대용량 문서 검색을 위한 의미 흐름 기반 클러스터링

박사준*

요약

사용자가 요구하는 정보를 정확하고 효과적으로 검색하는 작업은 더욱 불편해지고 있다. 문서의 클러스터링은 대용량의 문서 집합에서 효과적인 정보 검색을 위한 요구 기능이다. 본 논문에서는 문서의 검색 응용에 문서 단위로의 연산 보다는 문서 내의 의미 부분을 활용한다. 온톨로지를 활용, 문서내의 의미 흐름 기반으로 문서를 문단화하고 이를 클러스터링에 활용하는 기법을 제안한다. 의미 흐름 단위로 문서 분류를 수행하므로 의미 기반 클러스터링이 가능하다. 클러스터링을 수행하는데 사용하는 단위가 문서에서 문단으로 줄어든다. 따라서, 문단 기반의 검색이 가능하게 함으로써 사용자가 문서 내에서의 검색을 수행할 수 있다. Reuter-21578 문서 집합을 사용하여 실험한 결과 문단 기반 방식 보다 성능이 향상되었다.

Meaning-Flow Based Clustering for Document Retrieval in a Large Document Set

Sa-Joon Park*

ABSTRACT

It is inconvenient that users retrieve information from documents efficiently and precisely. Clustering documents is a function for efficient information retrieval from massive document set. In this paper, we use meaning particle unit rather than document unit for document retrieval. We present a method with using an ontology that makes a document into paragraphs based on meaning flow. As the process of classification is done based on meaning paragraph, it is possible to achieve meaning-based clustering. The processing unit of clustering is shrunk from a document to a paragraph. Therefore, paragraph-based retrieval makes it possible for user to retrieve information in a document. We performed some experiments by using Reuter-21578 document set and the results showed the performance of meaning-flow based clustering was better than the performance of documents-based clustering.

Key Words : Clustering, Meaning-Flow, K-Means, Retrieval, Paragraph

* 대구한의대학교 모바일콘텐츠학부 (✉ phdjoon@dhu.ac.kr)
· 제1저자(First Author) : 박사준 · 교신저자(Correspondent Author) : 박사준
· 접수일(2013년7월 17일), 수정일(1차 : 2013년 7월 31일), 게재확정일(2013년 8월 8일)

Ⅰ. 서 론

대용량 문서 집단으로 부터 사용자가 원하는 정보를 빠르고 정확하게 검색하는 것은 정보 검색의 기본적이면서도 중요한 해결 과제이다. 이러한 문제를 해결하기 위해 정보 검색과 데이터 마이닝 분야에서 문서 클러스터링 기법의 연구가 수행되었다[1]. 문서 클러스터링의 연구의 목표는 문서 검색에 있어 재현율과 정확율의 향상에 있다. 클러스터링된 문서 집합은 사용자의 질의에 빠르고 정확한 응답이 가능하게 하는데 중요한 역할을 수행하고 있다. 최근 생성된 대부분의 문서가 전자문서의 형태를 가짐에 따라, 문서의 클러스터링에 대한 연구는 더욱 중요한 의미를 띄고 있다[2].

문서 클러스터를 이용한 정보 검색에서는 "서로 관련 있는 문서들은 동일한 질의에 비슷하게 반응 한다"라는 가정 하에, 서로 관련 있는 문서들을 클러스터로 형성하고, 사용자 질의에 대해서 클러스터의 관련도에 따라 관련이 높은 클러스터에 있는 모든 문서를 검색 결과로 제시한다[3-4]. 클러스터 분석은 문서-문서 관련도를 측정하기에는 좋으나, 정보검색에서의 질의는 통계적으로 의미 있는 빈도 벡터를 얻기에는 너무 적은 단어들로 구성되기 때문에 질의-클러스터의 관련도를 계산하기에는 적합하지 않다[4]. 또한, 문서들 사이의 관련도는 단어의 출현 빈도에 따라 결정이 되므로 의미적으로 유사성을 기반으로 한 문서 검색은 아니라는 문제점을 가지고 있다.

본 논문에서는 문서의 의미를 함축하고 있는 문단을 활용하여 문서를 의미 흐름 기반으로 문서의 클러스터링을 수행하는 기법을 제안한다. 이의 기법은 다음과 같은 장점이 있다. 첫째, 문서의 의미 흐름 단위인 문단(Paragraph)을 기반으로 분류를 수행한다. 따라서, 보다 의미 중심적인 클러스터링이 가능하다. 둘째, 클러스터링을 수행하는데 사용하는 단위가 문서에서 문단으로 줄어듦으로 인해 비교 연산을 수행하는 영역의 차원이 줄어든다. 이는 클러스터링 시스템의 속도 및 정확도를 높여준다. 뿐만 아니라, 외부 온톨로지를 활용하여 의미를 확장 할 수 있는 여지를 높여준다. 셋째, 문단 기반의 검색이 가능하게 함으로써 사용자로 하여금 문서 내에서의 검색을 원활하게 수행할 수 있도록 해준다.

Ⅱ. 관련 연구

문서 클러스터링의 대표적인 기법들로는 응집 계층적 클러스터링(Agglomerative Hierarchial Clustering) 과 K-Means 기법이 있다. K-Means 기법이 응집 계층적 기법에 비해 빠르지만 클러스터링의 성능이 좋지 않다는 단점을 가지고 있다. 이를 극복하기 위해 다양한 형태의 연구가 수행되었다.

Cutting/Karger는 K-Means가 가진 효율성과 응집 계층적 클러스터링이 가진 정확도를 적절히 섞은 하이브리드 기법을 도입한 Scatter/Gather 검색 시스템을 제시했다. 해당 시스템은 질의에 대해 검색된 문서를 대상으로 클러스터링을 수행하는 동적 클러스터링 기법을 소개했다[5].

문서 집합의 클러스터링과는 반대로 K-Means를 활용하여 문서를 요약하는 기법도 제시되었다. Jain/Bewoor는 사용자의 질의를 바탕으로 문서내의 문장들을 K-Means 기법을 활용하여 클러스터링하고 요약하는 기법을 제시하였다[6].

하지만 상기의 방법들은 문서의 크기가 커질수록 K-Means는 효율성이 점차 떨어지고 응집 계층적 클러스터링은 그 성능이 떨어짐을 보이는 단점이 있다. 뿐만 아니라 문서의 유사도 측정 방식이 단어의 출현 빈도에 맞추어져 있으므로 인해 의미

적 유사도를 적절히 반영치 못한다는 문제점이 있다. 그리고 검색의 단위가 문서이므로 사용자가 원하는 정보를 검색하기 위해서는 검색된 결과 리스트의 문서를 재검색해야 하는 문제점이 있다.

본 논문에서는 의미 기반의 클러스터링 기법을 적용하여 이를 극복하려 한다. 문서를 효과적으로 문단화하기 위해 우리는 온톨로지를 활용하였다. 본 연구에서는 의미 기반으로 문단화된 문서에 K-Means 기법을 도입하여 효율성과 성능을 개선시켰다.

III. 의미 기반 클러스터링

관련 연구를 통해 대용량의 문서를 클러스터링하는데 있어 K-Means 기법이 효율적이며 클러스터링에 적용할 단위는 의미 응집적인 문단이 더욱 적합함을 알 수 있다. 그러나, 문서의 문단화를 위해 의미 흐름을 측정하기는 쉽지 않다는 문제점이 있다. [7]는 문서내의 단어의 출현 빈도를 기반으로 문장 간의 응집도를 측정하고 이를 문단화에 이용했다. 하지만 문서 작성자가 유사어를 활용한 경우 의미 단위의 추출은 더욱 어려워진다. 이의 해결을 위해서는 온톨로지의 도입이 요구된다.

본 장에서는 대용량의 문서를 클러스터링하기 위해 온톨로지를 활용, 문서들을 의미 흐름 기반으로 문단화 하고, K-Means 기법을 활용하여 효과적으로 문서를 클러스터링 하는 기법을 제안한다.

3.1 의미 흐름 기반 문단화

문서을 의미 기반으로 문단화하기 위해서는 문서 내 문장들 간의 유사도를 기반으로 수행한다. 문장들 사이의 의미 흐름을 분석하기 위해 WordNet을 온톨로지로 활용하였다. 본 연구에서 사용한 문장 간의 의미 유사도 측정은 다음과 같다.

문서(d)는 문장(s)들로 구성되고, 문장(s)는 단어(w)들로 구성된다. 따라서 문서 내 이웃한 문장 간의 유사도 측정은 다음과 같이 구할 수 있다.

$$d = \{s_1, s_2, \cdots\cdots, s_n\}$$
$$s_i = w_i = \{w_{i1}, w_{i2}, \cdots\cdots, w_{in}\}$$
$$s_{i+1} = w_j = \{w_{j1}, w_{j2}, \cdots\cdots, w_{jn}\}$$

$$Sim(s_i, s_{i+1}) = \frac{len(s_i)}{\sum_k \min(SD(w_{ik}, w_j))} \quad (1)$$

d : 문서
s_i : i번째 문장
w_i : i번째 문장의 단어 집합
s_{i+1} : $i+1$번째 문장
w_j : $i+1$번째 문장의 단어 집합
$len(s_i)$: i번째 문장의 길이,
i번째 문장을 구성하는 단어의 수

이웃한 문장 간의 유사도 $Sim(s_i, s_{i+1})$을 구하기 위해 사용된 SD함수는 Natural Language Toolkit의 short_distance 함수로서 WordNet에서 두 단어들 간의 최단 거리를 구한다. 본 논문에서는 WordNet을 온톨로지로 활용하므로 두 문장 내 단어들 간의 거리 측정을 통하여 문장의 유사도를 측정한다. $Sim(s_i, s_{i+1})$ 값은 이웃한 두 문장이 동일할 경우 1의 값을 가지지만 유사도가 떨어질수록 0에 가까운 값을 가진다.

문서에 대해 (1)에서 기술된 유사도 측정 방법을 이용함으로써, 문장 간의 유사도를 나열할 수 있다. 그리고 이러한 문장 간의 순차적인 유사도 나열은 문서의 의미 흐름(Meaning Flow)을 나타낸다. 다음은 문서 d에 대한 의미 흐름(MF)을 구하는 식이다.

$$MF(d) = \{Sim(s_1, s_2), \cdots\cdots, Sim(s_{n-1}, s_n)\} \quad (2)$$

상기와 같이 문서의 의미 흐름을 구한 후, 이웃한 문장들 간의 유사도가 급격히 떨어지는 부분을 추출함으로써 의미 흐름별 문단화를 수행한다. 의미 흐름별 문단화 $Para(d)$는 유사도의 변화를 측정하기 위해 $MF(d)$의 평균(AVR)과 표준 편차(STD)를 활용하였다.

$$Para(d) = \{i : Sim(s_i, s_{i+1}) \leq AVR(MF(d)) - \alpha * STD(MF(d))\} \quad (3)$$

즉, 이웃한 문장 간의 유사도가 문서내의 평균 문장 간의 유사도보다 표준편차 범위 밖으로 떨어질 경우 이를 의미별 문단화 지점으로 선택하였다.

3.2 의미 흐름 문서 클러스터링

의미 흐름 문서 클러스터링은 문서를 의미 흐름으로 문단화하여 클러스터링을 수행하는 방법이다. 문서내의 의미 흐름별 문단을 활용하여 클러스터링을 수행할 경우 크게 두 가지의 클러스터링 방법을 고려해 볼 수 있다. 첫 번째 방법으로 가장 의미 응집도가 높은 문단을 기준으로 문서 단위로 클러스터링을 수행하는 것이다. 두 번째 방법으로는 문서내의 모든 문단들을 기준으로 클러스터링을 수행하는 방법으로 문서가 클러스터에 존재할 확률적 결과값을 산출할 수 있다. 본 연구에서는 첫 번째 방법을 기반으로 클러스터링을 수행하였다.

문서의 클러스터링은 K-Means 클러스터링 기법을 기반으로 수행한다. 3.1에서 제안한 방법을 통해 문서를 의미 흐름 기반으로 문단화하고 K-Means 기법을 활용하여 문서를 클러스터링한다.

우선 K-Means 기법을 활용하기 위해 문단을 벡터화한다. 문단을 벡터화하는데 있어서, 단어의 중요도는 빈도수를 바탕으로 한 TF-IDF(Term Frequency - Inverse Document Frequency)를 이용한다[8-9]. 본 기법에서는 문서 길이의 영향을 제한하기 위해 정규화된 tf, idf를 사용하였다.

$$tf(t,d) = \frac{f(t,d)}{\max\{f(w,d) : w \in d\}} \quad (4)$$

$$idf(t,D) = \log\frac{|D|}{|d \in D : t \in d|} \quad (5)$$

K-Means 기법에 적용될 문단의 추출은 다음과 같다.

$$d_i = \{p_{i1}, p_{i2}, \cdots\cdots, p_{in}\}$$
$$C(p_{ik}) = \frac{\sum\{tf(t,d_i) * idf(t,D) | t \in p_{ik}\}}{|p_{ik}|} \quad (6)$$

d_i는 3.1절의 의미 기반 문단 분할 기법(수식 3)을 활용하여 문단으로 분할된 문서이다. p_{ik}는 문서내의 분할된 문단이다. $C(p_{ik})$는 문단의 의미 응집도를 나타내며 문단 내 단어들의 tf-idf값의 총합을 문단의 크기 $|p_{ik}|$로 정규화했다. $|p_{ik}|$는 문단 내 문장의 수를 의미한다. 문단 내에 의미성 있는 단어가 많이 포함될수록 의미 응집도는 높아지게 된다.

의미 기반 문서 클러스터링에서는 문서에서 의미 응집도 $C(p_{ik})$가 가장 높은 문단들을 K-Means 기법을 이용하여 클러스터링을 수행한다.

본 방법론에서 사용한 K-Means은 MA(Maqueen Approach) 방법을 사용한다[10]. 첫 단계에서는 자료에서 k개의 관측값을 랜덤하게 선택하여 초기값으로 결정하여 초기 클러스터링의 중심을 형성한다. 다음 단계에서는 초기값으로 설정되지 않은 값들을 가장 가까운 초기값이 중심이 되도록 클러스

터링을 수행하여 클러스터를 형성한다. 형성된 클러스터의 중심을 다시 계산하여 관측값과의 거리를 계산하여 가까운 중심의 클러스터로 관측값을 이동하여 새로운 중심을 구한다. 클러스터링의 중심의 이동이 임계값 이하가 될 때까지 이 과정을 반복한다.

Ⅳ. 실험 및 평가

의미 흐름 기반 클러스터링의 성능을 검증하기 위해 Reuter-21578 문서 집합을 이용하였다. Reuter-21578 문서 집합에서 문서 크기순으로 2800개의 문서를 추출, K-Means 기법을 이용하여 의미 흐름 기반 클러스터링을 수행(K=50, 반복 횟수는 20을 적용)하고 7개의 질의어('earn', 'acq', 'grain', 'wheat', 'corn, 'coffee', 'silver')를 이용하여 클러스터링의 성능을 평가하였다. <그림 1>은 문서 기반 클러스터링(Document- Based Clustering)과 의미 기반 클러스터링(Meaning-Based Clustering)에 대해서 Reuter-21578 문서 집합을 대상으로 재현율을 비교한 결과이다.

실험 결과에서 보듯이 의미 흐름 기반 클러스터링의 재현율은 60~80%를 기준으로 등락을 보였으며 대부분의 경우에 있어 의미 기반의 클러스터링이 문서 기반에 비해 더 좋은 결과를 보였다. 특히 'silver' 그룹에서는 2배 이상의 좋은 성능을 보였다.

K-Means 기법에서 K 값은 클러스터링을 수행할 개수를 나타내며, K 값에 따라서 문서 집합의 클러스터링 성능이 영향을 받는다. <그림 2>은 본 실험에서 K 값을 50부터 10씩 변화시켜가면서 클러스터링의 성능을 평가한 결과, K 값이 70일 경우가 가장 나은 결과를 보여주고 있다.

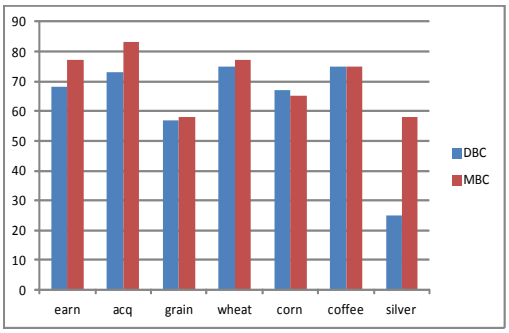

그림 1. 문서 기반 클러스터링(DBC) vs. 의미 기반 클러스터링(MBC) 성능 비교
Fig. 1. The performance comparison between Document-Based Clustering and Meaning-Based Clustering

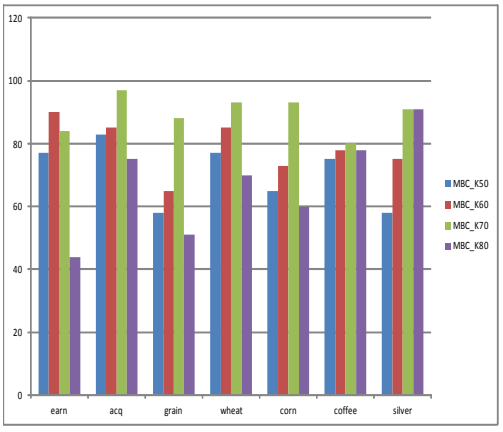

그림 2. K값에 따른 클러스터링 성능 비교
Fig. 2. The performance comparison of clustering by K

Ⅴ. 결론 및 향후 연구 방향

대용량의 문서에서 정확도가 높은 결과를 도출하기 위해서 문서를 의미 흐름 기반으로 문단화하고 문장 간의 유사도를 측정하였다. 측정된 유사도를 기반으로 하여 K-Means 기법을 이용하여 클러스터링을 시도하였다. Reauter-21578 문서 집합을 이용하여 성능을 실험하였다. 문서 기반 클러스터

링에 비해 성능이 향상되었음을 실험 데이터에서 확인할 수 있었다.

본 시스템의 특징은 문서의 의미 흐름 단위인 문단(Paragraph)을 기반으로 분류를 수행한다. 클러스터링을 수행하는데 사용하는 단위가 문서에서 문단으로 줄어듦으로 인해 비교 연산을 수행하는 영역이 줄어든다. 이는 클러스터링 시스템의 속도 및 정확도를 높여준다. 뿐만 아니라, 외부 온톨로지를 활용, 의미를 확장 할 수 있는 여지를 높여준다. 문단 기반의 검색이 가능하게 함으로써 사용자로 하여금 직접적인 검색을 수행할 수 있도록 해준다.

현재 클러스터링을 수행함에 있어서 가장 유사도가 높은 문단들을 기준으로 문서 단위의 클러스터링을 수행하였다. 사용자 질의에 대해 보다 정확한 검색 결과를 도출하기 위해서는 확률적 기반의 클러스터링이 요구된다. 따라서 향후 문서내의 문단들을 기준으로 의미별 클러스터링을 수행하여 확률적 기반의 클러스터링과의 연계에 대한 연구를 수행할 예정이다. 또한 다양한 방법의 K-Means 방법을 적용하여 더 좋은 성능을 발휘하는 방법을 연구할 계획이다.

참고문헌

[1] T. Radecki, "A model of a document-clustering-based information retrieval system with a Boolean search request formulation", *SIGIR 1980*, pp 334-344, 1980.

[2] M.Steinbach, G.Karypis, and V.Kumar, "A comparison of Document Clustering Techniques", *KDD Workshop on Text Mining*, 2000.

[3] A. Leusik. "Evaluating Document clustering for interactive information retrieval", *CIKM 2001*, pp 33-40, 2001.

[4] K. Lee, "A document ranking model based on vector space retrieval and cluster analysis in information retrieval", *KAIST, Ph.D.dissertation*, 2001.

[5] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections", *SIGIR'92*, pp. 318-329, 1992.

[6] H. J. Jain, M. S. Bewoor, and S. H. Patil, "Context Senstive Text Summarization Using K Means Clustering Algorithm", *IJSCE*, Vol 2. Issue 2. pp. 301-304, 2012.

[7] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird, "Learning collection fusion strategies", *Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.172-179, 1995.

[8] G. Salton, *Automatic information organization and retrieval*, McGraw-Hill, 1968.

[9] K. Spark Jones, "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, Vol. 28, pp. 11-21, 1972.

[10] J.B. Macqueen, "Some Methods for Classification and Analysis of Multivariate Observations", *Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281 - 297, 1967.

감사의 글

이 논문은 2011년도 대구한의대학교 기린연구비 지원에 의한 것 임.

저자소개

박사준(Sa-Joon Park)

1990년 중앙대학교 전자계산학과 (학사)

1994년 중앙대학교 대학원 컴퓨터공학과(공학석사)

2004년 중앙대학교 대학원 컴퓨터공학과(공학박사)

2005년-현재 대구한의대학교 모바일콘텐츠학부 교수

※ 관심분야 : 시맨틱 웹, 인공지능, 모바일콘텐츠