닝을 실시하기 위해서는 정보분석과 마찬가지로 분석목적에 따라 문서수집 적정 범위를 결정해야 한다.

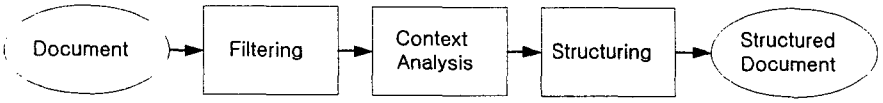
만약 고객과의 상담내용을 분석하여 서비스(제품, 용역 등) 개선을 목적으로 한다면 고객 언급사항, 이메일, 웹게시물(Q&A), 전화통화내용(문서) 등과 내부 영업사원들이 수집한 정보 등을 중점 수집해야 할 것이다.

웹에서 특정 기술분야(또는 학술분야)의 정보를 수집할 때에는 적절한 사이트로 수집범위를 제한하지 않으면 안된다. 적절한 사이트를 선정하기 위해서는 ①웹사이트 평가표에 의해 사람이 직접 평가하는 방법과 ②인터넷 검색엔진에서 적용하고 있는 사이트 랭킹 기술을 이용하는 방법 ③두가지 방법을 적절히 절충하는 방법에 대한 연구가 요구된다. 웹문서 수집을 위해서 Web Crawler 또는 로봇에이전트를 사용하는 방법은 이미 보편화되어 있다.

2) 문서 전처리(Document pre-processing)

이 단계에서는 각 문서를 텍스트마이닝에 적합한 형태로 변환한다. 텍스트 자체가 자연어로 되어있기 때문에 언어학적 관점에서의 자연어 처리과정은 필수적이다. 자연어 처리에는 다국어를 처리하는 문제와 한글을 처리하는 문제가 있다.

<그림 3> 텍스트 문서 Pre-Processing 과정



- 정제과정(Filtering) : 이 단계에서는 먼저 텍스트마이닝에 필요 없는 단어 또는 기호를 정제해야한다. 웹문서에서는 이미지파일, HTML태그, 스크립트 등을 제거해야 한다. 웹문서의 처리과정에는 문서의 URL과 타이틀을 추출하는 과정이 추가된다.

- 정규화과정(Normalization) : 문장의 정확한 의미 파악을 위해서 각 단어의 어간(Stem)을 파악하고, 동의어를 할당해야 한다. 특히 한글 처리를 위해서는 문장에서 최

소의 의미단위를 추출해 내는 형태소 분석(morphological analysis) 단계와 통사 구조를 파악하는 구문 구조 분석(syntactic analysis) 단계, 의미 구조를 추출하는 의미 분석(semantic analysis) 단계, 그리고 문장들 사이의 관계를 분석하는 문맥 분석 단계(discourse analysis) 등 네가지의 분석 단계를 거쳐야 한다. 이들 단계 중 형태소 분석 단계는 입력으로 받은 자연어에서 최소의 의미 단위인 형태소를 추출하여 자연어 분석의 최소 단위를 제공하는 단계로서 자연어 이해 시스템에서 가장 기본이 되는 단계이다.

이러한 전처리 과정을 통해 분석과정에 적합한 최적의 데이터 상태를 만들어 분석의 질을 향상시킬 수 있는데, 전처리 작업은 실제 분석에 소요되는 시간보다 더 많이 걸릴 수 있으며 수집한 데이터를 잘 이해하는 일이 필요하다.

3) 텍스트마이닝

텍스트마이닝은 비구조적인 텍스트위주 문서에 다양한 텍스트 분석기법을 적용함으로써 지식을 발견해내는 기술로서 '독서'와 매우 유사한 과정을 거치게 된다. 독서를 하기 위해서는 먼저 무엇을 읽을 것인가를 결정하고, 중요한 내용물(Entity)을 인지하고, 다음 내용물간의 관계를 규명한다. 끝으로 새로운 정보를 다른 Article이나 지식과 결합하는 과정을 거치게 된다.

텍스트마이닝에 적용하는 주요 분석기법은 군집(Clustering)과 분류(Categorization)이며, 분류분석 수행에 앞서 군집화를 먼저 수행시켜 전체 문서집합의 개요를 획득하고 분류를 위한 판단기준을 얻어낸다. 즉, 군집은 분류의 준비단계로서 사용자가 i)분류해 낼 항목(Category)을 명확히 정의하고 ii)각 항목에 따른 훈련문서를 선정하여 학습시키는 과정에 군집결과를 이용하는 것이다.

가. 텍스트 군집화 (Text clustering)

텍스트 군집화는 텍스트의 집단을 내용의 유사도에 따라 여러개의 소집단으로 분할하는 과정이다. 또한 각 소집단을 계속 세부적으로 분할할 경우 구성된 텍스트는 계층적 구조를 형성하게 되고 이를 계층적 텍스트 군집화(Hierarchical text clustering)라 한다.[4)]

텍스트 군집화는 정적 군집화와 동적 군집화로 분류된다. 정적 군집화는 군집 개수를 미리 정해 놓고, 군집 개수만큼의 대표 텍스트를 지정하고, 나머지 텍스트는 유사도가 최대인 대표 텍스트가 속한 군집에 포함시키는 방법이다. 동적 군집화

는 각 텍스트에 대하여 태표 텍스트와의 유사도가 임계 유사도보다 클 경우 해당 군집에 텍스트를 포함시키고, 그렇지 않으면 새로 군집을 생성하는 방법이다.

군집화는 데이터에 대한 기반지식 없이 분석 초기에 행하여 결과를 분석할 수 있다는 장점이 있으며, i)중복 혹은 유사한 문서를 제거하고 ii)다른 문서의 주제와 다른 주제를 가진 문서를 구별하고 iii)대량의 문서집합의 개요를 획득하는 데 적용할 수 있다. 각 클러스터에 특성추출을 행한 결과로 나타나는 단어들에서 전체 데이터가 포함하는 주제나 성격이 무엇인지 감지할 수 있으며 실제 분류해 내어야 할 기준을 얻을 수 있다.

군집을 통해 생성된 각 클러스터 내에는 공통 특징을 공유하여 서로 높은 유사도를 가진 문서들도 존재하기 마련이다. 특정 임계값 이상의 유사도를 갖는 문서들은 그 클러스터의 성격을 잘 나타낸다고 볼 수 있으며, 다른 클러스터와 구별되는 특성이 되기 때문에 분류작업의 준비단계인 샘플링(sampling) 과정에 각 클러스터 내의 문서들을 활용한다. 즉, 학습 문서를 선정하는 데 있어서 주제를 내포하는 문서를 개별적으로 선택하기에 앞서 서로 긴밀하게 뭉쳐진 문서들 중에서 우선적으로 선정하는 것이다.

군집 단위(클러스터) 별로 다음 사항들을 정리하여 관리할 필요가 있다..

- 클러스터의 길이
- 포함되어 있는 문헌의 수
- 문서들의 하이퍼링크 목록
- 클러스터 내의 대표 키워드를 출현빈도에 의해 정렬한 목록

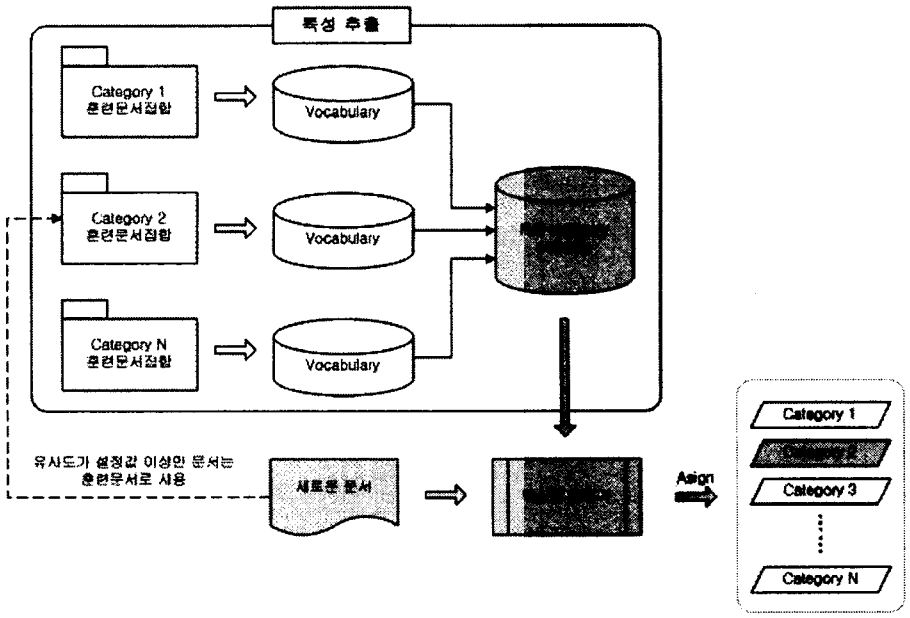
나. 텍스트 분류

텍스트 분류(Text categorization)란 텍스트의 내용에 따라 미리 정의해놓은 범주를 부여하는 과정이다. 'Categorization'은 'Classification'으로도 사용되는데 '범주에 할당'한다는 의미 가 부각되기 때문에 Categorization이 더욱 보편적으로 사용된다.[5]

군집(clustering)과 달리 분류(Categorization)를 수행하기 위해서는 각 항목을 위한 학습데이터를 사용자가 선정하여 훈련시키는 과정이 필요하다. <그림 5>는 위에서 선정된 각 훈련문서에서 특성을 추출해 내어 특성벡터(feature schema)를 구성하는 기본 학습과정을 보이고 있다[6]. 특성벡터에는 추출된 각 특성의 성격 및 발생빈도와 발생위치에 따른 값과 가중치, 그리고 그 외의 부수적인 값들이 함께

부여되어 분류의 근거가 된다. 따라서 학습(training) 과정이 진행될수록 특성벡터의 크기도 증가하게 되는데, 이는 분류기(classifier)의 지식이 점차적으로 확장되어 감을 의미하는 것이다. 결과적으로, 분류기는 새로운 문서에 대해 어떠한 항목(category)으로 지정할 것인지에 대한 정보를 특성벡터를 통하여 얻음으로써, 자동으로 분류를 수행하게 된다.

<그림 4> 분류도구에서 사전구성을 위한 학습과정 3)



일반적으로 텍스트마이닝 분류 결과는 미리 정의한 분류항목(category)에 대한 관련도의 점수치(score)로 나타난다. 점수치는 각 항목에 대한 상대적 혹은 절대적 수치 값으로 나타낼 수 있다. 이렇게 추가적으로 발생된 정보(점수치)는 텍스트 문서를 저장하는 데이터베이스의 별도 필드에 기입이 된다. 텍스트마이닝 시스템을 디자인할 때는 가장 높은 관련도를 갖는 1순위 항목으로 지정하거나, 2순위 항목과 점수 차이가 작을 경우 분류를 이중으로 하거나(중출), 또는 사람이 개입하여 분류를 인위적으로 지정하는 방법을 채택할 수 있다. 또한 절대적 신뢰수치가 현저하게 낮은 문서를 처리하기 위해서는 수록 내용의 가치가 떨어지기 때문에 문서를 삭제하거나 현재의 분류항목을 수정해 새로운 분류를 추가할 수도 있다. 최윤정 등은 분류의 정확도를 높이기 위해서 텍스트마이닝과 데이터마이닝을 통합하는 방법을 제안하였다.3)

어떤 방식으로 텍스트마이닝 시스템을 디자인할 것인가는 마이닝을 실시하는 목

적과 수집 분석하고자 하는 텍스트 문서의 유형에 따라 달라져야 한다.

<표 1> 텍스트마이닝 분류결과 예 (Total Score의 평균값: 442.07)

Rank Score / Data Id	1	%	2	%	3	%	4	%	5	%	Total Score	Assign	분석 결과
634227	F	97	A2	1	B2	1	B3	1	B1	0	726.33	F	유효
1114467	F	74	B3	11	A3	6	A1	5	B2	3	662.8	F	유효
1678389	B1	29	A3	28	A1	17	F	15	B2	10	514.42	B1	무효
10389	A3	74	F	9	B1	7	A2	6	A1	4	139	A3	무효

(자료) 참고문헌 (3)

다. 텍스트 요약

텍스트 요약(Text Summarization)은 문서의 전체 내용을 반영할 수 있는 일부 내용을 추출하는 과정이다. 정보 검색 결과 건수가 많을 경우 전체 내용을 보기 전에 초록을 보게 되는데, 이러한 초록을 컴퓨터가 생성해주는 것이라고 할 수 있다.

텍스트 요약 기법은 크게 표면수준접근(Surface level approach), 개체수준접근(Entity level approach), 화법수준접근(Discourse level approach) 3가지를 들 수 있다. 표면수준접근은 텍스트에 포함되어 있는 단어, 위치, 주제에 근거하여 텍스트의 일부 내용을 추출하는 방법이다. 개체수준접근은 텍스트를 구성하는 단어를 그래프 또는 트리의 형태로 표현하는 방법이고, 화법수준접근은 요약을 추출하는 자체보다는 통신을 목적으로 하는 접근 방법으로 텍스트를 표준화 양식으로 변환하는 과정이다. 일반적으로 텍스트 요약은 3가지 중 2가지 이상의 기법을 조합해서 이용한다. 텍스트 요약 정밀도는 요약한 결과를 전문가의 직감에 의해 평가하는 방법과 미리 수작업에 의해 작성한 요약문과 요약프로그램에 의해 작성한 요약의 유사도를 평가하는 방법이 있다.

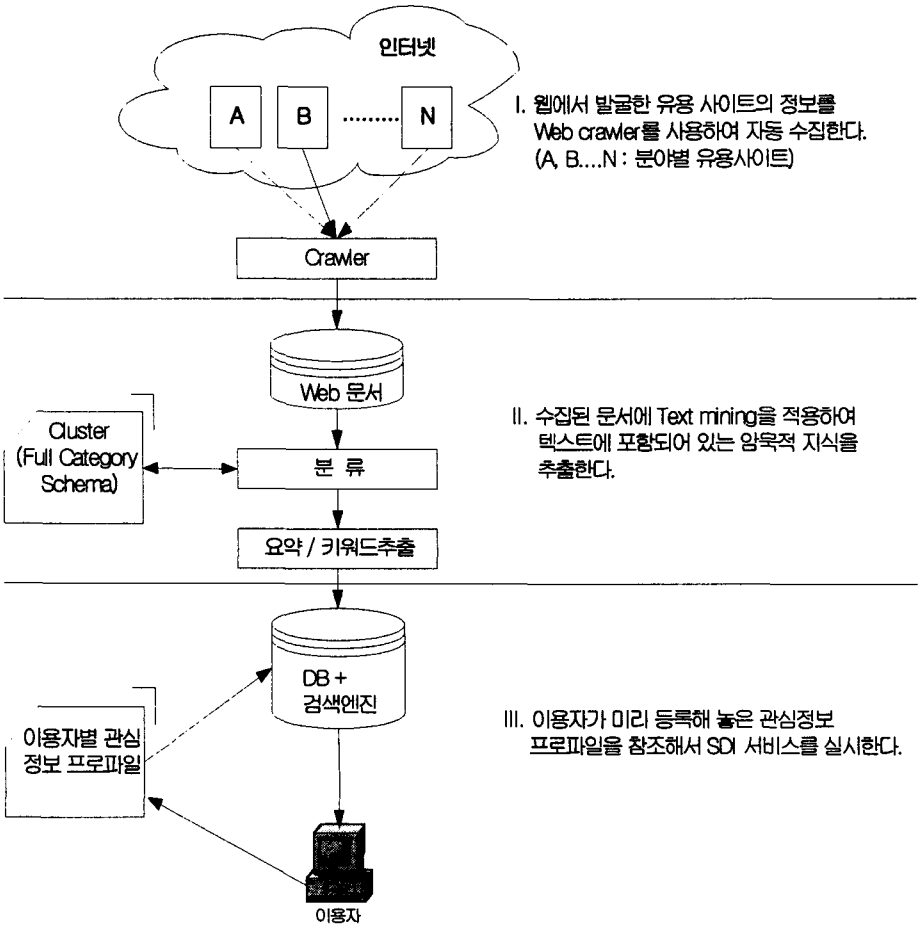
III. 텍스트마이닝의 응용

1. 기술정보분석 과 지식공학시스템(검토중)
2. 정보검색시스템(검토중)
3. 전자상거래 등 응용분야(검토중)

Ⅳ. 텍스트마이닝을 적용한 정보수집 및 분석 모델

분야별로 유용한 웹사이트에서 가치 있는 정보를 선별하고, 분석 재가공하여 새로운 유용한 정보를 생성하고, 이를 필요로 하는 사람에게 주기적으로 전달하는 모델(<그림 5>)을 앞에 나열한 텍스트마이닝 등의 정보처리 기술을 적용하여 제안하다. 여기서 제안하는 모델은 직관적 정당성에 근거하는 것으로 구현가능성과 효과에 대해서는 향후 연구가 요구된다.

<그림 5> 분야별 웹문서 수집 및 분석 모델



직관적 정당성의 근거를 나열하면 다음과 같다.

- 특정 분야에 대하여 상대적으로 유용한 사이트가 존재한다. 인터넷 검색엔진에서 'digital library'라는 키워드로 검색을 하면 130만 사이트가 검색이 되고, 이중에

는 각 학문 분야별 Web reference가 제공하는 사이트가 포함되어 있다. 예를 들자면 DLNET(Digital Library Network for Engineering and Technology)는 미국 NSF의 지원을 받아 구축한 공학기술분야 전체에 대한 Site Link를 구축해 놓은 사이트이다. 이는 분야별로 유용한 사이트를 발굴할 수 있다는 것을 의미한다.

유용사이트를 발굴하는 방법으로는 ①분야별 유용사이가 될 확률이 높은 사이트를 사전에 조사하거나 이미 구축된 사이트 정보를 이용하여 후보 사이트를 정하고, 이 상트를 대상으로 페이지 랭킹을 부여하는 방법과, ②전 페이지를 대상으로 특정 분야와의 유사도가 높은 분야를 일일이 선정하는 방법이 있을 수 있는데 실효성에 의문이 있다.

- 유용한 사이트에 대한 객관적 정당성은 검색엔진의 중요 사이트별 랭킹 기술을 도입하면 확보될 수 있다. 특히 구글 홈페이지 랭킹 방법은 분야별 유용 사이트를 발굴하는데 효과적일 것이다. 구글의 랭킹 방법은 ①어떤 사이트가 다른 사이트에서의 인용 회수(Back link)가 많은 사이트가 중요하다. ②링크를 요구하는 사이트의 중요도에 따라 차별화 하여야 한다. 즉 Yahoo와 같이 중요한 사이트에서 인용이 된다면 링크가 오직 한개 뿐일지라도 그 사이트는 살펴볼 만한 가치가 있다.

사이트별로 중요도를 평가하는 방법으로는 ①구글사와 같은 이미 웹사이트 랭킹 프로그램을 개발한 기관에 의뢰하는 방법과 ②랭킹 알고리즘을 자체 개발해 사용하는 방법이 있을 수 있다. 앞의 방법은 구글사와의 협의가 필요하기 때문에 불확실한 점이 있다. 후자의 경우는 구글을 직접 개발한 Sergey Brin이 1998년에 발표한 "The pagelink citation ranking" 등의 논문에서 기본 알고리즘을 참조할 수 있기 때문에 비교적 용이하게 접근할 수 있을 것이다.

- 이러한 방식으로 홈페이지를 평가하는데 소요되는 시간이 그다지 길지 않아 현실적인 솔루션이 될 수 있다. 실제 2,600만 페이지를 평가하는데 단 몇 시간이면 가능하다고 한다.

- 텍스트마이닝에 대해서는 기본적인 기능과 응용사례를 전술한 바 있다. 실효성은 위 모델에 대한 프로토 타입을 개발하는 시점에 입증이 되겠지만, 컴퓨터를 이용한 마이닝 기술의 발전은 그 결과를 낙관할 수 있게 해준다. 이미 SAS, SPSS와 같은 통계 분석 프로그램을 개발한 기업들도 축적해 놓은 데이터마이닝 기술을 응용하여 본격적으로 텍스트마이닝 분야에 뛰어 들고 있고 인터넷을 통해 구할 수 있는 텍스트마이닝 툴의 개수가 이미 70여개를 넘어서고 있다. 기술적 가능성을 토

대로 시장 확보가 가능하다는 판단을 전제로 할 것이다.[7]

- SDI 서비스는 이미 보편화 되어 있는 기술로서 도입이 용이하다. 단지 이용자가 스스로 작성한 프로파일의 신뢰도가 낮은 문제가 있다. 관심주제 보다 너무 포괄적인 키워드(예를 들면 '컴퓨터')를 등록해 놓거나 채택한 키워드가 보편적인 어휘가 아닐 경우 검색건수가 매우 적어지는 문제 등은 해결 과제이다.

V. 결론 및 향후 연구

본 연구에서는 텍스트마이닝의 기본 개념과 응용 사례를 살펴보았으며, 또한 이를 토대로 분야별 기술정보 수집 및 분석 모델을 제시하였다.

텍스트마이닝은 웹문서와 같이 비구조적 텍스트 문서를 분류하고 분석하여 문서에서 새로운 지식을 추출해내는 기술로서 최근 그 응용 범위가 넓혀지고 있다. 텍스트마이닝 기술을 적용하여 텍스트 형태의 답안을 비교하여 컨닝을 하였는지 밝혀내는 제품도 상용화 되었다.

그러나 텍스트마이닝이 자연어 처리에 기반을 두고 있기 때문에 의미 파악을 위한 계산이 복잡하여 처리시간이 과다하게 소요되고, 분석후 오류율도 많이 개선할 여지를 가지고 있다. 또한 2개 이상의 언어를 동시에 인식하고 궁극적으로는 다국어를 동시에 처리할 수 있도록 성능이 개선되어야 한다.

각 웹사이트의 규칙성을 감안하여 파싱을 할 경우 비구조적 텍스트 문서를 구조화 하는게 가능하다. 텍스트마이닝 시스템이 이와 같이 domain knowledge를 어떻게 잘 이용해서 parsing efficiency를 개선할 수 있을지, 그리고 보다 더 compact 한 intermediate form 을 만들어 낼 수 있을지에 대한 연구가 필요하다. 현재까지는 전문가들을 대상으로한 텍스트마이닝 툴이 있을 뿐이지만, 향후에는 KMS (Knowledge Management System) 의 일부로서 비전문가들도 쉽게 사용할 수 있는 툴이 개발되어야 한다. 자연어 질의 처리가 이에 해당된다. 또한 agent 의 형태로 존재하며 사용자의 프로파일을 학습해서 텍스트마이닝을 효과적으로 수행해주는 형태도 생각해볼 수 있다.

4장에서 제안한 '분야별 웹문서 수집 및 분석 모델'은 프로토타입의 설계과정을 거치면서 직관적 가능성을 실제 입증하기 위한 연구가 진행되어야 한다. 각 요소기술의 과제, 현실적 제약, 실제 적용의 한계 등에 대해서는 추가적인 실증적 연구가 필요하다.

또한, 텍스트마이닝은 자연어로 된 텍스트 문서를 대상으로 하여, 텍스트의 구조화/분류, 군집화 및 요약(중요 문장 추출) 등을 통하여 비구조적 텍스트로부터 관계를 추출하여 지식을 발견하는 것으로, 데이터베이스화(=구조적 텍스트화)하는 수준으로 한계가 있을 수 있다.

따라서, 기술정보분석에 활용하기 위해서는 텍스트마이닝, 데이터마이닝을 포함하는 지식공학시스템 적용을 검토하여야 할 것으로 보이며, 웹 등의 정보를 분류하고 중요 기술을 경보(Alert)하고 분석하여 심층적인 기술정보분석에의 활용은 심도있게 검토해야 할 것이다.

[참고문헌]

Morris, Steven et al.(2002), "DIVA: A Visualization System for Exploring Document Databases for Technology Forecasting", Computers & Industrial Engineering, Sep., 43-4, 841-862.

Yoon, Byung-Un, Yoon, Chang-Byung, Park, Yong-Tae(2002), "On the development and application of a self-organizing feature map-based patent map", R&D Management, 32-4, 291-300.

문영호 외(2000), 「온라인 DB검색을 통한 기술분석시스템 구축」, 산업자원부.

최윤정(2002), 「웹 컨텐츠의 분류를 위한 텍스트마이닝과 데이터마이닝의 통합 방안 연구」, 한국인지과학회논문지, 제13권, 제3호, pp.33-46.

설성수(2002), "기술분석의 고도화", 「기술혁신학회지」, 5권 3호, 260-276.

1) Ah-Hwee Tan (1999), Text Mining : The state of the art and the challenges

2) IBM Systems Journal, vol.40, n0.4, 2001, p.968-969

3) 최윤정, 웹 컨텐츠 cml 분류를 위한 텍스트마이닝과 데이터마이닝의 통합 방안 연구, 한국인지과학회논문지, 제13권, 제3호, pp.33-46.

4) 조태호, 텍스트마이닝의 개념과 응용, 지식정보인프라, 2001.6, pp.76-85.

5) Berry de Bruijin, Int. Journal of Medical Informatics, no.67, 2002, pp.7-18.

6) Dorre J., P. Gerstl, R. Seiffert, "Text Mining: Finding Nuggets in Mountains of Textual Data", in Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

7) Text Mining, Review of TPAC Technologies for ONR, 2002.8, pp.1-3.