질의 확장을 이용한 병렬 정보 검색

정유진[0]
한국외국어대학교 컴퓨터공학과
chungyj@hufs.ac.kr

Parallel Information Retrieval with Query Expansion

Yoojin Chung[0]
Dept. of Computer Engineering, Hankuk University of Foreign Studies

요 약

이 논문에서는, PC 클러스터 환경에서 질의 확장을 사용하는 정보 검색 시스템 (IR)을 설계하고 구현한 내용을 기술한다. 이 정보 검색 시스템은 문서 집합을 저장하고, 문서 집합은 역색인 파일 (IIF)로 색인되고, 랭킹 방법으로 벡터 모델을 사용하며, 질의 확장 방법으로 코사인 유사도를 사용한다. 질의 확장이란 사용자가 준 원래의 질의에 연관된 단어를 추가하여 검색 효율을 향상시키는 것이다. 여기서 제안하는 병렬 정보 검색 시스템에서는 역색인 파일은 여러 개로 분할되는데 lexical 분할 방법과 greedy 분할 방법을 사용한다. 사용자의 질의가 들어오면 질의 확장을 하여 여러 개의 단어로 이루어진 확장된 질의가 만들어지는데 이 확장된 질의를 구성하는 단어들은 각 단어와 연관된 IIF를 가지고 있는 노드에 보내어져서 병렬로 처리된다. 실험을 통하여, 병렬 IR 시스템의 성능이 질의 확장과 IIF의 두 가지 분할 방법에 의해 어떻게 영향을 받는지 보인다. 실험에는 표준 한국어 테스트 말뭉치인 EKSET과 KTSET을 사용하였다. 실험에 따르면 greedy 분할 방법이 lexical 분할 방법에 비해 20%정도의 성능 향상을 보였다.

1. 서 론

일반적으로 IR 시스템은 문서 집합을 저장하고 색인하며, 질의가 주어지면 키워드 기반의 검색을 한다. 이 논문에서 제안하는 질의 확장을 이용한 병렬 IR 시스템은 빠른 네트워크 카드로 연결된 PC 클러스터에서 구현되었다. 매우 큰 문서 집합의 효율적인 질의 처리를 하기 위해 특별한 색인 기법들이 사용되어야만 하는데, 여기서는 문서들이 역색인 화일 [2] 을 사용하여 색인된다. 질의 확장이란 사용자가 준 원래의 질의에 단어들을 추가하여 검색 효율을 향상시키는 기법이다. 여기서 사용하는 질의 확장 방법은 동시등장 기반의 유사도 계산을 사용한다. 제안하는 시스템의 개발 환경인 PC 클러스터에서는 여러 노드가 있으므로 이들 노드들에 정보를 분할하여 저장하는 것이 이상적인데 이 논문에서는 전체의 역색인 파일 (IIF)을 한 노드에서 구한 후 이를 여러 개로 분할한다. 분할 방법으로 이전 연구 [4] 에서 제안하였던 lexical 분할 방법과 greedy 분할 방법을 사용한다. 질의 확장을 이용한 병렬 정보 검색에 대한 연구는 저자가 아는 한 이전에는 없었다.

이 논문을 다음과 같이 구성되어 있다. 2절에서는 질의 확장을 이용한 병렬 정보 검색 시스템에 대한 개관을 설명하고 3절에서는 실험과 결과를 기술한다. 마지막으로 결론을 기술한다.

2. 질의 확장을 이용한 병렬 정보검색 시스템

PC 클러스터 기반 병렬 정보검색 시스템은 주 노드와 종속 노드로 구성된다. 주 노드는 사용자와 인터페이스하는 기능과 종속 노드와 자료를 주고받는 기능을 하며 주 노드는 종속 노드의 역할도 수행한다. 주 노드는 질의가 들어오면 질의를 구성하는 단어들을 분할 정보에 따라 해당 단어를 처리할 종속 노드를 결정하여 해당 단어를 그 종속 노드로 보낸다. 그런 다음, 종속 노드가 보내준 색인어 역파일을 모아서 AND 또는 OR 연산을 수행하고 결과를 순위화한다. 종속 노드는 사용자와의 인터페이스를 전혀 고려하지 않으며, 주 노드로부터 작업지시를 받아서 질의를 처리한다. 종속 노드는 각각이 자신의 데이터베이스를 가지고 있으면서, 주 노드로부터 질의어 리스트를 받아들여 해당 색인어에 관련된 색인어 역파일을 하드디스크로부터 가져와서 주 노드로 보낸다.

2.1 네트워크 구조

이 논문에서 제안하는 병렬 정보 검색 시스템은, SCI 기반의 8개의 노드로 이루어진 PC 클러스터 [4] 환경에서 구현되었다.

2.2 질의 확장

정보 검색에서는 단어 대 단어의 유사도가 유용한데, 질의 확장에서도 검색 효율을 향상시키기 위해 다양한 유사도 계산 방법이 제안되어져 왔다 [5]. 이 논문에서는, 질의 확장을 위해 추가할 단어를 선택하기 위해 코사인 유사도를 사용한다. 코사인 유사도란 동시등장 기반 유사도의 일종으로 임의의 두 단어에 대해 각 단어의 빈도수와 동시등장 빈도수의 함수로 계산된다 [8].

질의 확장 시스템은 아래와 같은 순서로 수행된다.

(1) 질의가 입력되면 미리 구축된 역파일을 사용하여 결과 문서들을 정렬하여 1차 검색을 수행한다.

(2) 질의에 추가할 확장 대상 용어를 선정한다; 확장 대상 용어 선정을 위해서는 여러 개의 용어로 구성되어 있는 질의와 확장하려는 하나의 용어간의 유사도 측정이 필요하다. 질의는 여러 개의 용어로 구성되므로, 각각의 질의 용어 벡터를 합한 질의 벡터를 정의하여, 질의와 확장하려는 하나의 용어간의 유사도는 이 질의 벡터와 확장 대상 용어 벡터와의 유사도를 측정하여 구한다 [12]. 이렇게 하여 다른 많은 단어들과 가장 높은 유사도 값을 가지는 용어를 확장 대상 용어로 선정한다.

(3) 질의 확장 대상 용어들과 1차 검색된 문서 내에서 같이 나타난 모든 용어들에 대해 질의-용어 유사도를 구하고 내림차순으로 정렬하여 질의 확장을 위한 확장용어 후보 목록을 작성한다.

(4) 유사도 값에 의해 역정렬된 확장 용어 후보들을 목록에 있는 순서대로 추가하여 질의를 확장한다.

2.3 두 개의 IIF 분할 방법

이전 연구에서 [4], 순차 IR 시스템의 수행 시간을 분석하여 IR 시스템 수행에 있어 가장 시간이 많이 드는 부분이 디스크 접근 시간 (전체 질의 수행 시간의 약 45%) 임을 알게 되어, 디스크 접근을 병렬화하는 방법으로 IIF을 분할하여 각 노드에 분배, 저장하고자 IIF을 분할하는 두 가지 방법을 제안하였다.

분할을 할 때, 디스크 입출력과 그 이후의 일련의 순위화 과정이 가능한 모든 노드들에서 균등하게 수행되도록 IIF를 분할할 수 있으면 최대로 병렬 시스템을 사용하는 효과를 얻을 수 있다. 이러한 목표를 얻고자 여기서는 두 가지 분할 방법을 사용한다. 첫째 방법은, lexical 분할 방법이라 부르는데 이것은 단순히 IIF 에 있는 각 단어에 대한 정보를 단어별로 사전순으로 각 노드에 나누어주는데, 모든 단어에 대한 정보가 다 나누어질 때까지 되풀이하여 나누어준다. 특별한 분할 알고리즘을 전혀 적용시키지 않고 색인어들을 분산하는 방법으로써 greedy 분할 방법의 비교 대상으로 사용된다. 두 번째 방법은, greedy 분할 방법이라고 부르는데, 이 방법은 3절의 실험 결과에서 기술된 대로, lexical 분할 방법보다 훨씬 성능이 좋다. 이 방법은 같은 질의에 같이 나타날 확률이 낮은 단어들을 동일한 노드에 할당하려고 노력한다 [4]. 이 방법은 사용자의 질의를 효율적으로 병렬처리 하기 위해 색인어를 분산하는 방법으로 두 색인어의 동시 등장 가중치를 사용하고 있다. 동시 등장 가중치는 두 색인어가 동시에 등장한 문서의 수이다. Greedy 분할 방법은 서로 동시에 나타날 확률이 낮은 색인어, 즉 동시 등장 가중치가 낮은 색인어끼리 하나의 노드에 모이도록 하는 방법이다.

3 실험과 평가

실험에서는 두 개의 표준 한국어 테스트 말뭉치를 사용한다. 하나는 최초의 한국어 테스트 말뭉치인 KTSET [10] 이고 다른 하나는 가장 큰 한국어 테스트 말뭉치인 EKSET [11] 이다.

사용자가 준 원래의 질의를 구성하는 각 단어와 한번이라도 어떤 문서에서 동시 등장한 적이 있는 모든 단어에 대하여 질의와 각 단어간의 유사도를 계산하여 단어들을 유사도 값에 따라 내림차순으로 정렬한다. 실험에서는 확장된 각 질의가 24개의 단어를 가지도록 원래의 질의에 위의 정렬된 단어들 중에 높은 순위를 가지는 단어를 추가한다.

이전 연구에서는 [4], 500,000 건의 한국어 신문으로 구성된 말뭉치와 500개의 질의를 사용하여 lexical 분할 방법과 greedy 분할 방법의 성능을 비교하였는데, 이 때 각 질의는 24개의 단어를 가지도록 다음과 같이 인공적으로 만들어졌었다. 즉, 질의를 만드는 첫 단계는 문서 집합에서 500개의 문서를 샘플링한다. 샘플된 각 문서로부터 가장 중요하다고 여겨지는 24개의 단어를 선택한다. 단어의 문서에서의 중요도는 단어의 빈도에 비례하고 문서 빈도에 반비례하다고 가정한다. 이러한 방식으로 인공적으로 만든 질의는 구성하는 단어들간에 연관성이 약하여 실험의 결과가 greedy 분할 방법이 lexical 분할 방법에 비하여 단지 3.7% 정도의 성능 향상을 보였다. 또한, 질의 확장을 사용하지 않는 이러한 체계에서는 웹 상에서 일반적으로 사용되는 4개 이하의 짧은 질의로 병렬 시스템을 사용하는 이점을 살릴 수 없다. 아래의 실험에서, 코사인 유사도를 사용하는 질의 확장을 적용한 병렬 정보 검색은 greedy 분할 방법과 연계하여 검색 효율을 상당히 향상시키고 또한 짧은 질의로도 병렬 시스템의 장점을 잘 이용한다는 것을 보일 것이다. 또한 이전 연구의 실험은 표준 말뭉치를 사용하지 않아 정확도를 계산하기 어려우나 이 논문에서 제안한 질의 확장 방법은 평균 정확도를 KTSET에서는 대략 10%, EKSET에서는 대략 20% 향상한다 [8] 고 알려져 있다.

3.1 EKSET에서의 실험

이 절에서는 표준 한국어 테스트 말뭉치의 하나인 EKSET을 사용하여 greedy 분할 방법과 lexical 분할 방법의 성능을 비교한다. EKSET의 문서의 개수는 23113 개이고 질의의 개수는 46개이다. 각 질의는 평균 2.15개의 단어로 구성된다. 이 실험에서는 앞에서 설명한 방법으로 각 질의를 확장하여 테스트 질의는 24개의 단어로 구성되도록 하고 8개의 노드로 이루어진 PC 클러스트를 사용한다. EKSET에서 greedy 분할 방법과 lexical 분할 방법을 비교한 실험 결과는 Table 1에 나와 있다.

Table 1. EKSET에서의 실험 결과 (단위: 초)

	Lexical 분할 방법	Greedy 분할 방법	향상 비율 (%)
평균 질의 처리 시간	0.79	0.65	17.7
디스크 접근 시간과 로칼 IR 연산 시간 합의 평균값	0.535	0.43	18.3

Table 1을 보면, greedy 분할 방법이 lexical 방법에 비해 전체 질의 처리 시간에서 대략 17.7% 의 성능 향상을 보인다. 전체 질의 처리 시간 중 디스크 접근 시간과 로칼 IR 연산 시간에서만 두 방법이 차이가 나므로 이 시간들만을 따로 측정하였다. 이 측정을 나타내는 Table 1의 둘째 열을 보면 greedy 방법이 대략 18.3%의 향상을 보이는데 이는 이전 연구 [4]의 결과 (5.7% 향상)과 비교하면 검색 효율이 상당히 개선되었다는 것을 알 수 있다. 이것으로, greedy 분할 방법과 연계한 코사인 유사도를 사용한 질의 확장 기법이 병렬 시스템의 병렬성을 잘 활용함을 알 수 있다.

3.2 KTSET에서의 실험

이 절에서는 표준 한국어 테스트 말뭉치의 하나인 KTSET을 사용하여 greedy 분할 방법과 lexical 분할 방법의 성능을 비교한다.

Table 2. KTSET에서의 실험 결과 (단위: 초)

	Lexical 분할 방법	Greedy 분할 방법	향상 비율 (%)
평균 질의 처리 시간	0.0344	0.0283	17.6
디스크 접근 시간과 로칼 IR 연산 시간 합의 평균값	0.0233	0.0187	19.5

KTSET의 문서의 개수는 1000 개이고 질의의 개수는 30개이다. 각 질의는 평균 3.2개의 단의로 구성된다. 실험 방법은 3.1절에 기술된 것과 같다. KTSET에서 greedy 분할 방법과 lexical 분할 방법을 비교한 실험 결과는 Table 2에 나와 있다

Table 2를 보면, greedy 분할 방법이 lexical 방법에 비해 전체 질의 처리 시간에서 대략 17.6% 의 성능 향상을 보인다. 디스크 접근 시간과 로칼 IR 연산 시간만을 고려해 보면 greedy 방법이 대략 19.5%의 향상을 보인다.

4. 결론

이 논문에서는, 고속 네트워크 카드로 연결된 PC 클러스터 환경에서 구현된 병렬 정보 검색 시스템에 질의 확장을 적용한 효과를 lexical 분할 방법과 greedy 분할 방법을 사용하여 살펴보았다. 두 개의 한국어 표준 테스트 집합을 사용하여 실험하였다. 병렬 정보 검색 시스템에 질의 확장을 적용하여 다음과 같은 3 가지 효과를 얻었다. 첫째, 3절의 실험 결과에서 보듯이 이전 연구 결과에 비하여 상당한 검색 효율의 향상을 얻었다. 이는 코사인 유사도를 사용하는 질의 확장 방법이 greedy 분할 방법과 협력하여 병렬 시스템의 병렬성을 잘 이용한다는 것을 의미한다. 둘째, 웹 상에서 일반적으로 사용되는 짧은 질의도 여러 노드로 이루어진 병렬 시스템의 병렬성을 이용할 수 있다. 셋째, 코사인 유사도를 사용하는 질의 확장 방법을 사용함으로써 KTSET에서 대략 10%, EKSET에서 대략 20%의 평균 정확률 향상을 얻을 수 있다 [8].

References

1. Park, S.H., Kwon, H.C.: An Improved Relevance Feedback for Korean Information Retrieval System.Proceedings of the 16th IASTED International Conference on Applied Informatics, IASTED/ACTA Press, Garmisch-Partenkirchen, Germany (1998) 65-68
2. Frakes, W., Baeza-Yates, R.: Information retrieval data structures & algorithms. Prentice- Hall (1992)
3. Cormack, G.V., Clarke, C.L.A., Palmer, C.R., Kisman, D.I.E.: Fast Automatic Passage Ranking (MultiText Experiment for TREC-8). The proceedings of the Eighth Text Retrieval Conference (TREC-8), Gaithersburg, Maryland (1999) 735-741
4. Chung, Y.J., Kwon H.C., Chung, S.H., Ryu, K.R.: Declustering Web Content Indices for Parallel Information Retrieval, Lecture Notes in Artificial Intelligence 2109 (2001) 346-350
5. Xu, J., Croft, W.B.: Query Expansion Using Local and Global Document Analysis, The proceedings of the 19thACM SIGIR International Conference on Research and Development in Information Retrieval, Zurich (1996) 4-11
6. Salton, G.: Automatic Text Processing. Addison-Wesley Publishing Company (1989) 313-319
7. Rijsbergen, C.J.V.: A Theoretical Basis for the Use of Cooccurrence Data in Information Retrieval, Journal of Documentation 33, 106-119
8. Kim, M.C., Choi, K.S.: A Comparison of Collocation-based Similarity Measures in Query Expansion, Information Processing and Management 35 (1999) 19-30
9. Qiu, Y, Frei, H.P.: Concept Based Query Expansion, The Proceedings of the 16th ACM SIGIR International Conference on Research and Development in Information Retrieval, Pittsburgh (1993) 160-169
10. Kim, S.H., Seo, E.K., Lee, W.K., Kim, M.C., Kim, Y.H., Kim, J.K.: The Development of Test Collection for Automatic Indexer, Journal of the Korean Society for Information Management, 11(1), 81-102
11. Kang, H.K. Choi, K.S.: Two level Document Ranking Using Mutual Information in Natural Language Information Retrieval, Information Processing and Management, 33(3). 289-306
12. 김명철, "공기기반 용어간 유사도를 이용한 정보검색 질의확장 비교연구", 한국과학기술원 전산학과, 박사학위 논문, 1999