문서 분류를 위한 문장 응집도와 주어 주도의 주제어 추출

안희국° 노희영
강 원 대 학 교 컴 퓨 터 과 학 과
creadelp@mail.kangwon.ac.kr, young@mail.kangwon.ac.kr

Sentence Cohesion & Subject driving Keywords Extraction for Document Classification

Heui-Kook Ahn° Hi-Young Roh
Dept. of Computer Science, Kangwon National University

요 약

문서분류 시 문서의 내용을 표현하기 위한 자질로서 사용되는 단어의 출현빈도정보는 해당 문서의 주제어를 표현하기에 취약한 점을 갖고 있다. 즉, 키워드가 문장에서 어떠한 목적(의미)으로 사용되었는지에 대한 정보를 표현할 수가 없고, 문장 간의 응집도가 강한 문장에서 추출되었는지, 아닌지에 대한 정보를 표현할 수가 없다. 따라서, 이 정보로부터 문서분류를 하는 것은 그 정확도에 있어서 한계를 갖게 된다. 본 논문에서는 이러한 문서표현의 문제를 해결하기위해, 키워드를 선택할 때, 자질로서 문장의 역할(주어)정보를 추출하여 가중치 부여방식을 통하여 주어주도정보량을 추출하였다. 또한, 자질로서 문장 내 키워드들의 동시출현빈도 정보를 추출하여 문장 간 키워드들의 연관성정도를 시소러스에 담아내었다. 그리고, 이로부터 응집도 정보를 추출하였다. 이 두 정보의 통합으로부터 문서 주제어를 결정함으로서, 문서분류를 위한 주제어 추출 시 불필요한 키워드의 삽입을 줄이고, 동시 출현하는 키워드들에 대한 선택 기준을 제공하고자 하였다.

실험을 통해 한번 출현한 키워드라도, 문장을 주도하는 주어로서 사용될 경우와 응집도 가중치가 높을 경우에 주제어로서의 선택될 가능성이 향상되고, 문서분류를 위해 좀 더 세분화된 키워드 점수화가 가능함을 확인하였다. 따라서, 선택된 주제어가 문서분류의 정확도에 있어서 향상을 가져올 수 있을 것으로 기대한다.

1. 서 론

컴퓨터와 인터넷의 발달로 점차 전자문서의 사용이 일반화되고, 사용량이 증가되고 있으며, 효율적으로 문서 정보를 관리, 이용하기 위해 문서분류를 하게 된다. 문서분류는 문서를 미리 정의된 하나 이상의 범주, 클래스로 분리시키는 일로서, 수작업으로 문서마다 범주를 지정해 줄 경우, 사람의 노력, 시간, 비용 면에서 심각한 어려움을 초래하게 된다. 따라서, 이를 자동으로 분류해주는 시스템의 요구가 증대되고 있다. [1]

자연언어 문장을 미리 정의된 집합으로부터 주제별로 분류하는 Text categorization (TC = Text Classification)은 규칙학습에 의한 선형모델(linear model), 의사결정트리(decision tree), k-최근접 이웃(k-nearest neighbor), 신경망(neural network), 유전자 알고리즘(genetic algorithm)등의 기계학습 패턴인식 방법들이 이용되고 있다. 이러한 연구들의 대부분은 문서에 나타난 키워드들을 벡터로서 나타내는 bag-of-words표현법을 사용하고 있으며, 단어의 의미(context)나 순서 등을 고려하지 않고, 주제와 연관된 단어들을 추출하기위해 자질로서 출현빈도나 패턴을 고려하고 있다. 따라서, 주제어 추출 시, 특정 키워드가 문서에서 중요하게 사용됨에도 불구하고, 낮은 빈도 값을 가질 경우, 누락되거나, 중요하게 사용되지 않음에도 불구하고, 높은 빈도값을 가질 경우, 선택되는 일이 발생된다. 그 결과 문서분류 시 잡음(noise)으로서 작용하게 되어, 분류기 설계에서 전체적인 일반화 성능에 정확도의 하락을 가져오게 된다.[2] 현재 이러한 문서의 표현문제를 해결하기 위한 여러 방법들이 연구되고 있는데, 주제어 추출 시 발생되는 잡음을 줄이기 위해, 제목과 문장의 중요도를 계산하여 중요도가 높은 문장에서 나타난 키워드를 주제어로 선택하는 방법이 연구되어 성능 면에서 향상을 얻은 바 있다.[3] 또한 출현하는 키워드의 연관성을 고려하기위해 키워드의 출현빈도정보와 공기(Co-occurrence)정보로부터 주성분분석 방법을 이용해 정량적으로 문서의 주제어를 추출하는 방법이 연구된 바 있다.[4][5]

이러한 연구들은 제목과 문장 내에 출현하는 키워드들이 서로 의미적(2차적)으로 연결되어있을 경우를 고려하지 않고 있어서, 제목과 문서 내에서 출현하는 키워드 벡터값이 서로 상이할 경우, 옳지 않은 결과를 도출할 수가 있다. 따라서, 중요문장을 추출 할 때, 제목과의 연관성뿐만 아니라, 문장 간의 관계에서 해당키워드의 문장 내 중요도를 고려할 필요성이 있다. 이를 위해 본 논문에서는 실제 문장 내에서 주어가 문장의 주체로서 의미를 주도한다는 점에 착안하여 키워드가 주어로 사용되었는지에 대한 정보로부터 키워드의 주어주도 정보량을 추출하였다. 또한, 문서 내 키워드의 응집도 정보를 얻어내기 위해 Kimoto&Iwadera의 시소러스를 변형하여 키워드 응집도 정보를 추출하고, 이를 통합하므로서 문서분류를 위한 주제어로서의 신뢰도를 높이고자한다.

2. 관련 연구

2.1 한국어의 문형구조

문장(sentence)은 단어들이 일정한 형식과 규칙에 맞게 배열(조합)되어 구성되고, 문형은 문장을 구성하는 요소들의 배합되는 문장의 틀(frame)을 말한다. 따라서, 문형구조를 밝히는 것은 문장을 구성하는 요소들의 문법적 관계, 의미적인 결합관계를 파악하는데 영향을 미치게 된다. 국어의 문장은 문장을 구성하는 요소들의 순서가 상당히 자유롭기 때문에 전형적인 문형을 확정하는 것은 쉬운 일이 아니다. 하지만, 가장 기본이 되는 형식인 "주어+서술어(모든 문장의 모체형식)"구조로부터 한 성분이 생략되거나, 다른 성분이 첨가된 형식으로 나타나게 된다. 특히, 국어의 모든 문장을 포괄할 수 있는 전형적인 문장 구조의 형식(기초형식)은 다음과 같이 분류할 수가 있다.

(1) 무엇이 무엇이다. (때가 봄이다.)
(2) 무엇이 어찌한다. (봄이 온다.)
(3) 무엇이 어떠하다. (날씨가 따뜻하다.)

이로부터 기본문형과 변화문형으로 나눠지게 된다.[6] 본 논문에서는 문형구조로부터 문장을 구성하는 요소들의 의미적 결합관계 정도를 파악하기위하여 한국어의 기초형식 중에서 공통으로 나타나는 "주어"를 중심으로 주제어를 추출하고자한다. 한국어가 갖고 있는 교착어의 특징상 "어근+어미"의 구조에서 명사가 어근으로서 출현이 가능하기 때문에 서술어는 본 논문에서 고려하지 않고, 단일문장의 관점에서 주어가 문장의 의미적주체로서 작용하기 때문에 주어를 중심으로 하는 의미체(semantic body)를 형성하고, 그로부터 주어주도 정보량을 추출하고자 한다.

2.2 주제어 추출기법

문서분류 시 문서의 표현은 개개의 문자대신 각 단어를 중요 특성(feature)로 고려하게 되며, 정보검색분야에서는 단어를 인덱싱 용어로 사용하기 위해 문서에서 단어의 순서는 큰 문제를 일으키지 않는다고 가정한다. 따라서, 별개의 단어는 자질이 되고, 문서내에서 단어 빈도수값-TF(w,d), 단어(w)가 한번이상 나타난 문서의 수-DF(w), 이를 혼용한 TF-IDF로 표현되기도 한다.[2] 이러한 전통적 문서표현방법은 기본적으로 문서 내 출현위치라든가 문장내에서의 역할이나 중요도를 고려하지 않기 때문에 이를 보완하기 위한 방법들이 연구되고 있다. 문서요약방법을 적용하여 주제어를 추출하는 방식이 있었는데, 이는 문장 내에서 단어가 출현했을 때, 단어가 포함된 문장이 제목문장과의 연관성을 고려하여, 문장의 중요도를 결정하고, 그에 따라 단어에 가중치를 부여하는 방식을 통해 주제어를 추출하였다.[3] 그리고, 단어가 동시에 출현했을 때, 해당 단어 간의 관련성정보를 주성분분석 방법을 통해 파악한 후, 주제어를 추출하는 방법이 있었다. 하지만, 이러한 방법들은 문장내에 키워드의 중요도를 고려하지 않고 있으며, 문서 내에서 문장간 관련성정보를 통해 해당 단어의 가중치 정도를 구분하는 데는 한계를 갖게 된다. 따라서, 본 논문에서는 주어주도 정보량과 의미체간의 관련정보를 시소러스에 저장하고 이로부터 추출한 응집도 정보를 이용하여 주제어를 추출하고자한다.

3. 자동 문서분류 시스템구조

본 논문에서 제안하는 전체 문서분류시스템의 구조는 동적 시소러스와 문서정보추출 에이전트, 폴더정보추출 에이전트, 문서분류기로 구성된다. 동적 시소러스는 기존의 문서들로부터 학습에 의해 키워드 벡터들의 관련정보를 갖게 되고, 폴더정보추출에이전트에게 문서분류 시 주제와 관련된 키워드를 제공하는 역할을 한다.

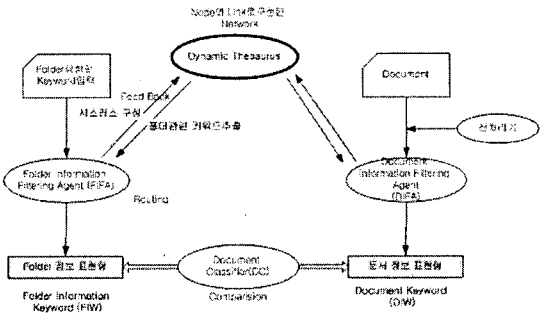


그림 1. 전체 문서분류 시스템 구성도

문서정보추출 에이전트(DIW)는 문서로부터 주제어를 추출하기위해 전처리과정을 거친 후, 응집도와 주어주도 정보계산을 한 후, 문서분류를 위한 주제어를 제시하게 된다.

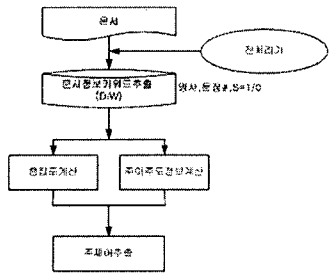
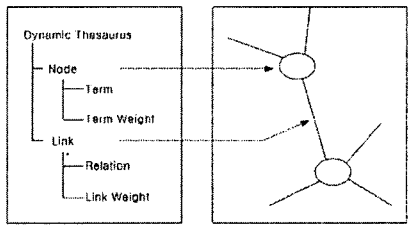


그림 2. 주제어추출 과정

3.1 응집도정보 추출

응집도 정보는 문서 내 키워드벡터들의 상호 관련성정보를 표현하는 것으로 본 논문에서는 문서 내 키워드들의 출현빈도정보와 문장 내 키워드들의 동시출현빈도 정보를 시소러스 구조에 담아 추출하였다.



시소러스	용어 정보	의 미
Term	Keyword	시소러스를 구성하고있는 용어이다.
Term Weight	Keyword Ranking	용어가 갖고있는 정보값으로 그 용어의 출현횟수에 대한 정보를 갖는다.
Relation	Relation Information	용어들간의 관계를 표현하는 관계로 한문장에 동시출현시 성성된다.
Link Weight	Strength of Relation	링크가 갖고있는 정보값으로 링크가 누적된 횟수에 대한 정보를 갖는다.

그림 3. 응집도 추출을 위한 시소러스 구성

위 구조로부터 응집도가 높은 키워드를 추출하기위해 본 논문에서는 다음과 같은 전략을 사용하였다.

⇕ 노드가중치를 계산한다.

$$\text{노드가중치}(TF) = \frac{\text{노드빈도값}}{\text{전체출현노드누적값}} \quad (식1)$$

⇕ 노드당 링크갯수 정보를 계산한다.

$$\text{링크가중치}(LF) = \frac{\text{링크빈도값}}{\text{전체출현링크누적값}} \text{(식2)}$$

⇕ 노드당 (⇧+⇩)의 값을 정량화한 후 값이 높은 노드를 선택한다.

$$\text{노드변환값} = \frac{TF+LF}{2} \text{(식3)}$$

⇕ 동일한 노드 변환값을 가질 경우, 노드가중치가 높은 노드를 선택한다.

3.2 주어주도 정보추출

문서는 문단과 문장으로 구성되어있고, 문장을 하나의 의미체로 볼 때, 의미체간의 담화구조, 의미적 연결 구조를 통해 최종적으로 문서의 의미가 도출되게 된다. 문단간의 관계나 문장 간의 의미적 연결 구조를 파악하는 것은 의미망(semantic network)라든가 담화구조를 정확히 파악해야하는 문제이므로 본 논문에서는 처리대상을 문장으로 하고, 연결 구조를 키워드의 출현정보로 제한하며, 한 문장의 주어가 문장의 전체 의미를 주도하므로, 자질로서 단어의 출현정보뿐만 아니라, 주어정보를 함께 추출하여 이용한다.

문장 내에서 추출된 키워드는 다음과 같이 주어를 상위레벨로, 그 밖의 키워드들은 하위레벨로 트리를 형성한다.

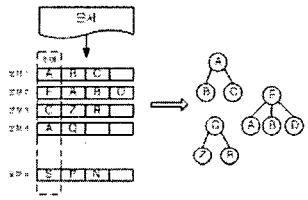


그림 4. 주어중심의 트리구성

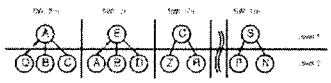


그림 5. 레벨구분및 트리병합

SW(Sentence Weight)는 문서 내 문장의 중요도를 표현한 값으로 동일한 명사가 서로다른에서 주어로서 문장을 이끌 경우, 그림 5와 같이 트리병합을 하고, 문장에 중복 가중치를 부여한다. 본 논문에서는 문장 내에서 레벨1의 노드에대해서는 0.5의 가중치를 부여하고, 레벨 2에는 0.5/n의 가중치를 부여하여 주어와 그 밖의 키워드들에 대한 가중치를 부여하였다.

전체 각 노드들이 갖고 있는 주어주도 정보는 다음과 같이 계산된다.

$$\text{레벨1노드정보량} = \frac{L1\,Node\,\#}{Sentence\,\#} \times 0.5 \text{(식4)}$$

$$\text{레벨2노드정보량} = \frac{L1\,Node\,\#}{Sentence\,\#} \times \frac{0.5}{L2\,node\,\#} \text{(식5)}$$

각 키워드별 주어주도 정보량은 위의 (식4)와 (식5)의 합으로 구해진다.

3.3 주제어 추출

문서분류를 위한 주제어는 단어 당 응집도 정보와 주어주도 정보를 더해서 가장 높은 주제어 정보 값을 갖는 키워드를 문서의 주제로 제시하게 된다.

4. 시스템의 기능

그림 5의 간단한 예제를 바탕으로 식(1), 식(2), 식(3)을 통하여 추출되는 응집도가 높은 키워드는 다음그림과 같다.

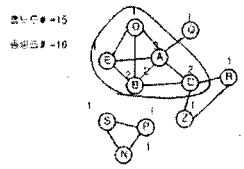


그림 6. 응집도정보 추출결과

식(4), 식(5)을 통하여 추출된 주어주도 정보량과 응집도 정보량을 각각 1로 정량화한 후 문서정보추출 에이전트는 다음과 같은 주제어후보를 제시하게 된다.

	응집도	주어주도정보량	+주제어후보값	+주제어후보Rank	+출현빈도정보
A	0.194	0.233	0.427	1	3
C	0.129	0.167	0.296	2	2
B	0.145	0.100	0.245	3	2
E	0.080	0.100	0.180	4	1
S	0.065	0.100	0.165	5	1
Q	0.049	0.067	0.116	6	1
D	0.080	0.033	0.113	11	1
N	0.065	0.050	0.115	7	1
P	0.065	0.050	0.115	7	1
R	0.065	0.050	0.115	7	1
Z	0.065	0.050	0.115	7	1
총합	1.00	1.00	2.00		

그림 7. 추출된 주제어후보값

결과로부터 출현빈도정보보다 본 논문에서 제시한 방법이 문장 내 역할과 문장 간 응집도 정보에 따라 좀 더 세분화하여 주제어를 추출 가능함을 확인할 수 있다.

5. 결론 및 향후 연구과제

본 논문은 문서 분류 시 문서정보추출 에이전트의 성능을 향상시키기 위해 응집도 정보와 주어주도 정보를 이용하여 주제어를 추출하였다. 방법으로는 키워드들의 빈도정보와 문장 내 공기정보를 이용하여 시소러스를 구성하고, 그로부터 키워드의 문서 내 응집도 정보를 추출하였다. 한국어의 기초문형에서 나타나는 주어가 문장을 주도한다는 점에 착안하여 키워드의 주어주도정보량을 추출하였고, 이러한 정보들로부터 주제어를 추출한 결과 단순히 출현빈도 정보만을 사용할 때보다, 좀 더 세분화하여 키워드들의 랭킹을 결정할 수 있었으며, 그로부터 문서분류 시 정확도에 있어서 향상된 결과를 가져올 수 있을 것으로 기대한다.

향후 연구과제로는 추출된 표현정보로부터 문서분류를 하였을 때, 정확도의 변화를 실험하고, 실험을 통해 상위레벨과 하위레벨에 가중치를 부여할 때 좀 더 적정한 수치를 찾아내는 작업이 필요할 것이다. 추가적으로 문서 내 키워드들의 연관정도를 좀 더 정확히 파악하기위해 연구대상을 의미사전과 문장 간 담화구조 등으로 확장하고, 이로부터 주제어를 추출하는 연구가 필요할 것이다.

참고문헌

[1] Lewis, D. D., Evaluating and Learning in Information Retrieval, Ph. D. Thesis, 1992.

[2] 김영택 외 공저, "자연언어처리", 생능출판사, 2001

[3] 고영중, 박진우, 서정연 "문장중요도를 이용한 자동문서 범주화" 정보과학회 논문집 2002. 6.

[4] 안희국, 노희영, "동적시소러스와 유전자 알고리즘을 이용한 개별화된 전자메일 분류시스템", 한국정보과학회 춘계 학술발표논문집(B) 제29권 1호, pp.472-474, 2002

[5] 이창범외 4인 "주성분 분석을 이용한 문서 주제어 추출" 정보과학회 논문지 2002. 10

[6] 유승국 "현대국어의 문형에 관한 연구(문장의 형식구조를 중심으로)" 2001. 중앙대학교 대학원 박사학위논문