웹컨텐츠의 분류를 위한 텍스트마이닝 시스템 설계 및 구현

최윤정[0] 박 승 수
이 화 여 자 대 학 교 컴 퓨 터 학 과
{ cris, sspark}@ewha.ac.kr}

Design and Implementation of Text Mining System for Web-Contents

Yun-Jeong Choi[0] Seoung-Soo Park
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

데이터마이닝기능은 문서의 구조화되지 않은 텍스트문서 보다는 일반적인 데이터베이스 등의 구조화된 자료를 주대상으로 한다. 정보화의 과정속에서 많은 기업이나 조직들이 과거의 시스템을 데이터베이스로 구축하는 경우가 많아졌지만, E-business, E-commerce가 활성화되면서 보유하고 있는 데이터베이스 기반이 아닌 무작위의 새로운 데이터가 생성되기도 한다. 고객을 위한 창구로서 활용되는 게시판이나 웹 검색사이트에서 초기수집한 데이터는 이러한 비구조적 데이터의 좋은 예이다.

본 연구에서는 비구조적인 텍스트위주 문서에 숨어있는 정보들을 발견하기위한 텍스트마이닝 시스템을 설계하고, 이를 검색회사가 보유한 웹컨텐츠 문서와 문서집합에 대해 분석도구를 적용하는 어플리케이션을 구현해 보았다. 대규모의 문서집합에 분석도구를 이용함으로써 빠른 문서처리가 가능하고 많은 양의 문서들을 다룰 때의 시간비용을 최소화 시킬 수 있다. 특히, 새로운 카테고리에 따른 분류 및 재조직이 가능할 뿐 아니라 검색결과의 품질을 향상시킬 수 있는 방법이 될 수 있다. 또한 텍스트마이닝 과정을 통해 발견한 지식과 특징들을 기반으로 반구조화된 파일이나 데이터베이스로 변환도 가능하다. 이러한 결과에 일반적인 데이터마이닝기법이나 OLAP 적용을 위한 전처리과정으로 활용하여 규칙발견이나 의미 있는 새로운 결론을 얻을 수 있을 것이다.

1. 서론

대부분의 기업과 웹사이트에 있어 기존의 데이터베이스 기반이 아닌 무작위로 드나드는 사용자들의 동선들로부터 생성되는 데이터처럼 데이터베이스 구조를 가지지 않았지만 상당한 잠재적 가치를 지니고 있는 텍스트 데이터가 있다. e-mail이나 웹 검색결과, 관련문서들을 고려해볼 때, 이러한 대규모의 텍스트 데이터들을 사용자의 필요에 의해 개인적으로 데이터베이스로 구성하여 사용하기가 쉽지 않고, 정보들이 문서에 함축되어 있기 때문에 간단한 단어라도 찾기가 어렵다. 따라서 이러한 과정에 텍스트마이닝 기법이 적용될 수 있다.

텍스트마이닝은 문서정보마이닝, 즉 텍스트로부터 숨겨져 있거나 흥미있는 지식을 발견하고, 특성추출기법을 기반으로 방대한 양의 문서집합에서 지식탐사를 위해 패턴을 이끌어내는 과정이라고 할 수 있다. 또한 최근, 많은 기업들에서 데이터간의 관계, 패턴을 탐색하고 모형화 하여 기업의 의사결정에 적용하기 위해 시도되고 있는 데이터마이닝은 구조화되지 않은 텍스트문서보다는 일반적인 데이터베이스와 같은 구조화된 자료에 초점이 맞춰져 있다. 따라서 데이터마이닝 작업을 위해서는 적용될 데이터가 정확하고 표준화되어야 하며, 구조와가 잘 되어진 후에야 비로소 데이터마이닝 적용이 가능할 것이다. 이 부분은 데이터마이닝의 전처리 단계에서 텍스트마이닝 기법인 특성추출, 분류과정을 통해 얻은 결과를 기반으로, 데이터마이닝 작업에 적합한 구조적인 형태로 변환함으로써 보다 쉽게 해결할 수 있다. 따라서 텍스트마이닝은 가치있는 텍스트정보의 데이터마이닝을 위한 전처리 단계에서 효과적인 도구로 활용될 수 있으며 문서에서 정보를 추출하고, 주제별로 문서를 구성하는데도 사용될 수 있다. 또한 문서집합에서 주요주제를 찾고, 검색도구를 통해 관련문서를 검색하는 시스템을 구현하는데 사용할 수 있다.[2][5]

본 논문은 다음과 같이 구성되어 있다. 2장에서는 일반적인 텍스트마이닝의 분석도구의 종류와 특징을 간단히 살펴보고 3장에서는 인터넷환경에서의 텍스트마이닝 시스템 설계과 적용결과를 반구조화일과 데이터베이스로 이끌어내기 위해 제안한 시스템 구성 설계를 보인다. 4장에서는 웹컨텐츠의 분류를 위한 텍스트마이닝 시스템구현과 실행결과를 보인 후 5장에서 결론을 내린다.

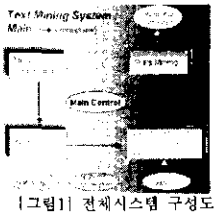
2. 관련연구

텍스트마이닝 기술체계는 자연어처리기법, 시각화, 데이타베이스기술, 기계 학습, 그리고 데이타마이닝 분야를 포함하고 있다. 텍스트마이닝에서 개발된 많은 기술과 도구는 문서의 텍스트에서 정보나 지식을 발견하고 추출하는데 사용된다. 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.[3][6]
첫째, 작성된 문서의 언어를 자동으로 찾아내는 언어식별도구, 둘째, 문서를 미리 정의된 카테고리나 주제등을 자동으로 지정해주는 주제분류도구, 셋째, 사전 데이타를 이용하여 문서에서 미리 정의된 용어 없이 중요한 용어나 이름, 약어, 장소 등의 항목을 자동으로 인식하는 특성추출도구, 넷째, 유사한 문서집합을 자동으로 그룹이나 '클러스터'로 나눌수 있는 클러스터링도구, 다섯째, 문장을 분석하여 문서의 요약정보를 추출하는 요약도구등이 있다. 각 분석도구를 적절히 이용하여 문서에서 정보를 추출하고, 주제별로 문서를 구성할 수 있다. 또한 문서집합에서 주요주제를 찾고, 검색도구를 통해 관련문서를 검색하는 시스템을 구현하는데 사용할 수 있다.[4][7]

본 연구에서 사용한 텍스트마이닝 툴인 IBM의 인텔리전트 마이너의 분석기법은 NLP기법인 N-gram과 frequency에 기반하고 있으며, 분석도구를 간단히 살펴보면, 특성추출도구는 문서를 분석하고 문서내용에 대한 중요도에 따라 단어의 등급을 매긴 후, 문장에서 각 단어의 등급들을 모두 더해 문장의 총등급을 매긴다. [표 1]은 관계성을 추출하는 예를 보이고 있다. 주제분류도구는 각 카테고리의 샘플데이터에 특성추출도구가 선행되어야 하며, 분류의 결과는 카테고리 이름의 목록과 각 문서에 대한 신뢰도 레벨로 나타난다. 문서는 하나 이상의 카테고리에 지정될 수 있고, 신뢰도가 낮으면 보통 분류자가 최종 결정을 내릴 수 있도록 문서를 따로 구분해 놓아야 한다. 새로운 카테고리를 정의해야 할 경우 카테고리에 대한 예를 모아서 학습프로세스를 재실행해야한다.[4]
클러스터링에는 최대차이지점에서 모음을 클러스터로 나누는 하향식 접근 방법과 비슷한 문서들을 그룹으로 계속 추가하는 상향식 접근방법이 있는데, 이 두 방법들을 모두 사용하여 문서집합을 다른 시각으로 볼 수 있으며 다른 견해를 얻을 수 있다. 이러한 분석도구들을 여러가지 방식으로 조합하여 반복적으로 수행함으로써 사용자가 마이닝 솔루션을 작성할 수 있다.

3. 시스템 설계

전체 시스템구성은 [그림1], 접근방식은[그림2]와 같다. 텍스트마이닝 진행순서는 어떠한 구조도 갖지 않은 텍스트 문서나 구조의 의미가 모호한 문서를 중간단계의 형태로 변환하는 텍스트정제 과정과 그 중간형태로부터 패턴이나 지식을 추론해내는 지식발견 과정으로 진행한다. 그리고, 그 결과를 기반으로, 이후에 적용될 분석도구에 적합한 구조로 변환하여 좀더 나은 결과로 이끄는 것을 목표로, [그림3]의 텍스트마이닝 모델을 설계하고,' 데이터 준비' 와 분류, 학습 등 텍스트마이닝의 전 단계를 활용할 수 있는 어플리케이션을 구현해보았다.
데이터준비과정과 분석적용단계는 입력데이터의 형태와



[그림1] 전체시스템 구성도 [그림2] 텍스트분석 접근방식

성격에 따라 각기 다른 처리가 요구된다. 지난연구[1]에 이은 KD Nugget 사이트의 로이터 뉴스데이터를 대상으로 한 연구에서는 한 파일당 약 2000건에 이르는 기사의 형태로, 특정 tag를 중심으로 split하는 기능이 선행되었다. 이번 데이터는 검색회사가 보유한 방대한 량의 URL집합들로 약38000건에 달하는 URL에 해당하는 온라인문서를 수집해온 후, 이를 URL, TITLE, CONTENTS로 구분하는 기능이 요구되었다.
이미 정의된 그룹으로 분류된 각 자료들의 특성을 인식한 후 새로운 자료를 그들의 그룹에 할당하는 '분류'과정를 위해 클러스터링을 연결시켜 문서의 개략적인 이해를 도와 카테고리선정을 돕고, 학습을 통해 사전을 구성한후, 신뢰도 5순위까지 결과를 얻어내고 재학습에 의한 재분류가 가능하도록 하였다.[그림3,4]

[그림3] 시스템 적용과정 [그림4] 사전구성을 위한 학습과정

4 시스템 구현

본 응용프로그램은 Windows NT와 Solaris 2.6상에서 인텔리전트마이너(IM4T)와 델파이를 사용하여 구현하였다. IM4T의 텍스트분석도구는 [표1]에서처럼 UNIX나 DOS의 커맨드라인 형식으로 프로그램에서 구현하기에 용이한 function을 제공한다.

```
command: "imzxrun -b 2 -f C -x n -o outfile 199380.txt
<IMZ ID>199380.txt</IMZ ID>
<IMZ TITLE> Top 10 Wacky Internet Events Of 1999 </IMZ TITLE>
<IMZ CONTENT>
NC 2 America PLACE
NC 1 Apple UNAME
NC 1 Blair Witch Project ORG
NC 1 Local Education Outreach Program ORG
NC 2 National Science and Technology Week ORG
NC 1 Forget Super Bowl UNAME
NC 1 Fox News ORG
NC 1 New York City PLACE
NC 1 President Clinton PERSON
NC 1 Web OTHER
NC 2 Y2K UNAME
</IMZ CONTENT>
```

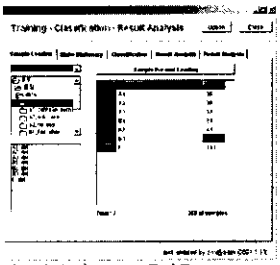
[표 1] 인텔리전트마이너의 특성추출도구 예

4.1 동작 시나리오

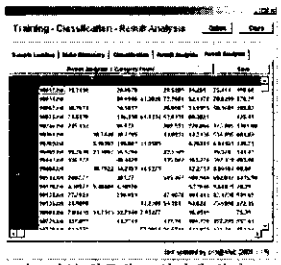
1	url에 해당하는 온라인문서 수집, 약 38000건
2	[표2]과 같이 <imgage>,<script>,<tag>가 제거된 URL,TITLE,CONTENTS형식의 문서로 재저장

준비	3	분석도구로의 indirect input을 위한 전체 data의 파일이름을 갖는 full dataset 파일생성
분석도구적용	1	언어식별도구 : 사이트의 주요언어 구분
	2	클러스터링도구 : 대규모 문서집합에 대한 개요 획득
	3	특성추출도구:이름,용어,관계등을 추출, 주제분류도구에서 정의해야할 항목으로 사용할 수 있는 정보획득
	4	2와 3을 통한 카테고리 선정 및 학습을 위한 sampling (sample data,training data,validation data) **주제사용자의 관심사에 따라 항목추가
	5	주제분류도구 : 신뢰도가 낮으면 분류자가 최종 결정을 내릴 수 있도록 문서를 따로 구분
해석		결과해석 및 post-processing
활용		비구조적인 텍스트문서들에 텍스마이닝을 적용하여 얻은 지식과 분류결과를 기반으로 반구조적인 형태를 갖는 파일로 변환하고 데이터마이닝 적용

4.3 실행화면

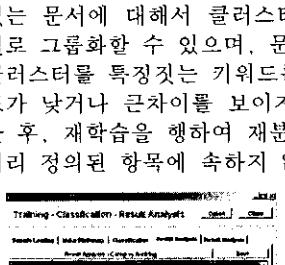


|그림5| 각 sample문서를 loading

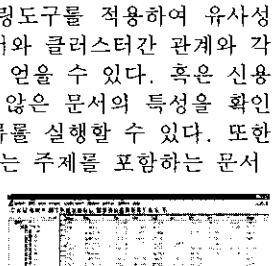


|그림6| 항목별로 얻어낸 화면

분류과정은 간단히 다음과 같이 정리해 볼 수 있다. 각 카테고리별 샘플메시지세트를 수집하여 [그림4]의 학습과정을 위한 학습데이터로 사용한다. [그림5]는 정의된 분류항목과 학습데이터를 loading하여 건수를 표시한다. [그림6,7]은 IM4T의 주제분류도구를 이용한 결과의 분석을 위한 화면으로, 분류결과를 순위별 및 항목별로 나타낸다. 문서마다 분류항목과 신뢰도가 나타나는데, 주제분류과정 중 문서가 임계값을 넘는 신용도가 없다는 것은 결정을 내리기 위한 충분한 증거를 찾지 못한 경우이며, 직접 메시지를 읽어 분류해야 함을 의미한다. 남아있는 문서에 대해서 클러스터링도구를 적용하여 유사성별로 그룹화할 수 있으며, 문서와 클러스터간 관계와 각 클러스터를 특징짓는 키워드를 얻을 수 있다. 혹은 신용도가 낮거나 큰차이를 보이지 않은 문서의 특성을 확인한 후, 재학습을 행하여 재분류를 실행할 수 있다. 또한 미리 정의된 항목에 속하지 않는 주제를 포함하는 문서



|그림7| 순위별로 얻어낸 화면



|그림8|데이터마이닝적용(PolyAnalyst)

URL: http://www.usatoday.com/news/index/clinton/clin328.htm
TITLE: MSDW Online Trade stocks on the go with your Palm(tm)!
CONTENT: uBid YOU set the price on thousands of products. Career Builder Find the perfect job for you Search the site the Web ···White House intern Monica Lewinsky is trying to help cover the mounting costs of her legal defense and also allay her family's fears about her future by working at the office of her attorney, William Ginsburg, her father says. "She's working in Bill's Washington office," Bernard Lewinsky said

|표2| 온라인문서를 수집하여 만든 입력데이터

n. Data	1.	2.	3.	4.	5.	6.	7.	Total
1. 10017.txt	F. 75.444		B3. 34.255		A1. 31.2418		B2. 29.5895	A3. 20.0675 ... 190.60
2. 10031.txt	B3. 92.4178		A3. 90.8648		B2. 72.7064		F. 70.0299	B1. 44.2028 ... 370.25
3. 10057.txt	F. 60.3686		B3. 31.8955		A1. 30.2021		B2. 26.6667	A3. 20.5012 ... 169.03
4. 10075.txt	A3. 146.148		B3. 80.3921		A1. 77.8129		B1. 64.4434	B2. 62.8138 ... 425.41
5. 10246.txt	F. 473.865		B2. 289.531		B3. 226.664		A1. 215.112	A3. 96.672 ... 1301.80
6. 10361.txt	F. 554.995		B3. 14.3136		B2. 11.0994		A2. 10.7448	A3. 10.4705 ... 601.62
7. 10389.txt	A3. 110.087		B1. 11.0505		F. 6.84401		B3. 6.26311	A2. 5.46303 ... 139.71
8. 10409.txt	A3. 56.5794		A2. 23.1807		B2. 22.6309		A1. 19.7838	F. 19.379 ... 141.43
9. 10443.txt	F. 302.429		B3. 161.276		A1. 136.772		B2. 135.062	A3. 68.6429 ... 803.98
10. 10488.txt	B1. 14.5273		A3. 14.2153		B3. 12.2713		A2. 10.7522	F. 9.09184 ... 60.98
11. 10512.txt	A1. 2407.77		F. 692.017		B2. 545.407		B3. 506.948	A3. 283.77 ... 4435.90

|표3| 구조적인 text 문서로 저장

들이 자주 나타나면 새로운 항목을 정의해야 하며, 새로운 예를 모아서 [그림4]의 학습프로세스를 재실행한다. 결과결과해석에서는 마이닝도구를 적용하여 이끌어낸 결과를 해석하고 시각화기능을 적용한다. 준비한 데이터를 재선택, 임의추출, 통합, 여과하거나 변환하여 반복수행을 통하여 놓쳐버릴 수 있는 특징들을 찾아낸다.

5. 결론

본 연구에서는 검색사이트회사가 가진 미분류데이타들의 대상으로 문서내에서와 문서집합에 대해 분석도구를 적용하여 지식을 발견하는 과정을 설계 및 구현해 보았다. 반복적인 샘플테이터 적용과 다양한 분류로의 작업이 시도되었으며, 최종적으로 서비스대상으로 적합한 것과 부적합한 것을 가려내어 검색결과의 품질을 향상시킬수 있었다. 즉, 컨텐츠내용에 불법자료나 성인자료가 있는 것으로 높은 신뢰도의 결과를 보일경우 불가로 판정을 내리고, 전체 5순위의 신뢰도차이가 0.3%미만일 경우 결과를 무시했다. 단, 1위가 shop, 2위 not legal이면 not legal한 shop이라 판단했다. 또한, 1-2순위편차가 5% 미만이고 1-2위가 not lega l, legal을 포함하면 결과를 무시했다. 최종결과는 전체 38000건의 데이터중 21436건이 적합, 나머지는 부적합으로 나타났고, size를 고려하여 xml tag대신' ,' 으로 구분한 text로 저장했다. 이에 엑셀이나 [그림8]와 같은 다른 마이닝도구에 적용하는 작업들이 가능하여, 마이닝 수행과정을 통해 변환한 형태는 다른 서비스를 위해 유용하게 활용할 수 있었다.

6. 참고문헌

[1] 최윤정, "인텔리전트 마이너를 이용한 텍스트마이닝 시스템의 설계 및 구현", 한국정보과학회,춘계학술발표논문,2000
[2] Dö rre J., Gerstl, P., and Seiffert, R., "Text Mining:Finding Nuggets in Mountains of Textual Data" Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999
[3] Lee Hing Yan, "Text Mining-Knowledge Discovery from Text", Trend in Knowledge Discovery from Databases, 29th June 1999
[4] Kevin K., "Mining Online Text", Commun. ACM 42,(Nov. 1999)
[5] IBM white paper for Intelligent Miner for Text
[6] Larsen, B., Aone, C., "Fast and Effective Text Mining Using Linear-Time Document Clustering", Proceedings of the Fifth ACM SIGKDD International Conference on KDD, 1999
[7] Witten,I.H., Frank,Eibe, DATA MINING, Morgan Kaufmann Publishers, 1999