용어 선별 기법에 의한 유사 문서 판별 시스템

장성호[0], 강승식
국민대학교 컴퓨터학부, 첨단정보기술연구센터
{mpquake, sskang}@cs.kookmin.ac.kr

Text Similarity Decision System by Term Selection Method

Sung-Ho Jang, Seung-Shik Kang
School of Computer Science, Kookmin University and AITrc

요 약

대부분의 정보 검색 시스템은 문서 내에서 추출된 모든 용어를 이용해서 문서간 유사도 계산이나 문서 분류, 문서 클러스터링 등에 활용한다. 그러나 실질적으로 문서 내의 모든 용어를 추출해야만 이러한 정보 검색 시스템을 활용할 수 있는 것은 아니며, 오히려 용어 빈도수 같은 가중치가 낮은 용어를 용어 추출에서 제외시킴으로써 모든 용어 추출로 인해서 발생하는 시간과 공간을 많이 소비하는 문제를 해결할 수 있다. 또한 정확하고 자동적인 문서 분류를 위한 문서 클러스터링보다 유사 문서 검색의 활용은 검색 효율의 증가를 가져 올 수 있다. 본 논문에서는 유사 문서 판별 시스템을 이용해 용어 추출의 효율성을 실험하였으며, 모든 용어를 추출한 경우보다 중요 용어만 추출한 경우에 더 좋은 성능을 보였다.

1. 서론.

최근 정보 검색의 추세는 검색된 문서의 정확성뿐만 아니라 검색된 문서와 유사한 문서들을 같이 검색해주는 추세이다. 이것은 검색된 문서와 유사한 문서가 검색어와도 관련이 있을 수 있기 때문에 질의에 대한 검색 결과에도 도움을 줄 수 있기 때문이다. 이러한 문서와 문서간의 관련성을 확인하기 위해 문서간의 유사도를 계산하여 상호 관련성 여부를 확인한다[6,7,8,9].

유사 문서 추출 기법과 비슷한 정보 검색 기법은 문서 클러스터링과 문서 분류가 있다. 문서 클러스터링은 더 좋은 성능을 보이고 있는 문서 분류의 단점인 분류 체계의 수동 작성이 필요하지 않기 때문에 앞으로 정보 검색 시스템에서 반드시 필요한 분야이다. 그러나 문서 클러스터링은 문서간 클러스터링 작업을 위한 너무 많은 시간과 공간을 요구하는 문제점이 있다[3,4,5].

일반적으로 정보 검색 시스템은 문서에 나타난 모든 용어를 추출해서 문서간 유사도 판별에 사용하고 있지만 실질적으로 유사도 판별에 중요한 작용을 하는 용어는 빈도가 높거나 혹은 기타 여러 가지 용어 가중치 기법을 이용해서 높은 가중치를 가지는 용어만이 문서간 유사도에 영향을 주는 편이다. 그러나 문서에 나타난 모든 용어보다 그 문서를 나타낼 수 있는 중요 문서 주제어라고 판정된 용어만으로 문서간 유사도를 판별한다면 유사도 계산에 드는 시간과 공간을 절약할 수 있다.

본 논문에서는 유사 문서 판별 시스템을 이용해서 문서에 나타난 모든 용어를 추출해서 유사도 계산을 한 것과 중요한 용어라고 판단되는 용어만을 추출해서 유사도 계산을 한 것을 비교해서 모든 용어 추출 시의 효율성을 실험해 본다.

2. 문서 표현 방법

대부분의 정보 검색 시스템에서 문서를 표현하는 방법으로 용어 리스트를 이용한 방법과 용어와 용어 가중치 쌍의 리스트를 이용하는 두 가지 방법이 있다. 용어 리스트만으로 문서를 표현할 경우 다음과 같이 나타낼 수 있다.

$$D_i = \{ t_{1i}, t_{2i}, t_{3i}, \cdots, t_{ni} \}$$

이때 t_{ji}은 문서 내에서 추출된 용어만을 이용한 단순 용어로 나타내거나 여러 가지 구(Phrase) 추출 기법을 이용해서 추출된 구조적인 구(Syntactic Phrase)로 나타낼 수 있다[10]. 용어와 용어 가중치 쌍으로 문서를 표현하는 방식은 다음과 같다.

$$D_i = \{ \{ t_{1i}, w_{1i} \}, \{ t_{2i}, w_{2i} \}, \cdots, \{ t_{ni}, w_{ni} \} \}$$

t_{ni}은 단순용어나 구조적인 구(Syntactic Phrase)가 될 수 있으며, w_{ni}은 t_{ni}의 가중치 값이 할당된다. 가중치 값은 용어의 경우 일반적으로 tf-idf 방식이 활용된다. 본 논문에서는 용어 리스트만을 이용해서 문서간 유사도 계산에 활용한다.

3. 문서간 유사도 비교 방법

3.1 용어만을 이용한 비교 방법

추출된 용어만으로 문서간 유사도 비교를 할 경우에는 문서간에 공통으로 출현한 용어 수와 각 문서의 총 용어 수를 이용한다. 공통 용어의 수와 각 문서에서 추출된 용어 수의 관계로 임의의 두 문서가 서로 유사한지를 판단하기 위해서 문서간 관계를 보면 첫 번째로 각 문서에서 추출된 용어 수와 공통 용어 수가 모두 같은 경우이다. 이 경우는 두 문서가 동일 문서일 가능성이 높은 문서들로 생각할 수 있다. 두 번째로 공통으로 출현한 용어 수가 한 문서의 총 추출 용어 수와 같고 다른 문서의 총 추출 용어 수보다는 작은 경우이다. 이 경우는 한 문서가 다른 문서의 일부분으로 이루어진 문서일 가능성이 높으며, 유사 문서 관계로 판단해도 문제가 없다. 세 번째로 공통으로 출현한 용어의 수가 하나 이상이고 두 문

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

서의 추출 용어 수보다 작은 경우이다. 이 경우는 임계치보다 큰 경우 유사 관계로 판단할 수 있다. 마지막 네 번째는 공통으로 출현한 용어의 수가 하나도 없는 경우인데, 이 경우는 유사 문서 관계가 아니라고 판단된다.

그러므로 첫 번째, 두 번째의 경우와 세 번째의 일부가 유사 문서 관계로 생각할 수 있다. 특히 세 번째의 경우 다음과 같이 문서간 공통 용어 수와 각 문서의 전체 용어 수의 비교 방식에 따라서 결과가 달라진다.

$$S_{D_i,D_j} = \frac{n_{i,j}}{MIN(n_i, n_j)} \quad (1)$$

$$S_{D_i,D_j} = \frac{n_{i,j}}{MAX(n_i, n_j)} \quad (2)$$

$$S_{D_i,D_j} = \frac{2\,n_{i,j}}{n_i + n_j} \quad (3)$$

$$S_{D_i,D_j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}} \quad (4)$$

여기서 n_{ij}는 문서 d_i와 문서 d_j의 공통 문서 개수이며, $MIN(n_i, n_i)$과 $MAX(n_i, n_i)$는 각각 문서 d_i와 문서 d_j의 추출 용어 수의 최소값과 최대값을 나타낸다. 식(3)과 식(4)는 각각 Dice 상관계수와 Jaccard 상관계수를 이진 용어 가중치를 이용한 경우이며, 본 논문에서는 직관적으로 생각할 수 있는 첫 번째 방식을 문서간 유사도 판정에 사용하였다.

3.2 용어 가중치를 이용한 유사도 계산 기법

일반적으로 단순히 추출된 용어의 수만으로 문서 비교를 하는 것보다 추출된 용어와 그것의 가중치 값을 이용해서 문서 비교를 하는 것이 좀 더 좋은 성능을 나타낸다. 용어 가중치 할당 방식으로 가장 많이 사용하는 TF-IDF 가중치 방법과 주제어 가중치 방법을 사용하였다. TF-IDF 가중치 방법은 임의의 문서에서 출현한 용어의 출현 빈도수(TF)와 모든 문서 중에서 출현한 용어가 있는 문서수(DF)를 이용해서 용어의 가중치를 할당하는 방식이며, 주제어 가중치 방법은 빈도수뿐만 아니라 품사 정보와 격 정보 등 어절 단위의 용어 특성과 문장을 단위로 하는 용어의 구문론적 기능, 문서내에서 문장의 위치 및 역할에 의한 용어의 특성 등을 이용하여 용어의 가중치를 할당하는 방식이다[1].

그리고 문서간 유사도 방식을 위해 많이 사용되는 Dice 상관계수, Jaccard 상관계수, cosine 상관계수 등이 있는데[2,3], 본 논문에서는 코사인 상관계수를 사용하였다. 코사인 상관계수는 다음과 같다.

$$S_{d_i,d_j} = \frac{\sum_{k=1}^{t}(\omega_{i,k} \times \omega_{j,k})}{\sqrt{\sum_{k=1}^{t}\omega_{i,k}^2} \times \sqrt{\sum_{k=1}^{t}\omega_{j,k}^2}} \quad (5)$$

여기서 $\omega_{i,k}$는 문서 d_i의 k번째 용어의 가중치이며, $\omega_{j,k}$는 문서 d_j의 k번째 용어의 가중치이다.

4. 유사 문서 판별 기법

유사 문서 판별을 위해서 첫 번째로 출현 빈도에 따라서 용어를 선별해서 추출한 후 식(1)을 이용해서 문서간 유사도를 비교하였다. 이것은 실제 빈도가 높은 용어만으로 문서간 유사도를 비교해 봄으로서 빈도가 높은 용어의 문서 대표어로서의 변별력을 확인해 보기 위한 것이다. 또한 용어 출현 빈도수에 따라서 개수로 용어를 추출할 경우 용어 선택의 모호성이 발생하기 때문에 이러한 점을 방지하기 위해서 출현 빈도에 따라 선별하였다.

두 번째로 대표 용어 추출 비율에 따라서 용어를 선별한 후 식(1)을 이용해서 문서간 유사도 비교를 하였다. 여기서 대표 용어에 대한 판단은 용어 가중치 할당 기법 중 주제어 가중치 할당 방식을 활용하였다. 즉, 가중치 값이 높은 순으로 비율별 추출을 하였다.

세 번째는 두 번째 방식과 비슷하게 대표 용어 개수에 따라서 용어를 선별한 후 식(1)을 이용해서 문서간 유사도를 비교하였다. 두 번째와 세 번째는 문서별로 추출되는 용어의 수가 다르기 때문에 개수와 비율에 따라서 문서간 유사도 변화를 알아보기 위한 실험이다.

5. 실험 및 평가

실험에 사용된 문서는 6개의 국내 신문사 웹사이트(조선일보, 경향신문, 국민일보, 동아일보, 한겨레신문, 한국일보 등)에서 3일간의 기사를 무작위로 선정하였다. 총 문서 수는 383개였고, 평균적으로 문서당 132개의 용어가 추출되었으며, 실험 결과는 다음과 같다.

그림 1은 문서에 출현한 전체 용어를 추출한 경우와 문서내의 용어 빈도수가 2와 3이상인 용어만을 추출해서 각각 용어 일치율에 따라 비교한 것이다. 결과를 보면 용어 빈도가 올라갈수록 성능이 떨어지는 것을 알 수 있다. 이것은 용어 빈도수가 높아질수록 추출되는 용어의 개수가 작아지기 때문인 것으로 추정된다. 실제로 전체 문서 중 출현 빈도수가 3인 용어가 없는 문서가 발생하기도 하였다.

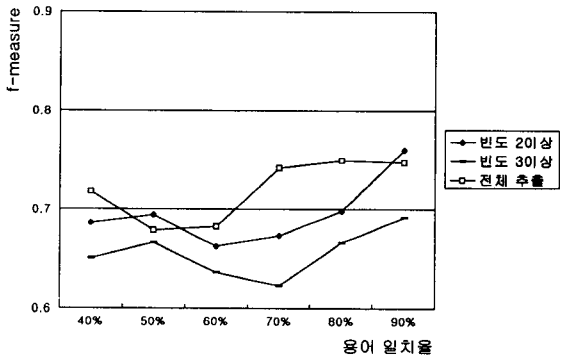

그림 1. 대표 용어 선택(빈도)에 따른 실험 결과

그림 2는 문서에 출현한 전체 용어를 추출한 경우와 주제어 가중치를 이용해서 중요 용어로 결정된 용어 중 백분율 순으로 추출해서 용어 일치율에 따라 유사 문서

로 판단한 경우를 비교한 것이다. 결과를 보면 10%만 추출한 경우에 성능이 가장 떨어지는 것을 알 수 있다. 이것은 평균 132개의 용어가 추출된 것을 감안할 때 10%를 추출했을 경우 13개 정도의 용어가 추출되는데, 이 정도의 용어 수로는 유사도 판단에 어려움이 있다고 판단된다. 그 외에는 대부분 비슷한 성능을 보였으며 20%~60% 추출한 경우가 평균적으로 좋은 성능을 보이고 있다.

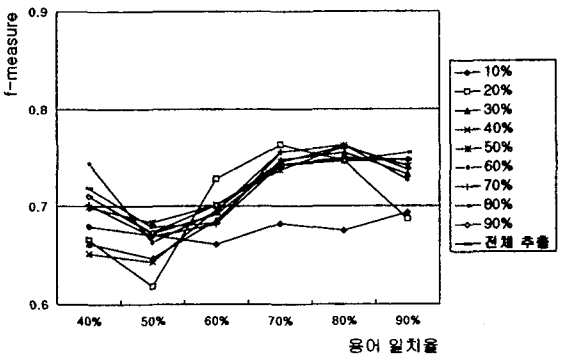

그림 2. 대표 용어 선택(백분율)에 따른 실험 결과

그림 3은 문서에 출현한 전체 용어를 추출해서 각 용어 일치율에 따라 유사 문서로 판단한 경우와 주제어 가중치를 이용해서 중요 용어로 결정된 순으로 개수 별로 추출해서 용어 일치율에 따라 유사 문서로 판단한 경우를 비교한 것이다. 전체적으로 큰 성능 차이를 보이지 않고 있으나, 40개~60개의 용어를 추출한 경우가 좋은 성능을 보인다.

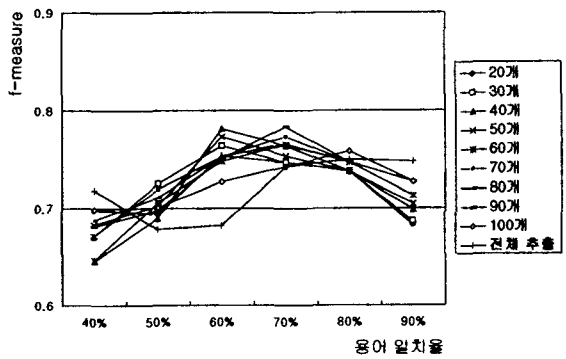

그림 3. 대표 용어 선택(개수)에 따른 실험 결과

6. 결론

유사 문서 판별 시스템은 문서 클러스터링과 유사하다. 그러나 문서 클러스터링처럼 클러스터 내의 문서간 관련성이 크지 않아도 되며, 시간과 공간을 보다 절약할 수 있다. 그렇기 때문에 정보 검색 시스템에서 질의에 대해 관련된 문서와 유사 문서를 찾아서 검색 결과에 활용할 수도 있다.

또한 기존의 정보 검색 시스템에서는 문서에서 출현한 모든 용어를 추출해서 검색 시스템에 활용한다. 그러나 문서내의 모든 용어의 추출로 인해서 시간과 공간을 많이 소비하는 문제가 발생하며, 용어 가중치를 이용하는 경우 낮은 가중치를 가지는 용어로 인해서 문서간 유사도 계산과 같은 문서간의 관계에 영향을 주는 경우가 발생한다.

실험 결과에서 나타나듯이 모든 용어를 추출한 경우보다 40~60개의 중요 용어만 추출한 경우에 비슷하거나 더 좋은 성능을 나타낸다는 것을 알 수 있으며, 이것으로 문서내의 모든 용어 추출이 반드시 필요한 것이 아니라는 것을 알 수 있다. 그러나 이것만으로는 모든 용어 추출의 효율성이 나쁘다고 말하기 어렵다. 실제로 빈도수가 낮은 용어도 문서의 주제어가 되는 경우도 있기 때문에 향후 과제로서 좀 더 다양한 영역의 문서에 의한 실험과 문서 분류나 문서 클러스터링 등과 같은 정보 검색 시스템에 기반 한 실험이 필요하다.

참고 문헌

[1] 강승식, 이하규, 손소현, 홍기채, 문병주, "조사 유형 및 복합명사 인식에 의한 용어 가중치 부여 기법", 한국정보과학회 가을 학술발표논문집, Vol.28, No.2, pp.196-198, 2001.

[2] 김영택, 자연언어처리, 생능출판사, 2001.

[3] Frakes, W. B. and R. Baeza-Yates, Information Retrieval, Prentice Hall, 1992.

[4] Murtagh, F., "Complexities of Hierarchic Clustering Algorithms: State of the Art", Computational Statistics Quarterly, Vol. 1, pp.101-113, 1984.

[5] Willett, P., "Recent Trends in Hierarchic Document Clustering: A Critical Review", Information Processing and Management, Vol. 24, No.5, pp.577- 597, 1988.

[6] Anderberg, M. R., "Cluster Analysis for Applications", New York: Academic, 1973.

[7] Can, F., and E. A. Ozkarahan, "Dynamic Cluster Maintenance", Information Processing & Management, Vol. 25, pp.275-291, 1989

[8] Dubes, R., and A. K. Jain, "Clustering Methodologies in Exploratory Data Analysis", Advances in Computers, Vol. 19, pp.113-227, 1980.

[9] Sibson, R. "SLINK: an Optimally Efficient Algorithm for the Single-Link Cluster Method", Computer Journal, Vol. 16, pp.328-342, 1973.

[10] Lewis, D. D. and Croft, W. B. "Term clustering of syntactic phrases", In J.-L. Vidick, editor, Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval, pp.395-404. 1990.