단어 및 단어쌍 별 빈도수를 이용한 문서간 유사도 측정

김혜숙, 박상철, 김수형
전남대학교 전산학과
전화 : 062-530-0430 / 핸드폰 : 011-9642-3788

Measurement of Document Similarity using Word and Word-Pair Frequencies

Hye Sook Kim, Sang Cheol Park, Soo Hyung Kim
Dept. of Computer Science, Chonnam National University
E-mail : hsfight@hanmail.net

Abstract

In this paper, we propose a method to measure document similarity. First, we have exploited single-term method that extracts nouns by using a lexical analyzer as a preprocessing step to match one index to one noun. In spite of irrelevance between documents, possibility of increasing document similarity is high with this method. For this reason, a term-phrase method has been reported. This method constructs co-occurrence between two words as an index to measure document similarity. In this paper, we tried another method that combine these two methods to compensate the problems in these two methods. Six types of features are extracted from two input documents, and they are fed into a neural network to calculate the final value of document similarity. Reliability of our method has been proved by an experiment of document retrieval.

I. 서론

현대 정보화사회에서 정보는 사람의 관리가 불가능할 정도로 쏟아져 나오고 있는데 이러한 정보를 보다 빠르게 분류하고 효율적으로 관리하며 해당 문서를 쉽게 검색할 수 있는 방안이 모색되어야 한다[1]. 또한 소프트웨어가 점점 복잡해지고 대형화됨에 따라 사용자의 요구가 매우 다양해지고 있기에 이런 요구 사항을 정확히 분석하여 효과적으로 개발 단계에 적용하기 위해 문서간 의존성, 즉 상·하위 문서간 연계성을 측정할 수 있는 방법에 대한 연구가 필요하다[2].

이에 본 논문에서는 3가지 색인추출 방법을 통한 문서간 유사도 측정에 대해 살펴보고자 한다. 첫째, 전처리 단계로 형태소 분석기를 통해 명사를 추출한 후 한 단어가 하나의 색인을 구성하는 단어색인 방법을 사용하여 유사도를 측정해 보았다. 이는 실제로 유사하지 않음에도 불구하고 한 단어에 의해 전체 문서의 유사도에 영향을 끼칠 가능성이 크게되는 단점이 있다. 둘째, 이러한 단점을 보완하고자, 추출된 명사를 대상으로 공기쌍을 형성하여 이를 색인어로 하는 구색인 방법을 이용하여 유사도를 측정하였다. 이를 통해 실질적인 두 문서간 관련성에 대한 좀 더 개선된 지표를 마련할 수 있었다. 마지막으로, 이 두 방법을 결합하여 두 가지 방법 사이에 존재하는 상호 문제점을 보완할 수 있도록 하였다. 이를 토대로 본 논문에서는 문서간 유사도 측정에 대한 처리 절차를 설계 및 검증하고, 이를 구현하였다.

II. 문서간 유사도 측정

2.1 전체적인 시스템 구조도

본 논문에서는 단어색인 방법과 구색인 방법에 의해

유사도를 측정하였고, 또한 이 두 가지 방법을 결합하여 임의의 두 문서간 유사도를 측정하였다.
그림 1은 문서간 유사도 측정 시스템에서의 색인 추출부, 특징 추출부, 유사도 측정부를 보여주고 있다.

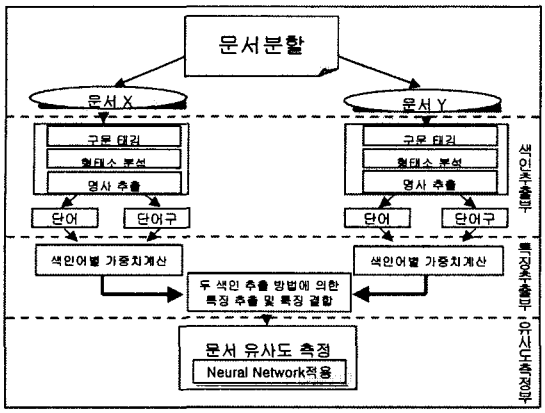

그림 1. 문서간 유사도 측정 위한 전체 시스템 구조도

2.2 색인 추출

(1) 단어색인 방법에 의한 색인 생성

두 문서간 유사도 측정을 위한 가장 기본적인 방법 중의 하나는 하나의 단어를 기본으로 색인을 생성하는 방법이다. 단어색인은 단일단어로 구성되어 적은 양의 데이터에서도 많은 색인을 추출할 수 있다는 장점이 있는 반면에 문맥 정보를 포함할 수 없다는 단점이 있다.

(2) 구색인 방법에 의한 색인 생성

구색인 방법은 여러 단어로 이루어진 하나의 구가 색인어가 되는 방법이다. 이는 공기 정보 등을 이용해 단어 쌍 등을 색인으로 보기에 문맥 정보를 어느 정도 반영할 수 있다는 장점이 있다[2]. 본 논문에서는 인접한 단어 사이의 공기 정보를 추출하기 위해서 Sliding Window기법을 사용하는데, 추출된 내용어의 순서열에 일정 크기(10)의 윈도우를 설정하고, 윈도우의 맨 앞의 내용어(Content word)와 다음 내용어들간의 쌍을 하나의 색인어로 추출한다.

2.3 특징 추출

특징 추출은 단어색인 방법, 구색인 방법, 이 두 가지 결합 방법에 대해 유사도에 영향을 끼칠만한 특징들을 추출하는 부분이다.

(1) 단어색인 방법을 통한 특징값

단어색인 방법을 통한 특징값 추출에 앞서 색인어 출현빈도에 기반한 특징값 추출을 위한 계산식(Cosine 가중치, 최소 가중치, 비율 가중치, 곱 가중치, Log가중치중 최소값, 최소가중치에 Log값) 6가지 경우를 고려해본 결과, Cosine 가중치와 최소 가중치가 분류율 면에서 15.04%와 14.89%의 가장 적은 오분류율을 얻어, 이 두 가지 계산식만을 특징값으로 추출하였다. 이외에도 두 문서간 일치하는 키워드 수, 상호정보에 대한 응용, 평균조건확률에 대한 응용과 같은 3가지 특징을 추가로 추출하였다. 단어색인 방법을 통해 추출된 특징값을 보면 표 1과 같다. 여기에서 tf 는 한 단어에 대한 문서내의 출현 빈도수를 말한다.

표 1. 단어색인 방법을 통한 특징값

	특징값
① Cosine 가중치	$\frac{\sum(tf_{ij}\times tf_{ik})}{\sqrt{\sum tf_j^2}\cdot\sqrt{\sum tf_k^2}}$
② 최소 가중치	$\sum MIN\left(\frac{tf_{ij}}{\sum tf_j},\frac{tf_{ik}}{\sum tf_k}\right)$
③ 두문서간 일치하는 키워드수	$N(k_{ij}=k_{ik})$
④ 상호정보에 대한 응용	$\frac{\sum(tf_{ij}\times tf_{ik})}{\sum tf_j+\sum tf_k}$
⑤ 평균조건확률에 대한 응용	$\frac{1}{2}\left(\frac{\sum(tf_{ij}\times tf_{ik})}{\sum tf_k}+\frac{\sum(tf_{ij}\times tf_{ik})}{\sum tf_j}\right)$

(2) 구색인 방법을 통한 특징값

단어 쌍을 색인어로 보는 구색인 방법은 단어색인 방법에 비해 문서의 내용을 더 잘 내포한다고 할 수 있다. 이를 통한 특징값 추출에 대해 살펴보기로 한다. 우선, 단어 쌍의 가중치를 계산하기 위해서는 단어 쌍의 문서내 출현 빈도뿐만 아니라 각 단어의 정보량 까지 고려해야 하는데 이 두 가지가 모두 고려된 특징은 다음과 같이 나타낼 수 있다.

$$\delta(X,Y)=\sum_{i\in p(X)\cap p(Y)}(\rho X(i)\times\rho Y(i))$$

$p(X)$와 $p(Y)$ 는 문서 x, y의 색인 파일을 나타내며, $\rho X(i)$ 와 $\rho Y(i)$는 문서 x와 y에 존재하는 i번째 색인의 단어쌍 별 빈도수와 각 단어의 정보량을 곱한 값을 의미한다. 이 특징값은 두 문서의 색인 파일에 공통으로 존재하는 단어 쌍의 가중치 값의 곱을 더하는 것으로 이 값이 클수록 유사도가 높게 된다[1].

2.4 유사도 측정

우선, 단어색인 방법에서는 문서내에 존재하는 명사 색인어에 대한 가중치(빈도수)를 바탕으로, 5개의 특징을 추출한 후 Neural Network을 적용하여 유사도를

측정하게 된다. 두번째, 구색인 방법에서는 두 문서간 형태소 분석을 통해 얻은 명사에 문장단위로 Sliding Window(size=10)를 씌워 단어쌍을 구성하여, 단어쌍별 빈도수로부터 계산된 가중치를 특징으로 유사도를 측정하였다. 또한 이들 두 가지 방법의 특징을 결합한 후 Neural Network을 적용하여 유사도를 측정해 보았다. 결합 방법에 의한 유사도 측정시 고려되는 Neural Network구조를 살펴보면 그림 2와 같다.

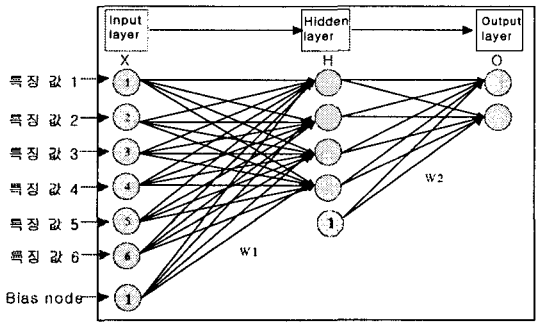

그림 2. 결합 방법을 통한 Neural Network 구조

이 구조에서 알 수 있듯이 출력값은 두 가지로, 하나는 임의의 두 문서간 동일함의 정도를 나타내는 값이고, 다른 하나는 동일하지 않음의 정도를 나타내는 값이다. 따라서 두 문서간 유사도 측정에서 동일함의 정도를 나타내는 값에서 월등히 큰 값이 나타나면 유사도가 그만큼 크다는 것을 의미한다.

III. 실험결과 및 분석

3.1 실험 문서 집단 구성

실험 데이터는 작가, 주제면에서 중복 없이 선정된 소설이나 수필을 대상으로 하였으며 하나의 문서 크기는 A4 용지 1장 규격으로 통일하였다. 문서 분할은 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, 9:1분할의 9가지 형태를 취했다. Neural Network 적용을 위한 데이터 구성은 Training data로 100개 문서쌍(동일 문서 50쌍, 다른 문서 50쌍), Test data로는 Training data에서 고려되지 않은 2500개 문서쌍 조합을 대상으로 하였다. 그리고 색인어 추출시 불용어 처리 부분은 제외하였다.

3.2 실험결과

Test data 2500개 문서쌍 조합에 의한 유사도 측정에 대한 실험결과는 표 2와 같다.

표 6. 유사도 측정에 대한 실험 결과

	동일문서여부	단어색인 방법		구색인 방법		결합 방법	
		O1	O2	O1	O2	O1	O2
Test1	O	**0.291**	**0.633**	0.899	0.082	0.891	0.090
Test2	X	0.021	0.973	0.169	0.798	0.015	0.981
:	:	:	:	:	:	:	:
Test52	O	0.994	0.003	**0.169**	**0.798**	0.998	0.002
Test53	X	0.198	0.753	0.169	0.798	0.133	0.835
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
Test2499	X	0.021	0.973	0.169	0.798	0.015	0.981
Test2500	O	0.665	0.184	**0.169**	**0.798**	**0.443**	**0.476**

O는 동일 문서를 X는 다른 문서를 나타내며, O1은 두 문서간 동일함의 정도를 O2는 동일하지 않음의 정도를 나타내는 값이다. Test 52번째 구색인 방법에 의한 유사도 측정값을 살펴보면 O1이 **0.169**로 O2값인 0.798보다 훨씬 작아 유사함의 정도가 작다는 것을 의미한다. 그러나 같은 문서를 대상으로 결합 방법에 의한 유사도 측정값을 살펴보면 O1값이 **0.998**로 큰 값이기에 두 문서가 상당히 유사함을 알 수 있다.

표 3은 표 2에서의 유사도 측정치에 대한 객관성 검증을 위해 50개 문서 대상 동일 문서 탐색에 대해 실험한 결과이며, 그림 4는 이를 도식화 한 것이다. 여기서는 상위 2%, 4%, 10% 안에 정답(동일문서)이 포함될 확률을 고려해 보았다.

표 3. 50개 문서 대상 동일 문서 탐색

	단어색인 방법		구색인 방법		결합 방법	
	정확률	백분율	정확률	백분율	정확률	백분율
상위2%	33/50	66%	23/50	46%	40/50	80%
상위4%	37/50	74%	29/50	58%	45/50	90%
상위10%	44/50	88%	34/50	68%	47/50	94%

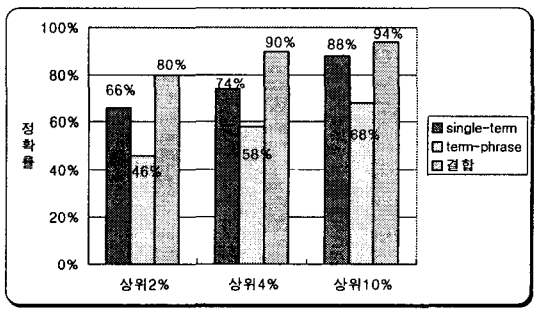

그림 3. 50개 문서대상 동일문서 탐색

단어색인 방법에 비해 구색인 방법으로 색인을 추출한 경우에 탐색효율이 상대적으로 낮았는데, 이는 두 문서간 같은 단어를 갖는 단어쌍의 빈도가 매우 적어

특징값이 대부분 0이라는 값을 갖기 때문이다. 즉, 문서간 탐색시 단어색인 방법에 비해 구색인 방법의 정확률이 떨어짐을 확인해 볼 수 있었다.
또한 그림 3에서 알 수 있듯이 결합 방법에 의해 유사도를 측정하여 문서간 탐색에 적용하였을때 정확률 면에서 가장 우수함을 알 수 있었다.

Ⅳ. 문서 검색에의 응용

본 장에서는 앞서 설계 및 검증한 문서간 유사도 측정에 관해 구현하였다. 구현 환경으로는 MS-Window 상에서 Visual C++을 이용하였다.
문서간 유사도 측정에 앞서 형태소 분석기를 거친 후 명사만을 추출하여 이에 따르는 빈도수를 구하는 작업이 우선되어야 한다. 추출된 단어에 대한 빈도수는 그림 4와 같다. 또한 구색인 방법에 의한 유사도 측정을 위해 단어쌍 및 빈도수, 가중치계산이 우선되어야 하며, 이에 대한 결과는 그림 5와 같다.

```
남자 11여자 19외국 1아름다움 4나라 2우리나라 5여자를 6멋 1한국 1
이상 1한마디 1고 1현실 1빛 1풀림 1계성 2로 1마음속 1뭘 1마지막 2
한가지 1처녀성 2욕심 1한말 1라 1문장 1생활 1학적 1반면 1전부 1
```

그림 4. 단어 및 단어빈도수 생성

```
남자 여자 25 0.61 남자 외국 1 1.02 남자 아름다움 1 1.02
여자 외국 2 1.00 여자 여자 8 0.88 여자 아름다움 4 0.95
외국 아름다움 1 1.02 나라 여자 2 1.00 나라 우리나라 1 1.02
```

그림 5. 단어쌍 및 빈도수생성과 이에 따른 가중치

그림 6과 7은 단어색인 방법과 구색인 방법에 의해 추출된 특징들을 보여준다.

파일이름	p_01_10c.txt	p_01_11c.txt
t_01_10c.txt	24,07 5 0,17 0,29 1,05 0	1,94 2 0,05 0,03 0,14 X
t_01_11c.txt	0,93 1 0,03 0,02 0,06 X	7,10 4 0,06 0,09 0,43 0
t_01_1c.txt	0,00 0 0,00 0,00 0,00 X	0,65 1 0,02 0,01 0,07 X
t_01_2c.txt	0,00 0 0,00 0,00 0,00 X	0,00 0 0,00 0,00 0,00 X
t_01_3c.txt	0,00 0 0,00 0,00 0,00 X	0,65 1 0,03 0,01 0,15 X
t_01_4c.txt	0,93 1 0,06 0,02 0,12 X	0,00 0 0,00 0,00 0,00 X
t_01_5c.txt	0,00 0 0,00 0,00 0,00 X	0,00 0 0,00 0,00 0,00 X
t_01_6c.txt	0,00 0 0,00 0,00 0,00 X	1,29 2 0,05 0,02 0,15 X
t_01_7c.txt	10,37 3 0,11 0,13 0,61 X	1,94 2 0,05 0,03 0,18 X
t_01_8c.txt	0,00 0 0,00 0,00 0,00 X	0,00 0 0,00 0,00 0,00 X
t_01_9c.txt	10,00 1 0,04 0,17 1,09 X	0,00 0 0,00 0,00 0,00 X
t_02_10c.txt	2,63 1 0,03 0,03 0,07 X	2,58 3 0,06 0,04 0,13 X
t_02_11c.txt	6,08 4 0,08 0,07 0,18 X	1,29 1 0,01 0,02 0,05 X

그림 6. 단어색인 방법에 의한 특징추출

	1	2	3	4	5
1	4.49 00.00 X0.00 X0.00 X0.00 X0.00 X0.00 X0.00 X0.00 X				
2	0.00 X0.00 X0.00 X0.00 X0.00 X0.00 X0.00 X0.00 X0.00 X				
3	1.01 X0.00 X0.00 X0.00 X0.00 X0.00 X0.00 X0.00 X0.00 X				
4	.....				

그림 7. 구색인 방법에 의한 특징추출

Ⅴ. 결론

인터넷은 정보의 보고라 불리울 만큼 수 많은 정보자원이 곳곳에 산재되어 있으며 하루에도 수 많은 정보들이 새로 부가 되고 있다. 이런 수 많은 정보를 지닌 문서들 사이의 유사도는 많은 분야에서 응용이 가능하다. 이를 위해 본 논문에서는 문서간 유사도 측정 방법에 대한 여러 가지 실험을 통하여 두 문서간 유사도를 측정해 보았고, 이에 대한 검증 과정을 통해 문서탐색시 정확률을 살펴보았다. 임의의 두 문서간 유사도를 나타내는 정량적인 수치 자체로도 두 문서가 어느 정도 유사한 지를 나타내 줄 수 있는 좋은 지표가 될 수 있다. 뿐만 아니라, 단어색인 방법, 구색인 방법, 결합 방법에 의한 유사도 측정을 상위 4% 기준으로 보았을때, 각각 74%, 58%, 90%로 단어색인 방법, 구색인 방법에 비해 결합 방법의 성능이 가장 우수했음을 알 수 있다.

이에 본 논문은 S/W 공학적 개발기법 적용시 각종 문서의 검색이나 수많은 문서들의 효율적인 분류와 관리에 중요한 기초 자료로 활용될 수 있을 것으로 사료된다.

참고문헌

[1] 정영미, "지식 분류의 자동화를 위한 클러스터링 모형 연구," 한국정보관리학회지, pp. 203-230, 2001.

[2] 박수용, 서정연, 김학수, 고영중, "유사도 측정 기법을 이용한 효율적인 요구 분석 지원 시스템의 구현," 정보과학회 논문지 제27권 제1호, pp.13-23, 2000.

[3] 김명철, "공기기반 용어간 유사도를 이용한 정보검색 질의 확장 비교 연구," 박사논문, 한국과학기술원, 1999.

[4] 이재윤, "동적 시소러스의 구축에 관한 실험적 연구," 석사논문, 연세대학교, 1994.

[5] 이재윤, 최보영, 정영미, "문헌 자동분류에서 용어 가중치 기법에 대한 연구," 제7회 한국정보관리학회 학술대회 논문집, pp. 41-44, 2000.

[6] 정영미, *정보검색론*, 구미무역 출판부, 1993.

[7] R. Ash , *Information Theory*, New York:Wiley - Interscience, 1965.

[8] Y. Karov and S. Edelman, "Similarity - based Word Sense Disambiguation," *Computational Linguistics*, Vol 24, No 1, pp41-60, March 1998.