문서 클러스터링 정보를 이용한 컬렉션 융합*

○
금기문, 남세진, 신동욱, 김태균
충남대학교 컴퓨터공학과

Collection Fusion using Document Clustering

○
K.M. Keum, S.J. Nam, D.W. Shin, T.K. Kim
Dept of Computer Engineering, Chungnam National University

요 약

본 논문에서는 여러 정보검색 엔진들이 분산되어 있는 환경에서 이 엔진들의 검색 결과를 효과적으로 취합하여 사용자에게 제시하는 컬렉션 융합 방안을 제안하고자 한다 이 방법은 우선 학습 질의어로 검색된 문서들의 클러스터링 정도를 이용하여 컬렉션에의 신뢰도를 측정하고 새로운 질의어가 입력되었을 때 각 컬렉션에서 검색된 문서의 유사도를 조정하여 융합하는 방법이다. 여기에서 각 컬렉션의 신뢰도는 미리 준비된 학습 질의어와 이 학습 질의어를 입력하여 검색된 문서들 사이의 유사도를 분석하여 측정한다. 이 신뢰도는 새로운 질의어가 입력되었을 때 각 컬렉션마다 문서들을 검색하고 이들 문서들을 어느 정도 신뢰할 것인가를 결정하는데 사용된다. 본 논문에서 제안한 방법은 학습과정에서 사람이 학습시킬 필요가 없는 비지도 학습에 기초하고 있다. 따라서 지금까지 지도 학습에 기초한 컬렉션 융합 방법과는 달리 인터넷과 같이 문서들이 동적으로 변하는 환경에서 쉽게 사용할 수 있다는 장점을 가진다.

1. 서론

인터넷과 정보 검색 시스템의 발전으로 다양한 컬렉션을 갖는 분산 정보 검색 시스템에서 각 컬렉션의 검색 결과를 하나의 결과로 취합하는 것은 정보 검색 연구 분야의 새로운 문제로 부각되었다. 이러한 문제를 해결하기 위해 여러 개의 독립적인 컬렉션의 검색 결과를 하나의 컬렉션에서 전체 문서를 검색한 것과 유사한 결과로 만들기 위해 검색 결과를 조합하는 것을 컬렉션 융합이라 한다. [1,2].

지금까지 컬렉션 융합에 관한 연구는 각 컬렉션에 관한 정보를 분산 환경에서 이용하는 방법과 컬렉션에 관한 정보를 이용하지 않는 방법이 있다. 전자의 경우에는 색인, IDF(Inverse Document Frequency)값 등의 컬렉션 정보를 중앙의 한 서버에서 관리하는 방법이 있다[3]. 그런데 이러한 중앙 집중식 방법은 중앙의 서버에 병목형상으로 전체 시스템의 성능 저하를 초래하는 문제점을 가진다. 또 다른 방법으로 분산 환경의 모든 서버들이 서로의 컬렉션 정보를 공유하는 방법[4]이 있는데 이 방법은 각 서버가 자신의 컬렉션에 관한 정보를 다른 서버들에게 전달하는 것으로 모든 정보를 보낸다면 정보 전송에 많은 비용이 필요하고 또 컬렉션이 동적으로 변하는 경우 이를 수행하기 어렵다.

컬렉션 정보를 이용하지 않는 방법은 학습 질의어를 이용하여 컬렉션의 특성을 파악하고 새로운 질의어에 대해 컬렉션의 관련도를 판단하는 방안이다 이 방법에서 지금까지 연구된 것은 Voorhees[1,2] 연구와 같이 주로 학습자가 질의 결과를 분석하여 컬렉션의 특성을 파악하는 것이다. 이는 지도 학습에 의한 융합 방법으로 인터넷에서와 같이 코퍼스가 동적인 경우 전문가가 학습을 반복하여 시켜야 하므로 적용하기 어렵다.

본 연구에서 제안하는 컬렉션 융합은 비지도 학습을 기초로 하는 방법으로 학습 질의어와 검색된

* 본 연구는 과학재단의 핵심전문 연구와 소프트웨어 연구센터의 기초과제로 수행되었다

문서의 유사도를 분석하는 것이다. 즉, 먼저 학습 질의어로 각 컬렉션을 검색하여 얻은 문서들과 학습 질의어와 어느 정도 클리스터링이 잘 되어 있는지를 계산하여 컬렉션의 신뢰도를 측정한다. 사용자 질의어가 입력되면 학습 질의어 중 관련된 몇 개를 이용하여 컬렉션들의 신뢰도를 계산하는데 이 신뢰도는 각 컬렉션에서 검색한 문서들을 최종적으로 랭킹(ranking)하는데 사용된다. 이 방법은 기본적으로 어떤 컬렉션에 대하여 질의어를 입력하였을 때 검색된 문서들이 실제 질의어와 잘 클러스터링되어 있는 컬렉션은 신뢰할만 하다는 가정에 기초하고 있다 또 이 방법은 학습하는데 사람이 개입할 필요가 없는 비지도 학습이므로 인터넷과 같이 검색 시스템의 컬렉션이 동적으로 변하는 경우에도 적용할 수 있는 장점이 있다 본 연구에서는 TREC컬렉션을 이용하여 제안된 방법의 검색 효율을 평가하였는데 실험결과 학습시키지 않은 기법보다 매우 우수한 효율을 보였다.

2장에서는 본 연구에서 제안한 비지도 학습에 의한 컬렉션 융합 방법을 설명하고 3장 본 연구의 타당성을 검토하며, 마지막으로 결론 및 향후 연구 과제로 끝을 맺는다

2. 비지도 학습에 의한 컬렉션 융합

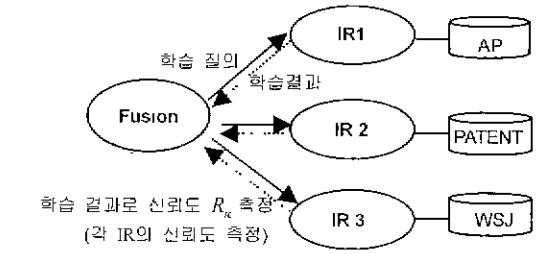
본 연구에서는 컬렉션의 신뢰도를 측정하는 방법으로 검색된 문서들과 학습 질의어 사이의 클러스터링 정도를 이용하는 방법을 제안한다. 즉 임의의 서버의 주제와 검색 정확도는 검색한 문서의 주제와 문서의 정확도를 가지고 판단할 수 있다. 이러한 판단 기준을 얻기 위해 벡터 공간 모델에서 학습 질의어에 대해 검색된 문서의 유사도를 이용한다. 임의의 학습 질의어에 대해 유사도가 우수한 검색 결과를 보이는 서버는 어느 정도 정확한 검색 결과를 보인다고 가정할 수 있다. 만약 사용자의 질의어가 이 학습 질의어와 비슷하다면 이 학습 질의어에서 얻은 유사도를 이용하여 각 컬렉션에서 얻은 결과의 신뢰도를 평가할 수 있다

그림1은 본 연구에서 제안한 비지도 학습에 의한 컬렉션 융합 과정을 보여준다.

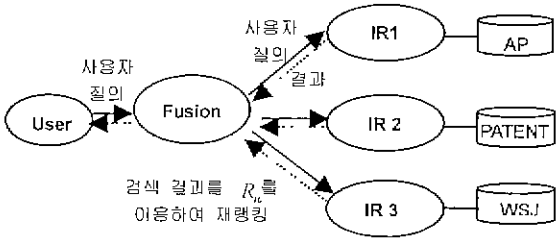
2.1 학습 질의어에 의한 컬렉션의 신뢰도 추정

학습 질의어 벡터 q_t와 컬렉션 I_c에서 검색된 문서 벡터 d_{jc}는 다음과 같이 구성된다고 할 때,

$$q_t = (tq_{1t}, tq_{2t}, ..., tq_{kt}, ..., tq_{mt})$$



(a) 학습 과정



(b) 분산 검색 과정

그림 1. 비지도 학습에 의한 컬렉션 융합

$$d_{jc} = (td_{1jc}, td_{2jc}, ..., td_{kjc}, ..., td_{mjc})$$

q_t와의 d_{jc} 유사도는 다음과 같이 구할 수 있다.

$$Sim(q_t, d_{jc}) = \frac{\sum_{k}^{n}(tq_{kt} \times td_{kjc})}{\sqrt{\sum_{k}^{n} tq_{kt}^{2} \times \sum_{k}^{n} td_{kjc}^{2}}}$$

이때 q_t로 I_c에서 검색된 문서의 수가 K개라면 q_t에 대한 I_c의 신뢰도 R_{tc} 다음과 같다.

$$R_{tc} = \sum_{j}^{K} Sim(q_t, d_{jc}) \times \alpha_{jc}$$

$$= \sum_{j}^{K} \left(\frac{\sum_{k}^{n}(tq_{kt} \times td_{kjc})}{\sqrt{\sum_{k}^{n} tq_{kt}^{2} \times \sum_{k}^{n} td_{kjc}^{2}}} \times \alpha_{jc} \right) \quad \text{(식 1)}$$

여기에서 R_{tc}는 q_t에 의해 검색된 문서들의 신뢰도를 나타내는 것으로 새로운 질의어로 각 컬렉션을 검색해 얻은 문서의 유사도를 변경하는데 이용된다.

α_{jc}는 q_t로 I_c에서 검색된 문서 리스트와 함께 반환하는 문서 d_{jc}의 유사도 S_{jc}를 반영하는 것으로 아래와 같이 각 학습 질의어들로 같은 컬렉션에서 검색된 문서들의 유사도 중 최대 값을 기준으로 계산된다

$$\alpha_{jc} = \frac{S_{jc}}{Max\{S_c\}} \quad (식\ 2)$$

2.2 사용자 질의어에 따른 컬렉션 신뢰도 결정

사용자 질의어 q_r이 다음과 같이 구성된다고 가정하자.

$$q_r = (tq_{1r}, tq_{2r}, ..., tq_{nr})$$

우선 학습 질의어 q_t에서 q_r의 용어를 포함하는 L개의 질의어 q_{lt}를 선택한다

q_r에 의해 I_c에서 검색된 문서 d_{jc}의 유사도 S_{jc}는 다음과 같이 새로이 S'_{jc} 계산된다.

$$S_{rc} = \sum_{l}^{L} Sim(q_{lt}, q_t) \times R_{lc} \quad (식\ 3)$$

$$S'_{jc} = S_{jc} \times \frac{S_{jc}}{Max\{S_c\}}$$

이와 같이 새로이 계산된 문서의 유사도에 따라 전체 컬렉션에서 검색된 문서의 순위를 조정한다

3. 실험 및 고찰

실험 환경은 네트웍 상에 분산되어 각각 자신의 컬렉션을 가진 검색 시스템들을 멀티쓰레드(multi-thread)방식을 이용하여 접근하였다.

본 연구의 실험에 이용한 컬렉션은 TREC 컬렉션의 WST와 PATENT, AP으로 이들을 색인하고 검색하였다. 실험은 크게 학습과정와 실험과정으로 나누어 진다. 학습과정을 위한 학습 질의어로는 TREC 질의어 중 51-100을 선택하였고 실험과정을 위해 TREC질의어 101-150을 선정하였다. 학습과정에서는 학습 질의어들에 대해 각 컬렉션에서 50개의 문서를 검색하여 클러스터링 정도를 측정하였고 실험과정에서는 각 질의어에 대해 컬렉션마다 300개의 문서를 얻어 유사도를 조정하였다 컬렉션 융합을 이용한 경우 평균 정확률은 0.5585 이었고 단순히 결합한 실험에서는 0.5162의 정확률을 얻었다. 문서 클러스터링을 이용한 컬렉션 융합 방법은 단순한 결합 방법 보다 우수한 결과를 보여 분산 환경 정보 검색 시스템을 위한 효과적인 방법임을 알 수 있다. 그림2는 컬렉션 융합을 한 경우와 컬렉션 융합을 하지 않은 경우의 정확율을 보여 준다

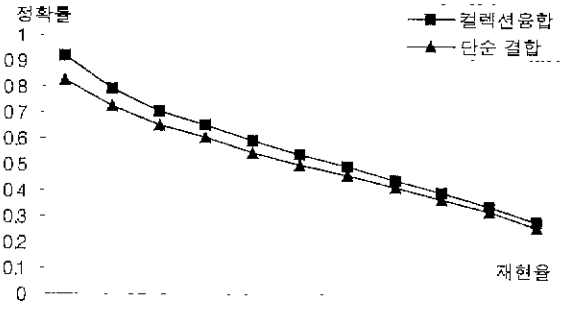


그림 2. 컬렉션 융합 실험 결과

4. 결론 및 향후 연구과제

본 연구에서는 분산 정보 검색시스템의 문제인 컬렉션 융합에 관한 연구이다.

본 연구에서 제안하는 방법은 실제 적용이 어려운 지도 학습 컬렉션 융합 방법을 개선한 비지도 학습 방법으로 학습 질의어와 검색된 문서 사이의 유사도를 계산하여 컬렉션에 대한 질의어의 관련도를 추정하였다. 새로운 질의어에 대해서는 학습 질의어 중 몇 개를 선택하여 컬렉션에서 검색된 문서의 유사도를 재조정하였다 본 연구에서 제안한 문서 클러스터링을 이용한 컬렉션 융합 방법은 타당성을 입증하기 위한 실험 결과에서 볼 수 있듯이 분산 환경의 정보 검색 시스템에서 매우 효과적인 방법임을 알 수 있다

본 연구는 분산 환경 디지털 도서관 같은 여러 컬렉션에서 효율적인 검색이 요구되는 분야에 필수적인 요소로 분산 환경에 적용하는 연구를 앞으로 진행할 것이며 인터넷 메타 검색 엔진에 적용할 것이다

참고문헌

[1] Ellen M Voorhees, Narendra K. Gupta Ben Johnson-Laird, "Learning collection fusion strategies", SIGIR '95, p172-179, 1995

[2] Ellen M Voorhees. Narendra K Gupta, Ben Johnson-Laird, "The collection fusion problem", TREC-3, 1994

[3] Alistair Moffat Justin Zobel, "Information Retrieval Systems for large document collections", TREC-3, 1994

[4] Charles L Viles. James C. French. "Dissemination of collection wide information in a distributed information retrieval system", SIGIR '95. p12-20. 1995

[5] Brian T. Bartell. Garrison W Cottrell. Richard K Belew. "Automatic combination of multiple ranked retrieval systems". SIGIR '94, p173-181, 1994

[6] Christoph Baumgarten. "A probabilistic model for distributed information retrieval". SIGIR '97, p 258-267. 1997

[7] James P Callan. Zhihong Lu. W Bruce Croft. " Searching distributed collections with inference networks ". SIGIR '95, p21-28, 1995