분산검색에서 부분정보를 이용한 컬렉션 선택 방법

이현숙[0] 맹성현 이만호
충남대학교 컴퓨터과학과
(hslee[0], myaeng, mhlee)@cs.cnu.ac.kr

A Method for Collection Selection using Incomplete Information in a distributed retrieval system

Hyun-Sook Lee[0] Sung Hyon Myaeng Mann-Ho Lee
Dept. of Computer Science, Chungnam National University

요 약

본 논문은 여러 컬렉션에 대해 검색을 수행하는 분산 검색시스템에서 질의어가 들어 왔을 때 질의어에 적합한 컬렉션을 자동으로 선택할 수 있도록 하는 컬렉션 선택 모델과 브로커 구조를 제안하였다. 각 컬렉션마다 과거 질의에 대해 검색된 결과 문서들을 색인하여 인접단어를 고려한 불완전 인덱스를 생성한다. 이러한 불완전 인덱스를 이용하여 컬렉션 선택하는 모델을 TREC 문서집합과 SMART 시스템을 이용하여 구현하였다.

1. 서 론

인터넷과 네트워크 서비스의 급속한 발전으로 인해 정보가 기하급수적으로 증가하고 있으며 이와 더불어 개별적인 컬렉션들이 증가하고 있다. 따라서 분산되어 있고 각각 다른 정보들을 저장하고 있는 여러 컬렉션들을 검색하고자 하는 분산검색 시스템의 필요성이 점차적으로 증가하고 있다.

분산검색 시스템은 일반적으로 사용자의 질의에 대해 모든 컬렉션에 검색을 요구하고 각 컬렉션이 개별적으로 검색한 결과를 받아서 통합한다. 이러한 방식은 특정 컬렉션이 질의어에 대해 유용한 문서를 포함하고 있지 않더라도 검색을 요구하게 되므로 불필요한 네트워크 트래픽을 일으키며 지역 자원의 낭비 및 사용자의 검색 시간 관점에서 많은 비용이 들게 된다. 그러므로 효율적인 분산검색을 위해 사용자들은 질의와 유사한 문서를 포함할 것으로 추정되는 컬렉션들을 선택함으로써 검색 대상을 제한해야 한다. 숙련된 사용자들은 과거의 경험이나 도움이 될 만한 참고자료를 이용하여 컬렉션을 선택할 수 있다. 그러나 많은 초보 사용자들은 어떤 컬렉션이 질의에 적합한지 알 수 없으므로 모든 컬렉션을 선택할 것이다. 그러므로 많은 컬렉션에 대해 검색을 수행할 경우에는 분산검색 시스템에서 자동적으로 질의어에 적합한 컬렉션을 결정할 수 있는 기술을 제공해야 한다. 컬렉션 선택에 관한 연구는 인덱스를 이용하는 방법, 확률모델을 이용하는 방법, 클러스터링을 이용하는 방법 등 다양하게 진행되어 왔다. 그러나 이러한 기존의 연구들은 컬렉션에서 특정 정보를 제공할 경우에 적용 가능하다는 한계점이 있다. 본 논문에서는 각 컬렉션마다 개별적인 검색 시스템이 지원된다는 가정 하에, 어느 분산검색 시스템에서나 사용 가능한, 적응력 높은 컬렉션 선택 전략을 제안하고 이러한 방법을 이용함으로써 얼마나 효율적인 검색이 이루어지는지 실험을 통해 검증하고자 한다.

2. 관련연구

분산 환경에서 컬렉션 선택 문제에 대한 접근 방법 중 가장 보편적인 방법은 컬렉션 인덱스를 이용하는 방법이다. 이 접근법은 각 컬렉션을 컬렉션 안에서 출현하는 단어의 리스트와 그에 해당하는 빈도수로 표현한다. 이것을 이용하여 색인을 하고 그 결과로 나온 데이터 구조를 컬렉션 인덱스라고 한다. 분산검색 시스템에서는 질의어가 들어오면 일반적인 문서 검색에서 인덱스를 이용하여 문서를 랭킹하는 것과 같은 방식으로 컬렉션 인덱스를 이용하여 컬렉션을 랭킹한다 [1].

그러나 분산검색 시스템에서 각 컬렉션에 해당하는 인덱스를 유지하는 것은 분산검색에 참여하는 컬렉션의 개수가 계속적으로 증가할 때 저장 공간의 문제를 일으킬 수 있다. 이에 따라 컬렉션들의 부분적인 정보들을 이용하여 컬렉션 인덱스를 구성하는 연구들이 많이 진행되어 왔다. 대표적인 연구로는 직접 색인 작업을 하지 않고 각 컬렉션으로부터 부분적인 통계정보만을 수집하여 컬렉션 인덱스를 생성하는 gGlOSS (a generalized Glossary-Of-Servers Server) 시스템이 있다. 그러나 컬렉션 인덱스를 이용하는 방법의 가장 큰 한계점은 분산검색에 참여하는 컬렉션에서 원문이나 통계정보를 제공하지 않으면 적용할 수 없다는 것이다 [1][2][4].

컬렉션 선택을 위한 다른 접근법으로 확률모델을 이용하는 방법이 있다. 확률모델을 이용한 컬렉션 선택은 컬렉션과 질의의 적합성 확률에 따라 각 컬렉션에서 몇 개의 문서를 검색할 것인지 결정함으로써 이루어진다. 이 때, 적합성 확률은 분산검색 시스템에서 각 컬렉션으로부터 랜덤하게 뽑아 낸 문서들을 대상으로 계산하여 컬렉션의 검색 길이를 추정한다 [3]. 적합성 확률을 구하기 위해서는 적합 문서와 부적합 문서를 미리 결정해야 하므로 사용자의 적합성 피드백을 요구한다. 그러나 분산검색에 참여하는 모든 컬렉션에 대해 적합성 피드백을 하는 것은 거의 불가능해 보이므로 자동적으로 구할 수 있는 방법이 필요하다.

클러스터링을 이용한 컬렉션 선택 방법은 컬렉션을 문서 검색과 같은 방식으로 랭킹하는 이전의 연구와는 달리 문서 클러스터링을 통해 컬렉션을 랭킹한다. 이 연구에서는 연관된 문서

들은 같은 질의에 대해 적절한 문서라고 가정하고, 클러스터링을 통해 생성된 토픽에 따라 문서를 그룹화한다. 질의가 들어오면 가장 적절한 토픽을 고르고 그 토픽을 포함하는 컬렉션에 대해 검색을 요청한다. 적절한 토픽의 선정은 토픽에 있는 단어 정보를 이용하여 질의를 가장 잘 생성할 수 있는 토픽을 결정함으로써 이루어진다 [4]. 클러스터링을 이용하는 방법은 컬렉션 선택에서 다른 연구들과 비교할 때 가장 좋은 성능을 보인다. 그러나 클러스터링을 위해서는 각 컬렉션의 문서들에 직접 접근할 수 있어야 하고, 각 컬렉션 안에 클러스터된 형태를 유지할 수 있어야 하므로 실제 분산검색에 적용될 가능성이 희박하다.

이 외에 학습질의를 이용하는 방법은 새로운 질의와 학습 질의의 유사도를 이용하여 각 컬렉션에서 검색할 문서의 개수를 추정한다 [5]. 이 연구는 다른 연구와 달리 컬렉션으로부터 정보를 요구하지 않고 학습 질의와 결과로 반환된 문서의 개수만 고려하여 컬렉션 선택을 하는 방법으로 TREC의 토픽을 질의어로 사용하여 실험하였다. 그러나 TREC의 토픽은 여러 단어들로 구성된 데 반해 인터넷 환경에서 대부분의 사용자들은 두, 세 개의 단어를 이용하여 질의를 하므로[6] 질의어 간의 유사도를 구하는 데 적절하지 않을 수 있다.

3. 컬렉션 선택을 위한 접근법

질의어와 가장 적합한 문서가 있는 컬렉션을 선택하기 위해서는 클러스터링을 이용하는 방법이 가장 정확한 것으로[3] 관련 연구에서 밝혀졌지만, 실제 환경에서는 문서에 직접 접근해야 한다는 한계 때문에 적용하기 어려우므로 질의어를 포함한 문서가 있는지 여부에 따라 컬렉션을 선택한다. 이를 위해서는 컬렉션에 출현하는 각 단어마다 용어 빈도수나 문헌 빈도수와 같은 통계정보가 요구된다. 그러나 통계정보를 제공하지 않는 컬렉션이 대부분이므로 본 논문에서는 모든 컬렉션이 통계정보를 제공하지 않는다고 가정한다. 대신 컬렉션 선택을 하기 위해 분산검색 시스템에서 질의어 집합을 이용하여 컬렉션들의 인덱스를 구성한다.

컬렉션 인덱스를 구성하기 위해 본 논문에서 제안한 방법은 과거 질의를 이용해 검색된 문서만을 대상으로 색인하여 해당 컬렉션의 내용을 표현하도록 하는 것이다. 이렇게 구성된 인덱스는 컬렉션의 모든 컨텐츠에 대한 완전한 인덱스는 아니지만 컬렉션의 컨텐츠를 반영하는 정보이다. 또한 불완전한 인덱스이기 때문에 기존의 컬렉션 인덱스와는 다른 구조를 가지며 컬렉션 선택 또한 독창적인 알고리즘을 이용한다.

3.1 컬렉션 선택 과정

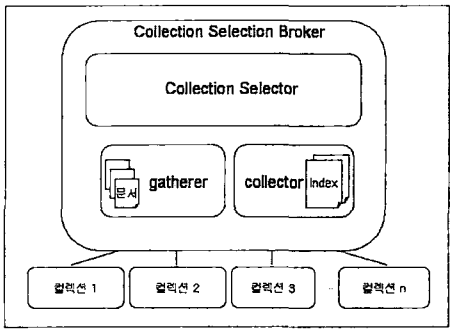
본 논문에서 제안한 불완전 인덱스는 과거 질의와 검색 결과 문서를 이용하여 구성한다. 불완전 인덱스를 구성하고 컬렉션 선택을 하는 과정을 개략적으로 살펴 보면 다음과 같다.

처음 단계에서는 과거 질의를 이용하여 컬렉션 i로부터 검색 결과를 수집한다. 불완전 인덱스를 구성하기 위해 각 질의마다 검색 결과 중 상위 m개의 문서를 선택하여 컬렉션 i의 샘플 문서집합을 구성한다.

두 번째 단계에서는 각 컬렉션마다 수집된 샘플 문서집합을 색인하여 인덱스를 구성한다. 불완전 인덱스는 완전 인덱스에 비해 매우 작은 인덱스를 구성하므로 인덱스에 없는 단어가 질의어에 포함되어 있을 경우에도 처리할 수 있도록 하는 다른 구조가 요구된다. 이를 위해 인덱스의 각 단어마다 가중치 외에 인접단어 리스트를 보유하도록 한다. 여기에서 인접단어란 특정 단어와 관계된 단어들을 의미한다. 본 논문에서는 특정 단어와 관계된 단어는 자주 같이 출현한다는 가정 하에 동시에 출현하는 빈도가 높은 단어들을 인접단어로 결정한다. 컬렉션 선택을 할 때 자신의 컬렉션 인덱스에서 질의어에 해당하는 단어를 찾지 못할 경우, 인접단어 리스트에 해당하는 단어들을 이용하여 컬렉션 선택을 하기 때문에 본 논문에서 인접단어의 역할은 매우 중요하다.

마지막 단계는 전 단계에서 구성한 불완전 인덱스를 기반으로 컬렉션을 선택한다. 인덱스에 질의어가 있는 컬렉션은 질의어를 포함하는 문서를 가지고 있기 때문에 선택 컬렉션 리스트에 추가한다. 없을 경우에는 질의어를 포함한 문서가 컬렉션에 존재하는지 여부를 알 수 없으므로 인접단어 정보를 이용한다. 즉, 인접단어 리스트 중 여러 인접단어들이 컬렉션 안에 나타나면 질의어도 그 컬렉션에 나타날 가능성이 클 것이라고 가정하고 그 가능성이 일정 임계값 이상일 경우에 컬렉션을 선택한다.

본 논문에서 제안한 컬렉션 선택을 위한 브로커 구조는 [그림 1]과 같다. 각 단계별로 첫 번째 단계에서는 수집기가, 두 번째 단계에서는 컬렉션 색인기가, 세 번째 단계에서는 컬렉션 선택기가 작업을 담당한다.



[그림 1] 컬렉션 선택을 위한 브로커 구조

3.2 컬렉션 선택 방법

컬렉션 선택을 위해서는 새로운 질의어 $Q=(q_1,q_2,...,q_n)$에 대해 컬렉션 DB_i에서 Q에 적합한 문서를 가질 확률 $P(DB_i|Q)$를 구한다. $P(DB_i|Q)$는 다음과 같이 구한다.

$$P(DB_i,Q)=\prod_{\{S_{ik}\neq 0 \& P(I_i,q_k)\neq 0\}}^{n_{k=0}} P(I_i,q_k)\times S_{ik}$$

$P(I_i,q_k)$는 I_i에서 q_k가 나올 확률 즉, 컬렉션 i에 q_k를 포함하는 문서가 하나라도 존재할 확률을 의미한다. S_{ik}는 q_k를 포함한 문서가 컬렉션 i에서 나오는 정도를 나타낸다. $P(I_i,q_k)$와 S_{ik}를 구하기 위해서는 다음과 같은 단계를 거친다.

1) 컬렉션 i의 인덱스 I_i에 q_k가 있는지 검사한다.

- 있을 경우 : $P(I_i,q_k)=1$, $S_{ik}=\dfrac{N_k(I_i)}{N(I_i)}$

($N(I_i)$: I_i에 있는 문서의 개수,

$N_k(I_i)$: I_i에서 q_k를 포함하는 문서의 개수)

2) I_i를 제외한 다른 컬렉션들의 인덱스에서 q_k의 인접단어에 해당하는 단어들을 찾는다.

- 인접단어 이용 : $P(I_i, q_k) = \frac{n}{l}$, $S_{ik} = \frac{1}{n} \times \sum_{j=1}^{n} \frac{N_j(I_i)}{N(I_i)}$

(l: 인접단어의 개수, n: I_i에 나온 인접단어의 개수,
$N(I_i)$: I_i에 있는 문서의 개수,
$N_k(I_i)$: I_i에서 인접단어 w_j를 포함하는 문서의 개수)

위와 같이 구한 $P(DB_i \mid Q)$가 임계값 h이상이면 질의 Q에 적절한 문서가 컬렉션 i에 나올 확률이 높으므로 컬렉션 i를 선택한다.

4. 실험 및 평가

본 논문에서의 실험은 TREC disk 1 문서집합에 대해 수행하였다. TREC disk 1의 컬렉션들은 7개의 컬렉션으로 나눌 수 있으며, 각 컬렉션마다 문서의 출처와 기간이 다양하다. 또한 컬렉션마다 포함하고 있는 문서의 개수와 각 문서의 길이가 다양한 이질적인 컬렉션들로 구성된다. 본 실험에서는 disk 1의 각 컬렉션마다 임의의 문서들을 선택하여 컬렉션 당 약 50Mbyte씩 7개 컬렉션을 생성하였다. 실험에 사용된 컬렉션에 대한 통계정보는 [표1]과 같다.

컬렉션	WSJ 87	WSJ 88	WSJ 89	AP	ZIFF	FR	DOE
크기 (Mbyte)	50	50.2	36.5	50.3	50.2	50	50.2
문서 갯수	19,767	19,927	12,390	16,913	16,062	4,760	63,330
유일단어의 총 개수	103,607	108,657	94,538	111,356	100,354	84,815	127,482

[표 1] 실험에 사용된 컬렉션 정보

이렇게 구성된 각 컬렉션의 문서들은 미국의 Cornell 대학에서 개발한 SMART 시스템을 이용하여 색인하였다 [7][8]. 각 컬렉션에 대해 불완전 인덱스를 생성하기 위해 사용한 질의어 집합은 TREC topics 51-100이고 모델 성능의 평가에는 TREC topics 101-150을 이용하였다.

[표 2]는 컬렉션 선택을 하지 않은 경우와 본 논문에서 제안한 방법으로 선택한 경우를 비교한 결과이다. 컬렉션 선택을 하지 않은 경우는 모든 컬렉션을 선택하였음을 의미한다. [표 2]를 보면 컬렉션 선택을 한 경우, 잘못 선택한 컬렉션이 현저히 감소하여 정확도가 증가함을 볼 수 있다.

	제대로 선택한 컬렉션(%)	잘못 선택한 컬렉션(%)	선택하지 못한 컬렉션(%)
컬렉션 선택	93	5	2
선택 안함	68	32	0

[표 2] 컬렉션 선택을 한 경우와 선택하지 않은 경우 결과

[표 3]은 불완전 인덱스에서 인접단어를 고려한 경우와 고려하지 않은 경우를 비교한 결과이다. 인접 단어를 고려하지 않으면 선택하지 못한 컬렉션과 잘못 선택한 컬렉션이 증가하는 것으로 보아 불완전 인덱스에서 인접단어를 고려하는 것이 효과적임을 알 수 있다.

	제대로 선택한 컬렉션(%)	잘못 선택한 컬렉션(%)	선택하지 못한 컬렉션(%)
인접단어 고려	93	5	2
고려 안함	83	11	6

[표 3] 인접단어를 고려한 경우와 고려하지 않은 경우 비교

5. 결론

본 논문에서는 각 컬렉션마다 컬렉션의 문서 전체가 아닌 과거 질의에 대해 검색된 결과 문서들만을 색인하여 불완전 인덱스를 생성하고, 이것을 이용하여 컬렉션 선택을 하는 방법을 제안하고 TREC 문서집합과 SMART 시스템을 이용하여 구현하였다. 실험을 통해 컬렉션 선택을 한 경우가 컬렉션 선택을 하지 않은 경우보다 더 좋은 성능을 보이며, 인접단어를 고려한 경우의 컬렉션 선택이 인접단어를 고려하지 않은 경우보다 정확도와 재현율에서 더 나은 성능을 보임을 알 수 있었다. 즉, 불완전 인덱스에서 인접단어가 컬렉션 선택에 기여함을 확인할 수 있었다.

본 논문에서 제안한 컬렉션 선택 모델은 계층적인 구조에서도 하나의 분산 검색시스템에서 수행한 것과 같은 컬렉션 선택 방법을 적용할 수 있으므로 향후 계층적인 분산검색 시스템에 이용할 수 있다. 본 논문의 컬렉션 선택 방법이 더 좋은 성능을 낼 수 있도록 하기 위해서는 인덱스의 가장 적절한 크기를 조정해야 한다. 또한 인덱스가 커지면 색인 단어 당 인접단어의 개수가 매우 증가할 것이므로 인접단어의 빈도수를 고려한 연구 또한 이루어져야 할 것이다.

참고문헌

[1] L. Gravano and H. Garcia-Molina. Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies. In Proc. 21th VLDB Conf., 1995

[2] J. P. Callan, L. Zhihong, and W. B. Croft. Searching Distributed Collections With Inference Networks. In Proc. 18th ACM SIGIR Conf., 1995

[3] C. Baumgarten. A Probabilistic Solution to the Selection and Fusion Problem in Distributed Information Retrieval. In Proc. 22th ACM SIGIR Conf., 1999

[4] Jinxi Xu and W. B. Croft. Cluster-based Language Models For Distributed Retrieval. In Proc. 22th ACM SIGIR Conf., 1999

[5] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird. Learning Collection Fusion Strategies. In Proc. 18th ACM SIGIR Conf., 1995

[6] Weiyi Meng, King-Lup Liu, Clement Yu, Xiaodong Wang, Yuhsi Chang and Napthtali Rishe. Determining Text Databases to Search in the Internet. In Proc. 24th VLDB Conf., 1998

[7] TREC(Text REtrieval Conference) http://trec.nist.gov/

[8] SMART
" http://pi0959.kub.nl:2080/Paai/Onderw/Smart/smart.html"