단어가중치 기반 문서간 유사도 측정에 관한 연구

김혜숙, 박상철, 김수형
전남대학교 전산학과

A Research of Documents Similarity Measuring Based on Word Weight

Hye-Sook Kim , Sang-Chul Park, Soo-Hyung Kim
Dept. of Computer Science, Chonnam Nat'l University
E-mail : hsfight@hanmail.net

요 약

사용자의 요구 사항을 정확히 분석하여 효과적으로 개발 단계에 적용하기 위해 문서간의 의존성 ,즉 상·하위 문서간의 연계성 등을 측정할 수 있는 방법에 대한 연구가 절실한데 이를 위해 적게는 두 용어가 얼마나 밀접한 관련이 있는가를 나타내는 용어간의 유사도 정보가 중요시된다. 이에 본 논문은 임의의 두 문서에 대한 다양한 유사도 측정방법을 통하여 최적의 유사도를 알아보고 두 문서간 유사여부를 검증하기 위해 Neural Network을 적용하였다. 이러한 유사도 측정과 검증 방법은 분산환경에서 입력되는 요구사항 문서들을 효율적으로 분류, 관리해 줄 수 있으며 사용자 요구사항 분석과 전체 Project 수행에 좋은 기초자료를 제공해 줄 수 있다

1. 서론

인터넷상의 정보와 다양한 서비스가 빠른 속도로 증가하고 있으며, 이를 만들고 사용하는 사람의 수 또한 증가하고 있다. 현대 정보화사회에서는 사람의 관리가 불가능 할 정도로 많은 정보가 쏟아져 나오고 있는데 이러한 정보를 보다 빠르게 분류하고 효율적으로 관리 할 수 있는 방안이 모색되어야 한다. 문서의 양과 사용자의 수가 증가함에 따라 자동문서 분류는 방대한 양의 데이터를 정리하는 사람들을 돕기 위해 중요한 도구가 되어가고 있다[1]. 또한 소프트웨어가 점점 복잡해지고 대형화됨에 따라 사용자의 요구가 매우 다양해지게 되는데, 이런 요구 사항을 정확히 분석하여 개발 단계에 효과적으로 적용하기 위해 서는 상·하위 문서간의 연계성 등을 측정할 수 있는 방법에 대한 연구가 절실하다. 이러한 문서간 연계성 측정을 위해서는 단어간의 유사도 정보가 중요시된다 [1].

본 논문은 임의의 두 문서에 대한 유사도 측정을 객관적 지표에 의해 정확한 수치로 나타내기 위해 전처리 단계로 형태소 분석기를 사용하였다. 두 문서 분할을 통한 형태소 분석기를 통해 추출된 명사를 통해 두 문서에 대한 문서내빈도수를 구하여 유사도 측정을 설계 및 검증하고, 이를 구현해 보고자 한다.

2. 관련 연구

일반적으로 단어를 통한 문서 유사도 비교 뿐 아니라 단어의 반복(reiteration)과 공기(collocation) 같은 언어현상을 이용한 많은 연구들이 정보 검색 분야에서 이루어졌다[2]-[8]. Palmer는 [9]에서 2층 구조의 TTC(two-tired clustering) 알고리즘을 이용하여 요구 분석서를 색인하고 클러스터링(clustering)하는 방법을 제안하였다. TTC 알고리즘은 먼저 각 요구 분석 문서에 속해 있는 동사들을 키워드로 하여 문서를 기능별로 분류하고, 이렇게 기능별로 분류된 문서들 사이에 코사인(cosine) 유사도를 측정하여 재분류하는 작업이다. 그리고 유의어 사전을 이용한 방법들은 문서간 유사도 측정의 정확률 면에서 개선된 성능을 보이고는 있으나 특정 영역마다 이런 사전을 구축하고

관리하는 일은 결코 쉬운 작업이 아니다. 정보검색에서 문서 검색 시 질의 수정 과정을 거치게 되는데 수정대상에 따라 검색 문헌 순위 재조정과 질의 확장이 있다. 질의 확장은 정보 획득원에 따라 지역정보와 전역정보로 나누어진다. 지역정보는 사용자 관련성 피드백 정보나 초기 검색된 상위 문헌으로부터 질의 확장에 필요한 정보를 획득한다. 반면, 전역정보는 미리 작성해 놓은 시소러스를 이용하는 방법으로 시소러스 구축 방법에 따라 수동 시소러스와 자동 시소러스가 있다. 시소러스란 용어간의 상하관계나 동의어, 관련 용어 정보를 포함하는 개념사전이며 자동 시소러스의 경우 , 두 용어가 얼마나 밀접한 관련이 있는가를 나타내는 용어간 유사도 정보가 더 중요시된다[10]. 분석되는 용어의 통계정보로는 용어의 문헌내빈도수, 문헌빈도수, 공기빈도수가 있고 분석 수준도 형태소수준, 구문수준, 의미수준으로 나누어 볼 수 있으며, 이들 모두 용어간 유사도를 구하기 위해 사용된다. 현재 이들 각각의 특징을 적용한 유사도 측정은 임의의 두 문서의 유사도, 또는 관련문서 검색에 많이 적용되어 오고 있다. 그러나 정보 검색 성능 향상에 어떤 유사도 측정 방법이 가장 좋은지에 대한 연구는 거의 이루어지지 않고 있다. 이에 본 논문은 두 문서간 유사도를 측정하고, 이를 이용한 두 문서사이의 검증 방법을 제안하고자 한다.

3. 문서간 유사도 측정 기법

3.1 전체 시스템 구조도

본 논문에서 제안하는 문서간 유사도 측정부는 요구 분석 시에 상위 문서와 하위 문서 사이의 연계성을 유지하기 위한 도구로 사용될 수 있으며, [그림1]과 같이 주제어 추출부와 유사도 측정부로 나뉠 수 있다. 주제어 추출부는 형태소 분석을 통한 명사만을 대상으로 하여 색인어를 추출하는 부분이고, 유사도 측정부분은 실제로 입력된 두 임의의 문서사이의 유사도를 측정하는 부분이다.

3.2 색인 추출

유사도를 측정하기 위해 두 임의의 문서로부터 색인어를 추출하게 되는데, 색인 추출방법은 형태에 따라 single-term방법과 term-phrase방법이 있다. 본 논문에서는 single-term 방법을 사용하는데 이는 단일 단어로 구성되어 적은 양의 데이터에서도 많은 색인을 추출할 수 있다는 장점이 있는 반면에 문맥 정보를 포함할 수 없다는 단점이 있다.

색인은 특정한 정보가 필요한 사람에게 그 정보의 위

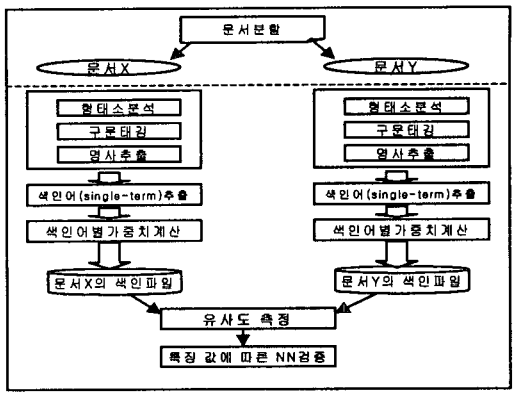

그림 1 전체 시스템 구조도

치를 지시해 주는 역할과 방대한 정보원으로부터 가장 유사한 내용의 정보 자료만을 선별해 주는 역할을 한다. 또한 문서의 내용을 분석하여 문서의 내용을 대표하는 용어인 색인어는 그 문서의 전체적인 내용을 나타내거나, 그 문서를 다른 문서들로부터 구별할 수 있도록 그 문서의 선택 단서가 되는 단어 또는 단어구를 의미한다[10]. 문장의 내용이나 특성을 잘 반영하는 단어를 내용어(content word: open-class word)라고 하며, 명사, 동사, 형용사 등에 해당되는 단어들을 말한다. 하지만 실제로 한국어에서 문서 분리도가 높은 단어들은 주로 개념을 표현하는 명사와 고유명사에 밀집되어 있기에 본 논문에서 유사도 비교를 위한 색인어는 명사만을 대상으로 하였다. 이렇게 추출된 색인어를 기반으로 유사도 측정시 어떤 유사도 측정 방법을 선택하는가에 따라 두 단어간의 유사도 값이 다르게 나타나기 때문에, 단어간의 유사도 측정 방법은 신중히 결정되어야 한다.

3.3 두 문서간 유사도 측정을 위한 색인의 생성

Salton에 의하면 특정한 단어가 문서 집단 속의 상호 관련 없는 문서들을 분리시키는 능력치가 큰 것이 좋은 색인어가 되고 나쁜 색인어 일수록 상호 관련 없는 문서들을 묶어 준다고 하였다.

다음은 본 논문에서 적용한 단어 가중치 측정 부분이다.

① 두 문서의 최소 단어 가중치

$$\sum MIN\left(\frac{fi}{\sum fi}, \frac{gj}{\sum gj}\right) \quad (1)$$

< f_i g_j : 두 문서에서의 단어 빈도수>

② 각각의 단어 빈도수를 최대 단어 빈도수로 나눈

가중치

$$\frac{freq_i}{\max freq} \quad (2)$$

< $freq_i$: k_i번째 단어에 대한 빈도수, max $freq$: 최대단어 빈도수>

자주 등장하는 단어는 문서의 유사도나 의미의 정보량을 적게 지니게 된다. 이를 반영하기 위한 식이라 할 수 있다.

③ 두 문서의 가중치 곱

$$\sum\left(\frac{f_i}{\sum f_i}\times\frac{g_j}{\sum g_j}\right) \quad (3)$$

④ 두 문서의 가중치에 log취한 가중치

$$\sum MIN(\log\frac{f_i}{\sum f_i}\times\log\frac{g_j}{\sum g_j}) \quad (4)$$

위의 식은 한 단어에 대한 가중치의 영향을 최소화 시켜보고자 각 단어에 대한 가중치에 로그를 취한 값이다.

4. 실험 및 평가

4.1 실험 방법 및 데이터 구성

먼저 50개의 문서 각각을 상위문서(Prototype)와 하위문서(Test)로 나누어 100개의 데이터 파일을 만들었다. 상위문서와 하위문서로 분할시 1대9, 3대7, 5대5, 7대3, 9대1 과 같이 각각 5가지 방법으로 분할 하였고, 각각의 경우에 대해 10개의 임의의 문서를 구성하였다. 그리고 상위50개 문서와 하위50개 문서로 분할된 문서에 대해 유사도를 측정하였다. 이렇게 5가지 형태의 분할을 해봄으로써 임의의 두 문서에 대해 문서유사도 측정시 문서길이에 의한 영향력을 추정해 볼 수 있었다. 총 2500번의 유사도 측정을 통해 가장 유사도가 높게 나온 하위 문서가 Prototype문서 중 이와 동일한 문서로 선택 되었다면, 이를 정답으로 간주하였다. 요구사항 문서의 길이는 A4용지 1장 규격으로 통일하였다. 색인 파일을 구성하기 위해 각각의 문서로 부터 한 단어를 추출하여 유사도를 측정하였다. 두 문서간의 유사도 측정치, 두 문서간의 일치하는 키워드 수, cosine유사도 측정치, 두 문서간의 일치하는 키워드에 대한 각각의 가중치에 threshold를 적용하여 이 threshold값을 넘는 키워드 수의 4가지 방법을 통해 이를 Neural Network에 적용하여 두 문서에 대한 일치 여부를 검증하였다. 최종적으로, 이를 통해 계산된 2가지 오류율(동일문서를 서로 다른 문서로 판단하는 경우, 서로 다른 문서를 동일한 문서로 판단하는 경우)을 평균하여 전체 오류율을 계산하고, 이를 통해 두 문서간의 일치 여부에 대한 전체 검증 성능을 평가해 보았다.

4.2 실험 결과

문서간 유사도 측정을 위해 3.3에서 언급한 4가지 방법을 사용하여 정확률을 비교한 결과는 [표1]과 같다. 정확률 이라 함은 지정된 상위 퍼센트 안에 정답이 포함될 확률을 의미한다.

	최소 가중치	빈도수 / 최대빈도수	가중치 곱	각가중치 log 취합
상위 5%	65 %	70 %	**80 %**	55 %
상위 10%	80 %	75 %	**85 %**	75 %

표 1 유사도측정방법에 따른 정확률 비교

이를 그래프로 나타낸 결과는 [그림2]와 같다.

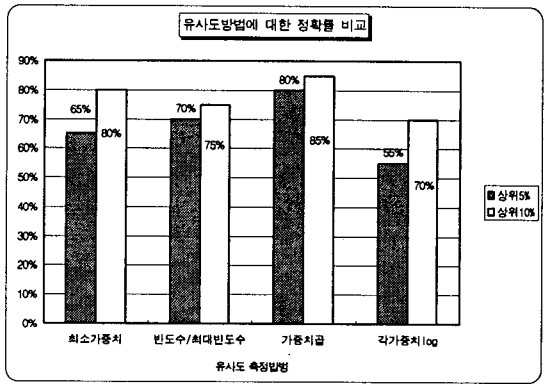

그림 2 유사도 측정방법에 따른 정확률 비교

상위 10%까지 고려할 경우, 가중치 곱을 적용한 경우의 정확률이 85%로 가장 좋은 성능을 보임을 알 수 있다.본 논문에서는 정확률 측정을 위해 총 50개의 문서중 training data로 30개, test data로 20개를 대상으로 하였다. 다음 결과는 test data인 20개만을 대상으로 조사한 결과이다. 위 결과는 사용자가 원하는 문서를 검색할 시에 해당 문서를 전부 검색하는 대신 각 문서 당 대략 10%의 노력으로 79% 정도의 만족을 얻을 수 있다는 것을 의미한다. [표2]는 다양한 유사도 측정 방법에 따른 오류율 비교이다.

가중치 곱을 적용한 경우의 오류율이 16.06%로 가장

	오류율(A : 동일문서 50개, B : 다른문서 2450개)		
1. 최소가중치	A->B	14 / 50 = 20 %	17.2 %
	B->A	157 / 2450 = 6.41 %	
2. 빈도수 / 최대빈도수	A->B	15 / 50 = 30 %	18.06 %
	B->A	150 / 2450 = 6.12 %	
3. 가중치곱	A->B	14 / 50 = 28 %	16.06 %
	B->A	101 / 2450 = 4.12 %	
4. 각가중치 log취합	A->B	15 / 50 = 30 %	18.20 %
	B->A	157 / 2450 = 6.41 %	

표 2 유사도 측정에 따른 오류율 비교

좋은 성능을 보였다. 다음은 이렇게 해서 구해진 유사도와 앞에서 언급이 되었듯이 이중 3가지 특징을 입력으로 하는 Neural Network에 적용하였을 때에 인식률을 [표3]에서 보여주고 있다.

	유사도값
feature1 (가중치곱 유사도)	0.95
feature2 (유사도, 일치하는키워드수)	0.95
feature3 (유사도, 일치하는 키워드 수, threshold 넘는 키워드 수)	0.925

표 3 특징값에 따른 NN의 인식 결과

또한 decision boundary 근처의 자료를 입력으로 삼았을 때의 Neural Network 인식률이 95%에서 100%까지 상당히 향상된 결과를 확인해 볼 수 있었다. 본 논문에 제시한 방법을 통해 측정한 문서 유사도 비교는 문서 검색이나 비교 분류 등과 같은 다양한 응용에 이용될 수 있을 것이다.

5. 결론 및 향후 과제

본 논문에서는 분산환경에서 입력되는 요구사항 문서들을 효율적으로 분류, 관리할 수 있고, 사용자 요구사항 분석과 전체 Project 수행에 좋은 기초자료를 제공해 줄 수 있는 문서 유사도 측정 방법에 대해 살펴보았다. 임의의 두 문서에 대한 유사도 측정을 객관적 지표에 의해 정확한 수치로 표현함으로 인하여 유사함의 정도를 파악할 수 있고, 이는 두 문서간의 표절의 좋은 참고 자료가 될 수 있을 것이다. 또한, 본 논문에서 적용된 방법은 언어 분석에서 비교적 하위 단계인 형태소분석만을 이용하여 유사도를 측정하였기 때문에 비교적 단순하다고 할 수 있다. 그러나 각각의 문서에 내포되어 있는 의미까지 포함하는 유사도 측정은 힘들다. 따라서 각 요구 분석 문서에 속해 있는 동사들까지 색인어로 추출하여 이들 사이에도 유사도를 측정한다면 좀더 객관적인 유사도 측정 기준이 될 것이다. 그리고 여러 문서에 대부분 공통적으로 많이 발생하기에 큰 의미를 주지 않는 단어들을 제거하는 불용어 처리를 추가시킨다면 보다 좋은 결과를 얻을 수 있을 것으로 보인다. 또한 유사도 측정을 위한 색인어로 명사 뿐 아니라 동사까지를 하나의 동일한 단어 쌍으로 본다면 더 좋은 결과를 얻을 수 있을 것이다.

[참고문헌]

[1] 유사도 측정 기법을 이용한 효율적인 요구 분석 지원 시스템의 구현, 정보과학회 논문지 제27권 제1호, pp.13-23, 2000

[2] Hearst M., "Multi-Paragraph Segment- ation of Expository Text," *proceedings of the ACL'94*, June 1994.

[3] Litman D. and Passonneau R., "Combining Multiple Knowledge Sources for Discourse Segmentation," *proceedings of the 33rd ACL*, May 1995.

[4] Jobbins A. and Evett L., "Text Segmentation Using Reiteration and Collocation," *Proceedings of the COLING-ACL'98*, pp.614-618, August 1998.

[5] Hajime M., Takeo H. and Manabu O., "Text Segmentation with Multiple Surface Linguistic Cues," *proceedings of the COLING-ACL'98*, pp.881-885, August 1998.

[6] Kim M., Klavans J. and McKeown K., "Linear Segmentation and Segment Significance," *proceedings of the 6th International Workshop of Very Large Corpora(WVLC-6)*, pp.197-205, August, 1998.

[7] Kozima H., "Text Segmentation Based on Similarity between Words," *proceedings of ACL'93*, pp.286-288, January 1993.

[8] Yaari Y., "Segmentation of Expository Texts by Hierarchical Agglomerative Clustering," *proceedings of RANLP'97*, pp.135-142, September, 1997.

[9] Maarek Y., Berry D. and Kaiser G, An Information Retrieval Approach For Automatically Construction Software Libraries, *IEEE Transaction On Software Engineering*, Vol. 17, No. 8, pp. 800-813, August 1991.

[10] 김명철(1999). "공기기반 용어간 유사도를 이용한 정보검색 질의 확장 비교 연구", 한국과학기술원 전산학과, 박사학위논문.