문서 클러스터링 기법을 활용한 병렬 정보 검색

강유경, 박세진, 류광렬, 정상화
부산대학교 컴퓨터 공학과 인공지능연구실, 병렬처리 연구실

Parallel Information Retrieval using Document Clustering Techniques

Yu Gyung Kang, Se Jin Park, Kwang Ryel Ryu, Sang-Hwa Chung
Dept. of Computer Engineering, Pusan National University

요 약

본 논문은 고품질의 정보를 신속하게 제공할 수 있으면서, cost-effective한 medium-grained 병렬 정보 검색 시스템을 제시하고 있다. 본 검색 시스템은 병렬 모델의 효율을 극대화하는 방안으로 문서 라이브러리를 작은 단위의 클러스터로 세분화하고 검색 시 클러스터 단위로 프로세서에 할당될 수 있게 하여 할당될 작업의 단위를 적절히 중규모화하였을 뿐만 아니라, 각 클러스터마다 독립적인 역색인 화일을 별도로 두어 순위 부여 계산 시 통신을 최소화 할 수 있도록 하였다. 또한, 기계 학습 기법을 이용하여 가능한 한 유사한 문서군이 되도록 클러스터링함으로써 불필요한 클러스터가 검색될 가능성을 최소화하여 성능을 높였다 본 검색 시스템은 분산메모리 MIMD 구조의 트랜스퓨터에서 구현되었으며, Connection machine에서 사용되는 Stanfill 방법과의 비교 실험을 통하여 계층적인 접근법의 성능을 비교, 평가하였다. 그리고 random 클러스터링 기법과 비교하여 기계 학습을 통한 클러스터링 접근 방법이 우수함을 보이고 있다

1. 서 론

인터넷과 정보 서비스 기술의 발달로 일반 대중에게 제공되는 정보의 양이 기하급수적으로 증가하고 있는 시점에서, 자신이 원하는 정보를 얼마나 신속하게 찾을 수 있는 가가 매우 중요한 이슈로 대두되고 있다. 현재 정보 검색 수요자들이 경험하는 병목현상은 아직 통신선로 상에 있다고 볼 수 있지만, 앞으로 초고속 정보통신망이 현실화되면 그 병목현상이 검색엔진 쪽으로 옮아 갈 것이 분명하다. 따라서, 사용자가 원하는 고품질의 정보를 신속하게 제공할 수 있는 검색 시스템의 실현이 시급히 요청되는 것이 현실이다.

현재 문서검색 기술의 주류는 통계적 접근방식을 기본으로 한 것으로 사용자가 지정한 키워드들을 포함한 문서들을 문서 라이브러리로부터 모두 찾은 뒤, 키워드의 중요도 및 등장빈도에 따라 순위를 부여하여 사용자에게 되돌려 주는 방식을 취하고 있다. 지정된 단어를 포함한 문서를 신속히 찾기 위한 수단으로는 역색인 화일이 널리 사용되고 있다. 역색인 화일을 이용하는 경우, 해당 문서를 찾은 뒤에는 지정된 단어의 중요도와 등장 빈도 등을 고려하여 문서들에 순위를 부여하게 된다. 이때 사용자가 원하는 고품질의 정보를 제공하기 위해 신뢰성이 높은 순위 부여 계산모델을 도입할 경우 파생되는 문제는 이를 위한 계산과정의 복잡성 또한 높아져서 계산시간이 늘어난다는 것이다. 이러한 상황에서 순위 부여에 소요되는 계산시간을 단축하기 위하여 효율적인 병렬처리 방안을 강구하는 것이 바람직하다.

이에 본 연구에서는 기계 학습 기법을 응용하여 문서 라이브러리를 가능한 한 유사한 문서군으로 세분화하고, 역색인 화일을 클러스터 레벨 및 문서 레벨의 두 단계로 계층화함으로써 불필요한 문서들에 대한 계산을 피하고 프로세서들 간의 통신도 최소화하였다.

본 검색 시스템은 분산메모리 MIMD 구조의 트랜스퓨터에서 구현되었으며. Connection Machine에서 제시된 Stanfill[6] 방법과 실험을 통하여 비교,분석하였다. 또한 기계 학습 기법이 검색 성능에 미치는 영향을 알아보기 위하여 기계 학습 기법을 이용한 경우와 random하게 클러스터링한 경우를 비교 실험하였다.

2. 관련연구

문서 검색의 병렬처리에 관한 종래의 연구로는 Stanfill 방법이 있다. Stanfill은 Connection Machine을 대상으로 한 fine-grained MPP 모델을 제안하였는데, 이 모델은 전체 문서 라이브러리에 대한 역색인 화일을 여러 개로 나누어 분산 저장시켜두고, 검색할 질의 단

◇ 본 연구는 한국 과학 재단 '97 핵심 연구 과제의 일환임
(과제 번호: 971-0901-013-2)

어에 대한 역색인 화일의 위치를 호스트 컴퓨터에 저장한 후 이것을 바탕으로 역색인 화일을 검색하게 된다

검색 시에 대상이 되는 문서들은 순서에 따라 미리 프로세서에 할당되어있으므로, 순위 부여 계산 시 프로세스들 간의 통신은 전혀 필요 없을 뿐 아니라 각 질의어 당 한 번의 디스크 접근만이 요구되므로 아주 빠른 시간 내에 검색이 가능한 장점이 있다.

그러나 각 질의에 해당되는 영역에는 해당 질의어 외에 다른 질의어 및 빈 공간도 포함되므로 저장 공간의 낭비를 가져 올 뿐 아니라 필요 이상의 영역을 읽어야 하는 문제점이 있으며, 전체 문서 라이브러리에 대해 하나의 역색인 화일을 유지하므로 수정이나 추가 등의 유지 보수가 필요한 경우 처음부터 다시 만들어야 하는 문제점이 있다.

3. 문서의 클러스터화

3.1. 클러스터링의 필요성

역색인 화일을 단순히 분할하여 이들을 프로세서에게 할당하는 방식은 단어별로 프로세서에게 할당되므로 순위 부여 계산 시 프로세서간의 과다통신을 유발시키게 된다. 반면, 문서 라이브러리를 작은 단위로 클러스터화하고 각 클러스터마다 그 내부의 문서만을 대상으로 하는 역색인 화일을 두는 것은 효과적인 통신 절감 방안이 된다.

문서 라이브러리를 클러스터화함에 있어 무작위로 클러스터링하는 것보다 서로 유사한 문서끼리 하나의 문서군을 이루게 하는 것이 1차 검색 시 관련 클러스터의 수를 줄일 수 있으므로 검색을 훨씬 효율적으로 할 수 있다

3.2. 문서의 표현

통계적 접근 방식에 의해 키워드를 기반으로 검색을 수행하는 시스템에서 각 문서의 특징은 그것이 어떤 단어들을 어떤 빈도로 포함하고 있는가에 의해 표현되어져야 한다. 즉 문서 라이브러리 내에 등장하는 단어의 종류가 모두 n가지일 경우 이들 단어의 순서를 미리 정해 둔 상태에서 임의의 문서는 다음과 같은 속성 vector V로 표현된다.

$$V=(tf_1, tf_2, tf_3, \cdots, tf_n)$$

여기서 tf_i는 i번째 단어가 문서 V내에 등장하는 빈도수이다. 그러나 단순히 단어의 빈도수를 속성값으로 취하게 되면, 단어의 빈도수는 문서의 크기에 비례하는 경향이 있으므로 서로 유사한 문서라 하더라도 문서의 크기가 현저히 다를 경우에는 상이한 문서처럼 보일 수 있다. 이러한 문제점은 문서의 크기로 정규화시킴으로써 해결할 수 있다. 따라서 본 논문에서는 SMART system[2]에서 제안되어진 다음과 같은 방식을 사용하였다.

$$V=(a_1, a_2, a_3, \cdots, a_n)$$

$$a_i=\frac{w_i}{\sqrt{\sum_{j=1}^{n} w_j^2}}$$

$$w_i=\ln tf_i+1$$

각 단어의 가중치는 그 단어의 logarithmic frequency[2]를 나타내며, 모든 단어들의 가중치의 합으로 각 단어의 가중치를 나눈 값을 속성값으로 하여 문서를 표현한다.

한편, 문서를 그에 등장하는 단어들의 vector로 표현하게 되면 속성의 수가 아주 많게 된다. 일반적으로 training data의 수에 비해 속성의 수가 많은 경우에는 training data에 overfitting하여 성능이 좋지 못한 classifier가 유도될 가능성이 높으므로, 중요한 속성들을 미리 선별하는 feature subset selection[1,4] 작업이 선행되는 것이 좋다. 이에 본 시스템에서는 feature subset selection작업을 통하여 선정된 단어만을 사용하여 기계 학습법을 적용하였다.

3.3. 문서의 클러스터링

본 논문에서는 클러스터링을 위해 decision tree 학습 알고리즘인 C4.5[3]를 사용하여 2160개의 문서를 training data로 하여 학습시켰으며, 문서 라이브러리를 125개의 클러스터로 나누었다

4. 병렬 정보 검색

본 검색 시스템은 계층적 역색인 화일 구조를 기반으로 하여 1, 2차 검색을 실시하고, 순위계산 모델로서 P-norm 모델[5]을 사용하여 문서 순위를 부여한다 순위 계산 작업을 마친 문서들은 루트 프로세서로 전송되어 히프 정렬에 의하여 정렬되고 사용자에게 결과를 되돌린다.

4.1. 질의 입력 및 분석

사용자로부터 입력된 질의는 호스트 컴퓨터에서 분석된 후 1차 역색인 화일 검색에 사용될 질의 단어 리스트로 구성되어 질의 큐에 저장된다.

4.2. 클러스터 레벨 역색인 화일과 1차 검색

클러스터링 기법에 의해 생성된 문서 클러스터가 n개라고 할 때, 각 클러스터 내에 있는 문서들을 모두 하나의 문서로 병합하여 n개의 새로운 문서를 구성한다. 이 n개의 문서들을 대상으로 만든 것이 클러스터 레벨 역색인 화일이며 그 구조는 그림 1과 같다

단어	클러스터 ID	관련 문서 수	클러스터내 단어빈도수

그림 1 클러스터 레벨 역색인 화일 구조

이러한 접근법은 질의어가 들어오면 이 질의어와 관련된 클러스터 ID를 쉽게 찾을 수 있게 한다. 또한 클러스터 내에서 질의어와 관련된 문서 수를 고려하여 2차 검색 후 프로세서에게 클러스터들을 할당함으로써, 각 프로세서의 부하를 균등하게 할 수 있는 이점이 있다

클러스터 내 단어 빈도수는 관련 문서 수와 클러스터의 크기 정보와 함께 클러스터의 점수를 계산하는데 사용된다. 클러스터의 점수는 클러스터 크기가 작고 관련 문서 수는 많으면서 클러스터 내 단어 빈도수가 큰 클러스터일수록 높다. 이 점수가 일정 임계치 이상인 클러스터만이 2차 검색에 참여할 수 있게 함으로써 2차 검색의 부담을 크게 줄일 수 있는 이점이 있다.

4.3. 문서 레벨 역색인 화일과 2차검색

기계 학습법에 의하여 문서 클러스터들이 만들어진 후, 각 클러스터 별로 독립적인 역색인 화일을 만든 것을 문서 레벨 역색인 화일이라고 하며 그 구조는 그림 2와 같다.

단어	문서 ID	단어 가중치

그림 2 문서 레벨 역색인 화일 구조

1차 검색에 의하여 클러스터들이 선정되면, 이 클러스터들의 문서 레벨 역색인 화일을 탐색함으로써 2차 검색이 이루어진다.

4.4. 순위 계산

2차 검색이 완료된 클러스터내의 각 문서들은 순위 계산 작업에 참여하게 된다. 본 시스템에서는 Boolean 모델을 확장하여 Vector Space 모델까지 포함하는 신뢰성이 높은 P-norm[5] 모델을 사용하였다.

4.5. 검색 결과 집계 및 검색 종료

각 프로세서에서 문서의 순위 계산 작업이 끝난 문서들은 루트 프로세서로 전송되고, 루트에서는 전송받은 문서들을 히프 정렬에 의해 정렬하고 검색을 종료하게 된다.

5. 실험

5.1. 실험 환경

본 검색 시스템은 1개의 루트 프로세서와 16개의 검색프로세서, 그리고 4개의 하드디스크로 구성된 분산 메모리 MIMD 구조의 다중 트랜스퓨터에서 구현되었다.

5.2. 실험 데이터 및 방법

실험 대상 문서로서는 신문기사에서 발췌한 9593개의 문서를 사용하였다. 본 연구에서는 계층적 접근법의 성능을 평가하기 위하여 Stanfill 방법과 비교, 실험하였으며, 또한 문서의 범주화기법이 시스템 성능에 미치는 영향을 알아보기 위하여 기계 학습 기법의 일종인 C4.5를 적용한 방법과 random하게 클러스터링한 방법을 비교, 실험하였다. C4.5를 적용한 방법에서는 9593개의 문서 중 2160개의 문서를 training set으로 사용하여 125개의 클러스터로 나누었으며, random하게 클러스터링한 방법은 무작위로 125개의 클러스터로 나누었다.

5.3. 실험 결과

그림 3은 세 방법에 대하여 질의어 수를 1에서 8개까지 증가시켰을 때의 평균 검색 시간에 대하여 보여주고 있다. 질의어는 1개의 질의어로 검색했을 때 가장 높은 점수를 가지는 상위 3개의 문서에 등장하는 단어들을 포함하도록 확장하였다.

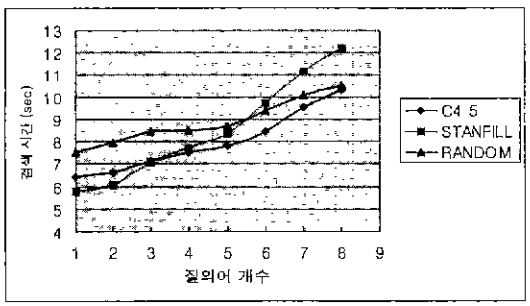

그림 3. 질의어 확장에 따른 검색 시간 비교

그림 3에서 알 수 있듯이 Stanfill 방법은 질의어 수가 적은 경우에는 그 성능이 우수하나 질의어 수가 많아지면 검색 시간이 크게 증가하는 것을 볼 수 있다. 계층적인 접근 방식은 질의어 수가 많아지더라도 검색 시간이 크게 증가되지는 않는다. 따라서 최근의 문서 검색 경향 중 thesaurus를 사용하는 방식에 입각하여 볼 때, 계층적인 접근 방식이 우수함을 알 수 있다. 또한 그림 3과 표 1에서 C4.5를 활용하여 클러스터링한 것이 random하게 클러스터링한 방법보다 검색 시간 및 정확도에서 더 우수하다는 것을 알 수 있다

그러나 Stanfill 방법의 경우, 질의어를 포함하는 모든 문서에 대하여 검색, 평가하는데 반하여 계층적인 방식에서는 1차 검색 후 그다지 좋지 못한 클러스터는 검색 대상에서 제외되므로 질의어를 포함하는 모든 문서에 대한 검색이 이루어지지 않는 위험이 따른다 이에 대한 평가로서 문서 검색의 결과로 돌려지는 상위 50개의 문서에 대한 정확도를 측정해 보았다 즉 각 방법에서 돌려지는 상위 50개의 문서가 Stanfill 방식의 결과로 반환되는 상위 50개의 문서와 일치하는 개수에 대한 비율을 구하였다. 그 결과는 표 1과 같다.

표 1 상위 50개 문서에 대한 정확도

질의어 개수	C4.5 (%)	RANDOM (%)
1	94.13	90 14
2	93.13	90.5
3	91.2	90
4	90.85	90.3
5	91.73	88.4
6	89.57	89.6
7	84	84.8
8	88 13	83.7

표 1에서 보면, 질의어 수가 확장됨에 따라 정확도 또한 조금씩 감소하는 경향을 보이는데, 질의어 수가 늘어나더라도 정확도를 일정 수준 이상으로 만들 수 있는 방안이 추가되어야 할 것이다 이는 1차 검색 후 2차 검색 대상 클러스터를 선정하는 단계에서 검색 시간 및 정확도의 tradeoff를 적절히 조절하는 작업을 통해서 구현될 수 있을 것이다.

6. 결론 및 향후 연구과제

본 논문에서는 medium-grained 병렬 정보 검색 모델을 제시하였다 클러스터 레벨과 문서 레벨의 계층적인 역색인 화일을 통하여 이루어지는 검색은 질의어 수가 늘어남에 따라 Stanfill 방법보다 더 빠른 시간 내에 검색이 가능하였으며, 이는 문서 검색의 흐름이 thesaurus를 활용하는 방향으로 나아가는 점을 고려하면 상당히 좋은 접근법이라 할 수 있다. 그러나 다양한 기계 학습법을 활용한 클러스터링 기법에 대한 연구 및 높은 정확도를 얻기 위한 방안등에 관한 연구가 잇달아야 할 것이다.

7. 참고 문헌

[1] Wettschereck, D. and Aha, D. W., "Weighting Features," *First International Conference on Case-Based Reasoning*, 1995

[2] Buckley, C., "Implementation of the SMART information retrieval system," Technical Report 85-686, *Cornell University*, 1985

[3] Quinlan, J R., *C4.5 Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, California, 1993.

[4] Sholom M W and Indurkhya N., *Predictive Data Mining*, Morgan Kaufmann Publishers, San Mateo, California, 1997.

[5] Smith, M.E., "Aspects of the p-norm model of information retrieval · syntactic query generation, efficiency, and theoretical properties," Ph.D. thesis, *Cornell University*, 1990

[6] Stanfill, C. and Thau, R., "Information Retrieval on the connection machine : 1 to 8192 Gigabytes," *Information Processing & Management*, pp.285-310, 1991.