텍스트마이닝 기술을 이용한 효율적인 검색시스템 알고리즘에 대한 연구

김제석* · 김장형**
*제주대학교

A Study of an Efficient Retrieval System Algorithm Using a Text Mining

Je-seok Kim* · Jang-Hyung Kim**
*Cheju National University
E-mail : Captanp@cheju.ac.kr

요 약

현재 네트워크 자원과 온라인 정보의 증가속도는 기존 정보시스템의 운용한계를 초과하고 있으며 서버의 처리속도나 네트워크 트래픽 해결을 위해 하드웨어 업그레이드와 네트워크 대역폭 확장으로 많은 문제가 제기 되고 있다.

본 연구에서는 많은 양의 온라인 데이터에서 원하는 문서의 위치를 빠르게 검색 할 수 있는 알고리즘을 연구함으로써 문서집합의 내용변화 또는 사용자의 관점변화에 적용한 최적의 검색내용을 검색할 수 있는 유기적 통합시스템 아키텍처를 제안한다.

ABSTRACT

Currently some problems are presented by the enlargement of network range and hardware upgrade for the solutions for network traffic and treatment speed of server processing, as well as the resource of networks and increasing speed of on-line information that is exceeding in operation limit of existing information systems.

The study proposes the Architecture, an organic unification system of optimized content for retrieval, which is adapted to variable points of view of users or content changes of document aggregation by the study of algorithm, which offers easy retrieval of the location of documents on a multitude of on-line data.

키워드
Text Mining, Retrieval System, Data Mining

Ⅰ. 서 론

정보화 사회에서는 인터넷과 이에 기반한 e-Business는 기존 산업의 전부분에 걸쳐 효율성과 생산성 증대를 위한 전략적인 도구로 그 중요성이 지속적으로 증대되고 있으며, 효율적 정보검색(information retrieval)은 각종 의사결정에 매우 중요하며 그 결과에 따라 개인이나 기업, 그리고 국가의 성패가 달라질 수 있다. 새로운 기업정보 자료들이 끊임없이 등록되고, 지난 자료들이 갱신, 수정되는 등 전 세계에 있는 수많은 기업에서 다양한 지식자산(Knowledge Asset)들이 지속적으로 생성, 저장, 재사용하는 정보 중 20%만이 활용성이 높은 정형 데이터로 구성되어 있고, 나머지 80%는 워드프로세서, e-mail, 프리젠테이션, 스프레드시트,PDF와 같은 복합문서와 인터넷 페이지

등의 비정형 텍스트 형태로 구성되어 있다[1]

검색 엔진들이 너무나 많은 정보를 검색해 주기 시작하면서 검색의 문제는 원하지 않는 정보들 사이에서 유용한 정보를 찾는 것으로 변화하였다. 최근 기업에서 잠재적인 정보를 발견해 내기 위해 많이 사용하는 데이터마이닝 기술 중 비구조적인 텍스트 문서로부터 정보를 찾아 지식을 발견하는 것이 텍스트마이닝이다. 그러나, 텍스트 마이닝은 정형화된 데이터를 위한 일반 데이터마이닝에 비하여 정보추출 능력이나 정확성 등이 많이 떨어지는 경향이 있다. 본 연구에서는 이러한 문제점을 해결하기 위하여 검색시 불필요한 불용어를 최우선적으로 제거하여 최적의 검색어를 추출하여 사용자 요구사항에 맞는 결과를 얻을 수 있게 가중치를 부여 유용한 정보를 상단에 랭킹시키므로 가장 유용한 정보를 검색할 수 있는 최적의 알고리즘 아키텍처를 연구하는데 최종 목표를 두고 있다.

II. 본 론

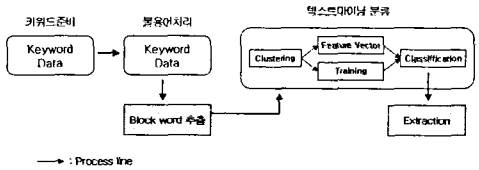
1. 텍스트 마이닝 기술에서 최적의 검색어 추출의 필요성

텍스트마이닝은 비구조적인 테이터안에서 테이터끼리의 관계와 패턴을 추출하여 그 내용을 자동 분류하고 새로운 지식을 생성하여 작업에 활용되고 있다. 오늘날 대부분의 정보들의 확실히 구조가 잡히지 않은 텍스트 형태로 존재하므로 그 내용을 정확히 파악하기 위해서는 내용끼리의 연관 관계와 패턴을 파악하여 정확한 정보의 추출과 불필요한 정보를 제거하여 보다 요구자의 필요한 정보를 손쉽게 검색함으로써 능률적인 일처리를 할 수 있다. 따라서 기존의 텍스트마이닝 기술체계는 주로 특성추출(feature extraction)를 통하여 특성백터(feature vector)를 생성하는게 특징인데,[2] 여러 번 나온 단어를 중요도가 높은 단어로 간주하고 가중치를 부여하여 정보와 지식을 발견하고 그 내용을 분류화, 군집화를 시켜 새로운 지식을 생성하는데 그 정확성이 접속사나 관사, 형용사등 우리말의 특성에 따라 검색결과가 다소 떨어진다. 그래서 본 연구에서는 정보를 찾고자하는 요구자의 검색어에서 접속사, 관사, 형용사등 불필요한 불용어를 제거하여 보다 정확한 검색어를 추출하고 기존의 텍스트 마이닝 기술을 이용하여 검색어의 연관관계가 높은 중요도에 가중치를 두어 상단에 랭킹함으로써 최적의 결과를 사용자에게 제공함과 동시에 능률적인 업무처리가 향상이 될 것이다. 그러므로 최적의 검색어를 추출하는 것은 기존의 불필요한 정보나 과다한 정보 검색결과로 정보 검색의 비효율을 개선하고 사용자가 요구한 키워드가 동음이어나 단어 중간 부분이 같을 때의 오 검색을 최소로 줄여주므로써 사용자 편익을 증진 시킬 수 있다.

2. 최적의 검색분류를 위한 마이닝시스템 설계

2.1 전체시스템 구조

본 연구에서 제안한 전체시스템을 설명하는 구성도로서 각 시스템의 내용은 단계별로 "키워드 준비" - "불용어처리" - "텍스트마이닝분류" - "결과출력" 순으로 처리한다.



(그림 1) 전체 시스템 구성도

요구자가 입력한 키워드에서 불필요한 불용어를 불용어 사전(불용어가 저장된 데이터베이스)을 이용하여 제거하고 정보가 저장된 데이터에서 키워드에 합당한 분류를 찾은 후 해당 그룹에서 검색, 추출하여 요구자에게 출력 결과를 나타낸다.

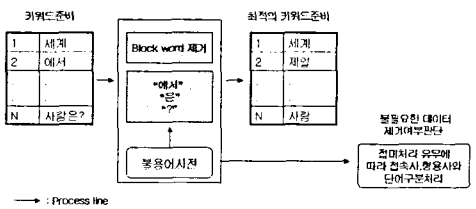
2.2 단계별 상세 구조

2.2.1 키워드 준비

사용자가 검색 대상에서 정확한 결과를 얻기 위해서는 검색결과의 오류를 최소화하고 특히, 철자 및 띄어쓰기 오류가 있는 키워드와 같이 불필요한 요소를 전 처리 과정에서 조정이 필요하며, 또한 입력 데이터가 자연어 또는 일반 단어 등 어떠한 키워드와 상관없이 처리 가능하도록 한다.

2.2.2 불용어처리

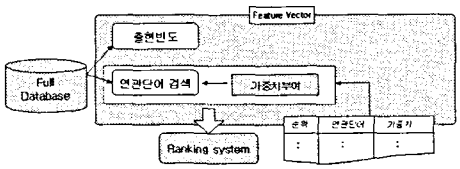
키워드로 입력된 데이터에서 저장된 불용어사전에서 불필요한 데이터를 키워드에서 제거한다. 단 여기서 단어와 접속사, 형용사의 중복 구분를 접미에 처리되는가에 유무에 따라 제거여부를 판단하는 알고리즘을 적용 후 키워드에서 유효한 단어를 추출한다.



(그림 2) 불용어 처리 구성도

2.2.3 텍스트마이닝 분류

이 과정에서는 현재 단어만 추출된 상태이므로 저장된 데이터베이스에서 다양한 텍스트 기법을 적용함으로써 최적의 지식을 발견해 낸다. 출현빈도에 따른 중요도 가중치를 적용하고 연관단어기법을 통한 최적 검색결과를 순위별로 표시한다.



(그림 3) 텍스트마이닝 분류

2.2.4 결과출력

텍스트마이닝에 의한 분류결과를 이용하여 가중치 가장 높은 데이터를 상단에 제시함으로써 사용자의 키워드에 적합한 최적의 결과를 얻을 수가 있다.

3. 시스템 구현 및 적용

기존 텍스트마이닝 기법을 이용한 실험과 본 연구에서 키워드에서 불용어 제거후 검색할 수 있는 시스템 구현을 응용프로그램인 Windows 2000server상에서 운용할 수 있는 웹프로그램인 ASP로 실험내용을 수행하여 비교 분석한다.

3.1 시스템 시나리오

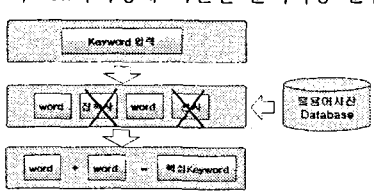
기존 방법	제안 방법
1단계:Keyword준비	1단계:Keyword준비
2단계:텍스트마이닝기	2단계:불용어 처리
법을 이용한 검색	3단계:텍스트마이닝기
3단계:가중치 부여	법을 이용한 검색
4단계:불용어처리	4단계:가중치 부여
5단계:검색결과	5단계:검색결과

본 실험에서는 뉴스기사, 일반문서등 저장되어 있는 파일에서 기존방법과 제안방법를 비교 검색 수행 하였다.

3.2 구축방법

3.2.1 불용어 처리 구현

- 한국어 형태소 분석 모듈 포함
- 불용어 사전을 이용한 불용어 추출
- 텍스트마이닝에 기반한 문서자동 분류구현

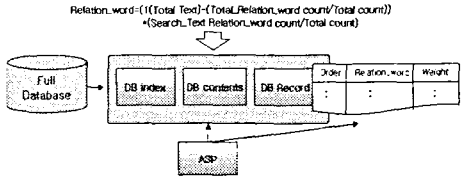


(그림 4) 불용어 처리(Block_word)

3.2.2 가중치 구현

- 연관단어를 찾아내어 가중치부여

(예:"자동차"는 "엔진"이라는 단어가 빈번하므로 가중치를 부여 순위에서 상단에 위치함)

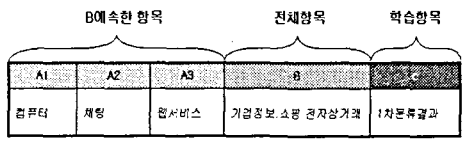


(그림 5) 가중치 구현(Relation_word)

가중치가 가장 높은 수치를 상단에 위치 하여 최적의 검색결과를 표시함으로써 고품질의 검색 기능을 수행하였다.

3.2.3 자동분류 구현

데이터 전체 집합에서 찾고자 하는 검색어를 분류시 텍스트마이닝에 의한 좀더 세부적으로 분류한다. 일반적으로 "채팅"이라 검색 시 전체문서에서 검색보다는 연관된 분류인 "컴퓨터"라는 카테고리에서 검색이 보다 빠르고 정확한 검색이 가능한 것처럼 세부적인 분류를 추가함으로써 보다 결과의 정확도와 검색속도를 개선할 수가 있다.



(그림 6) 자동분류 구현

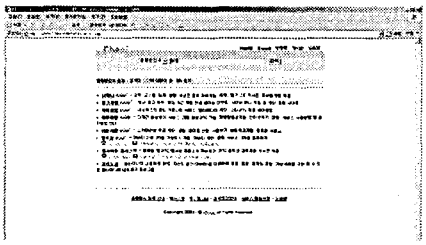
4. 실험결과

저장되어 있는 데이터 중 기존 방법을 통한 데이터 검색보다 월등히 빠른 시간 내에 검색 결과을 얻을 수 있었으며 제안방법에서 사용자가 생각했던 연관단어가 상당히 많이 나왔으며 오류정보검색도 어느 정도 향상이 되었다. 그러나 문서정보가 많을 경우 검색시간이 오래 걸려 분산 시스템 환경에서 처리한다면 좀 더 나은 속도 개선이 될 것이라 본다.

order	relation_word	weight	count	search_text
1	컴퓨터	0.76855	15	serch.text
2	키보드	0.38556	14	game11.text
3	마우스	0.28455	13	hard.html
4	모니터	0.15466	12	netthard.html
5	채팅	0.15444	11	theway.html
6	인터넷	0.15325	11	kserr.html
7	하드웨어	0.15125	10	dought.html
8	소프트웨어	0.14999	9	sys.html
9	CD	0.14555	9	ford.html

(그림 7) 가중치 결과 화면

텍스트마이닝 기반으로 ASP웹프로그램 제작한 사이트로 불용어 처리 구현하여 가중치 적용 가능하다.



(그림 8) 시스템 구축 및 검색 화면

4.1 기존방법과 제안방법의 비교

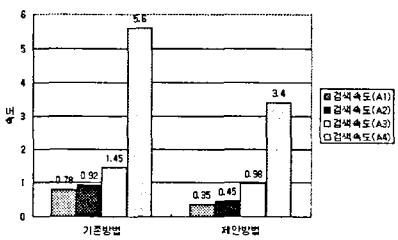
본 연구의 가장 특징적인 것은 검색어에서 불용어를 추출하여 검색결과를 보다 정확히 표현하고 검색속도를 개선하기 위함이다. 기존 텍스트마이닝기법은 자연어(Full text) 검색 시 전체문서에서 검색 후 연관단어를 찾아 가중치를 부여하여 불용어가 포함된 문서를 제거하는 방법으로 검색 시 많은 검색시간이 많이 소요되고 오검색이 많았다. 그러나 개선된 제안방법으로 검색 시 검색속도와 오검색이 많이 개선되었다. 다만 다량이 문서는 여전히 검색속도가 느렸다. 그건 기존 검색엔진 시스템에서 없는 연관단어에 가중치를 부여하여 관련단어가 많은 순서로 검색결과를 표현하기 때문이다. 또한 현재 사용중인 검색시스템은 검색속도는 빠르나 사용자가 원하는 검색결과가 중반이나 후반부분에 위치하는 단점이 있다. 본 연구에서 제안하는 방법으로 검색 시 보다 나은 검색 결과를 얻을 수 있다.

- 기존방법과 제안방법의 연관단어 측정비교

내용 / 방법	연관단어	가중치
기존방법	3.5	0.78
제안방법	3.6	0.79

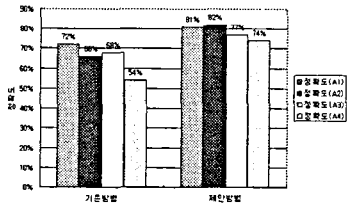
이부분에서는 기존방법과 연관단어 적용과 가중치는 큰 차이가 없으며 약간이 속도만 개선되었다.

- 기존방법과 제안방법의 검색속도 비교



기존방법에 비하여 검색속도가 개선 되었으며 약간의 소스코드를 변경 시 좀 더 많은 개선이 가능할 것 같다. 추가적으로 여러 번 실험 후 시행착오를 개선하여 기존보다 30%~40%정도 향상이 가능 하라 본다.

- 기존방법과 제안방법의 정확도 비교



검색시스템에서 가장 중요한 정확도가 기존방법보다는 20%~30%로 향상이 되었으며 불용어처리가 키워드에서 먼저 수행함과 동시에 패턴에 의한 세분화된 분류기준이 항목별이 근소한 수치를 가진 모호한 문서들이 분류 가능하여 보다 빠르고 정확히 검색이 가능하다는 증명이기도 하다.

III. 결 론

비구조적 문서가 다량이 포함된 웹 문서 검색 시 현재 연구되어 온 텍스트마이닝 시스템 보다 오류율과 속도 개선은 그동안 웹 중심의 자연어 텍스트문서를 자동 분석하기 위한 효율적 시스템에 새로운 접근이 이루어 졌으며 기존 검색 시스템의 여러 문제점과 한계들을 효과적으로 극복하고 사용자의 편익 증가 시킬 수 있다. 정보검색시 높은 서비스 만족감과 관련된 부가적인 비용을 줄여 효율적인 검색시스템 개발이 기대가 된다. 다만 앞으로는 시멘틱 웹기술을 이용한 지식검색 기능으로 한 단계 발전 할 수 있는 연구와 여전히 다량의 문서 검색 시 속도 개선 문제점을 있으므로 분산 환경시스템을 이용한 텍스트마이닝 연구가 진행이 되어야 할 것 이라 생각된다.

참고문헌

[1] 노나카 이쿠지로, Michael Polanyi, Delphi Group 1998.

[2] Yang, Y., "An Evaluation of Statistical Approaches to Text Categorization". Journal of Information Retrieval, 1999.

[3] Lee, H-Y, "Text Mining-Knowledge Discovery from Text", Trend in Knowledge Discovery from Databases, 29th June 1999.

[4] 구글 개발자들이 쓴 'The anatomy of large scale search engine' 논문

[5] 텍스트마이닝 기술을 적용한 대용량 온라인 문서데이타의 계층적 조직화 기법", 서울대 학위논문

[6] Soumen Chakrabarti , "Mining the Web Discovering knowledge from Hypertext Data" , MORGAN KAVFMANN PUBLISHERS