韓國 OA學會 論文誌
第4卷 第1號, 1999. 3.

99-4-1-2-1

자동 색인을 이용한 문서의 분류

신 진 섭*, 장 수 진**

Classification of Documents using Automatic Indexing

Jin-seob Shin* Soo-jin Jang**

요 약

본 논문은 단어들의 유사도를 이용하여 문서들을 자동으로 분류하는 새로운 방법을 제안한다. 단어들 중에서 의미있는 단어들을 찾아내기 위하여 자동색인 방법을 이용하였으며, 두 번째로 본 논문에서 제안한 확률 모델을 이용하여 각 단어들의 문서와의 연관관계를 분석하였다. 이를 토대로 분류를 가능하게 하기 위한 프로파일을 생성한다. 본 논문에서는 유전자 알고리즘과 신경망에 관련된 10개의 문서에 대하여 실험하여 유전자 알고리즘과 신경망에 해당하는 프로파일을 생성하였다.

Abstract

In this paper, we propose a new method for automatic classification of documents using the degree of similarity between words. First, we seek relevance terms using automatic indexing. Second, we found frequency in use words in documents and the degree of relevance between the words using probability model. Continuously, we extracted the set of words which is connected the relevance closely and created the profiles characterizing each classification. And, with the profile we finally classified them. We experimented on classifying two groups of documents. Some documents were about Genetic Algorithm. The others were about Neural Network. The results of the experiments indicated that automatic classification with word accordance of degree enable us to manage the retrieved documents structurally.

---

* 대전 보건대학 사무자동화과 전임강사
** 대전 보건대학 전산정보처리과 전임강사
논문접수 : 98. 12. 21. 심사완료 : 99. 2. 8.

Ⅰ. 서 론

현대는 기하급수적으로 늘어나는 다량의 정보로 인하여 다량의 정보를 저장, 관리하고 필요에 따라 사용자가 요구하는 정보를 빠른 시간 안에 정확하게 서비스할 수 있도록 만드는 정보관리 시스템이 널리 이용되고 있다. 정보 서비스 또는 정보 검색 시스템의 성공 여부는 정보 서비스의 정확성과 정보 추출의 속도에 의해서 결정된다.

정보검색을 빠르고 정확하게 하기 위해서는 읽어들인 자료들에 대해서 색인을 두어 관리해야 한다. 이러한 색인을 하는 방법은 색인 작업의 주체에 따라 수작업 색인(manual indexing)과 자동 색인(automatic indexing)으로 구분할 수 있다. 그러나, 수 작업 색인은 색인 할 전문 인력의 확보가 어려워지고 색인 작업에 소요되는 시간적인 지연으로 인해 엄청난 양의 정보 자료에서 신속한 검색이 불가능해짐에 따라 컴퓨터를 사용한 자동색인에 관한 연구가 시작되었다. 인공지능의 한 영역으로 발전한 자연어 처리로 인하여 자동색인은 많은 발전을 이루게 되었다.

따라서 본 논문에서는 방대한 양의 문헌 정보에서 효율적인 정보 검색을 위한 기존 방법을 연구하며, 효율적인 정보관리 및 검색을 위한 색인어를 이용한 정보 자동 분류 알고리즘을 제안하고 실험한다..

Ⅱ. 자 동 색 인

자동색인은 크게 두가지 방법으로 분류할 수 있다. 첫 번째는 어구의 출현빈도를 고려하는 통계적인 방법인데 이 방법으로는

① 단순빈도에 의한 추출법(Lhun의 모델)

② 확률을 이용한 모델 (2 Poisson 모델)

③ 분산을 이용하는 방법(Dennis-Salton 모델)

④ 문서를 n차 벡터로 표현하는 방법(Vector space 모델)

등을 들 수 있다. 또 다른 방법으로는 언어정보를 이용하여 문서의 의미를 바탕어로 하여 색인어를 추출하는데 다음과 같은 방식이 있다.

① 형태소만을 처리하는 방식

② 특정어구에 관련된 명사를 추출하는 방식

③ 명사구를 처리하여 추출하는 방식

이와 같은 방법으로 시소러스를 이용하여 개념의 상, 하위 관계를 고려하는 방식들도 개발되었다. 그리고 색인어에 문서와의 밀집도를 표현하는 방식에는, 색인어로서의 자격 여부만을 표현하는 이진색인과 색인어와 문서와의 밀집도를 가중치로 표현하는 가중치 색인이 있다.

본 논문에서는 통계적 방법과 언어학적 방법간을 비교 하였고, 그 결과는 간단히 표 1과 같다.

표 1. 자동 색인의 분류

색인의 종류		구 현	장 점	단 점
통계적 기법		단어의 빈도계산 불용어 제거	구현이 간단	정확도가 떨어짐 한국어에는 적용이 어려움
언어 학적 기법	형태소 해석을 이용한 기법	단어의 형태소 해석 빈도수 계산	구현이 간단 한국어에 적용가능	정확도가 떨어짐 구단위의 추출이 어려움
	구문해 석을 이용한 기법	단어의 형태소 해석 구문 해석 특정 의미 구를 선택	정확한 색인어 추출 구단위 색인어 추출	안어 해석결과의 애매성 구문해석기 구현이 복잡
	의미 해석을 이용한 기법	문장의 완전한 이해	가장 정확한 색인어 추출	현실적으로 각종 사전의 구성과 문장의 완전한 이해 불가능

1. 통계적 방법

통계적 방법을 이용한 자동색인은 단어의 출현 빈도를 근거로 구해진 단어 빈도(term frequency), 문헌

빈도(document frequency), 장서 빈도(collection frequency)를 구한 다음, 이것을 전체 문헌에서 구해진 해당 단어의 상대빈도, 단순빈도를 곱한 값을 기준하여 색인어를 추출한다.

통계적 방법을 이용한 자동색인은 구현이 간단하지만, 단어의 빈도에 의존한 색인어 추출이므로 색인어 추출의 정확성이 떨어진다. 그리고 첨가어인 형태를 따는 한국어에는 색인어가 될수 있는 명사에 보통 어미(보통조사, 보조사)가 붙어져 있어 어미를 분리하지 않고는 각각의 단어에 대한 빈도수 계산이 어렵다.

2. 언어학적 기법

언어학적 방법을 이용한 자동색인은 문헌 내의 문장을 분석하여 색인어를 추출하는 방법으로, 문장의 분석정도에 따라 형태소 해석, 구문해석, 의미해석 단계로 나누어 진다.

2.1 형태소 해석을 이용한 자동색인 방법

형태소 해석은 일반적으로 문장 중의 각 단어에 대해 형태소 해석을 행하여 단어의 원형을 복원하고, 접두어, 접미어, 조사, 보조사 등을 분리해낸다. 그 다음 앞 절에서 통계적 방법을 이용하여 가능한 색인어를 모두 추출한 후, 마지막 단계에서는 불용어 사전을 이용하여 필요 없는 색인어를 제거하는 방법을 취한다.

형태소 해석을 이용한 방법은 구현이 어느 정도 간단하고, 한국어에도 쉽게 적용하여 사용할 수 있는 장점이 있다. 그러나, 형태소 해석을 이용한 자동색인의 기본 원리는 통계적 방법을 그대로 사용했기 때문에 통계적 방법과 같이 색인어 추출의 정확성이 떨어진다. 그리고 이 방법은 형태소 해석을 각각의 단어에 대해 수행 한 다음, 하나의 단어에 대한 빈도수, 가중치를 계산하므로 구 단위의 색인어 추출이 어렵다.

2.2 구문 해석을 이용한 자동색인 방법

구문 해석을 이용한 자동색인은 문장 내의 특정한 구문의 의미를 가지는 단어, 단어구가 문헌의 내용을 나타낼 수 있다는 가정하에서 구문 해석을 자동색인에 이용한다. 구문해석을 이용한 자동색인 시스템은 먼저 형태소 해석을 하고, 그 결과를 가지고 구문 해석을 한 다음 구문적 의미를 지니는 특정한 단어, 단어구를 색인어로 추출한다.

구문 해석을 이용한 자동색인 방법은 형태소 해석을 이용한 자동색인보다는 훨씬 정확한 색인어가 추출돌 뿐 아니라 구 단위의 색인어도 훌륭히 추출할 수 있다.그러나, 구문 해석을 이용한 자동색인은 자연어 구문 해석 결과에서 나오는 필연적인 애매성과 단어 자체에서 나오는 애매성이 있고, 실제적으로 구문 해석기 구현이 매우 복잡하다는 단점이 있다.

2.3 의미 해석을 이용한 자동색인 방법

의미 해석을 이용한 자동색인 방법은 지식데이타베이스(knowiedge database)나, 지식 표현, 시소러스(thesaurus)등을 이용하여 문헌의 문장을 완전히 이해한 다음 전체 문헌에 맞는 색인어를 부여하는 방법이다. 현재로서는 지식 데이터베이스, 지식표현, 시소러스, 등의 구현이 매우 어렵고, 실제적으로 문장을 완전히 이해하기는 불가능 하므로 이와같은 방법을 사용하는 자동색인 시스템을 구현하기는 아직까지 많은 문제점이 있다.

Ⅲ. 색인 추출 알고리즘 설계

1. 통계적 기법

1.1 단순빈도와 상대빈도

색인어 선정을 위한 통계적 기준은 모두 단어의 출현 빈도에 근거하고 있다. 출현 빈도를 직접적으로 이용하는 기준은 단어의 빈도 산출방식에 따라 단순빈도와 상대빈도로 한다.

단순빈도는 단어가 어디에 출현했는가에 따라 단어빈도(Term Frequency, TF), 문헌빈도(Document Frequency, DF),장서빈도(Collection Frequency, CF)로 구분한다. 단어빈도(TF)는 색인대상이 되는

각 문헌i에 특정한 단어k가 출현한 횟수로

$$TF=fik$$

이다.

문헌빈도(DF)는 특정한 단어K가 출현한 문헌의 수로

$$DF=\sum_{i=1}^{n} bik$$

이며,fik>1일 때 bik=1,fik=0일 때 bik=0이다.

장서빈도(CF)는 특정한 단어k가 전체 문헌집단 내에 출현한 총 빈도로

$$CF=\sum_{i=1}^{n} fik$$

가 된다.

단순빈도는 문헌집단의 크기나 분석 대상 텍스트의 길이, 또는 단어의 사용 빈도를 전혀 고려하지 않은 것이므로 실제로 이것만을 색인어 선정 기준으로 사용하기가 어렵다. 따라서 이러한 요인을 고려한 상대빈도가 보다 더 적합한 기준으로 평가되고 있다.

상대빈도는 위의 단어빈도를 각각 문헌빈도, 장서빈도, 문헌길이(한 문헌에 출현한 단어의 총빈도를 의미함) 등으로 나누어줌으로써 빈도의 값을 표준화시킨 것이다. 이러한 상대빈도를 공식으로 표현하면 아래와 같다.

wik =fik /DF

wik=fik/CF

wik=fik/Pi(Pi는 문헌i내 단어의 총빈도)

Wik=fik/CF*Pi

공식에서 w_{ik}는 문헌 i에서 단어k가 갖는 주제어로서의 중요도를 의미하며 가중치라고 한다.

1.2 검색문헌의 적합성을 이용한 기준

역 문헌빈도나 문헌분리 등의 기준은 단어 k의 출현빈도를 산출할 때 특정한 질문에 대한 문헌들의 적합성을 고려하지 않았다 다시 말해 적합 문헌과 부적합문헌을 구별하지 않고 전체 문헌을 통틀어 출현빈도를 산출하였다.

이러한 앞의 방법은 기준과는 달리 문헌의 적합성 정보를 이용한 기준들은 단어의 출현빈도 뿐만 아니라 단어가 출현한 문헌의 유형(즉, 적합 문헌과 부적합 문헌)을 고려한 것이다. 지금까지 제시된 기준으로는 적합 문헌과 부적합 문헌의 분포를 직접 이용한 적합성 가중치(relevance weight) 공식과 단어의 탐색 실용성 이론에 근거한 가중치 공식이 있으며, 또한 결정 이론을 이용한 기준도 주목할 만하다. 적합성 가중치 공식중 대표적인 것으로는 단어 정확도(term precision)를 측정한 것이 있다. 특히 단어의 정확도가중치와 실용도가중치는 실험 결과 상당히 높은 검색 효율을 가져오는 것으로 나타나 있다. 그러나 적합성에 근거한 가중치는 색인작업이나 탐색작업에 앞서 적합 문헌과 부적합 문헌의 수를 알아내기가 쉽지 않으므로 실제로 적용하기에는 어려운 기준으로 보인다.

원래 가중치는 색인어와 탐색어에 모두 부여할 수 있으나 적합성을 이용한 가중치는 주로 탐색시 질문을 구성하는 탐색어에 부여할 목적으로 연구되었다는 점이 앞의 기준들과의 차이점이라고 하겠다.즉,질문Q가 Q=(qi,q2,..qm)과 같이 질문벡터로 표현될 때 탐색어 qk에 부여되는 가중치를 말한다 탐색어에 적합성가중치를 부여하여 검색한 결과는 가중치를 사용하지 않았을 때보다 검색효율이 월등히 높아진다는 실험결과가 보고된 바 있다.

적합성 정보를 이용하여 탐색어가 아닌 색인어에 가중치를 부여한 연구로는 시스템 이용자의 프로화일로부터 문헌의 유용성을 산출하여 가중치를 부여한 색인방식이 있다.

1.3 적합성 가중치

색인을 검색도구로 사용하는 검색시스템에서는 특정한 재현율 수준에서 높은 정확율을 가져오는 색인일수록 효과적인 색인이다. 정확률은 검색된 문헌속에 적합문헌이 많이 있을수록 높아진다. 따라서 색인어의 정확성은 이 색인어에 의해 검색된 문헌들 가운데 적합문헌이 얼마나 되는가를 나타내며 이런 의미에서 단어의 정확도가중치라는 개념이 사용되었다.

질문Q를 구성하는 단어 k에 대하여

N=문헌집단내 전체문헌수,

n=단어k를 색인어로 갖는 문헌수(즉 검색될 문헌 수),
R=질문Q에 대한 적합문헌 수,
r=단어k를 색인어로 갖는 적합문헌 수,라고 하면 단어k를 갖는 문헌의 적합성에 따른 분포는 표 2로 나타낼 수 있다.

표 2. 적합성에 따른 분포

	적합문헌수	부적합문헌수	
단어k가 부여된 문헌수	r	n-r	n
단어k가 부여되지 않은 문헌수	R-r	N-n-R+r	N-n
	R	N-R	N

위의 분포를 다양하게 이용한 가중치 공식은 아래와 같다

$$W_1 = \log \frac{r/R}{n/N} \quad \cdots\cdots (1)$$

$$W_2 = \log \frac{r/R}{(n-r)/(N-R)} \quad \cdots\cdots (2)$$

$$W_3 = \log \frac{r/(R-r)}{n/(N-R)} \quad \cdots\cdots (3)$$

$$W_4 = \log \frac{r/(R-r)}{(n-r)/(N-n-R+r)} \quad \cdots (4)$$

공식(1)과 공식(3)은 단어 k의 적합 문헌 집단내 분포와 전체 문헌 집단내 분포를 비교한 것인 반면 공식(2)와 공식(4)는 적합 문헌집단내 분포와 부적합 문헌집단내 분포를 비교하여 준 것이다.

위의 네 가지 가중치의 검색 성능을 비교한 결과 검색성능은 높은 것부터(4)-(3)-(2)-(1)의 순이었고 공식(1)과 공식(2),그리고 공식(3)과 공식(4)는 각각 서로 비슷한 수준의 성능을 보여 주었다. 마지막 두 가중치의 성능은 앞의 가중치 보다 훨씬 나은 검색결과를 가져왔다. 특히 공식(4)는 단어의 정확도 가중치라고 부른다. 단어 k의 정확성은 이 단어가 적합 문헌집단에는 색인어로 많이 부여되고 부적합 문헌집단에는 적게 부여될수록 높아지며 따라서 이 단어는 색인어나 탐색어로 적합하다는 것이다.

정확도 가중치는 다른 적합성 가중치와는 달리 문헌에 부여되는 색인어의 가중치로 사용할 수 있도록 아래의 공식을 제시하였다. 즉 문헌Di에서 단어k가 갖는 색인어로서의 중요도인 가중치 Wik는 공식(4)의 정확도가중치 Pk에 이문헌 i에서의 단어빈도 fik를 곱한 값이다.

$$Wik = fik * Pk$$

위의 가중치를 색인어 선정기준으로 사용하면, 많은 수의 적합문헌 속에 출현하고 적은 수의 부적합문헌 속에 출현한 단어일수록 높은 Pk를 갖게되며, 또한 색인대상 문헌속에 많이 출현한 단어일수록 높은 가중치 Wik를 갖게 된다.

1.4 결정이론

결정 이론은 일종의 분류를 위한 기본적인 이론으로서 모든 이용 가능한 정보들을 통합하여 수식화 한다. 결정이론에서 적용되는 대표적인 결정 규칙으로는 조건부 확률을 이용하는 Bayes rule 과 색인어로서의 자격 여부를 결정하기 위하여 상대적 분포도 값을 기준치로 하여 기준치 이하일 때 충족된 클래스로 분류하는 분류 법칙(Classification rule)이 있다.

2. 분류와 자동색인

2.1 알고리즘 설계

본 알고리즘은 여러개의 문서에서 각각의 색인을 자동 추출한 후에 추출된 각각의 문서의 색인들을 확률모델을 이용하여 분류하는 알고리즘이다.

2.1.1 분류 알고리즘

본 분류 알고리즘은 크게 인공지능에 관련된 두 종류의 문서를 이용하여 구축한 알고리즘이다.

〈 Classification Algorithm 〉

① 두 단어 m, n의 관계 분석

Rmn = -log(probability that their appearance s are just by accident)

② 단어의 연관성 추출

PwcDc * (1- Pwc)D-Dc * DCDc

Pwc 는 하나의 문서가 두 단어를 동시에 포함할 확률, Dc는 두 단어가 동시에 사용된 문서의 수,

D는 전체 문서의 수이다.

③ 하나의 문서의 두단어 포함 확률

Pwc = Pwm * Pmn

④ 단어들 사이의 관계 분석 결과

Rmn = -ln(PwcDc *(1- Pwc)D-Dc * DCDc)

2.1.2 키워드 분류 알고리즘

〈Algorithm〉

① 단어 빈도(term frequency, TF)계산

② 역 문헌 빈도(Inverse document frequency, IDF) 계산

③ IF (TF 〉 상위 임계치)

THEN 불용어로 처리

IF(TF 〈 하위 임계치)

THEN 색인 대상에 부적절한 것으로 간주

④ TF/IDF 계산

⑤ classification algorithm 적용

⑥ 색인 대상 리스트의 결합구 생성

⑦ 구 결합(phrase combination) 비교

⑧ 동의어,반의어(thesaurus)사전 비교

⑨ 결합구의 의미론적 비교

2.2 실험

본 실험은 "genetic algorithm"에 관련된 5개의 문서와 "neural network"에 관련된 5개의 문서를 혼합하여 실험하여 〈표 3〉과 같은 프로파일을 구축 하였다.

표 3. 유전자 알고리즘과 신경망의 프로파일

a. 유전자 알고리즘

Profile : Genetic	
Genetic	3.7
Algorithm	3.7
Search	2.2
Crossover	2.2
Population	1.8
Pool	0.8
Mutation	0.7

b. 신경망

Profile : Neuron	
Neuron	4.3
Neural	2.2
Network	2.2
Input	0.9
Node	0.7
Weight	0.7

Ⅳ. 결 론

자동색인에 관한 대부분의 연구는 색인 자체에 관한 이론의 정립 보다는 자동색인과 수작업 색인어간의 비교에 관한 것들이었다.

본 논문에서는 자동 색인과 분류 알고리즘을 접목하여 읽어들인 문서들을 자동으로 분류할 수 있는 새로운 방법을 제시하였다. 아직까지 범용적인 실험을 하지 못하였으나 앞으로 좀더 보완하면 다양한 분류에 이용되도록하는 연구가 뒤따라야 할 것이다.

참고문헌

[1] 김두홍,서혜란, "색인 및 초록 작성법", 구미무역(주)출판부, 1990

[2] 사공철,윤구호, "최신정보검색론", 구미무역(주)출판부, 1990

[3] 송미련, "자동색인방법과 자동색인시스템 성능", 정보관리연구,1980

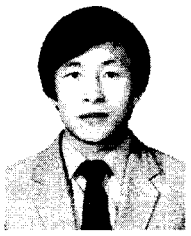
[4] 안현수, "한글문헌의 자동색인에 관한 실험적 연구", 연세대학교 석사학위논문, 1986

[5] 최기선, "한국어정보처리와 지능형 자동색인", 한국정보관리학회, 1991

[6] G. Salton, "The SMART Retrieval System", New Jersey : Prentice-Hall, 1971
[7] C.J.Van Rijsbergen, "Information Retrieval", Lon- don : Butterworths, 1979
[8] G.Salton, "Automatic Text Processing", Addison- Wesley Publishing Company, 1988

저 자 소 개

신진섭(申陳燮)
1986년 : 충남대학교 계산통계학과 졸업
1989년 : 건국대학교 대학원 전자계산학과 졸업, 공학 석사
1995년 : 건국대학교 대학원 전자계산학과 박사과정 수료
현 재 : 대전보건대학 사무자동화과 전임강사
관심분야 : 지능형 정보검색, 비주얼 프로그래밍



장수진(張秀鎭)
1985년 : 충남대학교 계산통계학과 졸업
1993년 : 충남대학교 대학원 계산통계학과 졸업, 이학 석사
1998년 : 한남대학교 대학원 전자계산학과 박사과정 입학
현 재 : 대전보건대학 전산정보처리과 전임강사
관심분야 : 소프트웨어 엔지니어링