문서의 키워드 추출에 대한 신경망접근

조태호, 서정현
연구개발정보센터, 정보시스템개발실
{jerry, tcjo}@ns.kordic.re.kr

Neural Based Approach to Keyword Extraction from Documents

Taeho C. Jo and Jerry Seo
KORDIC (Korea Research Development Information Center)

요 약

문서는 자연어로 구성된 비정형화된 데이터이다. 이를 처리하기 위하여 문서를 정형화된 데이터로 표현하여 저장할 필요가 있는데, 이를 문서 대용물 (Document Surrogate)라한다. 문서 대용물은 대표적으로 인덱싱 과정에 의해 추출된 단어 리스트를 나타낸다. 문서 내의 모든 단어가 내용을 반영하지 않는다. 문서의 내용을 반영하는 중요한 단어만을 선택할 필요가 있다. 이러한 단어를 키워드라 하며, 기존에는 단어의 빈도와 역문서 빈도 (Inverse Document Frequency)에 근거한 공식에 의해 키워드를 선택하였다. 실제로 문서내 빈도와 역문서 빈도 뿐만 아니라 제목에 포함 여부, 단어의 위치 등도 고려하여야 한다. 이러한 인자를 추가할 경우 이를 수식으로 표현하기에는 복잡하다. 이 논문에서는 이를 단어의 특징으로 추출하여 특징벡터를 형성하고 이를 학습하여 키워드를 선택하는 신경망 모델인 역전파의 접근을 제안한다. 역전파를 이용하여 키워드를 판별한 결과 수식에 의한 경우보다 그 성능이 향상 되었음을 보여주고 있다.

1. 서론

문서 데이터 베이스인 경우 문서의 내용은 물론 이를 정형적으로 표현한 인덱스를 각 문서와 연관하여 저장한다. 문서의 전체 텍스트는 자연어로 구성된 비정형 데이터이므로 컴퓨터가 이를 직접 처리할 수 없다. 문서를 컴퓨터가 직접처리 할 수 있는 형태로서 정형적인 데이터로 변환한 것이 문서의 인덱스이다. 문서의 인덱스는 주로 문서를 구성하는 단어들로 구성된다.

인덱싱이란 문서를 단어의 집합으로 변환하는 과정을 말한다. 문서의 모든 단어가 문서의 내용을 반영하는 것은 아니다. 그 중 스톱워드라 불리는 단어는 문법적 기능을 할 뿐 문서의 내용과는 무관하다. 영어의 전치사, 관사, 접속사, 대명사등이고 한국어에는 대명사, 관형사, 불완전 명사, 조사등이다. 스톱워드는 인덱스에서 제외하여야 한다. 대부분 명사 또는 동사가 문장의 내용을 반영한다. 이 단어들을 정보 키워드라 하고 인덱스에는 반드시 포함시켜야 한다 [1]. 스톱워드을 제외한 모든 단어가 정보 키워드가 되는 것은 아니다. 문서의 도메인 또는 내용에 의해 정보 키워드로서 인덱스에 포함하여야 하고, 불포함하여야 하는 단어도 존재한다. 예를 들어 컴퓨터에 관한 문서의 집합인 경우 컴퓨터라는 단어는 인덱스에 불필요한 단어 일 수 있지만, 다양한 내용의 문서인 경우는 컴퓨터라는 단어는 정보 키워드가 될 수 있다. 뉴스 기사에 있어서 정치에 관한 기사에 간헐적으로 등장하는 스포츠에 관한 단어는 정보 키워드가 될 수 없다. 전체 단어의 20%정도 만이 문서의 내용을 반영 한다.

문서로부터 내용을 실제 반영을 하는 키워드만을 추출함으로써 여러 가지 잇점이 발생할 수 있다. 정보검색에 있어서 실제의 내용을 나타내는 키워드 만을 질의와 매치함으로써 불필요한 문서의 추출을 방지할 수 있다. 문서의 자동 분류와 자동 요약 시 소수의 정보 키워드 만을 사용함으로써 문서 처리의 시간을 단축할 수 있다. 불필요한 단어로 인한 오분류 및 요약에서 불필요한 문장의 추출을 억제할 수 있다.

문서로부터 정보 키워드를 선택하는 것에 대한 연구가 진행되어 왔다. 1988년 G..Salton은 문서에서 단어를 추출하고 각 단어에 가중치를 부여하는 방법을 최초로 제안하였다 [1]. 1993년 F. Pereira 는 단어를 빈도에 근거하여 계층적으로 군집화하는 과정을 제안하였다 [2]. 1995년 Y.Yang은 통계적 방법으로 문서 자동 분류를 위한 특징으로서 단어를 추출하는 것을 제안하였다 [3]. E. D. Wiener는 석사 학위 논문에서 텍스트 분류에 신경망을 적용하기 위한 특징으로서 키워드를 선택하는 방법들을 열거하였다 [4]. 1997년 M.E. Maron은 정보검색에 있어서 문서의 관련성에 근거하여 키워드를 추출하는 방법을 제안하였다 [5]. 1998 년 Y. Tseng 은 다중언어의 키워드를 추출하는 방법을 제안하였다 [6]. 1999년 T. Hofmann은 문서로부터 키워드 추출에 있어서, 기존의 LSI (Latent Semantic Indexing) 을 개량한 PLSI

(Probability Latent Semantic Indexing) 기법을 제안하였다 [7]. 1999 년에 M. R. Brent 는 비교사 학습 모델을 이용하여 문서에 포함된 단어들을 군집화 하였다 [8]. S. Soderland는 rule 에 의해 문서의 키워드는 물론 중요 정보를 추출하는 WHISK 라는 시스템을 소개하였다 [9]. 2000 년 D. Freitag은 여러가지 기법을 조합하여 문서로부터 키워드와 중요 정보를 추출하는 방법을 제안하였다.

텍스트의 분류를 하기 위해서는 문서를 특징벡터로 표현되어야 하고 이를 위해 특징이 될 수 있는 키워드를 추출해야 할 것이다 [2][3]. 키워드는 문서 내의 빈도 (Term Frequency) 와 역 문서 빈도 (Inverse Document Frequency) 에 근거하여 키워드를 선택하는 방법을 개발해 왔다 [1][2][3][4][5][7][8].

문서 내의 단어에 대한 특징으로는 문서 내 빈도와 역 문서 빈도 뿐만 아니라 단어의 위치, 단어에 대한 다른 문서의 전체 빈도 등도 고려해야 할 것이다. 동일한 역문서빈도라 할 지라도 다른 문서의 전체 빈도는 상이할 수 있다. 영어의 경우 첫문장 또는 첫문단이 그 문서의 중요문장이 될 수 있으므로 단어의 문서내의 위치도 고려해야 할 것이다. 단어의 제목의 포함여부도 키워드 선택에 중요한 영향을 미친다. 그러한 요인들을 고려하여 수식을 표현할 경우 키워드 선택을 위한 공식이 복잡할 수 있다. 이 논문에서는 객관적으로 키워드와 그 외의 단어에 대한 정보를 추출하여 학습 패턴을 형성한 후 신경망 모델인 역전파에 학습시켜 키워드를 선택하는 방법을 제안한다. 단어의 문서 내 빈도와 역문서빈도 이외의 인자를 고려하고 이에 대해 신경망을 적용한 결과 정확도가 향상된 것을 제시할 것이다.

이 논문의 구성에 있어서, 다음절에는 신경망 모델인 역전파를 키워드 선택에 적용하기 위한 방법을 기술할 것이고 3 절에는 신경망 모델과 문서내 빈도와 역문서빈도에 근거한 전통적인 수식모델과 성능 비교를 한 결과를 제시할 것이다. 4 절에서는 현재 연구의 의의와 앞으로의 연구 계획을 결론으로서 제시할 것이다.

2. 키워드 선택에 대한 신경망 접근

이 절에서는 주어진 문장에 포함된 단어의 특징을 추출하여 신경망 또는 그 외 통계적인 패턴인식 모델을 접근할 있는 형태인 특징벡터를 형성하는 과정과 신경망 모델인 역전파를 적용하기 위한 입력노드와 출력노드의 정의에 대해 기술 하겠다.

단어의 특징 추출은 다음과 같다. 특징 추출 이전에 IDF와 ITF의 값을 설정하기 위해 충분한 개수의 표본 문서가 주어져 있음을 가정한다.

- TF (Term Frequency) : 주어진 문서내의 단어의 빈도
- IDF (Inverse Document Frequency): 표본 문서 중에서 그 단어를 포함하는 문서의 개수
- ITF (Inverse Term Frequency): 표본 문서 내에서 그 단어에 대한 총빈도
- T (Title): 단어의 문서제목에 포함 여부
- FS (First Sentence): 단어의 첫 문장 포함 여부
- LS (Last Sentence): 단어의 마지막 문장 포함 여부

위의 입력으로 주어지는 특징 벡터는 6차원으로 되어 있고 TF, IDF,그리고 ITF의 값은 정수로 주어지고 T, FS그리고 LS는 0과 1의 이진값으로 주어진다.

단어의 출력벡터로 주어 지는 특징은 다음과 같다.

- K (Keyword): 문서의 색인으로 포함되어 할 키워드
- NK (Non Keyword): 문서의 색인으로 제외되어야 할 단어

출력 벡터는 2차원으로 주어진다. 그리고 K와 NK의 값은 이진값으로 주어지고 동시 동일한 값이 주어질 수 없다.

위의 특징들을 신경망 모델인 역전파의 입출력으로 정의하여 그 구조는 그림 1과 나타낼 수 있다.

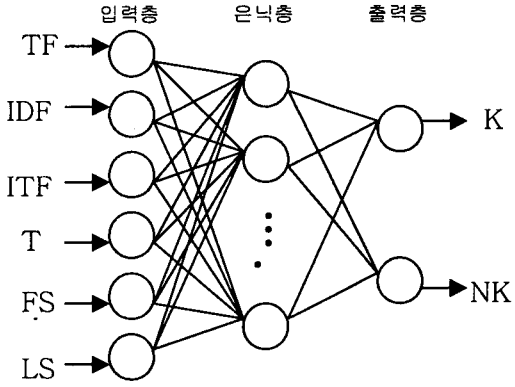

그림 1. 키워드 선택에 대한 역전파의 구조

위의 그림1에서 입력 노드의 개수는 6이고 출력 노드의 개수는 2로 정해진다. 은닉 노드는 임의로 정할 수 있다. 역전파의 학습 알고리즘은 문헌 [11]을 참조하고 이 논문에서는 생략하겠다.

3. 실험 및 결과

이 절에서 키워드 추출에 대한 신경망의 성능 결과를 제시하겠다. 뉴스기사를 도메인으로 하고 TF 및 IDF에 근거한 수식과 2절에서 제시한 6개의 인자에 의한 신경망 접근의 성능을 비교 분석하였다.

IDF와 ITF를 결정하는 표본 문서는 1000개의 영문 뉴스 기사를 www.newspage.com이라는 사이트에서 임의로 수집하였다.

학습패턴을 생성하기 위해 표본문서와는 별도로 동일한 사이트에서 뉴스기사를 수집하여 250개의 단어의 특징을 추출하였다. 각 단어는 수동적으로 뉴스기사의 내용을 기반으로 키워드와 비키워드의 범주를 설정하였다. 단어의 IDF와 TF는 표본문서를 참조하여 설정하였다.

테스트패턴은 다른 뉴스기사를 수집하여 학습 패턴과 동일한 방법으로 50개의 패턴을 생성하였다.

TF와 IDF에 근거한 2가지 공식을 이 논문에서 제안한 신경망의 성능에 대한 비교의 대상으로 설정하였다.

첫번째 공식은 1998년 삼성 SDS에서 개발한 지식경영시스템인 Kwave의 텍스트 분류 및 텍스트 요약에 필요한 키워드 추출에 사용되었으며, 다음과 같다 [12].

$$W_i = \frac{TF^m}{(IDF + ITF + 1)^n}$$

위의 공식에서 문서 내 빈도에 비례하고 역문서빈도와 역단어빈도에 비례하도록 설정되었다. 위의 공식의 m 과 n 의 값은 모수로 주어진다. 이 실험에서는 모수 m 과 n 을 모두

1로 설정하였다.위의 공식에 근거하여 순위선택, 임계선택, 그리고 비례 선택 3가지로 나타내며 그의 설명은 [13]을 참고하기를 바란다.

두번째 공식은 다음과 같고, 역시 TF와 IDF를 근거로 설정하였으며, 표본 문서의 개수를 또한 고려하였다 [14].

$$W_i = TF(\log_2 N - \log_2 IDF + 1)$$

위의 공식에서 N 은 표본문서의 개수를 나타내고, 단어의 빈도에 비례하고 역문서 빈도의 대수에 반비례하도록 설정되었다.

신경망 모델은 가장 보편적으로 사용되는 역전파를 채택하였고, 입력노드 6개, 은닉노드는 임의로 6개, 그리고 출력노드는 1개로 설정하였다. 학습률은 0.1로 설정하였고 학습횟수는 2500, 5000, 7500, 그리고, 10000회로 설정하였다.

위의 첫번째 공식과 두번째 공식에 대한 결과는 표1과 같이 나타내었다.

표1. 공식에 의한 성능 결과

선택율	첫번째공식	두번째공식
20%	0.68	0.7
40%	0.76	0.82
50%	0.74	0.8
60%	0.72	0.78
80%	0.68	0.7

표1에서 보는 바와 같이 키워드의 선택 방식은 비례선택을 채택하였으며, 단어를 위의 공식에 의한 가중치의 내림차순으로 정렬하여 전체 단어의 일정 비율을 선택한다. 첫번째 공식에서는 최고 76%의 정답률, 50개의 단어 중에 12개의 단어가 오분류 되었다. 두번째 공식에서는 82%의 정답률을 기록하였으며, 50개중에 9개의 단어가 오분류 되었다.

신경망에 의한 키워드 판별의 성능 결과는 표2와 같이 나타내었다.

표2. 신경망에 의한 성능결과

	2500	5000	7500	10000
0.9	0.58	0.58	0.58	0.62
0.8	0.7	0.68	0.74	0.68
0.7	0.72	0.7	0.74	0.76
0.51	0.8	0.84	0.92	0.84

신경망에 의한 성능 결과에서 임계값이 0.9인 경우 미분류가 많기 때문에 공식에 의한 경우보다 정답률이 낮다. 0.8인 경우는 첫번째 공식의 경우와 성능이 유사하고 0.7인 경우라도 첫번째 공식의 경우보다 향상이 되었으나, 두번째 공식의 경우보다 성능이 향상되지 않았다. 0.51인 경우 위에서 두가지 공식의 경우보다 성능이 향상되었고, 최고 92%의 정답률을 나타내었는데 50개 중에서 단지 4개의 단어만이 오분류를 나타낸 것이다.

5. 결론

이 논문에서 문서의 키워드 선택에 대해 학습을 통한 신경망의 접근 방법을 제안하였다. 단어의 문서내 빈도와 역문서빈도 뿐만 아니라 단어의 위치도 또한 고려하였다. 키워드 선택에 대하여 수식모델에 비해 신경망의 접근이 정확도를 향상 시키는 것을 제시하였다. 정확한 키워드의 선택은 문서를 키워드의 리스트로 인덱싱하여 저장하는 정보검엔진에 있어서 보다 불필요한 문서가 검색되는 것을 방지한다.

현재 연구에서는 소규모의 데이터로 키워드 선택에 대한 신경망의 성능을 검증하였는데, 단어 단위의 데이터에서 문서 단위의 데이터로 키워드 선택을 이 논문에서 제시한 수식 알고리즘 뿐만 아니라, k-NN, Bayesian Network, LSI (Least Semantic Index)등의 기법과 비교해야 할 것이다.

6. 참고 문헌

[1] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval", pp513-523, Information Processing & Management Vol 24 No 5, 1988.

[2] F. Pereira, N. Tishby, and L. Lee, "Distributional Clustering of English Words ", pp183-190, The Proceedings of 30 th Annual Meeting of the Association for Computational Linguistics, 1993.

[3] Y. Yang, "Noise Reduction in a Statistical Approaches to Text Categorization", pp256-263, The Proceedings of SIGIR 95, 1995.

[4] E. D. Wiener, "A Neural Network Approach to Topic Spotting in Text", Thesis submitted to the Faculty of the Graduate School of the University of Colorado, 1995.

[5] M.E. Maron, "On Relevance, Probabilistic Indexing and Information Retrieval ", pp 39-46, Readings in Information Retrieval edited by K. Sparck and P. Willett, 1997.

[6] Y. Tseng, "Multilingual Keyword Extraction for Term Suggestion", pp377-378, The Proceedings of SIGIR 98, 1998.

[7] T. Hofmann, "Probabilistic latent indexing ", pp50-57, The Proceedings of SIGIR 99, 1999.

[8] M. R. Brent, "An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery, pp71-105, Machine Learning 34, 1999.

[9] S. Soderland, "Learning Information Extraction Rules for Semi-Structured and Free Text ", pp233-272, Machine Learning 34, 1999.

[10] D. Freitag, "Machine Learning for Information Extraction in Informal Domains ", pp169-202, Machine Learning 39, 2000.

[11] S. Haykine, *Neural Networks* , Macmillan Colledge Publishing Company, 1994.

[12] T.C. Jo, "News Article Classification based on Representative Keywords in Categories", pp194-198, The Proceedings of CIMCA 99: Intelligent Image Processing, Data Analysis & Information Retrieval edited by M. Mohammadian, 1999.

[13] T.C. Jo, "The Categorization of News Articles with Informative Keywords", pp1136-1139, The Proceedings of ITC-CSCC 99, 1999.

[14] R. R. Korfhage, "Information Storage and Retrieval", John Wiley & Sons Inc, 1997.